The Sixth ACM SIGKDD International Conference on
Knowledge Discovery and Data Mining

Working Notes

Workshop on

Multimedia Data Mining

August 20, 2000
Boston, MA, USA

Chairs
Simeon J. Simoff, *University of Sydney*
Osmar R. Zaiane, *University of Alberta*

MDM/KDD 2000

Proceedings of the First International Workshop on

Multimedia Data Mining (MDM/KDD’2000)

August 20, 2000
Boston, MA, USA

Edited by: Simeon J. Simoff and Osmar R. Zaïane

in conjunction with

Sixth ACM SIGKDD International Conference on
Knowledge Discovery & Data Mining
August 20 - 23, 2000, Boston, MA, USA

Web Page:
http://www.cs.ualberta.ca/~zaiane/mdm_kdd2000/

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000 (S.J. Simoff, O.R. Zaïane, eds.)

The official workshop web site is:
http://www.cs.ualberta.ca/~zaiane/mdm_kdd2000/

An electronic version of the proceedings will be archived after the workshop at the ACM digital library archive site:
http://www.acm.org/sigkdd/proceedings/mdmkdd00/

Foreword

Multimedia computing, especially networked multimedia, in science, business, academia, medicine and government generates substantial amount of digital media data sets. Vinton G. Cerf, President of the Internet Society and Senior Vice-President of MCI Data Services compares the activities surrounding multimedia ideas with the metaphor of a disturbed ant hill, in which the inhabitants run hither and yon to discover the cause of disturbance and, perhaps, to do something about it. No wonder researchers and developers in multimedia information systems turn to data mining and knowledge discovery methods looking for techniques for improving the indexing and retrieval of necessary information out of these data sets. Furthermore, as the ultimate goal of the knowledge discovery process is turning data into knowledge, there is a need for methods, techniques and tools that on the one hand, extract patterns from such divergent data sets and transform them into useful information and knowledge, and, on the other hand, provide consistent framework for incorporation and use of discovered knowledge in the information systems.

This volume contains the papers selected for presentation at the First International Workshop on Multimedia Data Mining (MDM/KDD'2000) held in conjunction with the Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining in Boston, Massachusetts, USA on August 20th, 2000. The aim of the workshop is to bring together experts in analysis of digital media, state-of-art data mining and knowledge discovery in multimedia database systems, knowledge engineers and domain experts from various applied disciplines with potential in multimedia data mining. The papers in this volume describe recent advances both in theoretical and practical aspects of data mining in digital media representations. They are grouped in the following streams:

- Mining spatial multimedia data
- Mining audio data
- Mining image and video data
- Multimedia support for data mining

The papers are of particular interest to researchers, developers and users of advanced data analysis and mining methodologies.

KDD conference series is a very competitive forum and winning a space and time to conduct a workshop is not a trivial task. We would like to thank all those, who anticipated the need in a multimedia data mining workshop and continuously supported our efforts through all the stages - from the submission of competitive proposal to bringing the workshop to reality. There were 27 submitted papers from 12 different countries: Australia, Belgium, Canada, China, France, Germany, Japan, Malaysia, Switzerland, Taiwan, United Kingdom, and United States of America. All papers were extensively reviewed by three referees drawn from the program committee and external reviewers. Special thanks go to them for the final quality of selected papers depends on their efforts.

Simeon J. Simoff & Osmar R. Zaïane
July 2000

Workshop Chairs:

- Simeon J. Simoff, University of Sydney, Australia
- Osmar R. Zaïane, University of Alberta, Canada

Program Committee

- Max Bramer, University of Portsmouth,UK
- Alex Duffy, University of Strathclyde,UK
- Max J. Egenhofer, University of Maine, USA
- Tom Gedeon, Murdoch University, Australia
- William Grosky, Wayne State University, USA
- Howard J. Hamilton, University of Regina, Canada
- Jiawei Han, Simon Fraser University, Canada
- Odej Kao, Technical University of Clausthal, Germany
- Nik Kasabov, University of Ottago, New Zealand
- Raymond Ng, University of British Columbia, Canada
- Timothy K. Shih, Tamkang University, Taiwan
- Jaideep Srivastava , University of Minnesota, USA

External Reviewers

- Terry Caelli
- Hang Cui
- Mohammad El Hajj
- Randy Goebel
- Mark Haffey
- Kamran Karimi
- Mario Nascimento
- Tong Zheng

Acknowledgements

Special thanks and appreciation go to Osmar Zaïane for his tremendous organisational and intellectual effort and input in the paper collection, organisition of the review process and the camera ready version of the proceedings.

Program for MDM/KDD2000 Workshop

Sunday, August 20, 2000, Boston, MA, USA

8:45 - 9:00 Opening and Welcome

9:00 - 10:00 Session 1 (Mining Spatial Multimedia Data)

- 09:00- 09:20 Geo-Spatial Clustering with User-Specified Constraints
  Anthony K.H. Tong, Raymond T. Ng, Laks V.S. Lakshmanan, Jiawei Han
- 09:20- 09:40 Multi-level Indexing and GIS Enhanced Learning for Satellite Imageries
  Krzysztof Koperski and Giovanni B. Marchisio
- 09:40-10:00 Predicting Locations Using Map Similarity(PLUMS): A Framework for Spatial Data Mining. Sanjay Chawla, Shashi Shekhar, Weili Wu and Uygar Ozesmi

10:00 - 10:30 Coffee break

10:30 - 12:00 Session 2 (Mining Audio Data & Multimedia Support)

- 10:30-10:50 Learning Prosodic Patterns for Mandarin Speech Synthesis
  Yiqiang Chen, Wen Gao, Tingshao Zhu
- 10:50-11:10 Unsupervised Classification of Sound for Multimedia Indexing
  Bruce Matichuk, Osmar R. Zaïane
- 11:10-11:30 Effective Retrieval of Audio Information from Annotated Text Using Ontologies
  Latifur Khan and Dennis McLeod
- 11:30-11:50 Incorporating Domain Knowledge with Video and Voice Data Analysis in News Broadcasts. Kim Shearer, Chitra Dorai, Svetha Venkatesh
- 11:50-12:10 Multimedia Support for Complex Multidimensional Data Mining
  Monique Noirhomme-Fraiture

12:00 - 13:00 Lunch

13:00 - 15:00 Session 3 (Mining Image and Video Data)

- 13:00-13:20 A Self Organizing Map (SOM) Extended Model for Information Discovery in a Digital Library Context. Jean-Charles Lamirel, Jacques Ducloy, Hager Kammoun
- 13:20-13:40 Learning Feature Weights from User Behavior in Content-Based Image Retrieval
  Henning Müller, Wolfgang Müller, David McG Squire
- 13:40-14:00 When image indexing meets knowledge discovery
  Chabane Djeraba
- 14:00-14:20 Semantic indexing and temporal rule discovery for time-series sattelite images
  Rie Honda, Osamu Konoshi
- 14:20-14:40 Data Mining from Functional Brain Images
  Mitsuru Kakimoto, Chie Morita, Yoshiaki Kikuchi, Hiroshi Tsukimoto
- 14:40-15:00 Mining Cinematic Knowledge Work in Progress- An Extended Abstract
  Duminda Wijesekera and Daniel Barbara

15:00-16:00 Session 4 (Discussion)

- 15:00-15:15 Variations on Multimedia Data Mining
  Simeon J. Simoff
- 15:15-16:00 Discussion: What does Multimedia Mining encompass and what are the open issues

16:00-16:30 Coffee break

16:30 - 19:00 SIGKDD'2000 Conference Opening and awards

Table of Contents

Geo-spatial Clustering with User-Specified Constraints. [*]

Anthony K. H. Tung [†]
Simon Fraser U.
khtung@cs.sfu.ca

Raymond T. Ng
U. of British Columbia
rng@cs.ubc.ca

Laks V.S. Lakshmanan
IIT, Bombay & Concordia U.
laks@cs.concordia.ca

Jiawei Han
Simon Fraser U.
han@cs.sfu.ca

Abstract

Capturing application semantics and allowing a human analyst to express his focus in mining have been the motivation for several recent studies on constrained mining. In this paper, we introduce and study the problem of constrained clustering—finding clusters that satisfy certain user-specified constraints. We argue that this problem arises naturally in practice. Two types of constraints are discussed in this paper. The first type of constraints are imposed by physical obstacles that exist in the region of clustering. The second type of constraints are SQL constraints which every cluster must satisfy. We provide a prelimary introduction to both types of constraints and discuss some techniques for solving them.

1 Introduction

Cluster analysis, which groups data for finding overall distribution patterns and interesting correlations among data sets, has numerous applications in pattern recognition, spatial data analysis, image processing, market research, etc. Cluster analysis has been an active area of research in computational statistics and data mining, with many effective and scalable clustering methods developed recently.

These methods can be categorized into partitioning methods [KR90, NH94, BFR98], hierarchical methods [KR90, ZRL96, GRS98, KHK99], density-based methods [EKSX96, ABKS99, HK98], grid-based methods [WYM97, SCZ98, AGGR98], and model-based methods [SD90, Fis87, CS96, Koh82]. In the context of GIS, cluster analysis can be very useful in identifying groups of similar points on the map and performing detail analysis of each group. This can be useful for tasks like facilities planning since a facility can then be allocated to serve each group of objects separately.

Unfortunately, the task of planning the location of facilities is usually quite complicated since users could like to enforce some constraints when performing such a task. One possible constraint might be due to the existence of obstacles in the clustering region. Let us illustrate this with an example.

Example 1.1 *A bank manager wishes to locate 4 ATMs in the area shown in Figure 1a to serve the customers who are represented by points in the figure. In such a situation however, obstacles may exist in the area which should not be ignored. This is because ignoring these obstacles will result in clusters like those in Figure 1b which are obviously wrong. Since cluster C_1 for example is split by a river, some customers on one side of the river will have to travel a long way to the allocated ATM on the other side of the river.* □

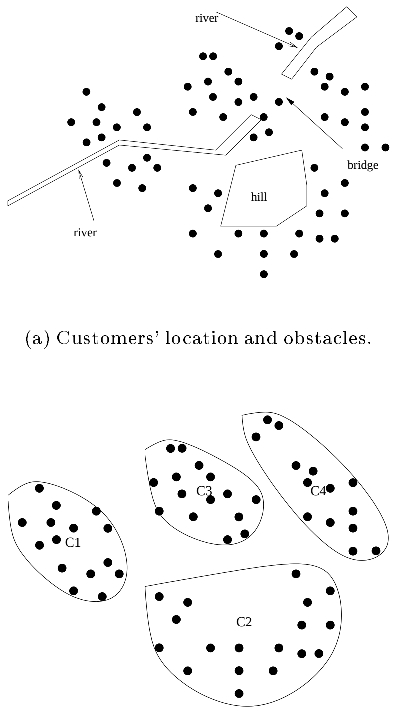
Besides constraints imposed by obstacles, users can also face constraints due to operational requirement as follows.

Example 1.2 Consider a package delivery company which is seeking to use a GIS to help de-

[*] Research was supported in part by research grants from the Natural Sciences and Engineering Research Council of Canada, and grants NCE:IRIS3 and NCE:GEOID from the Networks of Centres of Excellence of Canada.

[†] Person handling correspondence. Postal Address: Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, B. C., Canada.

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)



(a) Customers' location and obstacles.

(b) Clusters formed when ignoring obstacles.

Figure 1: Planning the location of ATMs

termine the locations for k service stations in a city. Suppose the GIS contains the information of customers based on the scheme: $customer(Name$, $AddrXcoord$,$AddrYcoord$, $MemberType$, $AvgMonthChg$). The company may formulate this location selection problem as an instance of the clustering problem, using the address fields $AddrXcoord$ and $AddrYcoord$ to define the distance function $df()$.

Suppose further that the company has two kinds of customers in consideration: *gold* customers, who need frequent, regular services, and *ordinary* customers, who require occasional services. In order to save the cost and provide good service, the manager may add the following constraints: (1) that each station should serve at least 50 gold customers; and (2) that each station should serve at least 5000 ordinary customers. With the constraints, this becomes an instance of the constrained clustering problem. □

As can be seen, the problem of constrained clustering is a very practical problem faced by the users who are not given ways to specify the type of clusters that they want to discovered. In view of this, we introduce the notion of constrained clustering in this paper and introduce some initial work which is being done to address these problems. The organization of the rest of this paper is as follows. In the next section, we will give an introduction to the problem of clustering with obstacles entities. We will described techniques which are used to improve the scalability of our algorithm when obstacle constraints are taken into consideration. In Section 3, we will look at the problem of clustering with SQL aggregate constraints and discuss some prelimary work on clustering with such constraints. We will conclude our paper with Section 4.

2 Clustering with Obstacle Entities (COE)

In order to solve the problem shown in Example 1.1, let us first formally defined the problem as follows.

Definition 2.1 *We are given a set P of n points $\{p_1, p_2, ..., p_n\}$ and a set O of m* **non-intersecting** *obstacles $\{o_1, ..., o_m\}$ in a two dimensional region, R. Each obstacle o_i is represented by a simple polygon with $o_i.nv$ sides and each vertex of the polygon is denoted as $o_i.v_j$, $1 \leq j \leq o_i.nv$. The distance, $df(p,q)$ between any two points, p and q is defined as the length of the shortest Euclidean path from p to q without cutting through any obstacles. To distinguish this distance from the direct Euclidean distance, we will refer to this distance as* obstructed distance *in this paper. Our objective is to partition P into k clusters $Cl_1, ..., Cl_k$ such that the following square-error function, E, is minimized:*

$$E = \sum_{i=1}^{k} \sum_{p \in Cl_i} d^2(p, m_i)$$

In order to solve the above problem, a trivial solution is to argue that obstacles in effect only cause a change in the distance function and thus can be hidden from the actual clustering algorithm by simply providing a different distance function call to it. However our work in [THH00] shows that a clustering algorithm which takes these obstacles into consideration can in fact be optimized to improve clustering efficiency.

In [THH00], we developed a clustering algorithm called COE-CLARANS to handle clustering with obstacles. COE-CLARANS is an improved version of CLARANS in [NH94] which is a k-medoid clustering algorithm. The CLARANS algorithm first randomly chooses k objects as the set of cluster centers, *current*. It then assigns the rest of the objects to the nearest cluster center and compute the square-error function E for the initial solution. A search is then done for a better solution by taking each cluster center following randomize order and trying to replace it with another randomly selected object not in *current*. If a better solution is found, i.e,. a lower value of E is computed for the new solution, *current* is set to the new solution and the whole process is repeated with the new *current*. For each cluster center, the attempt to find a better solution by center replacment is repeated *maxneighbor* times and the best solution is kept. If no better solution is found after *maxneighbor* attempts on all the k cluster centers, it is concluded that a local minima is reached. This process repeats *numlocal* times and the best local minima that is found will be output as the solution.

There are however certain issues which must be addressed in order to adopt CLARANS to cluster objects with obstacle constraints imposed. As can be seen, CLARANS is a generate-and-test algorithm which frequently recompute the square-error function E for testing a generated solution. To perform this operation, a scan must be done through the n objects to compute their distance from their cluster center. If the objects are stored in secondary storage, high I/O cost will be incurred. Furthermore, since the solution is generated by randomly picking another object to replace a cluster center, there is a good chance that it is not a better solution and thus does not justify the time spent on computing E. In the case of clustering with obstacles, such overhead is even higher as the obstacles have complicated the distance function. In order to overcome these problems, the following two approaches are adopted.

First, a pre-clustering step similar to those in BIRCH [ZRL96], ScaleKM [BFR98] and CHAMELEON [KHK99] is taken to group the objects into a set of *micro-clusters*. Ganti et. al. in [GGR99] gives an analogy to pre-clustering as follow:

> ”... if each data point is a marble on a table top, we replace clusters of marbles by tennis balls and then look for clusters of tennis balls.”

A *micro-cluster* is the tennis ball in the analogy. It is a group of points which are so close to each other that they are very likely to belong to the same cluster. To compress the data set, a point from each micro-cluster is selected to represent the micro-cluster. Since the size of these representative points is much smaller than the actual data set, they could be clustered using the COE-CLARANS algorithm in the main memory. To facilitate the clustering, information about the micro-cluster are stored together with the representative points. This information would include statistic like the number of points in the micro-cluster, the diameter of the micro-cluster, etc.

Second, to avoid the unnecessary computation of the square-error function E, an initial **lower bound** of E, E', is first computed. If E' is already higher than the best solution so far, then the generated solution can never be better than the best solution and thus can be abandoned without the need for E to be computed. To compute E', we underestimate the distance between the randomly chosen center o_{random} and the micro-clusters by using direct Euclidean distance instead of the obstructed distance. By doing so, each micro-clusters so formed will fall into one of the following categories:

1) p is correctly assigned to o_{random}.
Since the direct Euclidean distance between p and o_{random} must be shorter than the obstructed distance between p and o_{random}, we have underestimate the actual distance between p and o_{random}.

2) p is wrongly assigned to o_{random}.
Let o_i be the cluster center that p should rightfully be assigned to. Since p is assigned to o_{random} instead, the direct Euclidean distance between p

and o_{random} must be shorter than the obstructed distance between p and o_i which is computed before the iteration begins. Thus, we have underestimated the actual distance between p and o_i.

3) p is not assigned to o_{random}.
Since the obstructed distance of p to the rest of the $k-1$ cluster centers o_j is computed before the iteration begin, the distance used to compute E' must be correct.

As we can see, for all the three categories, we either underestimate or compute correctly the obstructed distance of a micro-cluster p to its nearest cluster center. As such E' must be a lower bound for the actual square-error function E.

By adopting the above two approaches, we are able to make our algorithm scalable for a large number of objects and a moderate number of obstacles. We illustrate the difference between clustering with obstacles and without obstacles in Figure 2. Further details of our work in this area can be found in [Hou99].

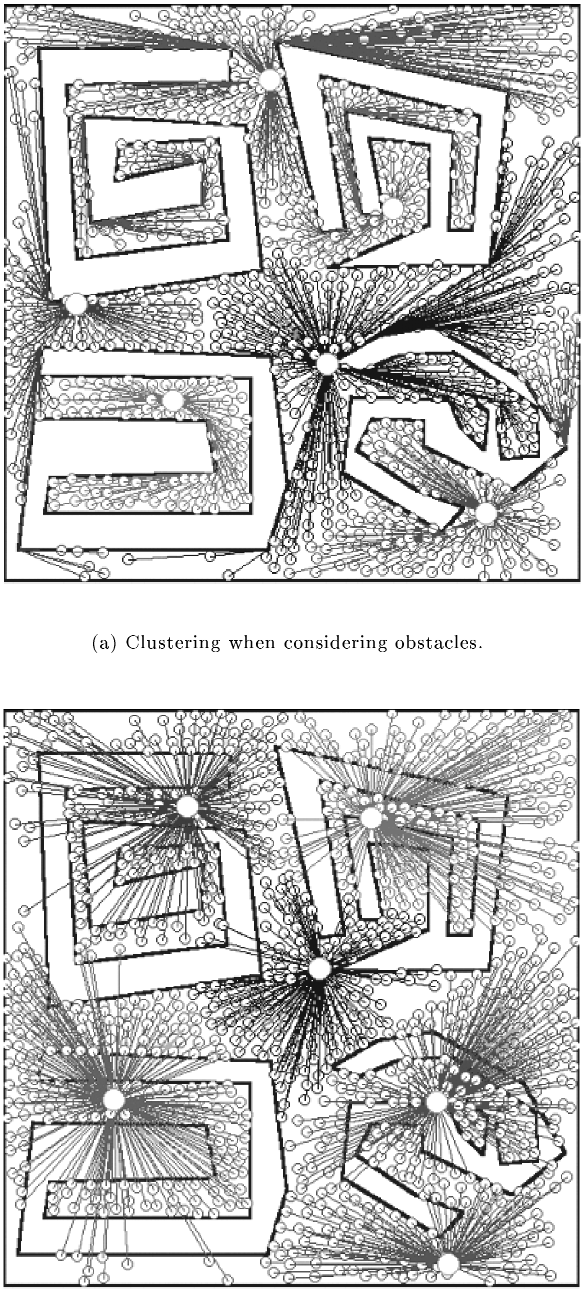
3 Clustering Under SQL Aggregate Constraints

In order to handle the type of constraints that we seen in Example 1.2, we look into the problem of clustering under SQL aggregate constraints in [TNLH00]. We define *SQL aggregate constraints* as follows.

Definition 3.1 (SQL Aggregate Constraints) Let each object p_i in the database D be associated with a set of m attributes $\{a_1, \ldots, a_m\}$. The value of an attribute a_j of an object p_i is denoted as $p_i[a_j]$.

Let the aggregate functions $agg_1 \in \{max(), min(), avg(), sum()\}$ and $agg_2 \in \{count()\}$. Let θ be a comparator function, i.e., $\theta \in \{<, \leq, \neq, =, \geq, >$ $\}$, and c represent a numeric constant. Given a cluster Cl, an SQL aggregate constraint on Cl is a constraint in one of the following forms: (i) $agg_1(\{p_i[a_j] \mid p_i \in Cl\})\ \theta\ c$; or (ii) $agg_2(Cl)\ \theta\ c$. □

While solving some of these SQL constraints can be rather complicated, a large number of them could however be reduced to a type of constraints called *existential constraints* defined as follows.



(a) Clustering when considering obstacles.

(b) Clustering when Ignoring Obstacles.

Figure 2: How Obstacles affect clusters.

Definition 3.2 (Existential Constraints) Let $W \subset D$ be any subset of objects. We often call them **pivot** objects. Let c be a positive integer. An **existential constraint** on a cluster Cl is a constraint of the form: $count(\{p_i | p_i \in Cl, p_i \in W\}) \geq c$. □

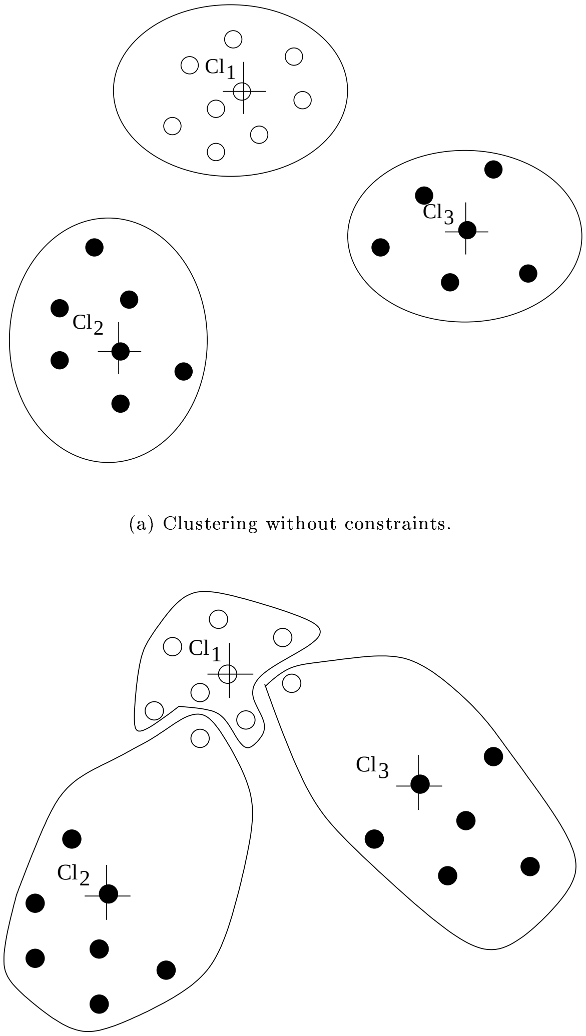
By examining the class of SQL constraints that we have defined, we can see that some of the SQL constraints can be easily reduced to an existential constraint. For example, "$count(Cl) \geq c$" is in fact a special case of existential constraints in which all objects are pivot objects. Similarly, a constraint like "$max(\{p_i[a_j] \mid p_i \in Cl\}) \geq d$" can also be reduced to an existential constraint in which the pivot objects are in the set $\{p_i | p_i[a_j] \geq d\}$ and each cluster must contain more than one pivot object.

Becauses of its importance, we focus on solving the constrained clustering involving in one existential constraint in [TNLH00]. More specifically, our problem definition is as below.

Definition 3.3 The Constrained Clustering (CC) Problem *Given a data set D with n objects, a distance function $df : D \times D \longrightarrow \Re$, a positive integer k, and* an existential constraints *EC, find a k*-clustering *$(Cl_1, \ldots, Cl_k)$ such that $DISP = (\sum_{i=1}^{k} disp(Cl_i))$ is minimized, and* each cluster Cl_i satisfies the constraint *EC, denoted as $Cl_i \models \mathcal{C}$.*

The "dispersion" or "square-error" of cluster Cl_i, $disp(Cl_i)$, measures the total distance between each object in Cl_i and some *representative* rep_i of Cl_i, i.e., $disp(Cl_i)$ defined as $\sum_{p \in Cl_i} df(p, rep_i)$. Typically, these representatives are the centroids or the medoids of the clusters which will minimize the dispersion of each cluster and thus their locations are good candidates for locating the facilities that serve the clusters.

With the introduction of an existential constraint, one major complication is that instead of being assigned to the nearest center, a pivot object might be assigned to a cluster center which is further away because of the need to satisfy the existential constraint. Let us consider the example shown in Figure 3a. In the figure, the hollow points represent pivot objects while the solid points are non-pivot objects. Without any constraint imposed on the clustering, a natural way to group the points is shown in Figure 3a. However if we impose a constraint that each cluster must at least contain one pivot point, then a solution could be in Figure 3b where one pivot point is "forced" to be in cluster Cl_2 and one in Cl_3 although both these points are actually nearer to the center of cluster Cl_1. Because of this, the constraint k-means algorithm which we introduce in [TNLH00] first tries to satisfy user-specified constraint before trying to refine the clusters by swapping objects between the clusters. In order for the clusters to be valid after the refinement, the swapping of the an object is only done if the change in membership of the object does not invalidate the user-specified constraint. More details of the algorithm can be obtained from [TNLH00].



(a) Clustering without constraints.

(b) Clustering with Constraints.

Figure 3: How an existential constraint affects clusters.

4 Conclusion

In this paper, we have introduced and studied the problem of having user-specified constraints in geo-spatial clustering. Even though constrained clustering problems arise naturally in practice, this appears to be the first attempt to tackle these problems. Two types of constraints are discussed in this paper. The first type of constraints are imposed by physical obstacles that exist in the region of clustering. The second type of constraints are SQL constraints which every cluster must satisfy. We discuss some techniques for solving these two types of constraints and hope that more work will be done in these area.

References

[ABKS99] M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *Proc. 1999 ACM-SIGMOD Conf. on Management of Data (SIGMOD'99)*, pages 49–60, Philadelphia, PA, June 1999.

[AGGR98] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*, pages 94–105, Seattle, Washington, June 1998.

[BFR98] P. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, pages 9–15, New York, NY, August 1998.

[CS96] P. Cheeseman and J. Stutz. Bayesian classification (AutoClass): Theory and results. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 153–180. AAAI/MIT Press, 1996.

[EKSX96] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. In *Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, pages 226–231, Portland, Oregon, August 1996.

[Fis87] D. Fisher. Improving inference through conceptual clustering. In *Proc. 1987 AAAI Conf.*, pages 461–465, Seattle, Washington, July 1987.

[GGR99] V. Ganti, J. Gehrke, and R. Ramakrishnan. Mining very large databases. *COMPUTER*, 32:38–45, 1999.

[GRS98] S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*, pages 73–84, Seattle, Washington, June 1998.

[HK98] A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. In *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, pages 58–65, New York, NY, August 1998.

[Hou99] J. Hou. *Clustering with Obstacle Entities.* M.Sc. Thesis, Simon Fraser University, Canada, December 1999.

[KHK99] G. Karypis, E.-H. Han, and V. Kumar. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling. *COMPUTER*, 32:68–75, 1999.

[Koh82] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[KR90] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis.* John Wiley & Sons, 1990.

[NH94] R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. In *Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94)*, pages 144–155, Santiago, Chile, September 1994.

[SCZ98] G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. In *Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)*, pages 428–439, New York, NY, August 1998.

[SD90] J.W. Shavlik and T.G. Dietterich. *Readings in Machine Learning.* Morgan Kaufmann, 1990.

[THH00] A. K. H. Tung, J. Hou, and J. Han. COE: Clustering with obstacles entities, a preliminary study. In *Proc. 4th Pacific-Asia Conf.on Knowledge Discovery and Data Mining (PAKDD'00)*, Kyoto, Japan, 18-20, Apr. 2000.

[TNLH00] A. K. H. Tung, R. Ng, L. Lakshmanan, and J. Han. Constraint-based clustering in large database. In *Submitted to ICDT'00*, Jun. 2000.

[WYM97] W. Wang, J. Yang, and R. Muntz. STING: A statistical information grid approach to spatial data mining. In *Proc. 1997 Int. Conf. Very Large Data Bases (VLDB'97)*, pages 186–195, Athens, Greece, Aug. 1997.

[ZRL96] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'96)*, pages 103–114, Montreal, Canada, June 1996.

Multi-level Indexing and GIS Enhanced Learning for Satellite Imageries

Krzysztof Koperski
Data Analysis Products Division of Mathsoft, Inc.
1700 Westlake Ave. N, Suite 500,
Seattle, WA, 98109-3044 USA
(206) 283-8802 Ext. 243
krisk@statsci.com

Giovanni B. Marchisio
Data Analysis Products Division of Mathsoft, Inc.
1700 Westlake Ave. N, Suite 500,
Seattle, WA, 98109-3044 USA
(206) 283-8802 Ext. 280
giovanni@statsci.com

ABSTRACT

Satellite technology produces data at an enormous rate. Most of the database research on the analysis of remotely sensed images concentrated on data retrieval and simple queries that involved spatial joins and spatial selections. For example, the Sequoia 2000 project [13] aimed at the retrieval of raster data, while the Sloan Digital Sky Survey [14] poses the need for the creation of multi-terabyte astronomy archive. The large scale systems for the analysis of remotely sensed images were specialized toward the detection of particular features like volcanoes [2], or proposed distributed and parallel data storage and query processing systems for handling of geo-scientific data retrieval queries [11]. The GeoBrowse project aims to provide infrastructure that would enable the analysis of large databases containing satellite images. Our work addresses two issues. One is the extraction of information that enables reduction of the data from multi-spectral images into a number of features. Second is the organization of the features that would allow flexible and scalable discovery of the knowledge from the databases of remotely sensed images. In this paper we present the concept of data mining system for the analysis of satellite images and preliminary results of the experiments with the collection of LANDSAT images.

Keywords

Remote Sensing, Image Databases, Bayesian Classification, Similarity Searches.

1. INTRODUCTION

Satellite data is used in many different areas ranging form agriculture, forestry, and environmental studies to transportation and mining. The applications include measurements of crop and timber acreage, forecasting crop yields and forest harvest, monitoring urban growth, mapping of ice for shipping, mapping of pollution, recognition of certain rock types, and many others. The United States Geological Survey web site [15] presents other applications that use the results of the satellite data analysis.

The *GeoBrowse* project aims at providing the infrastructure required for the analysis of satellite images. Most of the systems for analyzing remotely sensed images allow simple queries based on the date of image capture and location. Such systems also allow only simple analyses of single images. When we deal with large collections of remotely sensed images, the current systems do not scale well. Therefore new algorithms and new indexing methods are needed to enable the analysis of data produced by satellite systems.

In order to facilitate the analysis of large amount of image data, we propose to extract features of images. Large images are partitioned into a number of smaller and more manageable image tiles. In addition to faster extraction of segments the partitioning allows to fetch only the relevant tiles when only retrieval of part of the image is requested. Then these image tiles are processed in order to extract feature vectors. The *GeoBrowse* architecture distinguishes between three types of feature vectors: 1) *pixel level features*, 2) *region level features*, and 3) *tile level features*. Pixel level features store spectral and textural information about each pixel of the image. For example, the fraction of the endmembers, such as concrete or water, can describe the content of the pixels. Due to the large size, pixel feature vectors are used only for the extraction of other feature vectors and can be utilized in the refinement step of the queries. Region level features describe groups of pixels. Following the segmentation process, each region is described by its boundary and a number of attributes which present information about the content of the region in terms of the endmembers and texture, shape, size, fractal scale, etc. Image tile level features present information about whole images using texture, percentages of endmembers, fractal scale and others.

There are many similarities between data mining in the collections of photographic images and data mining in the collections of sattellite images. In both cases features, such as texture, or color histograms are used in the analysis. However, in the case of the remotely sensed images a user can use additional information, such as Digital Elevation Models (DEM), or land use maps, to enhance the search capabilities and improve the quality of the classification and prediction process.

In this paper we give an overview of the GeoBrowse system and present the results of similarity searches for different types of urban areas. For the experiments we used the LANDSAT image of Western Washington State. This image contains about 500MB of raw pixel information in 6 bands (3 visible range and 3 near infrared bands). The image was corrected for atmospheric and terrain distortions, and georeferenced. In order to enable work

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

with chunks of images that are feasible for the segmentation algorithm, the whole image was divided into 512 pixels × 512 pixels *image tiles*.

The remaining part of the paper is organized as follows. In Section 2 we present the architecture of the system. Section 3 describes the algorithm for the segmentation of multiband images and features that describe the regions. In Section 4 we present theresults of the similarity retrieval queries. Section 5 outlines the data mining methods for the analysis of remotely sensed images. The paper ends with conclusions and the description of future work.

2. ARCHITECTURE

We decided to use a database system for storage of images and their features. This way we may overcome limits related to the maximum size of files and benefit from indexing, query optimization, and partitioning features of the database. The image tiles and pixel level features are stored as BLOBs, each band in a separate column. The region and tile level features are stored in regular database tables, which can be easily accessed for further processing using GeoBrowse functions or by over 3000 function of S-PLUS software [12].

Spatial information about region level is stored in ESRI's Spatial Data Engine (SDE) together with the relevant GIS information. SDE provides open data access across local and wide area networks and the Internet using the TCP/IP protocol. It can retrieve data and perform spatial and geometric analysis with 14 topological searches, buffering, overlays and intersections, dissolve and clip, and topological data cleaning. Data stored in SDE can be also accessed from other ESRI products like ArcInfo, ArcView, and MapObjects, which provide alternative environment for the visualization of the query results.

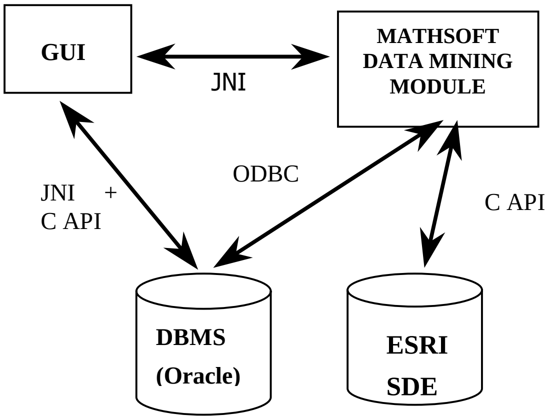


Figure 1. Architecture of GeoBrowse

A mining process or a similarity search is initiated by submitting a query written in a language similar to SQL-like data mining languages, such as DMQL [5] and GMQL [8]. In a query a user can specify the type of knowledge to be discovered; the set of data relevant to the mining process; and the thresholds to filter out uninteresting rules. Based on this query an SQL statement is constructed to retrieve the relevant data. If spatial conditions exist in the query the SDE is used for the processing, otherwise the data is retrieved directly from the database system. The data mining module processes the data and passes the information about the resulting tiles and regions to GUI, which in turn directly retrieves the images from the database.

Based on the classification model the data can be classified into a number of land cover classes and the resulting GIS map can be stored in the SDE for future use or a presentation by the GIS system.

3. SEGMENTATION AND FEATURE EXTRACTION

The segmentation process is using the function based on the algorithm presented in [7]. This function segments an input image into non-overlapping regions by minimizing an energy functional which trades off the similarity of regions against the length of their shared boundary. It starts by breaking the image into many small regions. The algorithm merges into one region the two adjoining regions that are the most alike in terms of the specified polynomial model given the length of the border between the two regions. Internally, the energy functional is evaluated using a Lagrangian parameter called lambda. Parameter is also called the scale parameter as it controls the coarseness of the segmentation where a small value of lambda corresponds to a finer segmentation with more regions and a large value corresponds to a coarse segmentation with fewer regions. Since the algorithm grows regions by merging alike regions, the value of lambda increases as the number of regions decreases. To achieve the segmentation uniformity between tiles the final value of lambda is set to be approximately the same for each image tile.

In the case of multi-band satellite images the values of the pixels are often correlated. Therefore, the Principal Component Analysis is performed based on a large sample of pixels from all tiles, all tiles are rotated to the same axes and the first three components are used for the segmentation of each image tile. After the segmentation the shape features such as eccentricity, orientation of the main axis, and invariant moments are extracted and stored in the database.

3.1 Texture Feature Extraction

We extract pixel level texture features based on Gabor wavelets. In the comparison study of texture based classification, Gabor features were judged to perform superior to other texture analysis methods, such as edge attribute processing methods, the circular simultaneous autoregressive model method and hidden Markov model methods [3]. In GeoBrowse for each pixel we extract eight features $a_i\big|_{i=0,7}$ using Gabor Filters with kernels rotated by $i\pi/8$. To achieve the rotation invariant features we find the values of the autocorrelation function $t_n = \sum_{i=0}^{7} |a_i| \left| a_{(i+n)\bmod 8} \right|$ [6]. To minimize the size of the pixel index, we have chosen to compute values of autocorrelation for $n = 0,2,4$. These values correspond to the 0°, 45°, and 90° difference in the orientation of Gabor kernels. Such shift should allow for distinguishing of urban road network, which usually are correlated within 90° rotation of the wavelet kernels. The extraction of other microfeatures such as frequency, orientation, is also possible [6] and we plan to perform more experiments with these features in the future.

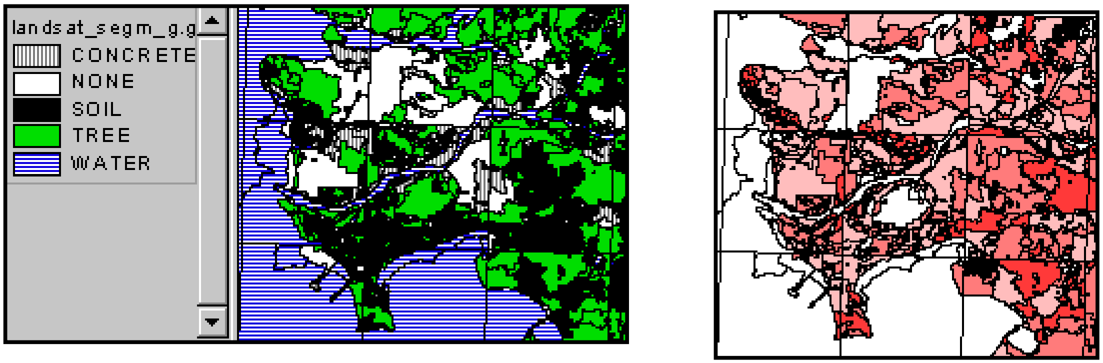


Figure 2.

Spectral Mixture Analysis region features (the prevalent endmembers).

Percentage of pixel that belong to one of the clusters.

3.2 Spectral Mixture Analysis Features

Spectral Mixture Analysis (SMA) [1, 4] enables the analysis of remotely sensed images using spectral endmembers such as concrete, water, soil, trees, etc. The pixels usually cover the area with the mixture of different endmembers. For example, in the urban areas we may find a mixture of concrete, trees, soil, grass, etc. The result of SMA represents the percentage of the contents of the endmembers within the area of the pixel. This way we can distinguish areas with different mixture of concrete, soil, water, and vegetation. The region and tile level features present the percentage of the area of a region a tile that is covered by particular endmembers. Region level texture and SMA features are presented in Figure 2.

4. SIMILARITY SEARCH

The GeoBrowse uses an SQL like query language that enables specification of the data mining task, features that are used in the mining process and further constraints. The system is capable of performing similarity searches based on any combination of features. A user can look for the most similar image tiles or the most similar regions based on a pattern tile or a region. GeoBrowse enables arbitrary weighting of the features. The values of the features can be adjusted to have the range [0, 1], they can be multiplied by a specific value, or they can remain the same.

In the case of region based searches we looked only for the regions with areas larger than 2000 pixels. The feature values were scaled to the range [0,1]. We compared the results of the similarity searches based on SMA features with the searches based on texture features and searches based on the combination of these two features. When only a single feature vector is used the results tend to have a high percentage of the areas, which could be classified as *false hits*. The selectivity of the SMA features seems to be quite high for urban patterns, but some rocks and crops have spectral signatures similar to the spectral signature of concrete and are classified as such. The selectivity of searches based on texture features is lower, but rotation invariance can be observed regardless of the orientation of the street networks. For example, the suburban area of New Westminster in the Greater Vancouver area is judged to be similar to East Vancouver, despite the fact that the main direction of the street network differs by about 30° for these two region. Figure 3 presents the result of the search for regions similar to downtown Seattle and Burnaby in British Columbia. Only regions in Puget Sound are shown. In the case of downtown Seattle the set of returned regions contained downtown areas of Vancouver, Burnaby, Bellingham, Bellevue, Tacoma, and Everett together with industrial areas of Renton, Tukwila and South Tacoma. Regions similar to Burnaby contain high-density residential areas with some small industrial and commercial pockets.

We compared the results of the tile similarity search with the region similarity search in the case when the tile containing the pattern region is treated as a pattern tile. In this case the returned tiles contained only about 40% of the top 20 most similar regions returned by region based similarity function. The features of the smaller regions tend to be overwhelmed by the overall features of the tile.

5. DATA MINING FUNCTIONALITY

In addition to the similarity search the GeoBrowse system will provide functionality for other types of the remotely sensed data analysis. This functionality will include the clustering of the data, building regression and classification models, prediction of land cover types, summarization of the data, etc.

5.1 Clustering

A user has an option to find clusters of image tiles based on any combination of feature vectors. Figure 4 shows the centroids (i.e., the image tiles located the most centrally in the feature space) for the four clusters. The clusters were found based on the relative content of endmembers in an image tile. In this case we may see that the image tiles that are the centroids of the discovered clusters represent mountain areas with large content of conifer trees; areas covered with deciduous trees; forested areas close to water; and urban areas close to water.

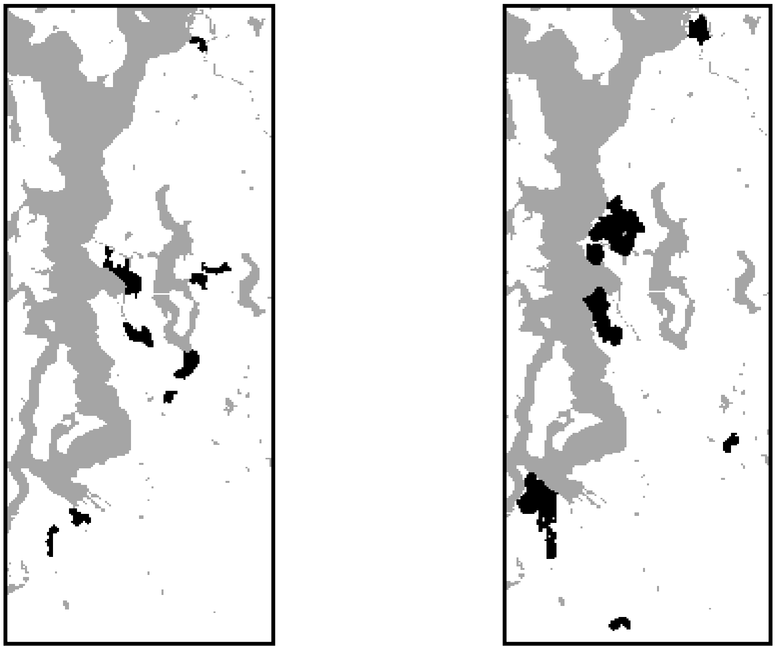


Figure 3.

Regions similar to downtown Seattle. **Regions similar to West Burnaby, BC.**

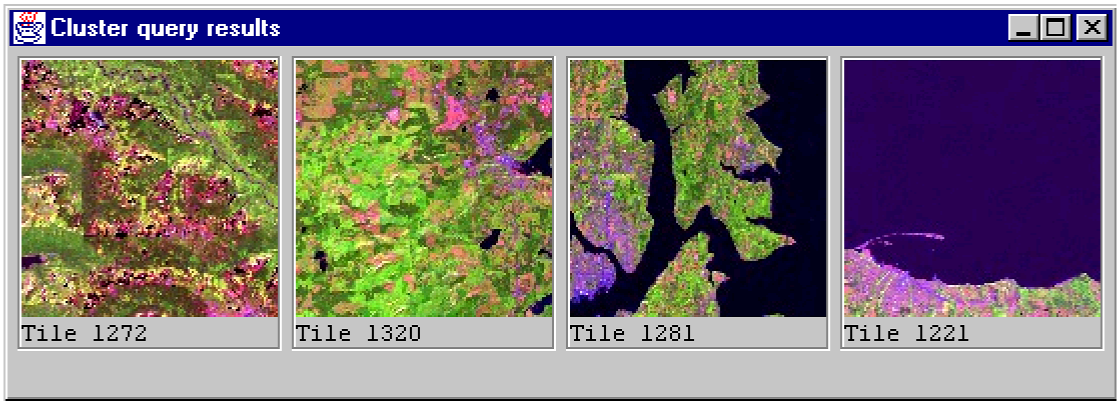


Figure 4. The centroid tiles of the clusters.

5.2 User Feedback Label Learning

In many cases it is very difficult to describe analytically the features of the objects that a user is looking for. Therefore the improvement of the description quality may play an important role in the image analysis. A method for interactive training of land cover labels using Naïve Bayesian classifiers is described in [10]. In that approach a user can interactively train Bayesian model to define a number of land cover classes, which can be based on textural or spectral properties of images. The training is done based on pixel level features, which are partitioned into a number of clusters. A user selects the pixels that belong to a new class and the pixels that do not belong there. Based on this information a model that estimates a posteriori probability of pixel's class membership is build. Using this model a user can find images with the highest probability of the defined class, or images with low or high separability of the classes. While the training is based on the pixel level features the retrieval is based on tile level features. Due to the nature of Naïve Bayesian classifier, which assumes the conditional independence of the attributes, it is possible to find out the probabilities of the pixel class assignment based on the aggregated information about all pixels in the image tile. Unfortunately the assumption of conditional independence is not always true. Therefore, Naïve Bayesian classifiers may perform well. We plan to add other classification methods, such as tree classifiers to improve user feedback label training. Because the classification process on the pixel level would be extremely expensive to compute we intend to perform experiments with the classification based on the region level features.

Building a classifier based on millions pixel features of the data would be a very time process. Instead of that we build the classifier based on region level features. In addition to spectral properties of the regions we can perform classification also based on shape properties and area of the regions, as well as auxiliary GIS information. For example, the spectral reflectance of concrete is very similar to spectral reflectance of different type of rocks. Additional information, such as Digital Elevation Models can be used to distinguish between these two types of land cover types.

6. FUTURE WORK AND CONCLUSIONS

We plan to perform experiments using multiple level spatial transformation methods for progressive refinement using more level than tile, region, and pixel levels. Multiscale image coding techniques, such as wavelets, can also be used for the analysis of images on multiple levels.

Such multilevel information can be combined with the auxiliary data in both vector and raster formats to enhance the data analysis capabilities of *GeoBrowse.* These auxiliary data can be used both during feature extraction process and during data mining process. We intend to do more experiments with other data mining methods such as regression, clustering and classification.

The quality of classification of land cover classes can be improved using time series of data, which can better differentiate between different types of crops due to the different times of crop growing seasons. We also plan to provide the functionality of multilevel presentation of the discovered knowledge. For example, the system should allow a user to see the generalized summary of the areas of particular crops by county, state, region, etc.

We designed, and we are in the process of implementing the *GeoBrowse* system for data mining of remotely sensed images. Three levels of feature are extracted from image tiles and used in the data mining process. In addition to simple queries based on simple properties, such as geographic location or acquisition date, a user can submit queries based on properties of images derived from feature vectors describing the images. The system also will allow for interactive training of the classification models that describe new types of objects. Scalability to the large databases is addressed through indexing of the feature vectors and by using scalable data mining algorithms in the query processing. Our region level indexing strategy enhances the data analysis and similarity search processes by allowing for the more refined classification of information derived from images.

7. REFERENCES

[1] Adams, J. B., M. O. Smith, and P. E. Johnson, Spectral Mixture Modeling: a New Analysis of Rock and Soil Types at Viking Lander 1. In *J. Geophys. Res*. 91:8113 – 8125, 1986.

[2] Fayyad, U. M., and P. Smyth. Image Database Exploration: Progress and Challenges. In *Proc. 1993 Knowledge Discovery in Databases Workshop*, Washington, DC, p. 14 – 27, 1993.

[3] Fountain, S. R., T. N. Tan, K. D. Baker. A Comparative Study of Rotation Invariant Classification and Retrieval of Texture Images. In *On-Line Proceedings of the Ninth British Machine Vision Conference* 1998. http://www.bmva.ac.uk/bmvc/1998/index.htm.

[4] Gillespie, A. R., M. O. Smith, J. B. Adams, and S. C. Willis, Spectral Mixture Analysis of Multispectral Thermal Infrared Images, In *Proceedings of the 2nd Thermal IR Multispectral Scanner Workshop*, JPL Publication 90-55:57 – 74, 1990.

[5] Han, J., Y. Fu, W. Wang, K. Koperski, and O. R. Zaïane. DMQL: A Data Mining Query Language for Relational Databases. *In Proc. of the Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, QB, pp. 27 – 34, 1996.

[6] Hayley, G. M., and B. M. Manjunath, Rotation Invariant Texture Classification using Modified Gabor Filters, In *Proc. of IEEE ICIP95*, pp. 262 – 265, 1994.

[7] Koepfler, G., C. Lopez and J. M. Morel, A Multiscale Algorithm for Image Segmentation by Variational Method, *SIAM Journal of Numerical Analysis*, vol. 31, pp. 282 – 299, 1994.

[8] Koperski, K. *A Progressive Refinement Approach to Spatial Data Mining*. Ph.D. Thesis. Simon Fraser University, 1999.

[9] Patel, J., et al. Building a Scalable Geo-Spatial DBMS: Technology, Implementation, and Evaluation. In *Proc. ACM-SIGMOD International Conference on Management of Data*, Tucson AZ, pp. 336 – 347, 1997.

[10] Schröder, M. Interactive Learning in Remote Sensing Image Databases In *IEEE Intern. Geoscience and Remote Sensing Symposium IGARSS'99*. Hamburg, 1999.

[11] Shek, E. C., R. R. Muntz, E. Mesrobian, and K. Ng. Scalable Exploratory Data Mining of Distributed GeoScientific Data. In *Proc. of The Second International Conference on Knowledge Discovery & Data Mining*, Aug. 2-4, Portland OR, pp. 32 – 37, 1996.

[12] *S-PLUS 2000 Programmer's Guide*, Data Analysis Products Division, MathSoft, Seattle, WA, 1999.

[13] Stonebraker, M., J. Frew, K. Gardels, and J. Meredith. The Sequoia 2000 Storage Benchmark. In *Proc. ACM-SIGMOD International Conference on Management of Data*, Washington, D.C., pp. 2 – 11, 1993.

[14] Szalay, A., P. Kunszt, A. Thakar, J.Gray, D. Slutz, and R. J. Brunner. Designing and Mining Multi-terabyte Astronomy Archives: The Sloan Digital Sky Survey. In *Proc. ACM-SIGMOD International Conference on Management of Data*, Dallas TX, pp. 451 – 462, 2000.

[15] U.S. Department of the Interior, U.S. Geological Survey, Landsat 7 data users and applications http://edcwww.cr.usgs.gov/l7dhf/L7MMO/L7applicatn.htm, 1999.

Predicting Locations Using Map Similarity(PLUMS): A Framework for Spatial Data Mining [*]

Sanjay Chawla
Vignette Corporation
Waltham, Massachusetts
chawla@cs.umn.edu

Shashi Shekhar
Computer Science
University of Minnesota
Minneapolis, MN 55455, USA.
shekhar@cs.umn.edu

Weili Wu
Computer Science
University of Minnesota
Minneapolis, MN 55455, USA.
wuw@cs.umn.edu

ABSTRACT

Spatial data mining is a process to discover interesting, potentially useful and high utility patterns embedded in spatial databases. Efficient tools for extracting information from spatial data sets can be of importance to organizations which own, generate and manage large spatial data sets. The current approach towards solving spatial data mining problems is to use classical data mining tools after "materializing" spatial relationships. However, the key property of spatial data is that of spatial autocorrelation. Like temporal data, spatial data values are influenced by values in their immediate vicinity. Ignoring spatial autocorrelation in the modeling process leads to results which are a poor-fit and unreliable. In this paper we will propose PLUMS(Predicting Locations Using Map Similarity), a new approach for supervised spatial data mining problems. PLUMS searches the space of solutions using a map-similarity measure which is more appropriate in the context of spatial data. We will show that compared to state-of-the-art spatial statistics approaches, PLUMS achieves comparable accuracy but at a fraction of the computational cost. Furthermore, PLUMS provides a general framework for specializing other data mining techniques for mining spatial data.

1. INTRODUCTION

Widespread use of spatial databases [14, 30, 33, 37] is leading to an increasing interest in mining interesting and useful but implicit spatial patterns[19, 24, 12, 29]. Efficient tools for extracting information from geo-spatial data, the focus of this work, are crucial to organizations which make decisions based on large spatial data sets. These organizations are spread across many domains including ecology and environment management, public safety, transportation, public health, business logistics, travel and tourism. [2, 15, 17, 21, 28, 34, 38].

Classical data mining algorithms [1, 10] often make assumptions(e.g. independent, identical distributions), which violate the first law of Geography: everything is related to everything else but nearby things are more related than distant things [5, 35]. In other words, the values of attributes of nearby spatial objects tend to systematically affect each other. In spatial statistics, an area within statistics devoted to the analysis of spatial data, this is called spatial autocorrelation [6]. Knowledge discovery techniques which ignore spatial autocorrelation typically perform poorly in the presence of spatial data. Spatial statistics techniques on the other hand do take spatial autocorrelation directly into account [3] but the resulting models are computationally expensive and are solved via complex numerical solvers or sampling based Markov Chain Monte Carlo(MCMC) methods [22].

In this paper we will propose PLUMS(Predicting Locations Using Map Similarity), a new approach for supervised spatial data mining problems. PLUMS searches the parameter space of models using a map-similarity measure which is more appropriate in the context of spatial data. We will show that compared to state-of-the-art spatial statistics approaches, PLUMS achieves comparable accuracy but at a fraction of the cost(two orders of magnitude). Furthermore, PLUMS provides a general framework to specialize other data mining techniques for mining spatial data.

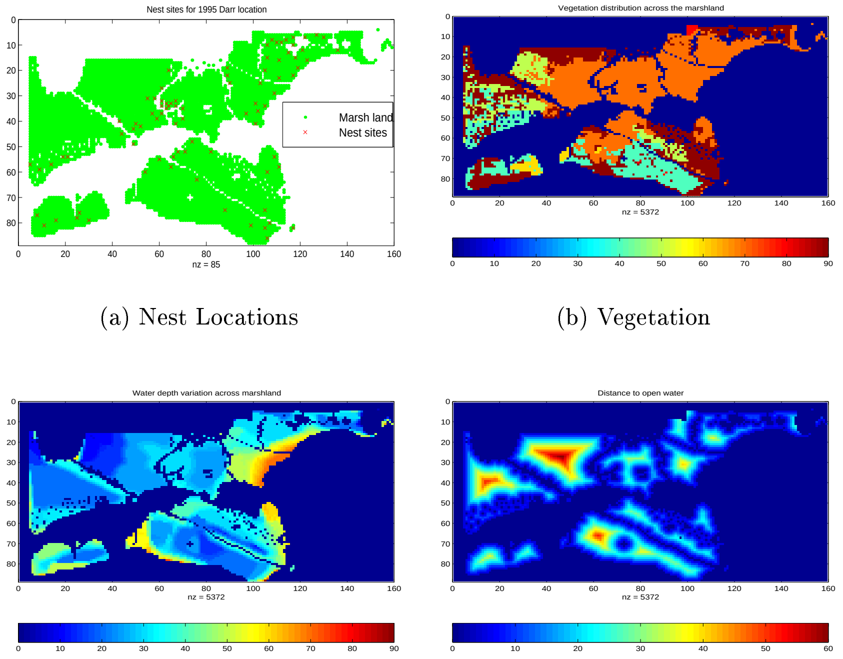
1.1 An Illustrative Application Domain

The availability of accurate spatial habitat models is an important tool for wildlife management, protection of critical habitat and endangered species. Since the underlying process governing the interaction between wildlife and environmental factors is complex, statistical models are built to gain some insight on the basis of data collected during field work. One of authors has been involved in the development of spatial model for the nesting locations of a marsh-nesting bird species [25, 26]. We will use this application, and the accompanying data, to explain the location predication problem and its unique aspects *vis-a-vis* classical data mining.

The learning and testing datasets that we will be used was collected in 1995 and 1996 from two wetlands(Darr and Stubble) located on the shores of Lake Erie in Ohio. For

[*]This work was supported in part by the Army High Performance Computing Research Center under the auspices of Department of the Army, Army Research Laboratory Cooperative agreement number DAAH04-95-2-0003/contract number DAAH04-95-C-0008, and by the National Science Foundation under grant 9631539.

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)


Nest sites for 1995 Darr location
Marsh land
Nest sites
nz = 85
Vegetation distribution across the marshland
nz = 5372
Water depth variation across marshland
nz = 5372
Distance to open water
nz = 5372
0
10
20
30
40
50
60
70
80
90
100
120
140
160

(a) Nest Locations

(b) Vegetation

(c) Water Depth

(d) Distance to Open Water

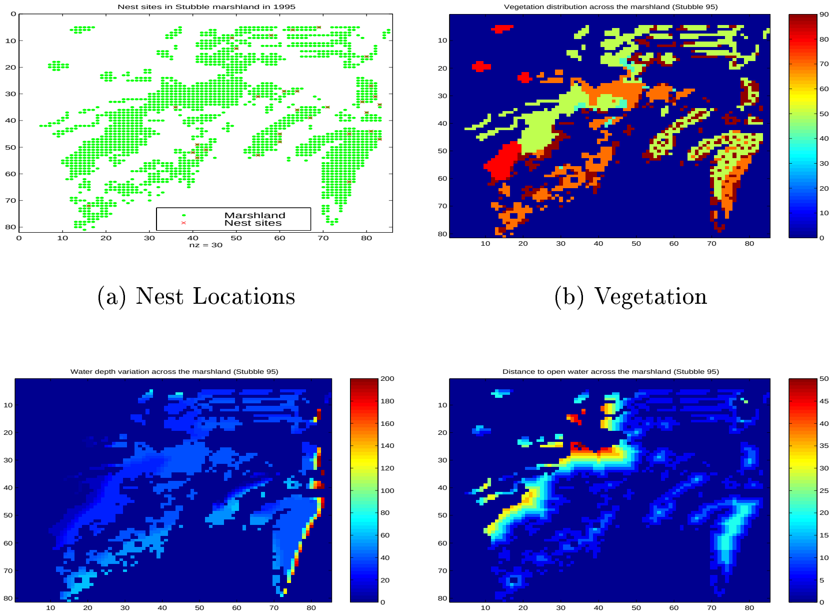
Figure 1: (a) Learning dataset: The geometry of the wetland and the locations of the nests, (b) The spatial distribution of *vegetation durability* over the wetland, (c) The spatial distribution of *water depth*, and (d) The spatial distribution of *distance to open water*.

the purpose of data collection, a local coordinate system was established for each wetland and a regular grid consisting of approximately 5000 cells was superimposed. The cells of the grid had square geometries of size 5 meters by 5 meters. In each cell the values of several structural and environmental variables were recorded, including *water depth, dominant vegetation durability index and distance to open water*. These three factors play the role of most significant explanatory variables. At each cell was also recorded the fact whether a bird-nest(red-winged blackbird) was present or not. The presence of the nest played the role of dependent variable. The geometry of the Darr wetland, locations of the nests and spatial distribution of the explanatory variables are shown in Figure 1. The corresponding maps for the Stubble wetland are shown in Figure 2.

One of the authors has applied classical data mining techniques like logistic regression[26] and neural networks[25] to build spatial habitat models. Logistic regression was used because the dependent variable is binary(nest/no-nest) and the logistic function "squashes" the real line onto the unit-interval. The values in the unit-interval can then be interpreted as probabilities. They concluded that using logistic regression the nests could be classified at a 24% rate better than random[25]. The use of neural networks actually decreased the classification accuracy[25] but led to a better understanding of the interaction between the explanatory and the dependent variable.

There are two important reasons why, despite extensive domain knowledge, the results of classical data mining are not "satisfactory". First, classical techniques, e.g. logistic regression, make assumption about independent distributions for the properties of each pixel, ignoring spatial autocorrelation. Figure 3(a) shows a spatial distribution consistent with assumption of classical regression. It looks like "white noise" as properties of pixel are generated from independent and identical distributions. Note that the maps of explanatory variable in Figure 1 have much more gradual variation indicating high spatial autocorrelation. Figure 3(b) shows a random distribution of nest locations which is quite different from the distribution of actual nests shown in Figure 1(a).

A second, more subtle but equally important reason is the objective function of classification measure accuracy. For a two-class problem the standard way to measure classification accuracy is to calculate the percentage of correctly classified objects. This measure may not be the most suitable for spatial data. Spatial accuracy is as important in this application domain due to the effects of discretization of continuous marsh into discrete pixels, as shown in Figure 4. Figure 4(a) shows the actual locations of nests and 4(b) shows the pixels with actual nests. Note the loss of information during the discretization of continuous space into pixels. Many nest location barely fell within the pixels labeled 'A' and were quite close to other pixels with label of no-nest. Now consider two predictions shown in Figure 4(c) and 4(d). Domain scientists prefer prediction 4(d) over 4(c), since predicted nest locations are closer on average to some actual nest locations. Classification accuracy measure cannot distinguish between 4(c) and 4(d), and one needs a measure of spatial accuracy to capture this preference.


Nest sites in Stubble marshland in 1995
Marshland
Nest sites
nz = 30
Vegetation distribution across the marshland (Stubble 95)
Water depth variation across the marshland (Stubble 95)
Distance to open water across the marshland (Stubble 95)

(a) Nest Locations

(b) Vegetation

(c) Water Depth

(d) Distance to Open Water

Figure 2: (a) The geometry of the wetland and the locations of the nests, (b) The spatial distribution of *vegetation durability* over the wetland, (c) The spatial distribution of *water depth*, and (d) The spatial distribution of *distance to open water*.

A simple and intuitive measure of spatial accuracy is the Average Distance to Nearest Prediction(ADNP) from the actual nest sites, which can be defined as

$$ADNP(A, P) = \frac{1}{K}\sum_{k=1}^{K} d(A_k, A_k.nearest(P)).$$

Here the A_k's are the actual nest locations, P is the map layer of predicted nest locations and $A_k.nearest(P)$ denotes the nearest predicted location to A_k. K is the number of actual nest sites. We now formalize the spatial data mining problem by incorporating notions of spatial autocorrelation and spatial accuracy in the problem definition.

1.2 Location Prediction: Problem Formulation

The Location Prediction problem is a generalization of the nest location prediction problem. It captures the essential properties of similar problems from other domains including crime prevention and environmental management. The problem is formally defined as follows:

Given :

- A spatial framework S consisting of sites $\{s_1, \dots, s_n\}$ for an underlying geographic space G.
- A collection of explanatory functions $f_{X_k} : S \rightarrow R^k, k = 1, \dots K$. R^k is the range of possible values for the explanatory functions.
- A dependent function $f_Y : S \rightarrow R^Y$
- A family $\mathcal{F}$ of learning model functions mapping $R^1 \times \dots R^K \rightarrow R^Y$.

Find : A function $\hat{f}^Y \in \mathcal{F}$.

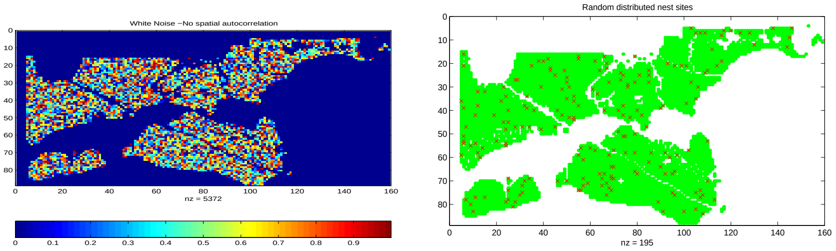
Objective : maximize similarity$(map_{s_i \in S}(\hat{f}^Y(f_{X_1}, \dots, f_{X_K})), map(f_Y(s_i)))$
$= (1-\alpha)\,\text{classification\_accuracy}(\hat{f}^Y, f_Y) +$
$(\alpha\,)\text{spatial\_accuracy}((\hat{f}^Y, f_Y)$

Constraints :

1. Geographic Space S is a multi-dimensional Euclidean Space [1].
2. The values of the explanatory functions, the f_{X_k}'s and the response function f_Y may not be independent with respect to those of nearby spatial sites, i.e. spatial autocorrelation exists.
3. The domain R^k of the explanatory functions is the one-dimensional domain of real numbers.
4. The domain of the dependent variable, $R^Y = \{0, 1\}$.

The above formulation highlights two important aspects of location prediction. It explicitly indicates that (i) the data samples may exhibit spatial autocorrelation and, (ii) an objective function i.e., a map similarity measure is a combination of classification accuracy and spatial accuracy. The *similarity* between the dependent variable f_Y and the predicted variable $\hat{f}^Y$ is a combination of the traditional accuracy" and a representation dependent "spatial classification" accuracy. The regularization term α controls the

[1] The entire surface of the Earth cannot be modeled as a Euclidean space but locally the approximation holds true.



(a) pixel property with independent identical distribution

(b) Random nest locations

Figure 3: Spatial distribution satisfying distribution assumptions of classical regression

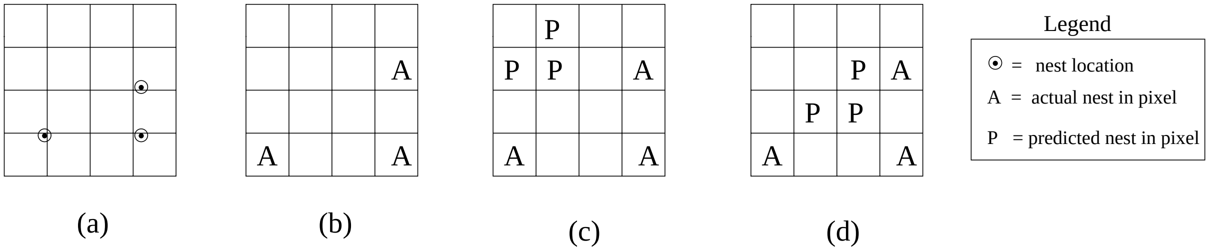


Figure 4: (a)The actual locations of nest, (b)Pixels with actual nests, (c)Location predicted by a model, (d)Location predicted by another mode. Prediction(d) is spatially more accurate than (c).

degree of importance of **spatial accuracy** and is typically domain dependent. As $\alpha \to 0$, the map similarity measure approaches the traditional classification accuracy measure. Intuitively, α captures the spatial autocorrelation dependent in the data.

The study of nesting location of red-winged black bird [25, 26] is an instance of the location prediction problem. The underlying spatial framework is the collection of 5mX5m pixels in the grid imposed on marshes. Explanatory variables, e.g. water depth, vegetation durability index, distance to open water, map pixels to real numbers. Dependent variable, i.e. nest locations, maps pixels to a binary domain. The explanatory and dependent variables exhibit spatial autocorrelation, e.g. gradual variation over space, as shown in Figure 1 and 2. Domain scientist prefer spatially accurate predictions which are closer to actual nests, i.e, $\alpha > 0$.

Finally, it is important to note that in spatial statistics the general approach for modeling spatial autocorrelation is to enlarge $\mathcal{F}$, the family of learning model functions(see Section 2.3). The PLUMS [2] approach(See Section 3) allows flexibility of incorporating spatial autocorrelation in the model, the objective function or both. Later on we will show that retaining the classical regression model as $\mathcal{F}$ but modifying the objective function leads to results which are comparable to those from spatial statistical methods but incur only a fraction of the computational costs.

1.3 Related Work and Our Contributions

Related work includes spatial statistics and spatial data mining.

Spatial Statistics: The goal of spatial statistics is to model the special properties of spatial data. The primary distinguishing property of spatial data is that neighboring data samples tend to systematically affect each other. Thus the classical assumption that data samples are generated from independent and identical distributions is not valid. Current research in Spatial Econometrics, Geo-statistics and Ecological modeling [3, 23, 13] has focused on extending classical statistical techniques in order to capture the unique characteristics inherent in spatial data. In Section 2 we will briefly review some basic spatial statistical measures and techniques.

Spatial Data Mining: Spatial data mining [9, 18, 19, 20, 29], a subfield of data mining [1, 10], is concerned with discovery of interesting and useful but implicit knowledge in spatial databases. Challenges in Spatial Data Mining arise from the following issues. *First,* classical data mining[1] deals with numbers and categories. In contrast, spatial data is more *complex* and includes extended objects such as points, lines, and polygons. *Second,* classical data mining works with explicit inputs, whereas spatial predicates (e.g. overlap) are often *implicit. Third,* classical data mining treats each input to be independent of other inputs, whereas spatial patterns often exhibit continuity and *high autocorrelation among nearby features.* For example, population density of nearby locations are often related. In the presence of spatial data the standard approach in the data mining community is to materialize spatial relationships as attributes and rebuild the model with these "new" spatial attributes [20, 19].

Our contributions: In this paper we will propose a new framework for spatial data mining. This framework con-

[2] An interesting piece of trivia is that there is actually a PLUM bird island just off the coast of Boston, Massachusetts.

sists of a combination of statistical model, a map similarity measure along with a search algorithm and a discretization of the parameter space. We will show that the characteristic property of spatial data, namely, spatial autocorrelation, can be incorporated in the statistical model or the objective function. We will also conduct experiments on the "bird-nesting" data to compare our approach with spatial statistical techniques. The rest of the paper is as follows. In Section 2 we will briefly review some important spatial statistical concepts. In Section 3 we will propose PLUMS, a new framework for spatial data mining. Experiments carried out to compare PLUMS and spatial statistical methods will be elaborated upon in Section 4. We will close in Section 5 with some comments and directions for future work.

2. BASIC CONCEPTS: MODELING SPATIAL DEPENDENCIES

2.1 Logistic Regression Modeling

Given an $n-$vector $\mathbf{y}$ of observations and an $n \times m$ matrix $\underline{X}$ of explanatory data, classical linear regression models the relationship between y and $\underline{X}$ as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon.$$

Here $\mathbf{X} = [1, \underline{X}]$ and $\beta = (\beta_0, \ldots, \beta_m)^t$. The standard assumption on the error vector ϵ is that each component is generated from an independent and and identical and normal distribution, i.e, $\epsilon_i = N(0, \sigma^2)$.

When the dependent variable is binary, as is the case in the "bird-nest" example, the model is transformed via the logistic function and the dependent variable is interpreted as the probability of finding a nest at a given location. Thus, $Prob(y = 1) = \frac{e^{X\beta}}{1+e^{X\beta}}$. This transformed model is referred to as **logistic** regression.

The fundamental limitation of classical regression modeling is that it assumes that the sample observations are independently generated. This may not be true in the case of spatial data. As we have shown in our example application, the explanatory and the independent variables show a moderate to high degree of spatial autocorrelation(see Figure 1). The inappropriateness of the independence assumption shows up in the residual errors, the ϵ_i's. When the samples are spatially related, the residual errors reveal a systematic variation over space, i.e., they exhibit high spatial autocorrelation. This is a clear indication that the model was unable to capture the spatial relationships existing in the data. Thus the model is a poor fit to the data. Incidentally the notion of spatial autocorrelation is similar to that of time autocorrelation in time series analysis but is more difficult to model because of the multi-dimensional nature of space. We now introduce a statistic which quantifies spatial autocorrelation.

2.2 Spatial Autocorrelation and Examples

There are many measures available for quantifying spatial autocorrelation. Each have their own strengths and weaknesses. Here we will briefly describe the Moran I measure.

In most cases the Moran's I measure (henceforth MI) ranges between -1 and +1 and thus is similar to the classical measure of correlation. Intuitively, a higher positive value is indicative of high spatial autocorrelation. This implies that like values tend to cluster together or attract each other. A low negative value is an indication that high and low values are interspersed. Thus like values are de-clustered and tend to repel each other. A value close to zero is an indication that no spatial trend (random distribution) is discernible using the given measure. The exact definition of MI is given in the Appendix.

All spatial autocorrelation measures are crucially dependent on the choice and design of the contiguity matrix W. The design of the matrix itself is predicated on determining "what constitutes a neighborhood of influence?" Two common choices are the four and the eight neighborhood. Thus given a lattice structure and a point S in the lattice, a four-neighborhood assumes that S influences all cells which share an edge with S. In an eight-neighborhood it is assumed that S influences all cells which either share an edge or a vertex. An eight neighborhood contiguity matrix is shown in Figure 5. The contiguity matrix of the uneven lattice(left) is shown on the right hand side. The contiguity matrix plays a crucial role in the spatial extension of the regression model.

2.3 Predicting Locations Using Spatial Statistics

We now show how spatial dependencies are modeled in the framework of regression analysis. This may serve as a template for modeling spatial dependencies in other data mining techniques. In spatial regression the spatial dependencies of the error term or the dependent variable are directly modeled in the regression equation [3]. Assume that the dependent values y_i' are related to each other, i.e. $y_i = f(y_j)\ i \neq j$. Then the regression equation can be modified as

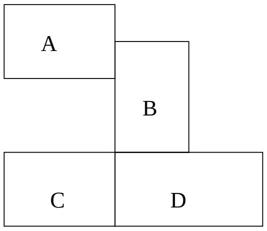
$$\mathbf{y} = \rho W\mathbf{y} + \mathbf{X}\beta + \epsilon.$$

Here W is the neighborhood relationship contiguity matrix and ρ is a parameter that reflects the strength of spatial dependencies between the elements of the dependent variable. After having introduced the correction term $\rho W\mathbf{y}$, the components of the residual error vector ϵ are now assumed to be generated from independent and identical standard normal distributions.

We will refer to this equation as the **Spatial Autoregressive Model(SAM)**. Notice when $\rho = 0$, this equation collapses to the classical regression model. The benefits of modeling spatial autocorrelation are many: (1) The residual error will have much lower spatial autocorrelation, i.e., systematic variation. With proper choice of W, the residual error should, at least theoretically, have no systematic variation. (2) If the spatial autocorrelation coefficient is statistically significant then it will quantify the presence of spatial autocorrelation. It will indicate the extent to which variations in the dependent variable (y) are explained by the average of neighboring observation values. (3) Finally, the model will have a better fit, i.e., higher R-squared statistic(See the Appendix for a dramatic example).

As in the case of classical regression, the SAM equation has to be transformed via the logistic function for binary dependent variables. The estimates of ρ and β can be derived using maximum likelihood theory or Bayesian statistics. We have carried out preliminary experiments using the spatial econometrics matlab package [3] which implements a Bayesian approach using sampling based Markov Chain

[3] We would like to thank James Lesage(http://www.econ.utoledo.edu/~lesage) for making the matlab toolbox available on the web.



(a) Map

$$\begin{array}{c|cccc|} & A & B & C & D \\ A & 0 & 1 & 0 & 0 \\ B & 1 & 0 & 1 & 1 \\ C & 0 & 1 & 0 & 1 \\ D & 0 & 1 & 1 & 0 \end{array}$$

b) Contiguity Matrix W

Figure 5: A spatial neighborhood and its contiguity matrix

Monte Carlo(MCMC) methods [23]. The general approach of MCMC methods is that when the joint-probability distribution is too complicated to be computed analytically, then a sufficiently large number of samples from the conditional probability distributions can be used to estimate the *statistics* of the full joint probability distribution. While this approach is very flexible and the workhorse of Bayesian statistics, it is a computationally expensive process with slow convergence properties. Furthermore, and at least for non-statisticians, it is very difficult to decide what "priors" to choose and what are the appropriate analytic expressions for the conditional probability distributions.

3. PREDICTING LOCATIONS USING MAP SIMILARITY(PLUMS)

Recall that we proposed a general problem definition for the Location Prediction problem, with the objective of maximizing "map similarity", which combines spatial accuracy and classification accuracy. In this section, we propose the PLUMS framework for spatial data mining.

3.1 Proposed Approach: Predicting Locations Using Map Similarity(PLUMS)

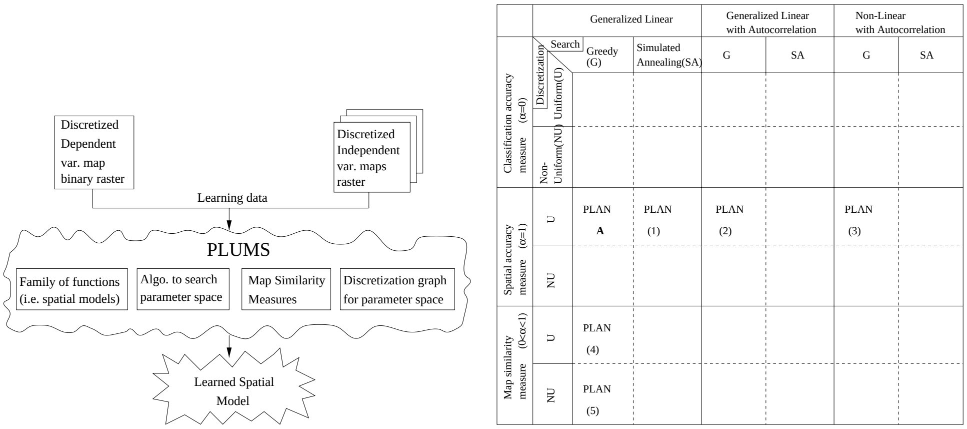
Predicting Locations Using Map Similarity(PLUMS) is the proposed supervised learning approach. Figure 6(a) shows the context and components of PLUMS. It takes a set of maps for explanatory variables and a map for the dependent variable. The maps must use a common spatial framework, i.e. common geographic space and common discretization, and produces a "learned spatial model" to predict the dependent variable using explanatory variables. PLUMS has four basic components, namely, a map similarity measure, a family of parametric functions representing spatial models, a discretization of parameter space, and a search algorithm. PLUMS uses the search algorithm to explore the parameter space to find the parameter value tuple which maximize the given map similarity measure. Each parameter value tuple specifies a function from the given family as a candidate spatial model.

A simple map similarity measure focusing on spatial accuracy for nest-location maps(or point sets in general) is the average distance from an actual nest site to the closest predicted nest-site. Other spatial accuracy and map similarity measures can be defined using nearest neighbor index [7], principal component analysis of a pair of raster maps [31] etc.

```
parameter-value-set find-A-local-maxima(parameter-value-set PVS, discretization-of-parameter-space SF,
        map-similarity-measure-function MSM, learning-map-set LMS) {
    parameter-value-set best-neighbor, a-neighbor;
    real best-improvement=1, an-improvement;
    while(best-improvement > 0) do {
        best-neighbor = PVS.get-a-neighbor(SF);
        best-improvement = MSM(best-neighbor,LMS) - MSM(PVS,LMS);
        foreach a-neighbor in PVS.get-all-neighbors(SF) do {
            an-improvement = MSM(a-neighbor,LMS) - MSM(PVS,LMS);
            if(an-improvement > best-improvement) {
                best-neighbor = a-neighbor; best-improvement = an-improvement;
            }
        }
        if (best-improvement > 0) then PVS=best-neighbor;
    } /* found a local maxima in parameter space */
    return PVS;
}
```

Algorithm 1: greedy-search-algorithm

A special case of PLUMS using greedy search is described in Algorithm 1. The function "find-A-local-maxima", takes a seed value-tuple of parameters, a discretization of parameter space, a map-similarity function and a learning data set consisting of maps of explanatory and dependent variables. It evaluates the parameter-value tuple in the immediate neighborhood of current parameter-value tuple in the given discretization. An example of a current parameter-value tuple in a red-winged-black bird application with 3 explanatory variables is (a,b,c). Its neighborhood may include the following parameter value tuples: (a+δ,b,c), (a-δ,b,c),(a,b+δ,c),(a,b-δ,c),(a,b,c+δ), (a,b,c-δ) given a uniform grid with cell-size δ discretization of parameter space. A more sophisticated discretization may use non-uniform grids. PLUMS evaluates the map similarity measure on each parameter value tuple in the neighborhood. If some of neighbors have higher values for the map similarity measure, the neighbor with highest value of map similarity measure is chosen. This process is repeated and it ends when no neighbor has a higher value of map similarity measure, i.e., a local maxima has been found. Clearly, this search algorithm can be improved using a variety of ideas including gradient descent [4, 11] and simulated annealing [32, 36] etc. A simple function family is the family of generalized linear models, e.g. logistic regression [22] with or without autocorrelation terms. Other interesting families include non-linear functions. In the spatial statistics literature many functions have been proposed to capture the spatial autocorrelation property. For example, Econometricians use the family of



(a) PLUMS Framework

(b) Space of Design Choice

Figure 6: (a)The framework for the location prediction process. (b)Space of Design Choice for PLUMS

spatial autoregression models [3, 23], Geo-statisticians [17] use Co-Kriging and Ecologists [16] use the Auto-Logistic models. Table 1 summarizes several special cases of PLUMS by enumerating various choices for the four components.

The design space of PLUMS is shown in Figure 6(b). Each instance of PLUMS is a point in the four dimensional conceptual space spanned by *similarity measure*, *family of functions*, *discretization of parameter space* and *external search algorithm*. For example, the PLUMS implementation labeled **A** in Figure **??** corresponds to the spatial accuracy measure(ADNP), generalized linear model(for the family of functions), a greedy search algorithm and uniform discretization.

4. EXPERIMENT DESIGN AND EVALUATION

Goals: The goals of the experiments are (1) to evaluate the effects of including the spatial autoregressive term, $\rho W\mathbf{y}$, in the logistic regression model and (2) compare the accuracy and performance of an instance of PLUMS with spatial regression models. The experimental setup is shown in Figure 7. The 1995 Darr wetland data was used as the learning set to build the classical and spatial models. The parameters of the classical logistic and spatial regression model were derived using maximum likelihood estimation and MCMC methods(Gibbs Sampling). The two models were evaluated based on their ability to predict the nest locations on the test data. Classification accuracy, which we describe next, was used to evaluate the two models. Then we compare these two models with PLUMS in terms of performance and spatial accuracy(ADNP).

Metric of Comparison for Classification accuracy: Classification accuracy achieved by classical and spatial logistic regression are compared on the test data. We use the Receiver Operating Characteristic(ROC) [8] curves to compare classification accuracy. ROC curves plot the relationship between the true positive rate(TPR) and the false positive rate(FPR). For each cut-off probability b, $TPR(b)$ measures the ratio of the number of sites where the nest is actually located and was predicted divided by the number of actual nest sites. The FPR measures the ratio of the number of sites where the nest was absent but predicted divided by the number of sites where the nests were absent. The ROC curve is the locus of the pair $(TPR(b), FPR(b))$ for each cut-off probability. The higher the curve above the straight line $TPR = FPR$ the better the accuracy of the model.

Metric of Comparison for Spatial Accuracy Spatial accuracy achieved by PLUMS, classical regression and SAM(Spatial Autoregressive Model) are compared based on ADNP(Average Distance to Nearest Prediction), which is defined as

$$ADNP(A, P) = \frac{1}{K}\sum_{k=1}^{K} d(A_k, A_k.nearest(P)).$$

Here the A_k's are the actual nest locations, P is the map layer of predicted nest locations and $A_k.nearest(P)$ denotes the nearest predicted location to A_k. K is the number of actual nest sites. The units for ADNP is the number of pixels in the experiment.

Result of Comparison between Classical and Spatial Regression (SAM) models: We use the 1995 Stubble wetland data to make comparison between the two models. The result is shown in Figure 8. Clearly, by including a spatial autocorrelation term, there is substantial and systematic improvement for all levels of cut-off probability on both the learning data(1995 Darr) and test data(1995 Stubble). However, the performance of SAM model is very slow and not scalable. The choice of contiguity matrix w is nontrivial, but very crucial to SAM model.

Result of comparison between PLUMS, Classical regression and SAM models: We carried out experiments to compare PLUMS with classical and spatial regression models. For this we also used the 1995 data acquired in the Stubble wetland. The results of our experiments are shown in Table 2. From the experiments it is clear that PLUMS(A) achieves similar spatial accuracy on test datasets as SAM, while it needs order of magnitude less computational time

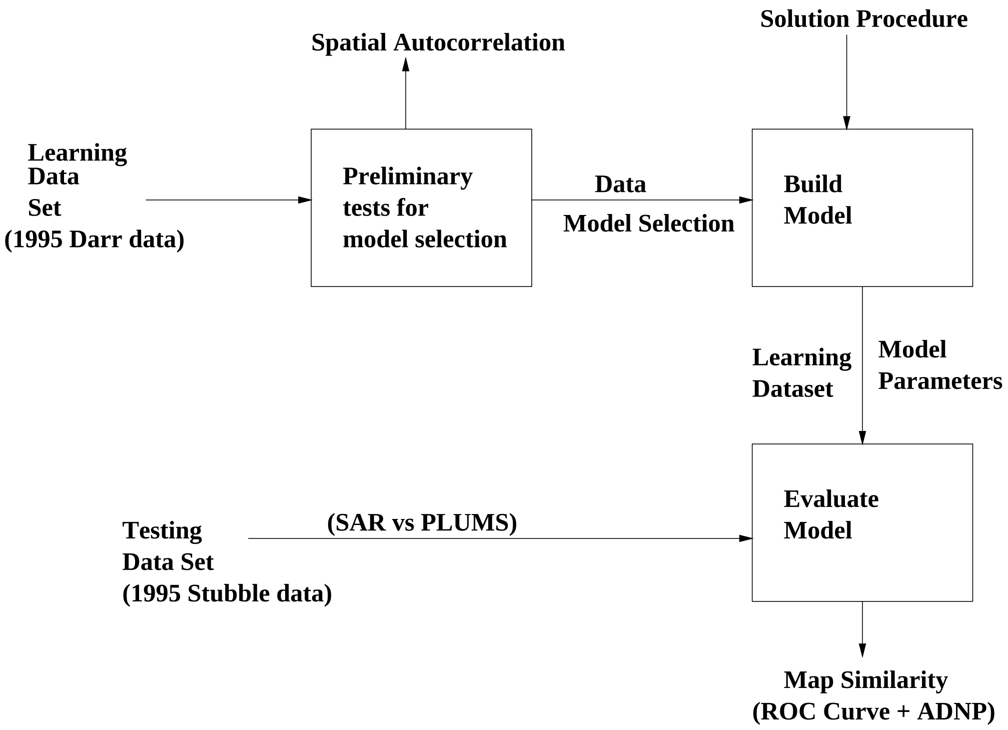

Solution Procedure
Spatial Autocorrelation
Learning Data Set (1995 Darr data)
Preliminary tests for model selection
Data
Model Selection
Build Model
Learning Dataset
Model Parameters
Evaluate Model
Testing Data Set (1995 Stubble data)
(SAR vs PLUMS)
Map Similarity (ROC Curve + ADNP)

Figure 7: Experimental Method for evaluation spatial autoregression

to learn.

The run-time for learning location prediction models for the three methods are shown in Table 2. We note that spatial regression takes two orders of magnitude more computation time relative to PLUMS using the public domain code [23] despite the sparse matrix techniques [27] used in the code.

Figures 9(a) is the ROC curves for the three models built using the Darr learning data and Figure 9(b) is the ROC curve for the Stubble test data. It is clear that by using spatial regression resulted in better predictions at all cut-off probabilities relative to PLUMS(A), a simple and naive implementation of PLUMS. Alternative smarter implementations of PLUMS enumerated in Figure **??** need to be explored to close the gap.

5. FUTURE WORK AND CONCLUSION

In this paper we have proposed PLUMS(Predicting Locations Using Map Similarity), a framework for spatial data mining. We have shown how spatial autocorrelation, the characteristic property of spatial data can be incorporated in the PLUMS framework. When compared with state-of-the-art spatial statistics method in predicting bird-nest locations, PLUMS achieved comparable spatial accuracy while incurring only a fraction of the cost. Furthermore, PLUMS provides a template for specializing other data mining techniques for spatial data.

Our future plan is to bring in other data mining techniques, including clustering and association rules, within the PLUMS framework. We also plan to investigate other search algorithms , new map-similarity measures and non-uniform parameter spaces and determine their dominance zones.

[1] 10,000 draws for Gibbs sampling, 1000 burn-outs

6. ADDITIONAL AUTHORS

Additional authors: Uygar Ozesmi (Department of Environmental Sciences, Ericyes University, Kayseri, Turkey, email: `uygar.ozesmi-1@tc.umn.edu`)

7. REFERENCES

[1] R. Agrawal. Tutorial on database mining. In *Thirteenth ACM Symposium on Principles of Databases Systems*, pages 75–76, Minneapolis, MN, 1994.

[2] P.S. Albert and L.M. McShane. A generalized Estimating Equations Approach for Spatially Correlated Binary Data: Applications to the Analysis of Neuroimaging Data. *Biometrics (Publisher: Washington, Biometric Society, etc.)*, 51:627–638, 1995.

[3] L Anselin. *Spatial Econometrics: methods and models.* Kluwer, Dordrecht, Netherlands, 1988.

[4] Vladimir Cherkassky and Filip Mulier. *Learning From Data Concepts, Theory, and Methods.* John Wiley & SONS Inc., 1998.

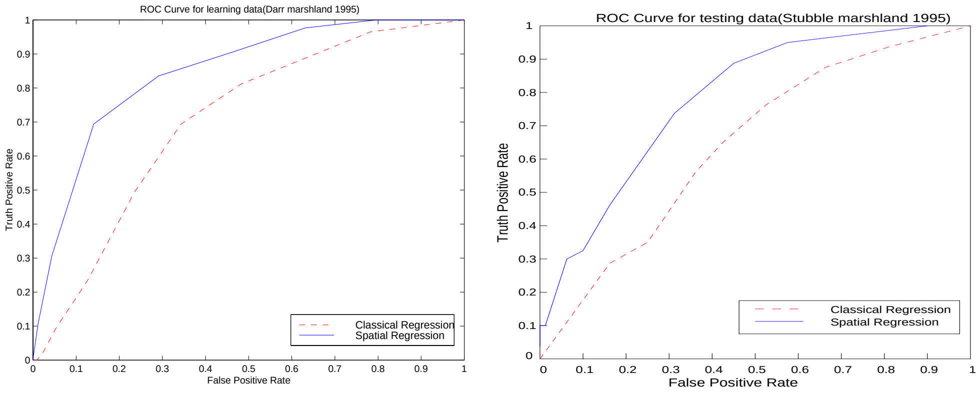
[5] P. Could. *The Geographer at Work.* Routledge and Kegan Paul, London, 1985.

[6] N.A. Cressie. *Statistics for Spatial Data (Revised Edition).* Wiley, New York, 1993.

[7] P.J. Diggle. *Statistical analysis of spatial point patterns.* Academic Press, 1993.

[8] J.P. Egan. *Signal Detection Theory and ROC analysis.* Academic Press, New York, 1975.

[9] M. Ester, H-P Kriegel, and J. Sander. Knowledge discovery in spatial databases. In *Advances in Artificial Intelligence, 23rd Annual German*



(a) Learning Data

(b) Test Data

Figure 8: (a) Comparison of the logistic and logistic with spatial autocorrelation on the 1995 Darr wetland learning data. (b) Comparison of the two models on the 1995 Stubble wetland testing data.

Conference on Artificial Intelligence, pages 61–74, Bonn, Germany, September 1999.

[10] U. M. Fayyad. Knowledge discovery in databases: An overview. In *Inductive Logic Programming, 7th International Workshop, ILP-97, Lecture Notes in Computer Scienc*, volume 1297, pages 3–16. Springer, September 1997.

[11] B. Flury. *A First Course in Multivariate Statistics (Section 7.5: Simple Logistic Regression)*. Springer, 1997.

[12] C. Greenman. Turning a map into a cake layer of information. *New York Times*, January 20th (http://www.nytimes.com/library/tech/00/01/circuits/arctiles/20giss.html) 2000.

[13] D. Griffith. Statistical and mathematical sources of regional science theory: Map pattern analysis as an example. *Papers in Regional Science (Publisher: Springer)*, (78):21–45, 1999.

[14] R.H. Guting. An Introduction to Spatial Database Systems. *Vary Large Data Bases Journal (Publisher:Springer Verlag)*, October 1994.

[15] M. Hohn and L. Gribko A.E. Liebhold. A Geostatistical model for Forecasting the Spatial Dynamics of Defoliation caused by the Gypsy Moth, Lymantria dispar (Lepidoptera:Lymantriidae). *Environmental Entomology (Publisher: Entomological Society of America)*, 22:1066–1075, 1993.

[16] F. Huffer and H. Wu. Markov chain monte carlo for autologistic regression models with application to the distribution of plant species. *Biometrics (Publisher: Washington, Biometric Society, etc.)*, 54(3):509–535, 1998.

[17] Issaks, Edward, and Mohan Srivastava. *Applied Geostatistics*. Oxford University Press, Oxford, 1989.

[18] E. Knorr and R. Ng. Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining. *IEEE TKDE*, 8(6):884–897, 1996.

[19] K. Koperski, J. Adhikary, and J. Han. Spatial data mining: Progress and challenges. In *Workshop on Research Issues on Data Mining and Knowledge Discovery(DMKD'96)*, pages 1–10, Montreal, Canada, 1996.

[20] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Advances in Spatial Databases, Proc. of 4th International Symposium, SSD'95*, pages 47–66, Portland, Maine, USA, 1995.

[21] P. Krugman. *Development, geography, and economic theory*. MIT Press, Cambridge, MA, 1995.

[22] J. LeSage. Regression Analysis of Spatial data. *The Journal of Regional Analysis and Policy (Publisher: Mid-Continent Regional Science Association and UNL College of Business Administration)*, 27(2):83–94, 1997.

[23] J.P. LeSage. Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, (20):113–129, 1997.

[24] D. Mark. Geographical information science: Critical issues in an emerging cross-disciplinary research domain. In *NSF Workshop*, Feburary 1999.

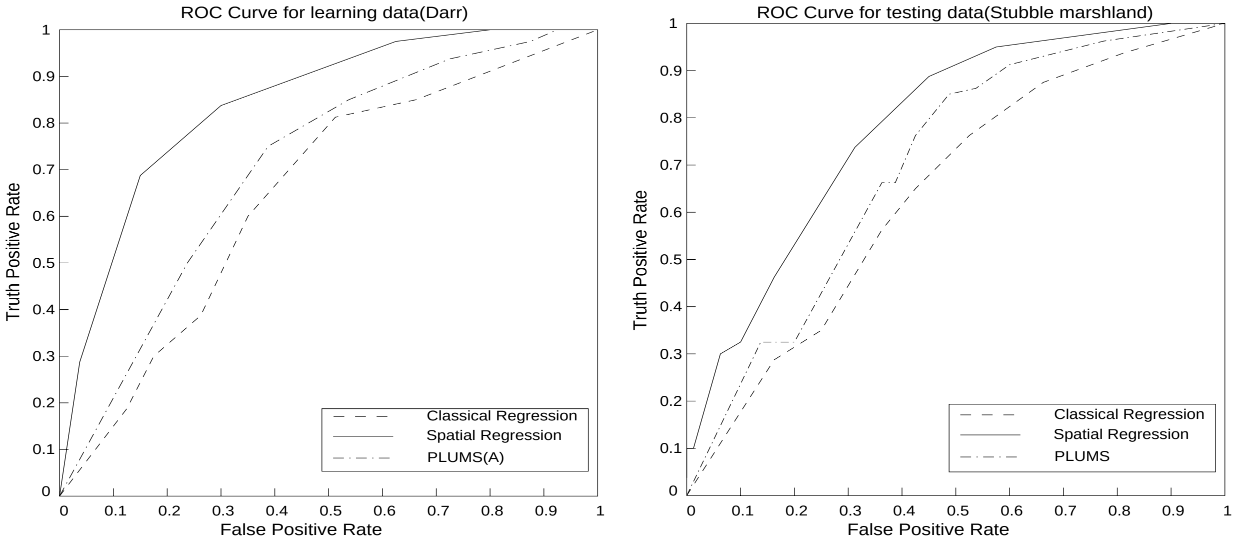
[25] S. Ozesmi and U. Ozesmi. An Artificial neural network approach to spatial habitat modeling with interspecific interaction. *Ecological Modelling (Publisher: Elsevier Science B. V.)*, (116):15–31, 1999.

[26] U. Ozesmi and W. Mitsch. A spatial habitat model for the Marsh-breeding red-winged black-bird(agelaius phoeniceus l.) In coastal lake Erie wetlands. *Ecological Modelling (Publisher: Elsevier Science B. V.)*, (101):139–152, 1997.

[27] R. Pace and R. Barry. Sparse spatial autoregressions. *Statistics and Probability Letters (Publisher: Elsevier Science)*, (33):291–297, 1997.

[28] R.J.Haining. *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge, U.K., 1989.

[29] John F. Roddick and Myra Spiliopoulou. A bibliography of temporal, spatial and spatio-temporal data mining research. *ACM Special Interest Group on*



(a) ROC curves for Learning Data

(b) ROC curves for Test Data

Figure 9: (a) Comparison of PLUMS(A) with other methods on the Darr learning data. (b) Comparison of the models on the test data.

Knowledge Discovery in Data Mining(SIGKDD) Explorations, 1999.

[30] S. Shekhar and S. Chawla. *Spatial Databases: Issues, Implementation and Trends.* (Under Contract)Prentice Hall, 2000.

[31] R. Schowengerdt. *Remote Sensing:Models and Methods for Image Processing.* Academic Press, 1997.

[32] S. Shekhar and B. Amin. Generalization by neural networks. *IEEE Trans. on Knowledge and Data Eng.*, 4(2), 1992.

[33] S. Shekhar, S. Chawla, S. Ravada, A.Fetterer, X.Liu, and C.T. Lu. Spatial databases: Accomplishments and Research Needs. *IEEE Transactions on Knowledge and Data Engineering*, 11(1), Jan-Feb 1999.

[34] S. Shekhar, T. A. Yang, and P. Hancock. An intelligent vehicle highway information management system. *Intl Jr. on Microcomputers in Civil Engineering (Publisher: Blackwell Publishers)*, 8(3), 1993.

[35] W.R. Tobler. *Cellular Geography, Philosophy in Geography.* Gale and Olsson, Eds., Dordrecht, Reidel, 1979.

[36] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 1997.

[37] M.F. Worboys. *GIS: A Computing Perspective.* Taylor and Francis, 1995.

[38] Y. Yasui and S.R. Lele. A Regression Method for Spatial Disease Rates: An Estimating Function Approach. *Journal of the American Statistical Association*, 94:21–32, 1997.

8. APPENDIX:SPATIAL AUTOCORRELATION

8.1 Moran's I measure

There are many measures available for quantifying spatial autocorrelation. Each have their own strengths and weaknesses. The two most well known measures are Moran's I and Geary's C measure. Here we will briefly describe the Moran I measure.

In most cases the Moran's I measure (henceforth MI) ranges between -1 and +1 and thus is similar to the classical measure of correlation. Intuitively, a higher positive value is indicative of high spatial autocorrelation. This implies that like values tend to cluster together or attract each other. A low negative value is an indication that high and low values are interspersed. Thus like values are de-clustered and tend to repel each other. A smooth surface will have a high spatial autocorrelation and a chessboard-like surface a high negative spatial autocorrelation. A value close to zero is an indication that no spatial trend (random distribution) is discernible using the given measure.

The formula for MI is

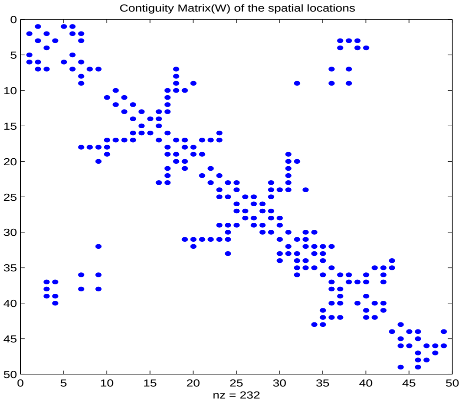
$$MI = \frac{n}{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}} \cdot \frac{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}$$

where n is the number of data points, $x_i's$ are the data values, $\bar{x}$ is the mean and W is the design or contiguity matrix. All spatial autocorrelation measures are crucially dependent on the choice and design of the contiguity matrix W.

8.2 Summary of different methods

We summarized all the methods that have been used to build the bird habitat model in Table 3.

8.3 Example of including the spatial autoregressive term



(a) The contiguity matrix of locations

	R-square	Moran I (residual)
Ordinary Regression	0.5521	0.23
Spatial Auto Regression	0.6518	0.04

(b) R-square and Moran I of residual

Figure 10: (a)Crime data set in 49 neighborhoods in Columbus Ohio [3],where number of crime incidents is dependent variable, and the explanatory variables include mean income and mean house value.(b)Coefficient of determination(R^2) and Moran I results of this data set via ordinary regression model and Spatial Auto Regression(SAM) model. Clearly the SAM model provides a better fit than the classical regression model.

PLUMS Component Choices	
Component	Choices
Map similarity	avg. distance to nearest prediction from actual, nearest neighbor index, ...
Search algorithm	greedy, gradient descent, simulated annealing, ...
Function family	generalized linear(GL) (logit, probit), non-linear, GL with autocorrelation
Discretization of parameter space	Uniform, non-uniform, multi-resolution, ...

Table 1: PLUMS Component Choices

Data set		PLUMS	Classical	SAM
Learning	spatial accuracy	16.90	47.16	13.96
Testing	spatial accuracy	19.19	41.43	19.30
Learning	Run-time(Seconds)	80	10	19420 [1]

Table 2: Learning time and spatial accuracies for learning and test data set

Method Name	Model Type	Spatial AC	Dependent Var. Type	Accuracy Measure	Solution Procedure
Linear Regression	Linear	No	Numeric	Total Square Error(TSE)	Closed Form
Neural Networks	NonLinear	No	Numeric/ Categorical	TSE	Gradient Descent Back-Propagation
Probit	Gen. Linear	No	Binary	TPR/FPR	Gradient Descent
Logit	Gen. Linear	No	Binary	TPR/FPR	Gradient Descent
SAM + Probit	Gen. Linear	Yes	Binary	TPR/FPR	ML/EM/Gibbs

Table 3: Different methods and their characteristics that have been used for building the bird habitat model.

Learning Prosodic Patterns for Mandarin Speech Synthesis

Yiqiang Chen Wen Gao
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China 100080
Email: yqchen@ict.ac.cn

Tingshao Zhu
Dept. of Computing Science
University of Alberta
Edmonton, Canada T6G 2H1
Email: tszhu@cs.ualberta.ca

ABSTRACT

Higher quality synthesized speech is required for widespread use of text-to-speech (TTS) technology, and prosodic pattern is the key feature that makes synthetic speech sound unnatural and monotonous, which mainly describes the variation of pitch. The rules that are now being used in most Chinese TTS systems are constructed by experts, qualitatively and with low precision. In this paper, we propose a combination of clustering and machine learning techniques to extract prosodic patterns from actual large mandarin speech database to improve the naturalness and intelligibility of synthesized speech. Typical prosody models are found by clustering analysis, some machine learning techniques including Rough Set, ANN and Decision tree are trained respectively for fundamental frequency and energy contours, which can be directly used in a pitch-synchronous-overlap-add-based (PSOLA-based) TTS system. The experimental results showed that synthesized prosodic features quite resembled their original counterparts for most syllables.

Keywords:

TTS, Pitch, Mandarin Speech Synthesis, Data Mining

1.INTRODUCTION

Text-To-Speech (TTS) technology is currently useful only in a limited number of applications because the quality of synthetic speech is not as good as people expected. Prosody, which includes the phrase and accent structure of speech, is one of important component for TTS system. In the field of speech signal process, pitch (fundamental frequency and F0) is the most mysteriously expressive of the prosodic phenomena, and the variation of pitch in speech can be used to express the speaker's intention, especially in Mandarin.

Although many researchers have proposed some prosodic variation patterns, the patterns are described qualitatively, or with great limitation. Wu [1][2] found that in Mandarin when syllables are combined, their tones changed to be continuous, and he gave some qualitative rules. Chu [3] uses some pitch patterns in Chinese speech synthesis system, including 14 kinds of pitch shapes of isolate syllables and 22 kinds of shapes of two-word phrases, but obviously only these shapes can't describe the variations of Mandarin to a large extent.

In recent years, some researchers intend to learn the variation patterns base on large speech database. Lee S. and Oh Y-H [4] describes the tree-based modeling of prosodic phrasing, pause duration for Korean TTS system. Ostendorf [5] describes a dynamical system model for generating fundamental frequency, which allows automatic estimation of parameter from labeled large speech database. Hu [6] proposed a template-driven generation of prosodic information for Chinese text-to-speech conversion. Ross KN. Chen [7] proposed a new RNN-based prosodic information synthesizer for Mandarin Chinese text-to-speech. Cai [8] establish a Chinese text to speech system and a prosody learning system based on NN.

Although these methods have made advances, they are still far away from reaching the goal of generating proper prosodic information for synthesizing speech with high naturalness. The drawback lies in their inability to elegantly invoke higher-level linguistic features in exploring the prosodic phrase structure of Mandarin speech. This motivates us to use a combination of clustering and ML techniques to learn prosodic variation patterns to improve the naturalness and intelligibility.

This paper is organized as follows. Sections 2 introduce TTS, Sections 3 discuss the clustering of prosodic pattern. The Training process and the prediction are described in Section 4, and some conclusions of our on-going research will be given in Section 5.

2. TEXT TO SPEECH

The three main steps in the TTS process are illustrated in Figure.1.

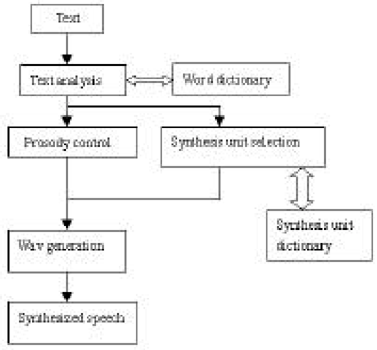


Figure 1 TTS system

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

Speech synthesized from text suffers from a major shortcoming - it sounds synthetic. It is found that synthetic speech significantly more difficult to comprehend than natural speech. There is broad consensus that poor prosody is the key feature that makes synthetic speech sound unnatural and monotonous. Prosody conveys the relative importance of words by making these words stand out acoustically (prominence), and helps listeners "chunk" the message by inserting acoustic punctuation marks (phrasing) as pauses, decreases in rate, and characteristic pitch movements. Speech with inappropriate prosody prevents listeners from understanding the spoken message. It will also undermine the credibility and effectiveness of the animated character that produces it.

Prosody in TTS systems involves three levels. First, text analysis components compute phrase boundary locations and prominence ("prosodic structure"). Second, acoustic prosodic components compute phoneme duration, fundamental frequency (or pitch) contours, and (optionally) contours for additional acoustic parameters such as amplitude or spectral tilt. Finally, signal processing components compute a digital speech wave that expresses the phoneme sequence having the desired timing and pitch contour.

The main goal of our work is to improve the prosodic structure of speech generated from text. The TTS system must analyze text and use this information to generate both a symbolic structure (e.g., locations and type of phrase boundaries and prominence of important words) and an acoustic waveform that expresses the desired meaning of each utterance.

This general problem of lack of an appropriate prosodic model was encountered in Mandarin TTS prosodic information synthesis. Mandarin Chinese is a tonal language. Each character is pronounced as a syllable. Only about 1300 phonetically distinguishable syllables comprise the set of all legal combination of 411 base-syllables and five tones. Each base-syllable is composed of an optional consonant initial and a vowel final. The word, which is the smallest syntactically meaningful unit, consists of one to several syllables. Because syllables are the basic pronunciation units in Mandarin speech, they are commonly chosen as the basic synthesis units in Mandarin TTS systems. Accordingly, the prosodic information that must be synthesized includes syllable pitch (F0) contour, syllable energy contour, syllable initial and final duration, as well as intersyllable pause duration. Among them, syllable pitch contour has the most important effect on naturalness of synthetic speech. So pitch contour synthesis is of primary concern in Mandarin TTS. The tone of syllable is mainly determined by its F0 contour. However, the pronunciation is usually highly context-dependent, It would seem that syllable F0 contour are various modification in continuous speech. Therefore, the F0 generation is not a trivial task. There are many methods have been proposed in the past, including rule-based methods [9][10], statistical model-based methods [4][6], and MLP-based methods [7]. In contrast to other researchers who employ a single technique, we present a combination of clustering and ML techniques, applied for the identification of relationship(s) between sound characteristics of speech pattern and linguistic features of the corresponding text.

We assume that the F0 contour in the continuos speech data are not variety randomly but can be obtained through modifying some classic F0 model with duration and mean. These classic F0 models can be obtained from the preprocessed actual F0 contours.

3.DATA PROCESSING

3.1 Speech Database

The speech corpora and the labeled speech corpora will be used for our acoustic prosody and signal processing efforts. The Speech Database that we are using is a Chinese speech synthesis database called CoSS-1. CoSS-1 includes the pronunciation of all isolate syllables, the 2-4 word phrases and some sentences. The number of isolate syllables with tone is 1268, and that of word phrase is 1640 and sentence is 210.

CoSS-1 records the speech wave and laryngograph synchronously. The sampling rate is 16000/s, and each sample is stored in two bytes. The sentence in the database covers almost the whole tone collocations in Chinese pronunciation.

The preprocessing mainly deals with the data from speech database directly, which extracts pitch, wraps the duration and normalizes and smooth and zero mean the pitch values to meet the requirement of cluster algorithm.

3.2 Pitch Extraction

To learn the patterns, the pitch should be calculated at first. There are many methods to extract pitch from speech wave, but the precision is very low [11]. Since we want to learn the patterns and use them to generate pitch after training, the accuracy is very important.

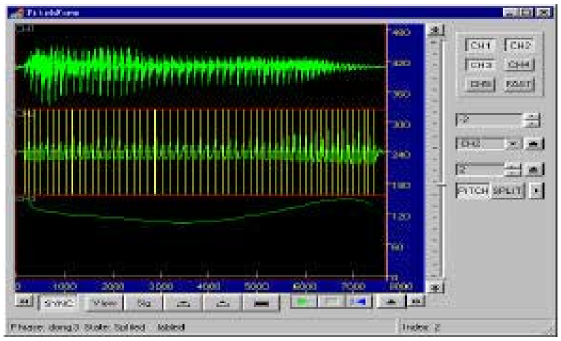


Figure. 2. A snapshot of Pitcher.

A tool called Pitcher is implemented to extract pitch from laryngograph. It works by annotating each cycle's beginning and ending point, then calculating the pitch. Let X_i be the beginning point of one cycle and X_j be the ending point, then the pitch of this cycle should be $16000/(X_j - X_i)$. Pitcher can also be used to split phrases and play the speech data. Figure 2 gives a snapshot of Pitcher.

Pitcher can be used to split a phrase and annotate each syllable.

The results of splitting and annotating are stored in database, then the algorithm can retrieve them for training and testing. To annotate pitches, you should firstly sign the reference cycle, then the beginning and ending point of the period. Pitcher deals with cycles one by one within the period to calculate pitches.

3.3 Time Wrapping and Normalization

The length of pitches that should acts as the training examples differs from each other significantly. A new algorithm is designed to wrap the pitches, which differs from the traditional time wrapping method DTW [11](Dynamic Time Wrapping) which is widely used in speech signal process. Figure 3 gives the new algorithm.

For the speech data we used, the pitches' value domain is between $50 - 260$. In this paper, the following equation is used to normalize the pitch value.

$$\textit{Normalized} = (\textit{Pitch} - \textit{min}) / (\textit{max} - \textit{min}) \qquad (1)$$

Where *max* is the maximum of all pitches' value and *min* is the minimum. *Pitch* stores the pitch to be calculated and *Normalized* is the normalized value.

3.4 Smoothing and filter: There are many filter algorithms in signal processing [11]. In this paper, the window design filter is presented to eliminating the large fluctuant data. Assume a sequence of observations $X = [x_0, x_1, \ldots, x_n]$, the windows width is m. we can gain another sequence states $Y = [y_0, y_1, \ldots, y_{n-m+1}]$ with the following formulation (2):

$$y_i = \sum_{j=i}^{i+m} x_j \bigg/ m \qquad (2)$$

3.5 Zero-mean: To avoid the effect of F0 energy, zero-mean method is proposed for each pitch fundamental frequency. The zero-mean method represents a sequence of observation $X = [x_0, x_1, \ldots, x_n]$ in terms of a sequence of states $Y = [y_0, y_1, \ldots, y_n]$ with the following equation (3):

$$y_i = x_i - \sum_{j=1}^{n} x_j \bigg/ n \qquad (3)$$

3.6Clustering method: The quantification of the similarity notion is important for clustering, and in our clustering, we use the following one:

$$Dist(X,Y) = \sqrt{\sum_{i=1}^{n}((x_i - \bar{x}) - (y_i - \bar{y}))^2} + \left|\bar{x} - \bar{y}\right| \qquad (4)$$

Where $X = (x_1, x_2, \ldots, x_n)$, $Y = (y_1, y_2, \ldots, y_n)$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. This method can calculate the similar more precise than the Euclidean distance method.

The ISODATA [12](Iterative Self-organizing Data) algorithm is chosen for our clustering, the main procedures are the following:

3.6.1Present the clustering parameters

C: number of expected classes; *MaxIterate*: the Max times for adjusting; ***MinSamples:*** the Min number of objects in one class; ***I****:* combination parameter; ***J:*** partition parameter.

3.6.2Choose initial cluster centers

Calculate the mean $\overline{x_i}$ and variance $s_i (i = 1,2,3,\ldots,n)$

Arbitrarily choose 2n+1 objects as the initial clustering centers ： $\overline{X} = \left(\overline{x_1}, \overline{x_2}, \overline{x_3}, \ldots, \overline{x_n}\right)$ and $\left(\overline{x_1}, \overline{x_2}, \ldots, \overline{x_i} \pm s_i, \ldots, \overline{x_n}\right), i = 1,2,\ldots,n,$

3.6.3Classify and adjust the objects based on K-means algorithm

If there is no re-distribution of the objects in any cluster happens or the max times ***MaxIterate*** for adjusting is achieved, the process terminates, otherwise, the adjusting will be repeated as following:

Deleting： if the number of objects in some class is less than ***MinSample***，then the class should be deleted, at the same time, the objects in that class will not be reused.

Partition： assume that m classes are generated after several times overlapping, and there must be one character in n of each class holding the Max variance. Let

$$S_{threshold} = \overline{s_{\max}} \bullet \frac{J}{1 + e^{-(m-C)}}$$

Where $\overline{s_{\max}}$ reprsent the mean of max variance of all the classes.

To each class, the max variance S_I of every character can be calculated，if $S_I > S_{threshold}$，then this class should be partitioned as following: $\left(\overline{x_1}, \overline{x_2}, \ldots, \overline{x_i} \pm s_i, \ldots, \overline{x_n}\right), i = 1,2,\ldots,n$

Combination: assume that m classes are generated after several times overlapping, and the min distance value between every two centers can be obtained. Let

$$D_{threshold} = \overline{D_{\min}} \bullet \frac{I}{1 + e^{-(m-C)}}$$

Where $\overline{D_{\min}}$ reprsent the mean of min distance of all the classes.

To every two classes, if the distance between their centers is less than $D_{threshold}$, then they are combined, and the center of new class should be recalculated. After clustering, there are 18 F0 pattern are classified, Figure 3 shows them:

After clustering, there are 18 F0 pattern are classified, Figure 3 shows them:

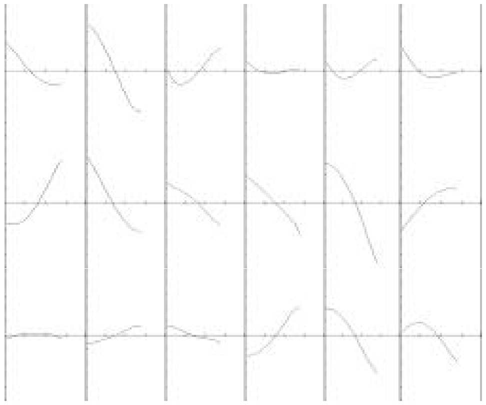


Figure 3: F0 patterns after clustering analysis.

The original pitch from sentences is discreted with extracted classic F0 models, and at the same time the original length and mean should be kept for future learning. The original prosody pattern is preprocessed into three parts: zero-mean F0 pattern, duration, and mean.

4. PROSODIC LEARNING

4.1 Linguistic Features

Our aim is to explore the relationship between the prosodic pattern of Mandarin speech and the linguistic features of the input text to simulate human's prosody pronunciation mechanism.

The Chinese Dictionary that we are using includes the spell, vowel, constant, tone, part of speech (POS), and some word syntax and semantic. From this dictionary and the existence of a text processing model, the lexical information (phonemic representations and lexical stress) and symbolic prosodic markers can be obtained. In this paper, after parsing, the linguistic features including the following:

The number of pitch in word
The sequence number serial number of pitch in word
Word class and POS
Is substantive or function word?
Is prediction or noun word?
The vowel, constant and tone of current pitch
The vowel, constant and tone of prior and post pitch

4.2 Feature Selection

Before training, the Rough set [13][14] is proposed to find the minimum attribute set. The rough set theory is based on indiscernibility relation. Suppose four finite, non empty sets R, A, V and f, where R is the universe, and A is a set of attributes, V is the value set of each attribute and f is a function map $f(U, A) \to V$. The indiscernible relation I is associated with every subset of attributes $P \in A$ and defines as: $I(P)=\{(r_i,r_j)\in U\times U{:}f(r_i,attr)=f(r_j,attr),\forall attr\in P\}$
Where $f\ (r_i,\ attr)$ is the value of attribute *attr* in object r_i. If $(r_i, r_j) \in I(P)$, then r_i and r_j are P-indiscernible.

Rough set can remove unnecessary attributes from the set A by considering redundancies and dependencies between attributes. Let P be a subset of A, and the initial P is the set A. If $I(P) \neq I(P-\{attr\})$, then we say that the *attr* can be moved from the set A. Thus the main features are selected by Rough set. The main features are used as input of ANN and the condition attributes of decision tree. We construct three ANN or decision trees respectively, they can predict the F0 model, the F0 mean and the F0 duration.

4.3 Training and Prediction

There are many kinds of neural networks, which can be used for learning. We intend to learn the mapping between the linguistic features and the F0 mean value. Since backpropagation network has implicit input layer and output layer [15], and it can also give very good result, thus it is chosen to be trained in our system.

In order to generate training and testing data, all the sentences are split firstly, calculating the pitches, wrapping the pitches to the same length, normalizing pitches' value and discrete the pitch. Then the pitch class, the linguistic parameters obtained by text parsing are labeled for neural net training and testing. For the network learning the F0 model, its input layer consists of 28 units, and the hidden layer consists of 34 units. There is only one unit in output layer. The input layer's units are described as Table 1.

Table 1: the definition of input layer

Number of units	Description
4	Length of word(1-4)
4	Pitch's location in word
5	Part of speech
6	Vowel/consonant
3	Tone of pitch
3	Tone of previous pitch
3	Tone of next pitch

The training of the F0 length and the F0 mean are as same as the training of F0 model. Then Three different neural networks are constructed to predict the F0 model, the F0 mean and the F0 length respectively.

Using the same linguistic parameters as condition attributes and the F0 model, the F0 mean and the F0 lengths as decision attribute, three different decision trees are constructed respectively. The C4.5 system [16], which has many advantages in building decision tree, is selected for our construction.

After training, the NN and the decision tree can be used to predict and generate the fundamental frequency. We compared two ways and results shows in Table 2.

OL means original Length of pitch
LPDT means Length predicted by decision tree
LPNN means Length predicted by ANN
OM means original mean of pitch
MPDT means mean predicted by decision tree
MPNN means mean predicted by ANN

Table 2: some predict result of NN and Decision tree

OL	LPDT	LPNN	OM	MPDT	MPNN
(zhi2) 24	17.5	27.8	261.4	185	246.4
(pai2) 45	37.5	29.1	234.4	175	239.2
(shi4) 32	27.5	36.2	251	265	235
(ran2)38	7.5	31.5	197	165	167.5
(qi4) 26	27.5	24.2	275.5	285	248.5
(re4) 35	32.5	24.1	277.3	255	255.3
(shui3)5	17.5	8.9	173	245	163.8
(qi4)26	27.5	24.4	211.6	185	206.7
(yan2)31	22.5	21.8	196	215	198
(jin4) 16	27.5	22.2	252.7	235	249.8
(an1) 24	17.5	32.2	249.5	315	239.3
(zhuang4)36	17.5	41.4	275.9	275	262.4
(zai4)28	42.5	30.1	233.3	305	277.3
(yu4)43	27.5	34	266	285	257.2
(shi4)6	37.5	25.5	173.7	185	189
(nei4)30	22.5	23.2	238.9	305	238.6
(shi3)12	27.5	14.7	140.9	185	144.2
(yong4)10	37.5	20.3	168.6	245	195.1

The experiment was taken on our labeled large speech database, the variation of the data between the original data and the predicted one was calculated as follow:

$$mean = \frac{\sum_{i=1}^{n} (prediction - original)}{n}$$

$$variation1 = \frac{\sum_{i=1}^{n} (prediction - original)^2}{n}$$

$$variation2 = \frac{\sum_{i=1}^{n} (prediction - original - mean)^2}{n}$$

$$unsimilarity = \frac{\sqrt{variation1} + \sqrt{variation2} + mean}{3}$$

Table3 shows the variations of predicted result:

Table 3: The variations of predicted result

variation	F0 model	F0 duration	F0 mean
Decision tree	2.6	5.9	33.8
ANN	1.8	3.3	11.7

The results show that the decision tree is not good at predicting the continuous attribute while ANN can do it well. Thus the decision tree is only be used to predict the F0 model while the BP was used to predict the F0 mean and F0 length. The linguistic features are reselected focus on each prediction respectively, and Table 3,4 shows the summary of our features selection:

Table4: Attributes of Decision tree for F0 model

Condition Attributes	Number of pitches in word (len)
	Series number of pitches(wordno)
	Part of Speech (type)
	Substantive or function word (xs)
	prediction or noun word(tw)
	Current tone, pretone, posttone
Decision	F0 model

Some rules for F0 model prediction：
type = 1 and tone = 2 and pretone = 3 and posttone = 2-> class 4
len = 3 and type = 1 and tone = 2 and posttone = 5 -> class 4
type = 4 and tone = 2 and pretone = 5 and posttone = 4 -> class 13
type = 6 and tone = 2 and pretone = 4 -> class 14

Table 5: the definition of input/output layer

Input layer Definition	Number of pitches in word (len)
	Series number of pitches(wordno)
	Part of Speech (type)
	Substantive or function word (xs)
	Prediction or noun word(tw)
	Consonant and tone (pycon, tone)
	pretone, posttone
Output of NN	Length of F0(discreted)

4.4Experiment

The F0 models are predicted by decision tree while the F0 duration and mean are predicted by ANN. The F0 model can be modified in accordance with F0 duration and F0 mean. The modification is based on a simple interpolation method. The experiment was taken on the CoSS-1 speech database, some experiment results are showed in figure 4.

zhi pai shi re shui qi yan jin an zhuang zai yu shi nei shi yong

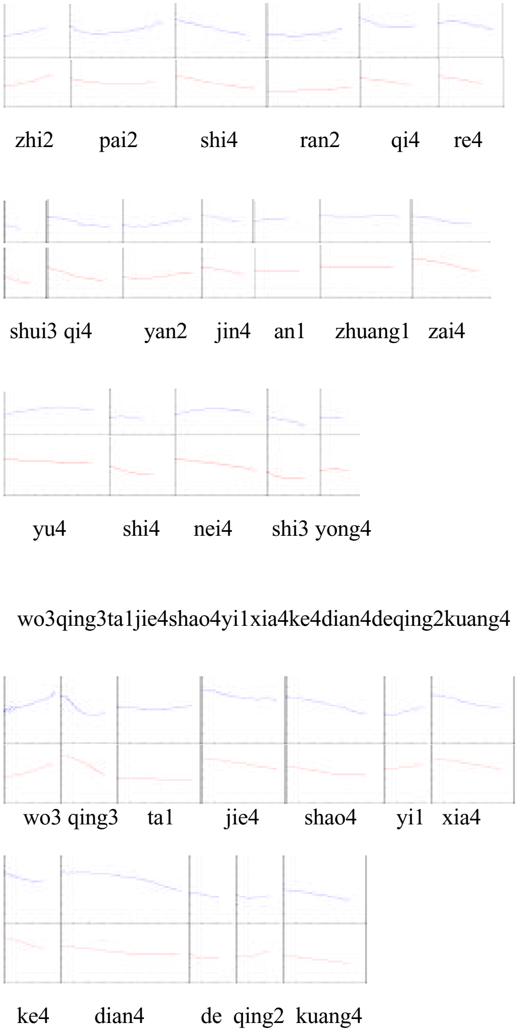


Figure 4 above is original pitch, lower is synthesis one

5.CONCLUSION

In this paper, we propose a combination of clustering and machine learning techniques to extract prosodic patterns from actual large mandarin speech database to improve the naturalness and intelligibility of synthesized speech. Typical prosody models are found by clustering analysis, some ML techniques including Rough Set, ANN and Decision tree are trained respectively for fundamental frequency and energy contours, which can be directly used in a pitch-synchronous-overlap-add-based (PSOLA-based) TTS system. The prediction result of ANN and Decision Tree can be combined to generate the fundamental frequency and energy contours. So, the effects of high-level linguistic features on prosodic information generation are well handled. The experimental results showed that synthesized prosodic features quite resembled their original counterparts for most syllables.

6.REFERENCES

[1] Zongji Wu, "The tone variation in mandarin", *Chinese grammar*. No. 6, pp.439-449, 1982.

[2] Zongji Wu, "The design of prosodic rule for improving the naturalness of the Marian TTS", *The research on Chinese language and words*, *Tsinghua University press,* pp.355-365, 1996.

[3] Min Chu, "Research on Chinese TTS system with high intelligibility and naturalness", *Ph.D thesis*, Institute of Acoustics, Academia Sinica, 1995.

[4] Lee S, Oh Y-H, "Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS system", *Speech Communication*, Vol.28, No.4, pp.283-300, 1999.

[5] Ross KN, Ostendorf M, "A dynamical system model for generating fundamental frequency for speech synthesis", IEEE *Transaction on speech and audio processing*, Vol. 7, No. 3, pp.295-309, 1999.

[6] Chung-Hsien Hu, Jan-Hung Chen, "Template-driven generation of prosodic information for Chinese concatenate synthesis", IEEE *International Conference on Acoustics, Speech, and Signal Processing*, Vol.1, pp.65-68, 1999.

[7] Sin-Horng Chen, Shaw-Hwa Huang, Yih-Ru Wang, "An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech", IEEE *Transaction on speech and audio processing*, Vol. 6, No. 3, pp.226-239, 1998.

[8] Cai Lianhong, Zhang Wei, Hu Qiwei, "Prosody learning and simulation for Chinese text to speech system", *Qinghua Daxue Xuebao/Journal of Tsinghua University*, Vol.38, No.S1, pp.92-95, 1998.

[9] L.S.Lee, C.Y. Tseng, and M. Ouh-Young, "The synthesis rules in a chinese text-to-speech system", *IEEE trans. Acoust., speech, signal Processing,* Vol. 37, pp. 1309-1320, 1989.

[10] C.H. Wu, C. H. Chen, and S. C. Juang, "An CELP-based prosodic information modification and generation of Mandarin text-to-speech", *in proc. ROCLING VIII*, pp. 233-251, 1995.

[11] L.Rabiner and B.Juang. "Fundamentals of Speech Recognition." *TsingHua University Publishing Company*. 1999.

[12] Bian Zhaoqi and Zhang Xuegong, "Pattern recognition", *TsingHua University Publishing Company*. 1999.

[13] Pawlak Z, "Rough classification", *International Journal of Human-Computer studies*, Vol.51, No.2, pp.369-383, 1999.

[14] Walczak B, Massart DL, "Rough sets theory", *Chemometrics &Intelligent Laboratory Systems*, Vol.47, No.1, pp.1-16, 1999.

[15] Wang Wei. Principle of Artificial Neural Network ---- rudiment and implement. *Beijing University of Aeronautics and Astronautics Press*, 1995

[16] J.Ross Quinlan, "C4.5: Programs for Machine Learning", *Morgan Kaufmann Publishers press*, 1993.

Unsupervised Classification of Sound for Multimedia Indexing[*]

Bruce Matichuk
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada
matichuk@cs.ualberta.ca

Osmar R. Zaïane
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada
zaiane@cs.ualberta.ca

ABSTRACT

Segmenting audio streams in a significant manner and clustering sound segments objectively, is a significant challenge due to the nature of audio data. This paper presents some preliminary work on clustering sound segments based on frequency and harmonic characteristics. New metrics for comparing the similarity of sound segments are also devised.

Keywords

Multimedia Data Mining, Sound Processing, Classification, Clustering, Similarity comparison

1. INTRODUCTION

Multimedia systems play an increasingly important role in our daily lives. Access to media through the Internet, and by a growing number of electronic media recording and playback devices, is influencing and changing our lives in a profound way. Navigating through the vast amounts of multimedia data being generated is becoming an overwhelming problem. Although automated techniques for accessing electronic media are being developed, the science is still in an embryonic stage. In the area of audio-data, there has been promising early research focusing on supervised learning techniques. The authors argue that unsupervised techniques are required to handle real world retrieval situations. The research presented investigates a clustering technique that provides an automated classification scheme for short sound samples. Recognition of short samples can be combined with well-known pattern recognition algorithms to provide a viable sound retrieval system. The research could be applied to synthetising speech and sound, automatic text to speech conversion and sound compression.

General sound recognition is a difficult problem requiring the division of samples into small recognizable sound segments that can be combined to recognize complex sound patterns. Sounds are composed of signals at multiple frequencies that can be graphed on a Time-Frequency Distribution graph (TFD). Real world objects vibrate with characteristic vibrations and thus produce sound waves with characteristic harmonics that allow people and systems to perform sound recognition. A harmonic is a component frequency of a complex wave that is an integral multiple of the fundamental frequency. A sound segment is analyzed by observing the frequency amplitude pattern that occurs in the segment. Standard pattern recognition techniques can be used to recognize a frequency amplitude pattern including, neural nets, belief nets, decision trees, distance metrics, etc.

The difficulty in analyzing compound sounds lies in the problem of dividing a compound sound into recognizable segments and classifying these segments. Consider the speech recognition problem. Speech can be viewed as a sequence of sound segments that can be recognized individually and therefore combined into compound segments representing phonemes and words. Other real-world sounds can be treated in a similar way. For example, consider the sounds of a dog barking or a child crying. Both represent sounds that can be described by collections of individually recognizable sound segments.

The first task of classification is to decide upon a classification scheme. With sounds, this is difficult because of the large number of variations possible for any given sound segment. Our hypothesis is that clustering techniques can be used to classify sound segments in an unsupervised way. This hypothesis is based on the observation that sound samples are composed of recurring sound segments and that the characteristics of these segments can be determined by clustering segments that seem similar. The segments within each cluster can be analyzed to determine each cluster's representative feature set.

The remainder of the paper is organized as follows. In Section 2, we introduce some related work to audio analysis. In Section 3, We present the methodology adopted for clustering sound segments. A similarity metric for clustering sound is presented in Section 4. Section 5 describes some preliminary experiments on frequency-based and harmonic-based clustering. Finally, we conclude our study and give some pointers to future work in Section 6.

[*] Work in Progress.
Research is supported in part by the Natural Sciences and Engineering Research Council of Canada.

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

2. RELATED WORK

Audio retrieval techniques are generally classified as content-based or browser-based [13]. Content-based retrieval allows the search through audio for a segment using some pre-defined criteria. Browsing allows a user to scan through audio based on some navigational parameters. Both methods require some kind of labeling of audio data that can be used in the search process. Audio analysis to support searching can vary with respect to the semantic nature of the data. Search tasks can be semantically simple such as "Find the next location of 1 second of silence." Search tasks can also be semantically difficult, such as "Find all occurrences of the word 'the'." Some techniques being used to index sounds use a generated feature vector as an index [12]. Some systems can analyze temporal patterns [6]. The main difficulty of currently researched techniques stems from the inability to detect "in-context" high-level semantics of sound. Recent research that uses cognitive models[13] appears promising but is limited by the small number of classifications used to segment sounds. No attempt is made to discover high-level structure in sound in an automated way. Research in [10] explores a broader range of sound classes but also does not attempt to deal with complex structure. Instead, heuristics are used to detect features within sound that can be used to classify sounds in some pre-determined way.

Most research in advanced audio processing has been performed in the context of speech recognition. However, the techniques used are finely tuned toward extracting human speech patterns and detecting pre-determined natural language structures. We are examining a technique that is able to automatically determine the features and the structure of sounds suitable for general recognition and audio retrieval.

3. METHODOLOGY

The real world is composed of objects that generate sounds that vary in pitch and volume. Pitch refers to the fundamental frequency of a wave. Volume refers to the overall energy of a wave. In some cases, pitch and volume convey important information that must be incorporated into the recognition process. Recognizing a song, for example, requires attention to pitch and volume. However, for sounds like parts of speech, recognition requires a mechanism that analyzes sounds in terms of harmonic components. The process of harmonic feature extraction involves several steps. First, convert the signal from the time domain to the frequency domain using the Fast Fourier Transform (FFT) [2]. This is a standard technique used in all sound recognition algorithms. Second, extract a small sample of points that represent local peaks. This allows the selection of features that are important to the recognition of a segment. Third, normalize the sample by choosing the largest sample value and setting it to a value of 1 and scale the other sample values relative to the maximum. This step provides volume independence for the sample. The resulting set of values represents a harmonic feature set that can be used to identify the original sample independent of the volume or pitch.

The kinds of clusters that are found in a signal will depend on the similarity technique used to compare segments. Similarity clustering that is based on harmonics will cluster parts of speech in a way that does not depend on pitch. The reason for this is that the relative strength of the frequencies in a signal are relative to the harmonic characteristics of a vibrating object and tend to be independent of pitch. Consider a person making the sound 'oh' with a low pitch versus a high pitch. The harmonic characteristics in both cases will be virtually identical although the frequency composition will be completely different. Similarity clustering using frequencies will cluster sounds that are similar in pitch regardless of the harmonics. Consider the sound 'oh' versus 'ee' at the same pitch. Both of these sounds will have closely matched frequencies at a given pitch and will thus be indistinguishable based on a frequency by frequency comparison.

A further consideration in sound analysis is the length of time of a sample segment. Shorter segments are more useful in describing complex sounds such as speech. FFTs over shorter segments, however, are less accurate. There is an ideal length for a variety of sound types and optimal length may even vary within a sound sample. Tests regarding timing are not reported in this paper but the toolkit that was developed allows for timing variations. All of our tests were performed using 1/10th-second time segments. The testing software we developed is able to handle arbitrary time segments, however, 1/10th-second intervals seemed to produce the best results. A future paper will report results on varying time segments.

3.1 Clustering Sound Samples

To investigate sound sample clustering, a test system was developed whereby a tester could easily record sounds and apply a clustering algorithm to the sample. Two techniques for comparing sound samples were investigated. One technique was devised to compare two samples on a frequency by frequency basis. Another technique was devised which allows the comparison of harmonic characteristics. A series of tests were performed to determine the effectiveness of the comparison algorithms and their capability for correct clustering. The test bed devised uses a graphical technique for examining the samples and the cluster constituent. A variant of the ROCK clustering algorithm[5] was chosen to cluster sound segments.

Although further investigation using other clustering techniques is required, the ROCK algorithm has some essential features that make it an attractive technique for clustering sound. ROCK does not merely classify items based on similarity alone. Rather, ROCK considers the number of neighbours that appear to be similar to an item a more important clustering metric than the degree of item similarity. Initially all items under consideration by the algorithm are treated as belonging to separate clusters. Based on a similarity "threshold" value, each "cluster" is assigned a list of neighbour clusters that seem similar. The two clusters that seem to share the most neighbours are combined into a single cluster. A repeated evaluation of all clusters is made, combining clusters that share the most neighbours, until a desired number of clusters is achieved or until all discovered clusters have no neighbours. Classification is made of new elements by counting the number of elements within a cluster that are similar to the new element and choosing the cluster with the largest neighbour count scaled relative to overall size of each cluster.

3.2 The Threshold Phase Transition Problem

Unfortunately, the ROCK algorithm is very sensitive to the value of the threshold. Above a particular value, clustering is poor and recognition is not very effective; most points are classified as disconnected. Below a certain threshold, clustering is also poor and recognition is not effective since most points are classified as neighbours. To alleviate these problems, reduced sensitivity to thresholds is required both in clustering and in later classification.

3.2.1 Trimming

To reduce the sensitivity of ROCK to the threshold value, "trimming" reduces the number of allowed neighbours. This allows for a varying value in the threshold while maintaining a reasonable number of neighbours to cluster. A neighbouring pair of clusters is any two clusters that share sufficient neighbour links that allow the classification of both clusters into a single cluster. The neighbour vector for any cluster is ordered according to link counts. The trimming algorithm works by eliminating neighbours from the most linked cluster until pre-determined maximum number of neighbours is achieved. It is important when doing this trimming that neighbours are trimmed in pairs. Without this reciprocal trimming, the rest of the ROCK algorithm will get "out-of-sync" and the algorithm performance will deteriorate. The ROCK algorithm requires all neighbour activity to work in pairs: one action for the link "to" node and a corresponding action for the link "from" node. Therefore, when trimming, for each neighbour that is eliminated from a cluster's neighbour list, the reciprocal entry in the corresponding cluster's neighbour list must also be eliminated. An assumption that was made in choosing the trimming value was to assume that each cluster was relatively similar in size; no cluster was significantly larger than all others. This implies that a good cluster trimming value is approximately the average cluster size. In our case we chose a trimming value calculated using the sample size divided by the desired cluster count.

3.2.2 Threshold Searching

Once the clustering algorithm has completed, the final phase of operation is the assignment of items to clusters. Certain threshold values will cause many items to be unclassifiable. In some cases, this might be acceptable, however, in cases where a classification is required, an alteration to the suggested classification technique is necessary. Consider, for example, the sounds 'aw', 'oh', 'ee', 'ay' and 'eh'. Suppose we are looking for five clusters. Below a certain similarity threshold, ROCK would likely classify new sounds, which were not in the original clustering, as belonging to none of these categories. Above a certain threshold, new sounds that are similar will all be placed in the first category.

The solution is to search for a useful threshold during each assignment. The threshold value must start low. Each cluster is checked for neighbour counts using a low threshold. If an insufficient assignment is made, the threshold value is increased until a similarity count for one of the clusters is above a certain value. To employ threshold searching our modification to the ROCK classification algorithm wraps the original ROCK classification routine with a threshold searching loop. Through each iteration, the threshold is multiplied by 0.9. Our algorithm searched for at least one neighbor link between the sample and one of the cluster samples. Note that the implication here is that the "searching" is not done manually. In fact the "searching" is done automatically within the classification loop.

4. SIMILARITY METRICS FOR CLUSTERING

4.1 Sound Sample Comparison

Sound sample analysis begins with the transformation of a sound segment into the frequency domain using a Fourier Transform. The result of the transform is a series of complex numbers representing the frequency components of the sample. Using an intensity function, each frequency component can be converted into a real number. This set of frequency intensity values for a sound segment can be analyzed, and can be transformed further into a set of characteristics representing the sound. In order to use the ROCK algorithm, a similarity metric was devised which distinguishes between sound samples. The similarity metric used in ROCK clustering must produce a value from 0 to 1: 0 indicating no match, whereas 1 indicates an exact match. The euclidean distance is often used as a similarity metric in clustering. However, another metric that produces similar results is a calculation S that sums up relative ratios. In this case, given two signals composed of frequency vectors X and Y, S can be calculated as follows:

$$S = \frac{\sum_i \frac{min(x_i,y_i)}{max(x_i,y_i)}}{d} \qquad (1)$$

where d represents the dimensionality of the data. The value d varies and is calculated using a counter. Pairs, which had values below a certain threshold, were discarded from the calculation. This was necessary since large numbers of low values do not contribute to the analysis.

4.2 Harmonic Filter Based Similarity Metric

To compare harmonics, a metric is desired that is independent from pitch or volume. To provide this comparison, a harmonic component vector is calculated which is comprised of the harmonics of a sample. Natural sounds other than noise consist of a number of harmonic frequencies above a base frequency. By comparing the intensities of the harmonics, one sound can be distinguished from another. Similar to the Frequency Based Metric, the Harmonic Based Metric uses the same metric as S in (1) where d represents the dimensionality of the data. The value d is calculated by adding up all pairs of harmonics where both x and y are above a given cutoff. Each x and y value represents a harmonic pair. The harmonics are ordered such that the first harmonic of sound x is compared with the first harmonic of sound y, etc. Harmonics are found by scanning the FFT of a sample and looking for peaks.

4.3 Calculating Harmonics

To calculate the harmonic intensities the following algorithm was deployed using the FFT sample:

1. Look for a value that is a maximum.

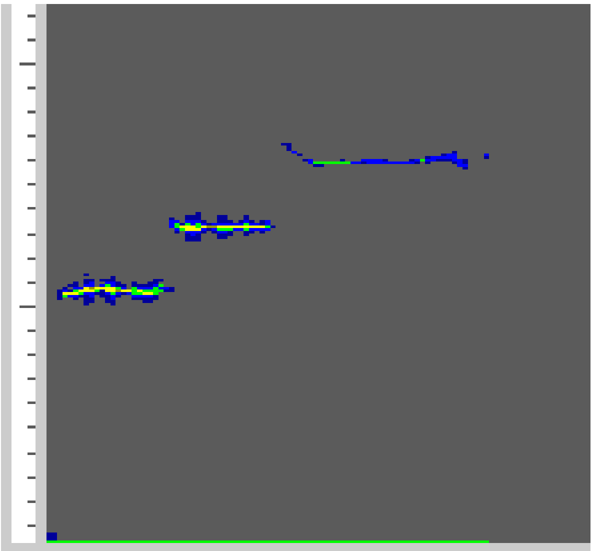


Figure 1: FFT of the recording used in a frequency clustering test.

2. For each successive value determine if the values are declining.
3. If values decline for two simultaneous measurements, record the last maximum and reset it to zero.

5. PRELIMINARY EXPERIMENTS

The test system was built using Microsoft Visual C++ 6.0 and Developer Studio. The recording code, Fourier transform code and the code that draws the frequency histograms was originally developed by a company called Relisoft Inc. based in Seattle. The code was modified to add: the ROCK algorithm, buttons to control sampling and clustering, and extensions to the draw code to allow for the drawing of cluster contents and frequency histograms with harmonic indicators.

The following tests were performed using frequency clustering and harmonic clustering. The first test involved whistling three distinct notes and searching for frequency clusters. The second involved recording the long vowel sounds “A” “E” and “U” and performing harmonic clustering.

In Figure 1 the FFT of the recording used in a frequency clustering test is displayed. The vertical axis indicates frequency. The horizontal axis is time. The tests performed called for 5 clusters from 50 samples. Using a theta value of 0.5, a cutoff of 0.4 and trimming each initial cluster to 10 neighbours, the results obtained are displayed in Figure 2. Figure 3 displays the results of the segment assigment for the entire original frequency test sample using the classification scheme derived from the frequency clustering results.

In Figure 4 the FFT of the recording used in a harmonic component clustering test is displayed. The tests performed also called for 5 clusters from 50 samples. Using a theta value of 0.25, a cutoff of 0.1 and without trimming, the results obtained are displayed in Figure 5. The divisions between phonemes can be visually distinguished by the varia-

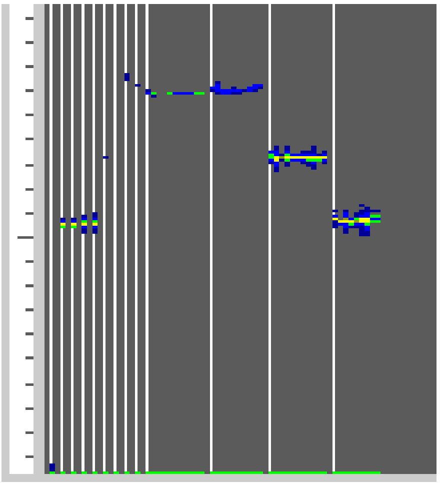


Figure 2: Clusters of frequency-based segment samples from original audio source.

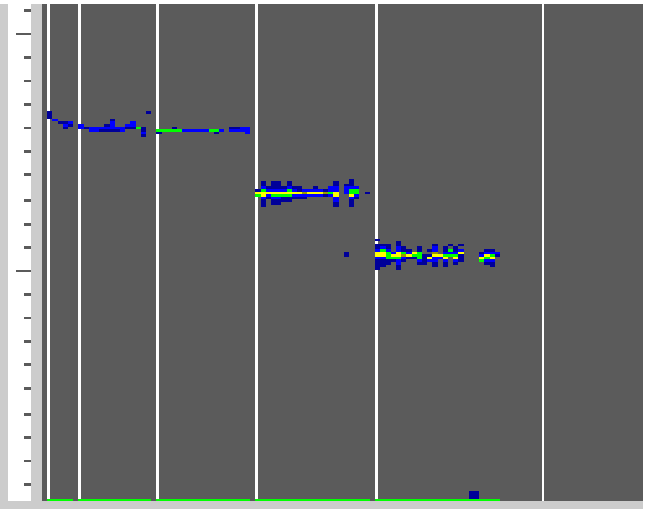


Figure 3: Frequency-based cluster assignments.

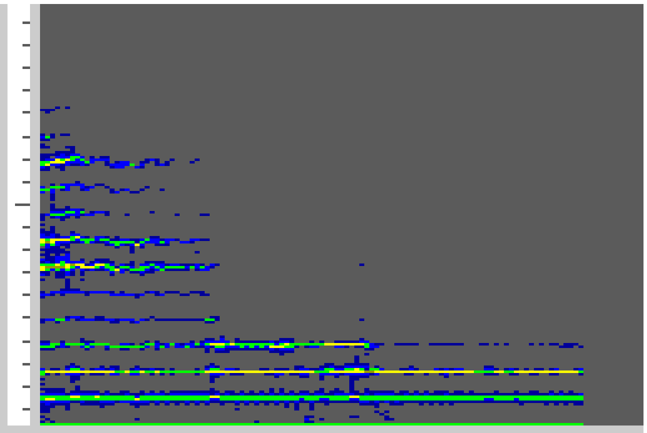


Figure 4: FFT of the recording used in a harmonic clustering test.

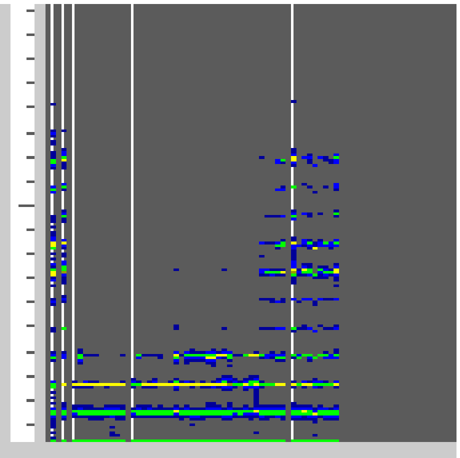


Figure 5: Clusters of harmonic-based segment samples from original audio source.

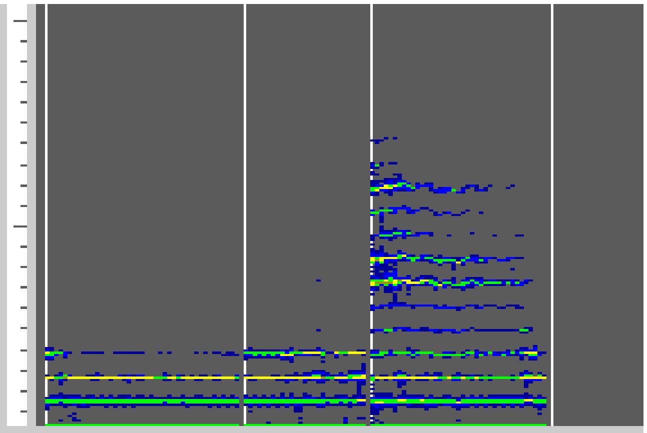
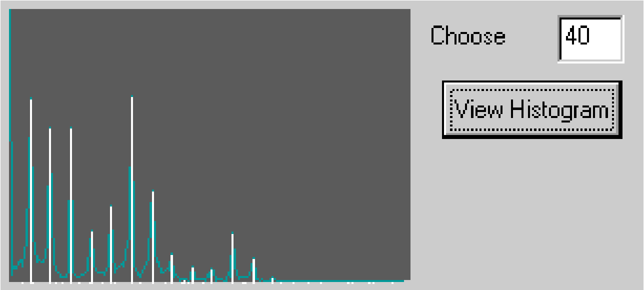


Figure 6: Harmonic-based cluster assignments.



Choose 40

Figure 7: Example of histogram graph.

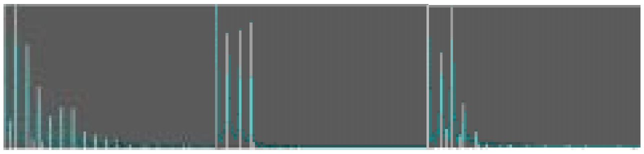


Figure 8: Histograms for the sounds "Aaa", "Ooo", "Eee".

tion in harmonic components. Figure 6 displays the results of the segment assignment for the entire original harmonic test sample using the classification scheme derived from the harmonic clustering results.

The test illustrated above was clearly very successful in that the clustering technique was able to precisely delineate the 3 vowel sounds.

Figure 7 is an example of the histogram graph. The white lines are used to indicate the locations of peaks. These peak values are used in the harmonic similarity equation.

The graphics in Figure 8 show the frequency histograms for the phrases "Aaa", "Ooo" and "Eee". The white spikes indicate peaks. The vertical axis indicates power. The horizontal axis is frequency.

Although a large number of tests were done using speech, the results reported here represent a small sample of those tests. A later paper will report on these tests and other tests using various kinds of sounds. The similarity technique that we are currently experimenting with does not deal very well with noise. Further work is required to allow noises to be included in the similarity metrics.

6. CONCLUSIONS AND FUTURE WORK

The ROCK algorithm is a very processing intensive technique for clustering. However, the algorithm is very intuitive and is general enough to apply to many different kinds of clustering problems. One major problem with ROCK is its sensitivity to the threshold value. The correct value to use is determined by the nature of the data and the nature of the similarity metric. To alleviate this problem we introduced the notion of "trimming" and the notion of threshold "searching". These techniques greatly reduced the sensitivity of the algorithm to the threshold value. Further research is required to determine if these techniques can be applied to broader problem sets. We were also able to determine that sound clustering is a viable technique for unsupervised classification of sounds.

Further work will focus on comparing ROCK to other clustering algorithms in the contexct of sound[14, 9, 3, 4, 7, 8, 11]. Other similarity metrics should be investigated including neural nets, decision trees and other pattern recognition algorithms. Additional similarity metrics are required to classify noise and inter-segment transitions [1]. Finally, clustering techniques that combine segments into complex groupings are required to classify higher order sounds. This would require multiple clustering passes that cluster groups corresponding to sound patterns such as rhythms, melodies or speech.

Further research will also investigate classification techniques that will allow segments to belong to multiple clusters. This can be achieved by applying a "fuzzy" similarity threshold value during classification in contrast to the current classification technique which is binary.

7. REFERENCES

[1] A. Czyzewski. Mining knowledge in noisy audio data. In *Proc. Second Int. Conf. on Knowledge Discovery and Data Mining (KDD96)*, pages 220–225, Portland, Oregon, August 1996.

[2] D. E. Dudgeon and R. M. Mersereau. *Multidimentional Digital Signal Processing.* Printice-Hall, 1984.

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. In *Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, pages 226–231, Portland, Oregon, August 1996.

[4] V. Ganti, J. E. Gehrke, and R. Ramakrishnan. CACTUS - clustering categorical data using summaries. In *Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99)*, San Diego, CA, 1999.

[5] S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. In *Proc. 1999 Int. Conf. Data Engineering (ICDE'99)*, pages 512–521, Sydney, Australia, March 1999.

[6] D. Hindus, C. Schmandt, and C. Horner. Capturing, structuring and representing ubiquitous audio. *ACM Transactions on Information Systems*, 11(4):376–400, Oct. 1993.

[7] A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. In *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, pages 58–65, New York, NY, August 1998.

[8] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2:283–304, 1998.

[9] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A survey. *ACM Comput. Surv.*, 31:264–323, 1999.

[10] K. Melih and R. Gonzalez. Audio retrieval using perceptually based structures. In *Proc. IEEE Intl. Conf. on Multimedia Computing and Systems*, pages 338–347, 1998.

[11] G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. In *Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)*, pages 428–439, New York, NY, August 1998.

[12] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-based classification, search and retrieval of audio. In *Proc. 1999 IEEE Multimedia Conf.*, pages 27–36, 1996.

[13] T. Zhang and C.-C. J. Kuo. Heuristic approach for generic audio data segmentation and annotation. In *Proc. 1999 ACM-Multimedia Conf.*, pages 67–76, Orlando, FL, October 1999.

[14] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'96)*, pages 103–114, Montreal, Canada, June 1996.

Effective Retrieval of Audio Information from Annotated Text Using Ontologies[1]

Latifur Khan and Dennis McLeod
Department of Computer Science and
Integrated Media Systems Center
University of Southern California
Los Angeles, California 90089
[latifurk, mcleod]@usc.edu

ABSTRACT

To improve the accuracy in terms of precision and recall of an audio information retrieval system we have created a domain-specific ontology (a collection of key concepts and their interrelationships), as well as a novel, pruning algorithm. Taking into account the shortcomings of keyword-based techniques, we have opted to employ a concept-based technique utilizing this ontology. The key problem in the retrieval of audio information is to achieve high precision and high recall. Typically, in traditional approaches, high recall is achieved at the expense of low precision, and vice versa. Through the use of a domain-specific ontology appropriate concepts can be identified during metadata generation (description of audio) or query generation, thus improving precision. In case of the association of irrelevant concepts to queries or documents there is a loss of precision. On the other hand, if relevant concepts are discarded, a loss of recall will ensue. Therefore, in conjunction with the use of a domain specific ontology we have proposed a novel, automatic pruning algorithm which prunes as many irrelevant concepts as possible during any case of query generation. By associating concepts in the ontology through techniques of correlation, this algorithm presents a method for the selection of concepts in the query generation. To improve recall, controlled and correct query expansion mechanism is proposed. This guarantees that precision will not be lost. Moreover, we present a way for the query generation in which domain-specific ontology can be used to generate information selection requests in terms of database queries in SQL. In trial implementations we have demonstrated that our ontology-based model outperforms keyword-based technique (vector space model) in terms of precision and recall.

Keywords

Metadata, Ontology, Audio, SQL, Precision, and Recall.

1 Introduction

The development of technology in the field of digital media generates huge amounts of non-textual information, such as audio, video, and images, as well as more familiar textual information. The potential for the exchange and retrieval of information is vast, and at times daunting. In general, users can be easily overwhelmed by the amount of information available via electronic means. The transfer of irrelevant information in the form of documents (e.g. text, audio, video) retrieved by an information retrieval system and which are of no use to the user wastes network bandwidth and frustrates users. This condition is a result of inaccuracies in the representation of the documents in the database, as well as confusion and imprecision in user queries, since users are frequently unable to express their needs efficiently and accurately. These factors contribute to the loss of information and to the provision of irrelevant information. Therefore, the key problem to be addressed in information selection is the development of a search mechanism which will guarantee the delivery of a minimum of irrelevant information (high precision), as well as insuring that relevant information is not overlooked (high recall).

The traditional solution to the problem of recall and precision in information retrieval employs keyword-based search techniques. Documents are only retrieved if they contain keywords specified by the user. However, many documents contain the desired semantic information, even though they do not contain user specified keywords. This limitation can be addressed through the use of query expansion mechanism. Additional search terms are added to the original query based on the statistical co-occurrence of terms [20]. Recall will be expanded, but at the expense of deteriorating precision [17, 24]. In order to overcome the shortcomings of keyword-based technique in responding to information selection requests we have designed and implemented a concept-based model using ontologies [12]. This model, which employs a domain dependent ontology, is presented in this paper. An ontology is a collection of concepts and their interrelationships which can collectively provide an abstract view of an application domain [5, 8].

There are two distinct questions for ontology-based model: one is the extraction of the semantic concepts from the keywords and the other is the indexing. With regard to the first problem, the key issue is to identify appropriate concepts that describe and identify documents on the one hand, and on the other, the language employed in user requests. In this it is important to make sure that irrelevant concepts will not be associated and matched, and that relevant concepts will not be discarded. In other words, it is important to insure that high precision and high

[1] This research has been funded [or funded in part] by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152.

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

recall will be preserved during concept selection for documents or user requests. In this paper, we propose an automatic mechanism for the selection of these concepts from user requests by addressing the first problem. This mechanism will prune irrelevant concepts while allowing relevant concepts to become associated with user requests. Furthermore, a novel, scalable disambiguation algorithm for concept selection from documents using domain specific ontology is presented in [13].

With regard to the second problem, one can use vector space model of concepts or more precise structure by choosing ontology. We adopt the latter approach. This is because vector space model does not work well for short queries. Furthermore, one recent survey about web search engines suggests that average length of user request is 2.2 keywords [4]. For this, we have developed a concept-based model, which uses domain dependent ontologies for responding to information selection requests. To improve retrieval, we also propose an automatic query expansion mechanism which deals with user requests expressed in natural language. This automatic expansion mechanism generates database queries by allowing only appropriate and relevant expansion. Intuitively, to improve recall during the phase of query expansion, only controlled and correct expansion is employed, guaranteeing that precision will not be degraded as a result of this process. Furthermore, for the disambiguation of concepts only the most appropriate concepts are selected with reference to documents or to user requests by taking into account the encoded knowledge in the ontology.

In order to demonstrate the effectiveness of our disambiguation model we have explored and provided a specific solution to the problem of retrieving audio information. The effective selection/retrieval of audio information entails several tasks, such as metadata generation (description of audio), and the consequent selection of audio information in response to a query. Relevant to our purpose, ontologies can be fruitfully employed to facilitate metadata generation. For metadata generation, we need to do content extraction by relying on speech recognition technology which converts speech to text. After generating transcripts we can deploy our ontology-based model to facilitate information selection requests. At present, an experimental prototype for the implementation of the model has been developed and implemented. As of today, our working ontology has around 7,000 concepts for the sports news domain, with 2,481 audio clips/objects of metadata in the database. For sample audio content we use CNN broadcast sports and Fox Sports audio, along with closed captions. To illustrate the power of ontology-based over keyword-based search techniques we have taken the most widely used vector space model as representative of keyword search. For comparison metrics we have used measures of precision and recall, and an F score that is the harmonic mean of precision and recall. Nine sample queries were run based on the categories of broader query (generic), narrow query (specific), and context query formulation. We have observed that on average our ontology outperforms keyword-based technique. For broader and context queries, the result is more pronounced than in cases of narrow query.

The remainder of this paper is organized as follows. In Section 2, we review related work. In Section 3, we introduce the research context in terms of the information media used (i.e., audio) and some related issues that arise in this context. In Section 4, we introduce our domain dependent ontology. In Section 5, we present metadata management issues that arise for our ontology based model in the context of audio information unit. In Section 6, we present a framework through which user requests expressed in natural language can be mapped into database queries in order to support index structure along with pruning algorithm. In Section 7 we give a detailed description of the prototype of our system, and provide data showing how our ontology-based model compares with traditional keyword-based search technique. Finally, in Section 8 we present our conclusions and plans for future work.

2 Related Works

Historically ontologies have been employed to achieve better precision and recall in the text retrieval system [9]. Here, attempts have taken two directions, query expansion through the use of semantically related-terms, and the use of conceptual distance measures, as in our model. Among attempts using semantically related terms, query expansion with a generic ontology, WordNet [15], has been shown to be potentially relevant to enhanced recall, as it permits matching a query to relevant documents that do not contain any of the original query terms. Voorhees [22] manually expands 50 queries over a TREC-1 collection using WordNet, and observes that expansion was useful for short, incomplete queries, but not promising for complete topic statements. Further, for short queries, automatic expansion is not trivial; it may degrade rather than enhance retrieval performance. This is because WordNet is too incomplete to model a domain sufficiently. Furthermore, for short queries less context is available, which makes the query vague. Therefore, it is hard to choose appropriate concepts automatically. The notion of conceptual distance between query and document provides an alternative approach to modeling relevance. Smeaton et al. [20] and Gonzalo et al. [7] focus on managing short and long documents, respectively. Note here that in these approaches queries and document terms are manually disambiguated using WordNet. In our case, query expansion and the selection of concepts, along with the use of the pruning algorithm, is fully automatic.

Although we use audio, here we show related work in the video domain which is closest to and which complements our approach in the context of data modeling for the facilitation of information selection requests. Key related work in the video domain for selection of video segments includes [1, 11, 16]. Of these, Omoto et al. [16] use a knowledge hierarchy to facilitate annotation, while others use simple keyword based techniques without a hierarchy. The model of Omoto et al. fails to provide a mechanism that automatically converts a generalized description into a specialized one(s). Further, this annotation is manual and does not deal with the disambiguation issues related to concepts.

3 Research Context: Audio

Audio is one of the most powerful and expressive of the non-textual media. Audio is a streaming medium (temporally extended), and its properties make it a popular medium for capturing and presenting information. At the same time, these very properties, along with audio's opaque relationship to computers, present several technical challenges from the perspective of data management [6]. The type of audio considered here is broadcast audio. In general, within a broadcast audio stream, some items are of interest to the user and some are not. Therefore, we need to identify the boundaries of news items of interest so that these segments can be directly and efficiently retrieved in response to a user query. After segmentation, in order to retrieve a set of segments that match with a user request, we need to specify the content of segments. This can be achieved using content extraction through speech recognition. Therefore, we present segmentation and content extraction technique one by one.

3.1 Segmentation of Audio

Since audio is by nature totally serial, random access to audio information may be of limited use. To facilitate access to useful segments of audio information within an audio recording deemed relevant by a user, we need to identify entry points/jump locations. Further, multiple contiguous segments may form a relevant and useful news item.

As a starting point both a change of speaker and long pauses can serve to identify entry points [2]. For long pause detection, we use short-time energy (E_n), which provides a measurement for distinguishing speech from silence for a frame (consisting of a fixed number of samples) which can be calculated by the following equation [18]:

$$En = \sum_{m=-\infty}^{m=\infty} [x(m)w(n-m)]^{2} = \sum_{m=n-N+1}^{m=n} x(m)^{2}$$

Where $x(m)$ is discrete audio signals, n is the index of the short-time energy, and $w(m)$ is a rectangle window of length N. When the E_n falls below a certain threshold we treat this frame as pause. After such a pause has been detected we can combine several adjacent pauses and identify what can be called a *long pause*. Therefore, the presence of speeches with starting and ending points defined in terms of long pauses allows us to detect the boundaries of audio segments.

3.2 Content Extraction

To specify the content of media objects two main approaches have been employed to this end: fully automated content extraction [10], and selected content extraction [23]. In fully automated content extraction, speech is converted to equivalent text (e.g., Informedia). Word-spotting techniques can provide selected content extraction in a manner that will make the content extraction process automatic. Word-spotting is a particular application of automatic speech recognition techniques in which the vocabulary of interest is relatively small. In our case, vocabularies of concepts from the ontology can be used. Furthermore, content description can be provided in plain text, such as closed captions. However, this manual annotation is labor intensive. For content extraction we rely on closed captions that came with audio object itself from fox sports and CNN web site in our case (see Section 7).

3.3 Definition of an Audio Object

An audio object, by definition and in practice, is composed of a sequence of contiguous segments. Thus, in our model the start time of the first segment and the end time of the last segment of these contiguous segments are used respectively to denote start time and end time of the audio object. Further, in our model, pauses between interior segments are kept intact in order to insure that speech will be intelligible. The formal definition of an audio object indicates that an audio object's description is provided by a set of self-explanatory tags or labels using ontologies. An audio-object O_i is defined by five tuple (id_i, S_i, E_i, V_i, A_i) where Id_i is an object identifier which is unique**,** S_i is the start time, E_i is the end time, V_i (description) is a finite set of tag or label, i.e., $V_i=\{v_{1i}, v_{2i}, \ldots ,v_{ji}, \ldots,v_{ni}\}$ for a particular j where v_{ji} is a tag or label name, and A_i is simply audio recording for that time period. For example, an audio object is defined as {10, 1145.59, 1356.00, {Gretzky Wayne}, *}. Of the information in the five tuple, the first four items (identifier, start time, end time, and description) are called *metadata*.

4 Ontologies

An ontology is a specification of an abstract, simplified view of the world that we wish to represent for some purpose [5, 8]. Therefore, an ontology defines a set of representational terms that we call *concepts*. Interrelationships among these concepts describe a target world. An ontology can be constructed in two ways, domain dependent and generic. CYC [14], WordNet [15], or Sensus [21] are examples of generic ontologies. For our purposes, we choose a domain dependent ontology. First, this is because a domain dependent ontology provides concepts in a fine grain, while generic ontologies provide concepts in coarser grain. Second, a generic ontology provides a large number of concepts that may contribute large speech recognition error.

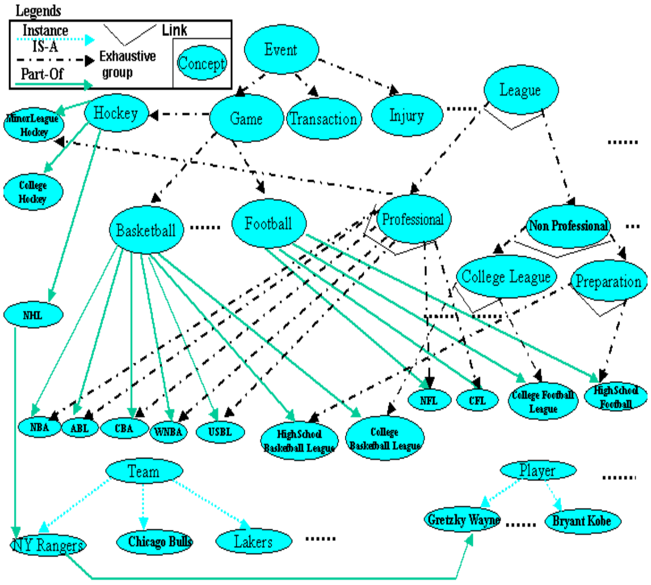


Figure 1. A Small Portion of an Ontology for Sports Domain

Figure 1 shows an example ontology for sports news. This ontology is usually obtained from generic sports terminology and domain experts. This ontology is described by a directed acyclic graph (DAG). Here, each node in the DAG represents a concept. In general, each concept in the ontology contains a label name and a synonyms list. Note also that this label name is unique in the ontology. Further, this label name is used to serve as association of concepts with audio objects. The synonyms list of a concept contains vocabulary (a set of keywords) through which the concept can be matched with user requests. Formally, each concept has a synonyms list ($l_1, l_2, l_3, \ldots, l_i, \ldots, l_n$) where user requests are matched with this l_i what we call *element* of list. Note that a keyword may be shared by multiple concepts' synonyms lists. For example, player "Bryant Kobe," "Bryant Mark," "Reeves Bryant" share common word "Bryant" which may create ambiguity problem.

4.1 Interrelationships

In the ontology, concepts are interconnected by means of interrelationships. If there is a interrelationship R, between concepts C_i and C_j, then there is also a interrelationship R′ between concepts C_j and C_i. In Figure 1, interrelationships are represented by labeled arcs/links. Three kinds of interrelationships are used to create our ontology: IS-A, Instance-Of, and Part-Of. These correspond to key abstraction primitives in object-based and semantic data models [3].

IS-A: This interrelationship is used to represent concept inclusion. A concept represented by C_j is said to be a specialization of the concept represented by C_i if C_j is kind of C_i . For example, "NFL" is a kind of "Professional" league. In other words, "Professional" league is the generalization of "NFL." In Figure 1, the IS-A interrelationship between C_i and C_j goes from generic concept C_i to specific concept, C_j represented by a broken line. The IS-A interrelationship can be further categorized into two types: *exhaustive group* and *non-exhaustive group*. An exhaustive group consists of a number of IS-A interrelationships between a generalized concept and a set of specialized concepts, and places the generalized concept into a categorical relation with a set of specialized concepts in such a way so that the union of these specialized concepts is equal to the generalized concept. For example, "Professional" relates to a set of concepts, "NBA", "ABL", "CBA", ..., by exhaustive group (denoted by caps in Figure 1). Further, when a generalized concept is associated with a set of specific concepts by only IS-A interrelationships that fall into the exhaustive group, then this generalized concept will not participate in the metadata generation and SQL query generation explicitly. This is because this generalized concept is entirely partitioned into its specialized concepts through an exhaustive group. We call this generalized concept a *non participant concept (NPC)*. For example, in Figure 1 "Professional" concept is NPC. On the other hand, a non-exhaustive group consisting of a set of IS-A does not exhaustively categorize a generalized concept into a set of specialized concepts. In other words, the union of specialized concepts is not equal to the generalized concept.

Instance-Of: This is used to show membership. A C_j is a member of concept C_i. Then the interrelationship between them corresponds to an Instance-Of denoted by a dotted line. Player, "Wayne Gretzky" is an instance of a concept, "Player." In general, all players and teams are instances of the concepts, "Player" and "Team" respectively.

Part-Of: A concept is represented by C_j is Part-Of a concept represented by C_i if C_i has a C_j (as a part) or C_j is a part of $C_{i.}$ For example, the concept "NFL" is Part-Of "Football" concept and player, "Wayne Gretzky" is Part-Of "NY Rangers" concept.

4.2 Disjunctness

When a number of concepts are associated with a parent concept through IS-A interrelationship, it is important to note that these concepts are disjoint, and are referred to as concepts of a disjoint type. When, for example, the concepts "NBA", "CBA", or "NFL" are associated with the parent concept "Professional," through IS-A, they become disjoint concepts. Moreover, any given object's metadata cannot possess more than one such concept of the disjoint type. For example, when an object's metadata is the concept "NBA," it cannot be associated with another disjoint concept, such as "NFL." It is of note that the property of being disjoint helps to disambiguate concepts for keywords during metadata or query generation phases. Similarly, concept "College Football", "College Basketball" are disjoint concepts due to their associations with parent concept, "College League" through IS-A. Furthermore, "Professional," and "Non Professional" are disjoint. Thus, we can say that "NBA," "CBA," "ABL," "College Basketball," and "College Football," are disjoint. Each of these league and its team and player form a boundary what we call *region* (see Figure 2). During annotation of concepts with an audio object we strive to choose a particular region. This is because an audio object can be associated with only one disjoint-type concept. However, it may be possible that a particular player may play in several leagues. In that case, we make multiple instances of the player. In other words, for each league he plays, we maintain a separate concept for him. This way we preserve disjoint-property.

Concepts are not disjoint, on the other hand, when they are associated with a parent concept through Instance-Of or Part-Of. In this case, some of these concepts may serve simultaneously as metadata for an audio object. An example would be the case in which the metadata of an audio object are team "NY Ranger" and player "Gretzky Wayne," where "Grezky Wayne" is Part-Of "NY Rangers."

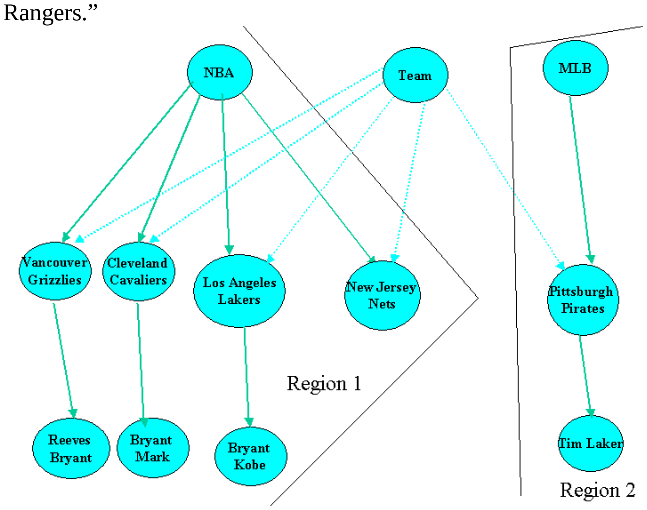


Figure 2. Different Regions of an Ontology

5 Metadata Acquisition and Management of Metadata

Metadata acquisition is the name for the process through which descriptions are provided for audio objects. For each audio object we need to find the most appropriate concept(s). Recall that using content extraction (see Section 3.2) we get a set of keywords which appear in a given audio object. For this, concepts from ontologies will be selected based on matching terms taken from their lists of synonyms with those based on specified keywords. Furthermore, each of these selected concepts will have a score based on a partial or a full match. It is possible that a particular keyword may be associated with more than one concept in the ontology. In other words, association between keyword and concept is one:many, rather than one:one. Therefore, the disambiguation of concepts is required. The basic notion of disambiguation is that a set of keywords occurring together determine a context for one another, according to which the appropriate senses of the word (its appropriate concept) can be determined. Note, for example, that base, bat, glove may have several interpretations as individual terms, but when taken together, the intent is obviously a reference to baseball. The reference follows from the ability to determine a context for all the terms. Thus, extending and formalizing the idea of context in order to achieve the disambiguation of concepts, we propose an efficient pruning algorithm based on two principles: co-occurrence and semantic closeness. This disambiguation algorithm first strives to disambiguate across several regions using first principle, and then disambiguates within a particular region using the second (see [13] for more details).

Effective management of metadata facilitates efficient storing and retrieval of audio information. To this end, in our model most specific concepts are considered as metadata. Several concepts of the ontology, for example, can become the candidate for the metadata of an audio object. However, some of these may be children of others. Two alternative approaches can be used to address this problem. First, we can simply store the most general concepts. But we may get many irrelevant objects (precision will be hurt) for queries related to specific concepts. For example, an audio object becomes the candidate for the concepts, “NHL,” “Hockey,” and “Professional.” We can simply store the general concept, “Professional” for this object. When user request comes in terms of specific concept, “NHL”, this object will be retrieved along with other irrelevant objects that do not belong to NHL (say, NFL, CFL, and so on). Therefore, precision will be hurt. Second, the most specific concepts can be stored in the database. Corresponding generalized concepts can then be discarded. In this case, recall will be hurt. Suppose, for example, an audio object becomes the candidate for the concepts "NHL", "Hockey", and "Professional." During the annotation process the object will only be annotated with the most specific concept, "NHL." In this case, the metadata of the audio objects stored in the database will be comprised of the most specific concepts. If query comes in terms of “Hockey” or “Professional”, this object will not be retrieved.

We follow the latter approach. By storing specific concepts as metadata, rather than generalized concepts of the ontology, we can expect to achieve the effective management of metadata. In order to avoid recall problem, user requests are first passed through ontology on the fly and expressed in terms of most specific concepts. Even so, the audio object, in the above example, can still be retrieved through querying the system by "NHL", "Hockey", and "Professional."

Here, we consider an efficient way of storing audio objects in the database: we maintain a single copy of all the audio data in the database. Further, each object’s metadata are stored in the database. Thus, this start time, and end time of an object point to a fraction of all the audio data. Therefore, when the object is selected, this boundary information provides relevant audio data that are to be fetched from all the audio data and played by the scheduler. The following self-explanatory schemas are used to store audio objects in the database: *Audio_News (Id, Time_Start, Time_End, ...), and Meta_News (Id, Label)*. Each audio object’s start time, end time and description correspond to Time_Start, Time_End, and Label respectively. Furthermore, each object’s description is stored as a set of rows or tuples in the Meta_News table for normalization purpose.

6 Query Mechanisms

We now focus specifically on our techniques for utilizing an ontology-based model for processing information selection requests. In our model the structure of ontology facilitates indexing. In other words, ontology provides index terms/concepts which can be used to match with user requests. Furthermore, the generation of a database query takes place after the keywords in the user request are matched to concepts in the ontology.

We assume that user requests are expressed in plain English. Tokens are generated from the text of the user's request after stemming and removing stop words. Using a list of synonyms these tokens are associated with concepts in the ontology through Depth First Search (DFS) or Breadth First Search (BFS). Each of these selected concepts is called a *QConcept*. Among QConcepts, some might be ambiguous. However, through the application of a pruning technique that will be discussed in Section 6.1 only relevant concepts are retained. These relevant concepts will then be expanded, and will participate in SQL query generation as is discussed in Section 6.2.

6.1 Pruning

Disambiguation is needed when a given keyword matches more than one concept. In other words, multiple ambiguous concepts will have been selected for a particular keyword. For disambiguation, it is necessary to determine the correlation between selected concepts based on semantic closeness. When concepts are correlated, the scores of concepts strongly associated with each other will be given greater weight based on their minimal distance from each other in the ontology and their own matching scores based on the number of words they match. Thus, ambiguous concepts which correlate with other selected concepts will have a higher score, and a greater probability of being retained than ambiguous concepts which are not correlated.

For example, if a query is specified by "Please tell me about team Lakers," QConcepts "Team," "Los Angeles Lakers," and major league baseball player, "Tim Laker" (of team "Pittsburgh Pirates") are selected. Note that selected concepts, "Los Angeles Lakers," and "Tim Laker" are ambiguous. However, "Los Angeles Lakers" is associated with selected QConcept, "Team" due to Instance-Of interrelationship. Therefore, we prune the non-correlated ambiguous concept, player "Tim Laker." The above idea is implemented using score-based techniques. Now, we would like to present our concept-pruning algorithm for use with user requests.

6.1.1 Formal Definitions

Each selected concept contains a score based on the number of keywords from the list of synonyms which have been matched with the user request. Recall that in an ontology each concept (QC_i) has a complementary list of synonyms ($l_1, l_2, l_3, ..., l_j ,...,l_n$). Keywords in the user request are sought which match each keyword on the element l_j of a concept. The calculation of the score for l_j, which we designate an *Escore*, is based on the number of matched keywords of l_j. The largest of these scores is chosen as the score for this concept, and is designated *Score*. Furthermore, when two concepts are correlated, their scores, called the *Propagated-score*, are inversely related to their position (semantic distance) in the ontology. Let us formally define each of these scores.

Definition 1: Element-score (Escore): The Element-score of an element l_j for a particular QConcept QC_i is the number of keywords of l_j matched with keywords in the user request divided by total number of keywords in l_j.

$$Escoreij \equiv \frac{\#\ of\ \ keywords\ of\ \ lj\ \ matched}{\|\ \#\ of\ \ keywords\ \ in\ lj\ \|}$$

The denominator is used to nullify the effect of the length of l_j on $Escore_{ij}$ and ensures that the final weight is between 0 and 1.

Definition 2: Concept-score (Score): The Concept-score for a QConcept, QC_i is the largest score of all its element-scores. Thus,

$$Score_{\,i} = max\ \ Escore_{\,ij}\ \ where\ \ 1 \le\ j\ \le\ n$$

Definition.3: Semantic distance (SD (QC_i , QC_j)): *SD(QC_i, QC_j)* between QConcepts QC_i and QC_j is defined as the shortest path between two QConcepts, QC_i and QC_j in the ontology. Note that if concepts are in the same level and no path exists, the semantic distance is infinite. For example, the semantic distance between concepts "NBA" and team "Lakers" is 1 (see Figure 2). This is because the two concepts are directly connected via a Part-Of interrelationship. Similarly, the semantic distance between "NBA," and "Bryant Kobe" is 2. The semantic distance between "Los Angeles Lakers," and "New Jersey Nets" is infinite.

Definition.4: Propagated-score (S_i): If a QConcept, QC_i, is correlated with a set of QConcepts *(C_j, C_{j+1},...,C_n)*, the propagated-score of QC_i is its own Score, $Score_i$ plus the scores of each of the correlated QConcepts' *(QC_k k=j, j+1, ..., n) $Score_k$* divided by *SD (QC_i, QC_k)*. Thus,

$$S_i = Score_i + \sum_{k=j}^{k=n} \frac{Score_k}{SD(QC_i,QC_k)}$$

$$= Score_i + \frac{Score_j}{SD(QC_i,QC_j)} + \frac{Score_{j+1}}{SD(QC_i,QC_{j+1})} + ... + \frac{Score_n}{SD(QC_i,QC_n)}$$

For example, in Figure 2 let us assume that values of $Score_i$ for "Los Angeles Lakers" and "Bryant Kobe" be 0.5 and 1.0 respectively. Furthermore, these concepts are correlated with a semantic distance of 1, and their Propagated-scores are 1.5 (0.5 + 1.0/1) and 1.5 (1.0+0.5/1) respectively. The pseudo code for the pruning algorithm is as follows:

```
$QC_1$, $QC_2$ , …, $QC_l$, …, $QC_r$ are selected with concept-score
$Score_1$,…,$Score_l$,…$Score_r$
Determine correlation of selected concepts ($QC_i$, $QC_j$, $QC_{j+1}$,
.., $QC_n$) and update their Propagated-scores using

$S_i = Score_i + \sum_{k=j}^{k=n} \frac{Score_k}{SD(QC_i,QC_k)}$
$= Score_i + \frac{Score_j}{SD(QC_i,QC_j)} + \frac{Score_{j+1}}{SD(QC_i,QC_{j+1})} + ... + \frac{Score_n}{SD(QC_i,QC_n)}$
Sort all QConcepts ($QC_i$) based on $S_i$ in descending order
//Find Ambiguous QConcepts and prune some of them
//which have low Propagated-score…
For a keyword that associated with ambiguous QConcepts,
    $QC_i$, $QC_j$, $QC_l$, … where $S_i > S_j > S_l$, ...
    Keep only $QC_i$ and discard $QC_j$, $QC_l$, …
//End of For Loop for a keyword.
Keep all specific QConcepts and discard corresponding
generalized concepts
For each QConcept that are not pruned
        Query_Expansion_SQL_Generation (QConcept)
    //see Figure 4
//End of For loop each QConcept
```

Figure 3. Pseudo Code for Pruning Algorithm

Using pruning algorithm (see Figure 3), for a user request, "team Lakers," at the beginning selected QConcepts are "Team", "Los Angeles Lakers" and "Tim Laker" (see Figure 2). Note that ambiguous concepts are "Los Angeles Lakers," and "Tim Laker." In Figure 2 the SD between concepts, "Team," and "Los Angeles Lakers" is 1 while the SD between concepts, "Team" and "Tim Laker" is 2. Furthermore, the Scores for concepts, "Team," "Los Angeles Lakers," and "Tim Laker" are 1.0, 0.5, 0.5 respectively. It is important to note that when two concepts are correlated with each other where semantic distance is greater than one, they will have a lower Propagated-scores, S_i and S_j compared to concepts with the same concept-scores and a semantic distance of 1. This is because for the higher semantic distance concepts are correlated in a broader sense. Thus, concepts which are correlated have a higher S_i in comparison with non-correlated concepts. Now, the Propagated-score for QConcepts, "Team," "Los Angeles Lakers," and "Tim Laker" becomes 1.75 (1.0+0.5/1+0.5/2), 1.5 (0.5+1.0/1), and 1.0 (0.5+1.0/2) respectively. Therefore, we keep the concept "Los Angeles Lakers" from among these ambiguous concepts and prune the other. Thus, the SD helps us to discriminate between ambiguous concepts.

Among selected concepts, one concept may subsume the other concept. In this case, we use specific concept for SQL generation. For example, if a user request is expressed in terms of "Please tell me about Lakers' Bryant," the QConcepts, team "Los Angeles Lakers," players, "Bryant Kobe", "Bryant Mark," "Reeves Bryant," are selected. Their concept-scores are 0.5, 0.5, 0.5 respectively. The latter three are ambiguous concepts. However, among these selected concepts, only "Bryant Kobe," and "Los Angeles Lakers" are correlated with a semantic distance of 1 (see Figure 2). Therefore, their propagated-scores S_i are high as compared to other concepts, in this case, 1.0, 1.0, 0.5, 0.5 respectively. Consequently, we throw away "Bryant Reeves" and "Bryant Mark." Furthermore, "Bryant Kobe" is a sub-concept of "Los Angels Lakers," due to a Part-Of interrelationship. In this case, we keep the more specific concept, "Bryant Kobe," and the SQL generation algorithm will be called for this QConcept only.

6.2 Query Expansion and SQL Query Generation

We now discuss a technique for query expansion and SQL query generation. In response to a user request for the generation of an SQL query, we follow a Boolean retrieval model. We now consider how each QConcept is mapped into the "where" clause of an SQL query. Note that by setting the QConcept as a Boolean condition in the "where" clause, we are able to retrieve relevant audio objects. First, we check whether or not the QConcept is of the NPC type. Recall that NPC concepts can be expressed exhaustively as a collection of more specific concepts. If the QConcept is a NPC concept, it will not be added in the "where" clause. On the other hand, it will be added into the "where" clause. Likewise, if the concept is leaf node, no further progress will be made for this concept. However it is non-leaf node, its children concepts are generated using DFS/BFS, and this technique is applied for each children concept. One important observation is that all concepts appearing in an SQL query for a particular QConcept are expressed in disjunctive form. Furthermore, during the query expansion phase only correct concepts are added which will guarantee that addition of new terms will not hurt precision. The complete algorithm is shown in Figure 4.

```
Query_Expansion_SQL_Generation ($QC_i$)
    Mark $QC_i$ is already visited
    If $QC_i$ is not NPC Type
        Add label of $QC_i$ into where clause of SQL as
        disjunctive form
    //Regardless of NPC type concept
    If $QC_i$ is not leaf node and not visited yet
        For each children concept, $QCh_l$ of $QC_i$ using
        DFS/BFS
            Query_Expansion_SQL_Genertaion ($QCh_l$)
```

Figure 4. Pseudo Code for SQL Generation

The following example illustrates the above process. Suppose the user request is "Please give me news about player Kobe Bryant." "Bryant Kobe" turns out to be the QConcept which is itself a leaf concept. Hence, the SQL query (for schema see Section 5) generated by using only "Bryant Kobe" (with the label "NBAPlayer9") is:

SELECT Time_Start, Time_End
FROM Audio_News a, Meta_News m
WHERE a.Id=m.Id
AND Label="NBAPlayer9"

Let us now consider the user request, "Tell me about Los Angeles Lakers." Note that the concept "Los Angeles Lakers" is not of the NPC type, so its label ("NBATeam11") will be added in the "where" clause of the SQL query. Further, this concept has several children concepts ("Bryant Kobe," "Celestand John," "Horry Robert," .... i.e. names of players for this team). Note that these player concepts' labels are "NBAPlayer9," "NBAPlayer10," and "NBAPlayer11," respectively. In SQL query:

SELECT Time_Start, Time_End
FROM Audio_News a, Meta_news m
WHERE a.Id = m.Id
AND (Label="NBATeam11"
    OR Label="NBAPlayer9"
    OR Label**=**"NBAPlayer10"...)

6.2.1 Remedy of Explosion of Boolean Condition

Since most specific concepts are used as metadata and our ontologies are large in the case of querying upper level concepts, every relevant child concept will be mapped into the "where" clause of the SQL query and expressed as a disjunctive form. To avoid the explosion of Boolean conditions in this clause of the SQL query, the labels for the player and team concepts are chosen in an intelligent way. These labels begin with the label of the league in which the concepts belong. For example, team "Los Angeles Lakers" and player "Bryant, Kobe" are under "NBA." Thus, the labels for these two concepts are "NBATeam11" and "NBAPlayer9" respectively, whereas the label for the concept "NBA" is "NBA."

Now, when user requests come in terms of an upper level concept (e.g., "Please tell me about NBA.") the SQL query generation mechanism will take advantage of prefixing:

SELECT Time_Start, Time_End
FROM Audio_News a, Meta_News m
WHERE a.Id=m.Id
AND Label Like "%NBA%"

On the other hand, if we do not take advantage of prefixing, the concept NBA will be expanded into all its teams (28), and let us assume each team has 14 players. Therefore, we need to maintain 421 (1+ 28 + 28 *14) Boolean conditions in the where clause of SQL query. This explosion will be exemplified by upper level concept like basketball.

7 Experimental Implementation

In discussing implementation we will first, present our experimental setup, and then we will demonstrate power of our ontology-based over keyword-based search techniques. We have constructed an experimental prototype system which is based upon a client server architecture. The server (a SUN Sparc Ultra 2 model with 188 MBytes of main memory) has an Informix Universal Server (IUS), which is an object relational database system. For the sample audio content we use CNN broadcast sports audio and Fox Sports. We have written a hunter program in Java that goes to these web sites and downloads all audio and video clips with closed captions. The average size of the closed captions for each clip is 25 words, after removing stop words. These associated closed captions are used to hook with the ontology. As of today, our database has 2,481 audio clips. The usual duration of a clip is not more than 5 minutes in length. Wav and ram are used for media format. Currently, our working ontology has around 7,000 concepts for the sports domain. For fast retrieval, we load the upper level concepts of the ontology in main memory, while leaf concepts are retrieved on a demand basis. Hashing is also used to increase the speed of retrieval.

7.1 Results

We would like to demonstrate the power of our ontology over the keyword-based search technique. For an example of keyword-based technique we have used the most widely used model-vector space model [19].

7.1.1 Vector Space Model

Here, queries and documents are represented by vectors. Each vector contains a set of terms or words and their weights. The similarity between a query and a document is calculated based on the inner product or cosine of two vectors' weights. The weight of each term is then calculated based on the product of term-frequency (*TF*) and inverse-document frequency (*IDF*). *TF* is calculated based on number of times a term occurs in a given document or query. *IDF* is the measurement of inter-document frequency. Terms that appear unique to a document will have high *IDF*. Thus, for *N* documents if a term appears in n documents, *IDF* for this term *=log(N/n) +1*. Let us assume query (Q_i) and document (D_j) have *t* terms and their associated weights are WQ_{ik} and WD_{ik} respectively for k = 1 to t. Similarity between these two is measured using the following inner product:

$$Sim(Q_i,D_j) = Cosine(Q_i,D_j)$$

$$= \frac{\sum_{k=1}^{k=t} WQ_{ik} * WD_{jk}}{\sqrt{\sum_{k=1}^{k=t} (WQ_{ik})^2 * \sum_{k=1}^{k=t} (WD_{jk})^2}}$$

The denominator is used to nullify the effect of the length of document and query and ensure that the final value is between 0 and 1.

7.1.2 Types of Queries

Sample queries are classified into 3 categories, with each category containing 3 queries. The first category is related to broad/general query formulation such as "tell me about basketball" which is associated with an upper level concept of the ontology. The second category is related to narrow query formulation such as "tell me about Los Angeles Lakers," which is associated with a lower level concept of the ontology. The third category is context query, in which a user specifies a certain context in order to make the query unambiguous, such as Laker's Kobe, Boxer Mike Tyson, and Team Lakers. The comparison metrics used for these two search techniques are precision, recall, and F score. We discuss precision, recall, and F score for individual queries.

7.1.3 Empirical Results

In Figures 5, 6, and 7, the X axis represents sample queries. The first three queries are related to broad query formulation, the next three to narrow query formulation, and the last three queries to context queries. In Figures 5, 6, and 7 for each query the first and second bars represent the recall/precision/F score for ontology-based and keyword-based search techniques respectively.

Although, the vector space model is ranked-based and our ontology-based model is a Boolean retrieval model, in the former case we report precision for maximum recall in order to make a fair comparison.

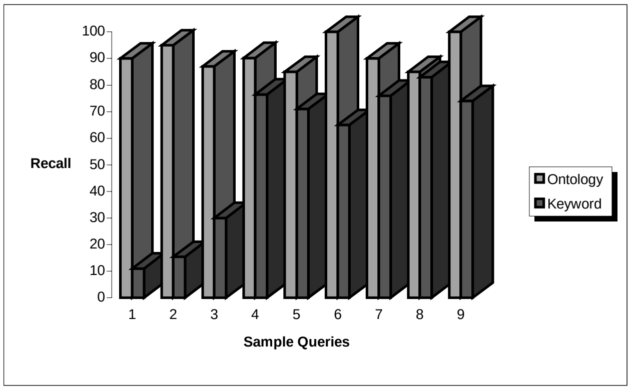


Figure 5. Recall of Ontology-based and Keyword-based Search Techniques

In Figure 5, the data demonstrates that recall for our ontology-based model outperforms recall for keyword-based technique. Note that this pattern is pronounced related to broader query cases. For example, in query 1, 90% verses 11% recall is achieved for ontology-based as opposed to keyword-based technique whereas for query 4, 90% and 76% recall are obtained. This is because in the case of a broader query, more children concepts are added, as compared to narrow query formulation or a context query case. Furthermore, in a context query case, it is usual for broader query terms to give context only. In an ontology-based model these terms will not participate in the query expansion mechanism. Instead, broader query terms will be subsumed under specific concepts. For example, in query 7, the user requests "tell me about team Lakers." Concepts referring to "team" will not be expanded. Therefore, the gap between the two techniques is not pronounced.

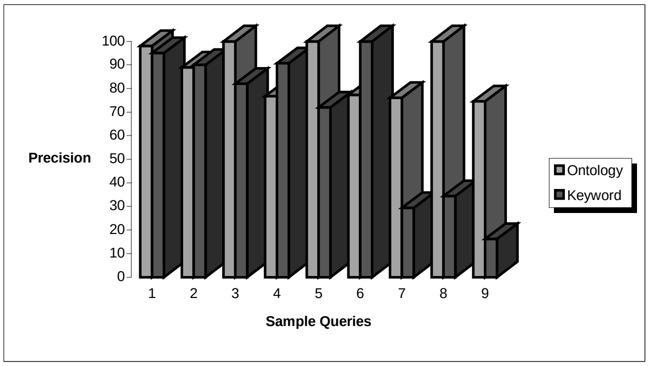


Figure 6. Precision of Ontology-based and Keyword-based Search Techniques

In Figure 6, for broader query cases, usually the precision of the ontology-based model outperforms the precision of the keyword-based technique. This is because our disambiguation algorithm disambiguates upper level concepts with greater accuracy compared to lower level concepts. For example, the disambiguation algorithm for metadata acquisition chooses the most appropriate region for each audio object. Recall that a region is formed by a league, its team, and its players. Thus if a query is requested in terms of a particular league, that is related to upper concept in this region, precision will not be hurt. However, the algorithm might fail to disambiguate lower level concepts in that region (e.g. players). For a narrow query formulation case, the precision obtained in the ontology-based model may not be greater than that obtained through use of the keyword-based technique. In query 4, the user requests "tell me about Los Angeles Lakers." In the ontology-based model the query is expanded to include all this team's players. It might be possible during disambiguation in metadata acquisition for some of these players to be associated with audio objects as irrelevant concepts; in particular when disambiguation fails. Some relevant concepts, such as other players, are also associated with these audio objects. Thus, for our ontology-based model these objects will be retrieved as a result of query expansion, leading to a deterioration in precision. In a keyword-based case, we have not expanded "Lakers" in terms of all of the players on the Lakers team. Therefore, we just look for the keyword "Lakers" and the abovementioned irrelevant objects associated with its group of players will not be retrieved. Thus, in this instance we observed 76% and 90% precision for ontology-based and keyword-based technique respectively.

In the case of the context query, it is evident that the precision of the ontology-based model is much greater than that of the keyword-based model. Since in the ontology-based model some concepts subsume other concepts, audio objects will only be retrieved for specific concepts. On the other hand a search using keyword-based technique looks for all keywords. If the user requests "team Lakers" the keyword-based technique retrieves objects with the highest rank when the keywords "team" and "Lakers" are present. Furthermore, in order to facilitate maximum recall, we have observed that relevant objects will be displaced along with irrelevant objects in this rank. Note that some irrelevant objects will also be retrieved that only contain the keyword "team." Thus, for query 7, levels of precision of 76% and 29% have been achieved.

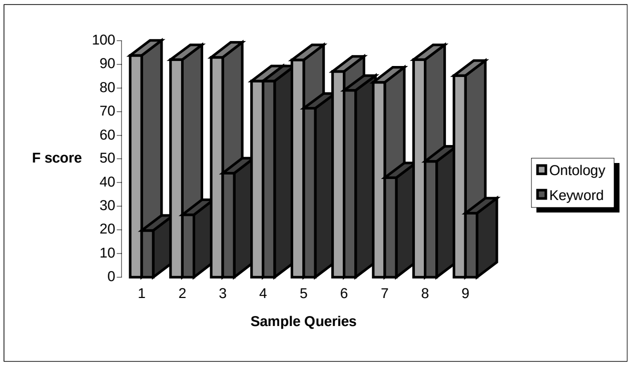


Figure 7. F score of Ontology-based and Keyword-based Search Techniques

Finally, the F score of our ontology-based model outperforms (or at least equals) that of a keyword-based technique (see Figure 7). For the broader and context query case, precision and recall are usually high for the ontology-based model in comparison with keyword-based technique. Therefore, F scores differences, for the ontology-based model are also pronounced. For example, for query 1, the F scores for ontology-based and keyword-based technique are 94% and 20% respectively. For the narrow query case, the F score of our ontology-based model is slightly better or equal to that of the keyword-based technique.

For example, in query 4, we observed a similar F score (83%) in both cases; however in queries 5 and 6 we observed that the F score of the ontology-based model (91%, 87%) outperformed the keyword-based technique, (71%, 79%).

8 Conclusions

In this paper we have proposed a potentially powerful and novel approach for the retrieval of audio information. The crux of our innovation is the development of an ontology-based model for the generation of metadata for audio, and the selection of audio information in a user customized manner. We have shown how the ontology we propose can be used to generate information selection requests in database queries. We have used a domain of sports news information for a demonstration project, but our results can be generalized to fit many additional important content domains including but not limited to all audio news media. Our ontology-based model demonstrates its power over keyword based search techniques by providing many different levels of abstraction in a flexible manner with greater accuracy in terms of precision, recall and F score. Although we are confident that the fundamental conceptual framework for this project is sound, and its implementation completely feasible from a technical standpoint, some questions remain to be answered in future work. These include detailed work on user studies and evaluation. In this connection, we are confident that we will ultimately be able to develop an intelligent agent that will dynamically update user profiles. This will provide a level of customization that can have broad application to many areas of content and user interest.

9 References

[1] S. Adali, K. S. Candan, S. Chen, K. Erol, and V. S. Subrahmanian, “Advanced Video Information System: Data Structures and Query Processing,” *ACM-Springer Multimedia Systems Journal*, vol. 4, pp. 172-186, 1996.

[2] B. Arons, “SpeechSkimmer: Interactively Skimming Recorded Speech,” in *Proc. of ACM Symposium on User Interface Software and Technology*, pp. 187-196, Nov 1993.

[3] G. Aslan and D. McLeod, “Semantic Heterogeneity Resolution in Federated Database by Metadata Implantation and Stepwise Evolution,” *The VLDB Journal, the International Journal on Very Large Databas*es, vol. 18, no. 2, Oct 1999.

[4] R. Baeza and B. Neto, *Modern Information Retrieval*, ACM Press New York, Addison Wesley, 1999.

[5] M. Bunge, *Treatise on basic Philosophy, Ontology I: The Furniture of the World*, vol. 3, Reidel Publishing Co., Boston, 1977.

[6] S. Gibbs, C. Breitender, and D. Tsichritzis, “Data Modeling of Time based Media,” in *Proc. of ACM SIGMOD*, pp. 91-102, 1994, Minneapolis, USA.

[7] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran, “Indexing with WordNet Synsets can Improve Text Retrieval,” in *Proc. of the Coling-ACL’98 Workshop: Usage of WordNet in Natural Language Processing Systems*, pp. 38-44, August 1998.

[8] T. R. Gruber, “A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition,” *An International Journal of Knowledge Acquisition for Knowledge-based Systems*, vol. 5, no. 2, June 1993.

[9] N. Guarino, C. Masolo, and G. Vetere, “OntoSeek: Content-based Access to the Web,” *IEEE Intelligent Systems*, vol. 14, no. 3, pp. 70-80, 1999.

[10] A. G. Hauptmann, “Speech Recognition in the Informedia Digital Video Library: Uses and Limitations,” in *Proc. of the Seventh IEEE International Conference on Tools with AI*, Washington, DC, Nov 1995.

[11] R. Hjelsvold and R. Midstraum, “Modeling and Querying Video Data,” in *Proc. of the Twentieth International Conference on Very Large Databases (VLDB’94)*, pp. 686-694, Santiago, Chile, 1994.

[12] L. Khan and D. McLeod, “Audio Structuring and Personalized Retrieval Using Ontologies,” in *Proc. of IEEE Advances in Digital Libraries, Library of Congress*, pp. 116-126, Bethesda, MD, May 2000.

[13] L. Khan and D. McLeod, “Disambiguation of Annotated Text of Audio Using Onologies,” to appear in *Proc. of ACM SIGKDD Workshop on Text Mining*, Boston, MA, August 2000.

[14] D. B. Lenat, “Cyc: A Large-scale investment in Knowledge Infrastructure,” *Communications of the ACM*, pp. 33-38, vol. 38, no. 11, Nov 1995.

[15] G. Miller, “WordNet: A Lexical Database for English,” *Communications of the ACM*,” vol. 38, no. 11, Nov, 1995.

[16] E. Omoto and K. Tanaka, “OVID: Design and Implementation of a Video-Object Database System,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 5, no. 4, August 1993.

[17] H. J. Peat and P. Willett, “The Limitations of Term Co-occurrence Data for Query Expansion in Document Retrieval Systems,” *Journal of ASIS*, vol. 42, no. 5, pp. 378-383, 1991.

[18] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.

[19] G. Salton, *Automatic Text Processing*, Addison Wesley, 1989.

[20] A. F. Smeaton and V. Rijsbergen, “The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System,” *The Computer Journal*, vol. 26, no. 3 pp. 239-246, 1993.

[21] B. Swartout, R. Patil, K. Knight, and T. Ross, “Toward Distributed Use of Large-Scale Ontologies,” in *Proc. of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems*, Banff, Canada, 1996.

[22] E. Voorhees, “Query Expansion Using Lexical-Semantic Relations,” in *Proc. of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 61-69, 1994.

[23] L. D. Wilcox and M. A. Bush, “Training and Search Algorithms for an Interactive Wordspotting System,” in *Proc. of ICASSP*, vol. 2, pp. 97-100, San Francisco, CA, 1992.

[24] W. Woods, “Conceptual Indexing: A Better Way to Organize Knowledge,” *Technical Report of Sun Microsystems*, 1999.

Incorporating Domain Knowledge with Video and Voice Data Analysis in News Broadcasts

Kim Shearer[*]
IDIAP
P.O. BOX 592
CH-1920 Martigny,
Switzerland
Kim.Shearer@idiap.ch

Chitra Dorai
IBM T. J. Watson Research
Center
P.O. Box 704, Yorktown
Heights
New York 10598, USA
dorai@watson.ibm.com

Svetha Venkatesh
School of Computing
Curtin University of
Technology
P.O. BOX U1987, Australia
svetha@cs.curtin.edu.au

ABSTRACT

This paper addresses the area of video annotation, indexing and retrieval, and shows how a set of tools can be employed, along with domain knowledge, to detect narrative structure in broadcast news. The initial structure is detected using low-level audio visual processing in conjunction with domain knowledge. Higher level processing may then utilize the initial structure detected to direct processing to improve and extend the initial classification.

The structure detected breaks a news broadcast into segments, each of which contains a single topic of discussion. Further the segments are labeled as a) anchor person or reporter, b) footage with a voice over or c) sound bite. This labeling may be used to provide a summary, for example by presenting a thumbnail for each reporter present in a section of the video. The inclusion of domain knowledge in computation allows more directed application of high level processing, giving much greater efficiency of effort expended. This allows valid deductions to be made about structure and semantics of the contents of a news video stream, as demonstrated by our experiments on CNN news broadcasts.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.2.4 [**Database Management**]: Systems—*Multimedia Databases*; H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms

Shot syntax, colour coherence vector, voice clustering

Keywords

Video annotation, domain knowledge, algorithm fusion

1. INTRODUCTION

Research into image databases and image indexing and retrieval has led to the creation of a number of useful tools for similarity retrieval for images [6, 9, 4, 16]. Application of these tools to video is possible, but the principles embodied in the tools do not yield a useful query system. Previous work on video indexing and retrieval [22, 10, 20, 23, 3, 9] has most commonly relied largely on one aspect of video, be it vision or sound, and has been restricted to low-level or undirected processing. The results of this processing are then used for classification, with the goal of detecting either video events, or some form of structure within video. Detection of events or structure permits a summary of the video to be formed, thus permitting more rapid user browsing by a restriction of the information or segments presented for browsing.

Examples of the form of summaries are the Video Icons of Tonomura and Abe [18, 19], the excellent work by Davis [5] on MediaStreams, general systems such as [14, 8, 17] and the scene transition graphs of Yeo and Yeung [21, 2]. These methods aim at presentation of video content in a condensed manner so that the extreme amount of information available may be scanned by the user in a more efficient manner. The scene transition graphs of Yeo and Yeung go slightly further than most earlier work in that they present a possibility for automated deduction of semantically related structure from a video stream.

In this paper we describe a collection of tools and their application to the detection of narrative structure in a news broadcast. In particular, these tools are used to break the broadcast into segments, each of which contains a single topic of discussion. These segments are classified further by labeling each individual shot as one of

- anchor person or reporter,
- footage with a voice over,
- sound bite,

[*] Corresponding author.

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

which gives a clear indication of structure within the video. This work differs from earlier work in that it employs not only low-level processing, but uses results from this processing, along with initial deductions about structure within video, to apply higher level processing in a directed manner. This allows a novel iterative approach to be used, with alternating processing and deduction employing progressively more complex computation as the interpretations become more finely focused. The summary produced from this work can then go further than simply presenting a representative sampling of video, by providing a summary based on the semantics of the content.

The aim of this work is to allow automated annotation of video, which will allow intelligent construction of summaries for large video databases. The particular target area is news broadcast and news magazine footage, such as that kept by major news companies. The annotation created will break the video into segments of homogeneous topic, and further label shots as anchor or report footage. A typical summary that might then be created would be a thumbnail of each anchor person or reporter present in a section of video. The user may then select the reporter who filed a story, rather than having to search for a representative frame which might be contained in the story required. Given the large volume of video data retained for such applications, and the volume captured at each moment, this could result in a large reduction in unproductive human time and lead to a scalable and efficient solution for content management in studios.

2. COMPONENTS AS TOOLS

A number of components may be employed in the analysis of video streams. These components are employed to assess similarity of shots within the video stream, along a number of axes. This similarity within the video stream is then used with a knowledge of shot syntax, and higher level processing, to deduce structure within the news video stream.

2.1 Detection of Anchor Segments

The concept of shot syntax was developed to describe the regular structure of camera parameters employed to capture a particular type of semantic content [2, 21]. The clearest example of regular shot syntax is in interviews. In an interview video it is generally the case that the interview will be introduced by the interviewer. There will then usually be either a shot of the interviewer and the interviewee, or a shot of the interviewee alone. Subsequent shots will be of either; interviewer, interviewee, a mid-range shot of the two people involved, or background footage. This repetitive structure is adopted for interviews as it has been found to be the best method of producing this type of program.

If the assumption can be made that such repeated structure will be present within a video stream of a particular program genre, then detection of repetition in shot settings provides a useful first pass for the grouping of shots into meaningful segments. News broadcast does in general adhere to such a structure, as shown in Figure 1. In this figure solid lines indicate required minimum paths through the syntax diagram, with dashed lines denoting optional paths. The regular structure displayed makes it useful to search for repetitions of anchor or reporter segments. That is, shots with one person addressing the camera, and this person presenting a particular segment of the program, therefore, appearing repeatedly. The term anchor shot will be used to refer to this type of shot, whether it is a shot of an actual anchor person, or a shot of a reporter who is the presenter for a particular story. A story presented (or anchored) by a reporter in the field generally represents a self contained sub–syntax of a larger report. Figure 2 shows a possible syntax for such a segment, the field report presented by a reporter is contained within the dashed line box. The shot syntax for this report is clearly similar to the syntax for a general report.

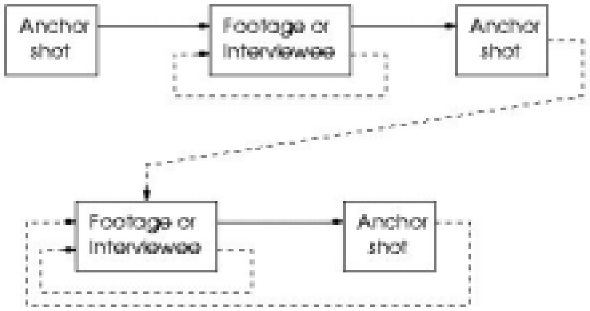


Figure 1: Shot syntax of a broadcast news program.

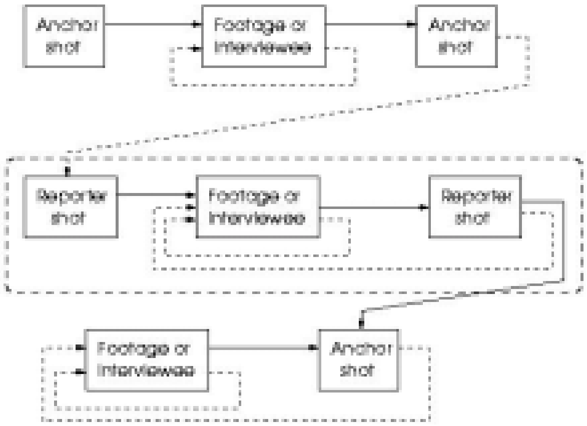
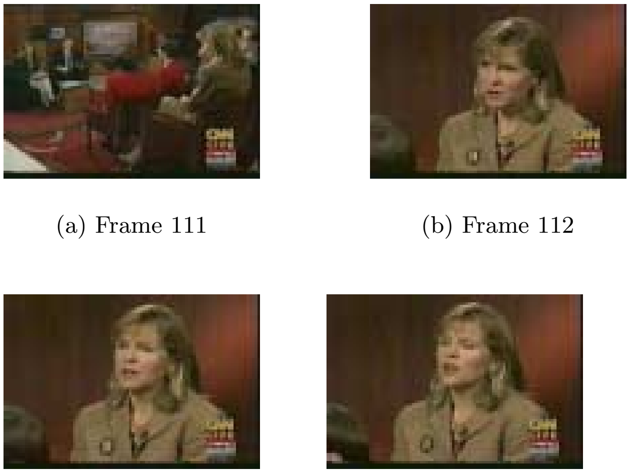


Figure 2: Typical syntax of a news program with a field report.

In our news video processing system, the search for anchor shots takes advantage of a property inherent to such shots. Anchor shots are intended to provide continuity for a news broadcast, which means that the intent of such shots is to present a consistent appearance to viewers. Therefore such shots are captured in a consistent location, with mostly consistent shot parameters. This visual consistency makes detection of repetitions of the anchor simple to detect. Reporters in the field also usually present a highly consistent appearance, however, this is less dependable due to outside factors.

Initially colour coherence vectors (CCVs) [11, 8] were used to detect similarity between frames sampled from a video



(a) Frame 111

(b) Frame 112

(c) Frame 113

(d) Frame 114

Figure 3: Facial rotation for which CCV performs poorly.

Table 1: Similarity measure using CCV for the video frames shown in Figure 3.

	111	112	113
112	5886		
113	25759	25559	
114	7839	4681	25544

Table 2: Similarity measure using spatial histograms for the video frames shown in Figure 3.

	111	112	113
112	71112		
113	71410	5374	
114	70844	8220	5454

to indicate anchor sections. However, CCVs perform poorly with a number of scenarios that occur frequently in news video. The main problem occurs with faces which dominate the frame, and rotate under studio lighting. In these cases the coherence of the colour regions can change dramatically for a small movement. This situation often occurs in anchor shots, where a reporter glances down at a page of notes, or to the left or right to pass to an interviewee or other reporter.

Simple colour histograms provide a useful indication of similarity, but as expected find too many shots to be similar. Using such a global measure allows too many frames of similar colour to be clustered as similar, and will also find frames within a shot that has a great deal of motion similar. For the task of separating anchor and reporter shots from other shots, it is acceptable that motion in the shot, such as the motion apparent in crowd scenes, cause frames to be found dissimilar. The goal is then that each anchor or reporter shot be found coherent (internally similar) and similar to other shots of the same reporter or anchor.

As a result a different similarity measure was employed in our system, where each frame is broken into 12 subframes, and a colour histogram is computed for each. Each histogram is quantized to 16 bins, and histogram difference Δ is a sum of the differences between values for each bin i. That is

$$\Delta = \sum_{i=0}^{15} |H_1[i] - H_2[i]| \qquad (1)$$

The histograms for spatially corresponding subframes are then compared, with the sum of the histogram differences for the subframes representing the distance between frames. The similarity values for the video frames in Figure 3 are given in tables 1 and 2. As can be seen from Table 1, the CCV algorithm finds that frame 111 is far more similar to frame 112 than frame 112 is to frame 113, and also that frame 114 is similar to frames 111 and 112 but not 113. This is due to the changes in colour values for the face and hair of the pictured person in frame 113 as the head tilts slightly. The size of areas containing a particular colour change dramatically with only small head movements. For the same four frames the histogram measure performs much more as expected, easily separating the frames correctly.

In addition to addressing the problem illustrated in Figure 3 the algorithm we employed has another useful property. While each shot of an anchor person or reporter is found to be coherent, most other shots are not. This is due to the sensitivity of the algorithm to overall fluctuations in colour and position of colour. Scenes which might seem likely to be found similar under a colour based measure, such as shots of a crowd, are in fact separated into numerous short pieces. This has the advantage of reducing the number of shots that are detected as repeated shots within a video stream, thus making the task of shot syntax analysis simpler.

There are of course other shots which will be repeated during a broadcast, such as the logo of the news station, advertisements which are repeated and footage used as a preview for stories in later programs. One tool which is often useful in distinguishing these shots from anchor shots is face detection. While face detection is only reliable in constrained applications, it is suitable for this problem. A search for faces in anchor shots will be assisted by the regular presentation of these shots, while advertisements are generally quite erratic and have few static, and therefore detectable, faces.

The face detection part of classification is performed using the CMU face detection software [13]. This is a neural network based face detector, in which neural networks are applied directly to each 20 by 20 pixel location in the image. In order to accommodate scaling transformations the image is presented to the system at actual size, and then repeatedly scaled down by a factor of 1.2 and again presented to the system. Training is accomplished on a set of face images, and non-face images, with false positives in the non-face images being used as negative examples in further training. A number of heuristics are used both to improve accuracy and to improve speed. This system is chosen as representative of the current state of the art in face detection, and its performance is easily sufficient for the given task.

Anchor shots exhibit the following properties which make face detection more reliable:

- the face is turned directly towards the camera,
- the face dominates the shot.

Face detection can therefore be restricted to searching for large faces. The majority of false detections that are artifacts of other parts of the image are small relative to the faces in anchor shots, so size can be used as an effective filter. Searching for only those faces which directly face the camera also simplifies the problem, further reducing the error rate.

Shots that repeat with a suitable shot syntax and have a consistently visible face are highly likely to be anchor shots. The assumption of temporal consistency can be used to further reduce error from face recognition by discarding faces that move rapidly or erratically. This will tend to discard footage of people addressing a crowd, but include field reporters. Reporters in the field will be less static than anchors in the studio, but all field reports in the data set tested were detected as dominant faces. Temporal consistency can also be applied to the colour histogram work by using an average histogram for each group of frames which are considered similar, to represent the matching attribute set. This limits the spread of a single group by preventing a chain of frames with small error from each other remaining part of a single group even though the error diverges further and further from a previous group.

Once these two steps of visual processing have been completed a first pass is performed to determine structure from shot syntax. This yields a preliminary label for each shot as either an anchor shot, or a non-anchor shot. To label the shots in finer detail the sound associated with the video is processed. This presents a difficult problem, as there is no simple method to ensure clean audio samples. While voice recognition in an environment for which extensive training samples are available, and voice samples are well separated can show good performance, this is not the case for this application.

2.2 Audio Analysis

To label the shots in finer detail, the audio associated with the video is analyzed. Much of the sound from news broadcast will contain noise of various forms, such as background noise for field reports. In addition, there are a number of behaviours presented by anchor people, which aid in keeping the flow of dialogue, that prevent clean segmentation of sound samples. One example is that the anchor person will often begin speaking before a field reporter or piece of footage has stopped, which aids flow but makes it impossible to separate one voice from another. In addition, the anchor will generally start speaking before the cut from one shot to another, or will start speaking just after the cut with sound from the previous segment continuing slightly past the cut. This means that most audio samples will contain multiple voices when segmentation of the audio stream is performed.

Previous work has suggested that four seconds is a suitable segment length for vocal samples to exhibit a consistent attribute profile [7], and this is the length employed in this work. Three methods of segmentation for sound were studied for comparison. Two methods attempt intelligent segmentation, the first using silence as an indicator for segmentation points and the second using cuts in the video. The final method employed was to simply cut the video every four seconds starting at the first frame. For each of the first two methods, sections longer than 4 seconds are cut into four second pieces, and segments shorter than 4 seconds are discarded. Segmentation based on silence detection performs significantly worse than either of the other methods, for reasons mentioned earlier. As there is little to choose between the performance of the two other methods, simple fixed time segmentation is used in our system for simplicity.

Audio classification is performed using formant frequency estimators [12, 15] and other low-level attributes as in [1], and k-means clustering. The most suitable number of clusters is chosen by minimizing total error, within a reasonable range. Thus at the end of audio processing, each four second audio segment is assigned an audio cluster label.

3. FUSION OF COMPONENT RESULTS

The three initial pieces of low-level processing are combined to determine the initial classification of shots as either anchor shot, voice over or sound bite using the following rules:

- Anchor shots will be repeated shots with a sequence of not more than 4 shots between, and a time between anchor shots of not more than 8 times the length of the anchor shot. They will also have a prominent face detected.
- Other shots will be initially classified as footage.
- Footage shots with vocal clustering similar to an anchor shot in the same grouping will be determined as voice over.
- Footage shots with vocal clustering dissimilar from any anchor shot in the initial grouping will be labeled as sound bite.

Table 3: Classification results.

	Total number	False positives	False negatives	Accuracy
Anchor shots	44	4	6	79%
Voice Over	54	2	8	82%
Sound Bite	28	4	4	75%

The first rule is also used to break the video stream into segments, with each segment containing a single story topic.

In practice the grouping of shots based on identification of anchor and reporter shots and duration between these shots detects 100% of the structure in the news video. The test set for this work contains two videos of approximately 50 minutes in length each, and includes a number of CNN news and magazine style programs. The structure detected represents a slight over segmentation, in that some reports have the anchor shot which introduces the segments, and the anchor shot concluding the segment discarded. This is due to the segment being anchored by a reporter, and thus exhibiting the shot syntax expected within the report (Figure 2), with the introductory and concluding segments being no more than a tie–in to the news program. It is deemed reasonable that these shots be discarded. The important feature of the segmentation is that no segment contains more than one topic, which could result in hiding of information from the user.

Table 3 gives a summary of the results from classification using the initial low-level processing and shot syntax. As can be seen, detection of shot syntax allows accurate classification of most of the video. The values in the accuracy column of Table 3 are calculated from the equation

$$\text{Accuracy} = \frac{Actual - F_{neg}}{Actual + F_{pos}} \qquad (2)$$

where *Actual* is the correct number of samples for the shot type, and F_{neg} and F_{pos} are the number of false negatives and false positives for the classification. The majority of the misclassifications are due to too few sound samples being available for accurate audio classification of a shot. The false negatives for the anchor shots are due partly to the lead and trailing shots of a long report being dropped as discussed earlier, and also to one group discussion having two presenters. The anchor shots for this section are detected as similar, but have no single dominant face. Further processing discussed in later sections in this paper could be used to improve detection to include this case.

4. DIRECTED APPLICATION OF HIGH LEVEL PROCESSES

Given this initial segmentation of the shots within the video stream into structured blocks, further processing may now be considered. The main additional processing is a more detailed face detection pass applied to the shots classified as footage. This allows *interview* shots to be more accurately detected.

Table 4: Interview shot detection.

	Total	False positives	False negatives
Sound bite	4	8	0
Interview	24	0	8

Allowing a greater range of sizes for a face increases both the time required, and the error rate for face detection. However, when footage is taken in the field it is less likely that an interview shot will show a dominant face front on. In this case greater care must be taken in assessing the results from the face detection algorithm. Results are examined closely for consistency of location and size of faces that are detected. Erratic size and or location can be sufficient to discard a face from consideration. Any shot which presents a single consistent face for the majority of the shot is labelled as a reporter.

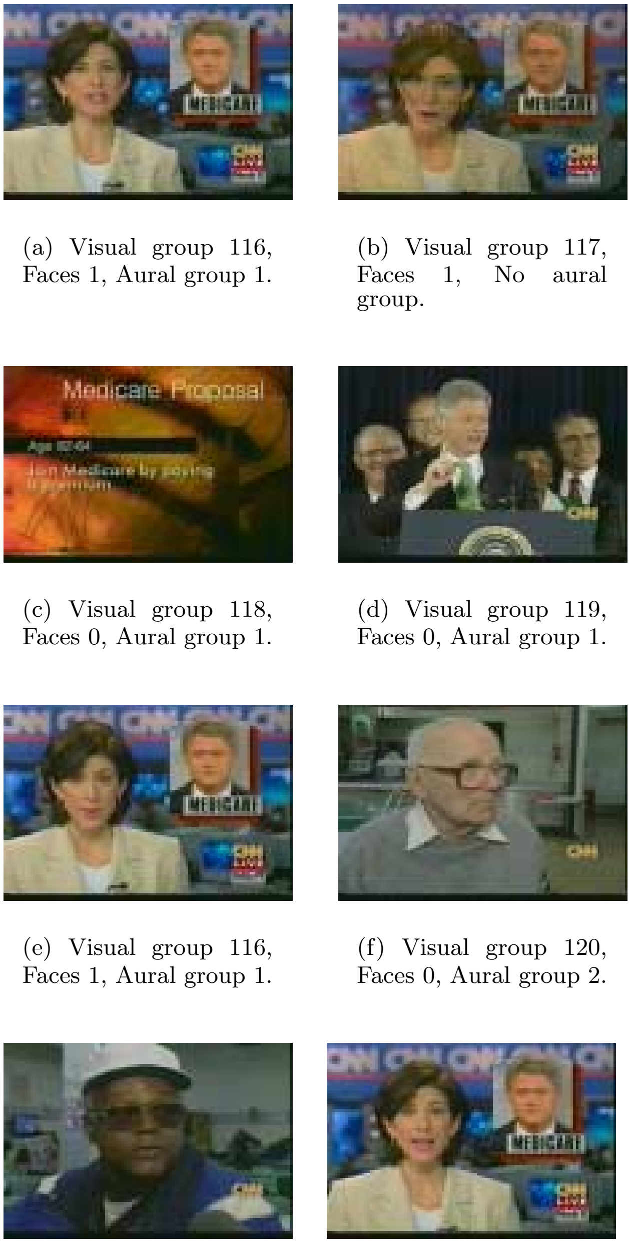
The result of this further classification applied to the sound bite shots is given in Table 4. These results indicate that the detection of faces in these shots is still less than perfect, however, two thirds of the interview shots were detected. Given this level of recognition further classification can be performed as determined by shot syntax.

Further processing could be employed to specifically search for faces that are not perpendicular to the camera, which could add to the accuracy of this second step. In particular shots which are likely to be part of an interview segment, and which have no dominant face, could be tested for two faces. This would help detect the interviewer and interviewee shots, which would add further weight to the classification of such shots. This is intended as future work.

5. RESULTS

Figure 4 shows the thumbnails for the shots from one segment of detected structure. The caption for each thumbnail gives the visual similarity group computed using segmented colour histograms, the number of faces detected using the CMU face detection software [13], and the similarity group from aural clustering for the shot.

The topic of the segment is a report on the public view of the Medicare bill recently introduced in the USA. There is an anchor shot (Figure 4(a)), followed by a shot of only one sample which coincides with a fade (Figure 4(b)). This shot would be discarded from consideration. There is then a shot of explanatory text (Figure 4(c)), which is correctly identified as a voice over. The next shot (Figure 4(d)) is of Bill Clinton addressing a group of reporters, this is identified as a voice over due to incorrect vocal clustering. No face was detected due to the mobility of the speaker around the stage. Figure 4(e) shows another anchor shot, which is correctly identified. Figures 4(f) and 4(g) are of "people on the street", interviewed about their views on the topic. They are correctly identified as separate pieces of footage, and labelled as sound bites. In both cases the camera parameters are too irregular to expect face detection. The final figure, Figure 4(h) is the closing anchor shot, and is identified as such.



(a) Visual group 116, Faces 1, Aural group 1.

(b) Visual group 117, Faces 1, No aural group.

(c) Visual group 118, Faces 0, Aural group 1.

(d) Visual group 119, Faces 0, Aural group 1.

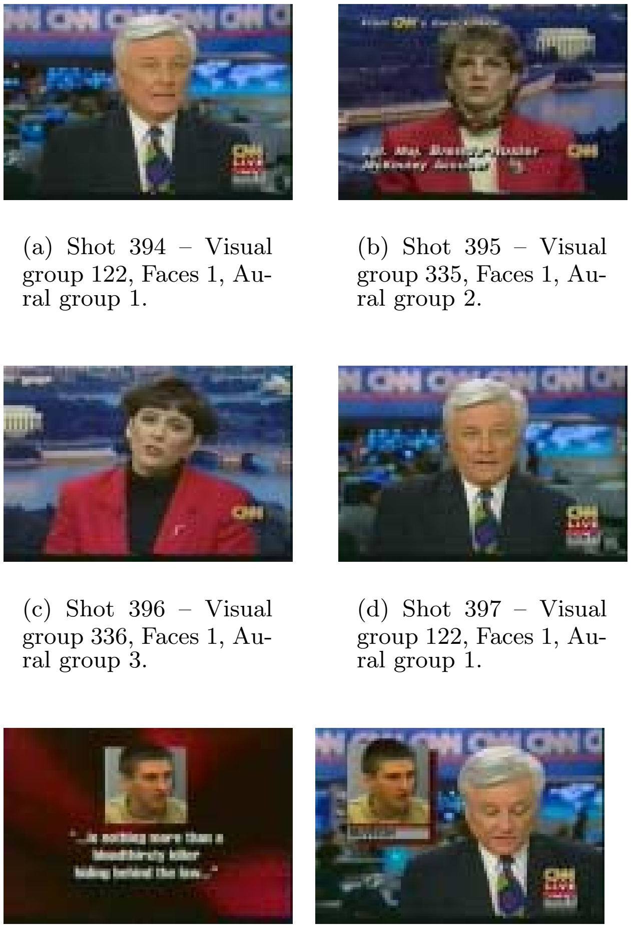
(e) Visual group 116, Faces 1, Aural group 1.

(f) Visual group 120, Faces 0, Aural group 2.

(g) Visual group 121, Faces 0, Aural group 3.

(h) Visual group 116, Faces 1, Aural group 1.

Figure 4: Structure in an example news program.


CNN
"...is nothing more than a bloodthirsty killer hiding behind the law..."

(a) Shot 394 – Visual group 122, Faces 1, Aural group 1.

(b) Shot 395 – Visual group 335, Faces 1, Aural group 2.

(c) Shot 396 – Visual group 336, Faces 1, Aural group 3.

(d) Shot 397 – Visual group 122, Faces 1, Aural group 1.

(e) Shot 398 – Visual group 337, Faces 0, Aural group 1.

(f) Shot 399 – Visual group 122, Faces 1, Aural group 1.

Figure 5: Thumbnails of a news report with male anchor.

As can be seen, the clip of Bill Clinton (Figure 4(d)) is classified as a voice over, rather than a separate piece of footage. This is in part due to the brevity of the shot, and in part due to the noise and length of pause in the spoken voice. Improved audio processing would perhaps reduce this difficulty. However, it must be assumed that many of the voices which occur in these shots will be unseen. While some people are regularly included in news bulletins (Bill Clinton as President), many others will be involved in news for only a brief period, corresponding to the time of a particular event and story. Moreover, the "people on the street" interviewed are intended to be random choices. This makes the task of separating such voices from each other more difficult. A further difficulty observed is that the anchor people will have numerous samples of their voice present, and any agglomerative classification method should associate these. The smaller groups of other voices, often with only a small number of samples, and the samples containing multiple voices, make it difficult to distinguish between outliers and separate samples.

Table 5: Vocal (dis)similarity for shots in Figure 5.

	395	396	397	398	399
394	0.747	0.720	0.054	0.102	0.142
395		0.340	0.958	0.887	0.907
396			0.931	0.860	0.881
397				0.167	0.228
398					0.005

An example where voice classification does work well is shown in Figure 5 and Table 5. This sequence of shots shows a male anchor person, Lou Waters, presenting a story on harassment, with two people interviewed (Figures 5(b) and 5(c)) and a commentary over a still (Figure 5(e)). Table 5 presents value of the distance measure used in audio similarity detection for the six shots. The values for the comparison of the two interviewees to the anchor person are clearly separable from those for the comparison of anchor person shots, with a range of $[0.72 - 0.958]$ compared to a range of $[0.005 - 0.228]$ for the similar shots. The voices of the two interviewees are quite similar, and could reasonable be clustered together, their dissimilarity value of 0.34 is classified by the system as similar.

Figure 5 also provides a further example of the frame similarity algorithm, with the shots in Figure 5(f) containing an extra image, but still being found similar to the earlier anchor shots. In addition to this the two shots of interviewees, although visually quite similar are correctly separated. Figure 5(f) also gives an additional example of the type of head movement which is misclassified by the CCV algorithm.

6. CONCLUSIONS

The process employed in this work combines a number of image and aural low–level processes that, in isolation, are unreliable for classification of video. The fusion of the results of these processes, together with knowledge of the shot syntax for a particular domain, leads to a reliable and high level structure labeling of the video. While the resulting classification is less than perfect, all significant structure is recognized, albeit slightly over segmented.

The segmentation produced separates shots into homogeneous story segments, and is able to identify the shots which contain anchor people and reporters. The ability to extract the shots containing reporters and anchors is particularly important, as this provides a powerful key for access to the video content. This gives a suitable starting point from which a summary may be produced without hiding information from the user.

Further processing, such as the proposed refinement of face detection, would allow extraction of more detailed structure. Detection of interviewer and interviewee shots in interview segments would allow not only the presenting reporter to be identified visually as a key, but also the interviewee.

Further visual processing in the form of text detection and recognition is a possible future extension. Improvement to the audio processing is also an avenue for increasing the accuracy of the system, and perhaps allowing further information to be extracted. Given key words recognized from audio, and text recognized from video such as can be seen in Figure 4(c), further fusion of results may be useful for improving recognition of these stages.

The inclusion of shot syntax as a model for structure within news video is a major advantage for detection of shot type. This allows the extension of simple attribute based indexing to deduction of semantic structure within video, and the separation of video into segments of homogeneous semantic content. Extraction of semantic segments and deduction of shot type from a video stream greatly increases the utility of a video warehouse. Currently research is being undertaken to examine how well the shot syntax concept generalizes to other forms of video. Interview and news footage have a very regular shot syntax, but there are other forms of video with regular shot syntax which might be detected using similar techniques, or by application of additional measures. Research is also being undertaken to determine methods for the deduction of shot syntax structure from samples of a particular video form. Such a process could be of great value in multimedia and video data mining.

7. ACKNOWLEDGMENTS

The authors would like to acknowledge the assistance of CNN for providing news footage that formed the base data set for this work.

8. REFERENCES

[1] T. Blum, D. Keisler, J. Wheaton, and E. Wold. Audio databases with content–based retrieval. In M. Maybury, editor, *Intelligent Multimedia Information Retrieval*, chapter 6, pages 113–135. The MIT Press, 1997.

[2] R. Bolle, B.-L. Yeo, and M. M. Yeung. Video query and retrieval. In *Advanced Topics in Artificial Intelligence*, volume 1342 of *Lecture Notes in Artificial Intelligence*, pages 13–24. Springer, December 1997.

[3] A. Cheyer and L. Julia. MVIEWS: multimodal tools for the video analyst. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 55–62. ACM, January 1998.

[4] J. M. Corridoni, A. Del Bimbo, and P. Pala. Image retrieval by color semantics. *Multimedia Systems*, 7(3):175–183, May 1999.

[5] M. Davis. Media streams: An iconic visual language for video annotation. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 196–202. IEEE, April 1993.

[6] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28(9):23–32, September 1995.

[7] H. Gish, M. Sui, and R. Rohlicek. Segregation of speakers for speech recognition and speaker identification. In *ICASSP–91*, pages 873–876. IEEE, IEEE, 1991.

[8] R. Lienhart, S. Pfeiffer, and W. Effelsberg. Scene determination based on video and audio features. In *Proceedings IEEE Multimedia 99*, pages 685–690, Firenze, June 1999. IEEE.

[9] W. Y. Ma and B. S. Manjunath. NeTra: A toolbox for navigating large image databases. In *Proceedings of the International Conference on Image Processing*, pages 568–571, 1997.

[10] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura. Video handling with music and speech detection. *IEEE Multimedia*, 5(3):17–25, July 1998.

[11] G. Pass, R. Zabih, and J. Miller. Comparing images using colour coherence vectors. In *Proceedings ACM Multimedia 96*, pages 65–74, Boston, November 1996. ACM.

[12] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Signal Processing Series. Prentice Hall, 1978.

[13] H. A. Rowley, S. Baluja, and T. Kanade. Neural network–based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.

[14] C. Saraceno and R. Leonardi. Audio as a support to scene change detection and characterization of video sequences. In *Proceedings of ICASSP 97*, pages 2597–2600. IEEE, IEEE Computer Society Press, 1997.

[15] K. Shearer, S. Venkatesh, and C. Dorai. Attribute based discrimination of speaker gender. Technical Report 4, Curtin University of Technology, GPO Box U1987, Perth 6001, Western Australia, November 1999.

[16] D. M. Shotton, A. Rodriguez, N. Guil, and O. Trelles. Analysis and content–based querying of biological microscopy videos. In *Proceedings of the 15th International Conference on Pattern Recognition*. IAPR, IAPR, 2000.

[17] S. Srinivasan, D. Petkovic, and D. Ponceleon. Towards robust features for classifying audio in the CueVideo system. In *Proceedings of ACM Multimedia 99*, pages 393–400. ACM, ACM, 1999.

[18] Y. Tonomura and S. Abe. Content oriented visual interface using video icons for visual database systems. In *IEEE Workshop on Visual Languages*, pages 68–73. IEEE, 1989.

[19] Y. Tonomura, A. Akutsu, K. Otsuji, and T. Sadakata. VideoMAP and VideoSpaceIcon: Tools for anatomizing video content. In *INTERCHI 93 Conference Proceedings*, pages 131–138, 1993.

[20] S. Tsekeridou and I. Pitas. Audio–visual content analysis for content–based video indexing. In *IEEE International Conference on Multimedia Computing and Systems*, pages 667–672. IEEE, IEEE, 1999.

[21] B.-L. Yeo and M. M. Yeung. Classification, simplification and dynamic visualization of scene transition graphs for browsing. In *Storage and Retrieval for Image and Video Databases VI*, pages 60–70. SPIE, December 1998.

[22] M. Yeung, B.-L. Yeo, W. Wolf, and B. Liu. Video browsing using clustering and scene transitions on compressed sequences. In *Proceedings of the SPIE*, volume 2417, pages 399–413. SPIE, 1995.

[23] S. J. Young, M. G. Brown, J. T. Foote, G. J. F. Jones, and K. S. Jones. Acoustic indexing for multimedia retrieval and browsing. In *ICASSP 97*, volume 1, pages 199–202. IEEE, IEEE, 1997.

Multimedia Support for Complex Multidimensional Data Mining

Monique Noirhomme- Fraiture
Institut d'Informatique, FUNDP
Namur, Belgium
mno@info.fundp.ac.be

ABSTRACT

ISO-3D project aims to develop tools in order to analyse and represent business information from large collection of data. The designed tools are mainly representation tools using 3D graphics, sound and animation.

In this paper, we will present two of the graphic representation tools which use also sound and animated picture. We will explain why and how we use sound and animation with graphical representation in this data mining approach.

Keywords

Symbolic Objects, Visualisation, Sound, Animation.

1. INTRODUCTION

Economical data usually depend upon time but are also explained by an important number of heterogeneous variables. When recorded systematically, like in portfolio management (for banks) or audience monitoring (for TV), they constitute rapidly huge data bases. Managers need tools for extracting daily knowledge from these data.

Symbolic Analysis is a kind of data analysis which is able to deal with complex multidimensional heterogeneous data and to summarise information. It is broadly accepted that graphic representation facilitates interpretation of results.

Development of multimedia techniques allow now other representation means than visualisation by graphics; we think here to sound and animation. Moreover, in order to take quick decisions, it is necessary to work without delay on data which are often stocked on another site. It is why network technology is used.

ISO-3D project aims to develop tools in order to analyse and represent business information from large collection of data integrating also those new techniques. The prototype is running in a client/server architecture called Infobus, based on more recent network technology. The designed tools are mainly representation tools using 3D graphics as well as sound and animation.

In this paper, we will present two of the graphic representation tools which use also sound and animated picture. Before that, we will explain what is Symbolic Analysis in order to understand the data in input of our representation process for introduction of multimedia support and we will develop some arguments for the use of sound in data exploration.

2. COMPLEX OBJECTS CALLED SYMBOLIC OBJECTS

The standard methods of statistical data analysis accept as input, « individuals » by « variables » matrix. Each cell (i,j) of such an array contains the value taken by individual i for variable j. The value is said to be « atomic » in the sense that it is not a list or a set of values.

The Symbolic Data Analysis [BOCK&00] extends the input data structure to « individuals » by « variables » arrays where the value taken by an individual on a variable may be non-atomic, but possibly a set of values, intervals of values or a probability distribution.

For example this values can be

1. A set of quantitative values : [Age = {15, 22, 45, 47}].

It means that the age of family members are 15, 22, 45 and 47.

2. A set of categorical values : [TV preference = {RAI1, R4}]

(notice that standard quantitative and categorical values are special cases of 1 and 2).

3. An interval : [Age = [15, 47]] which means that age in the family is between 15 and 47.
4. A set of weighted categorical values : [TV preference = {RAI1 (0.3), R4 (0.7)}] which means that 30 % of the family has daily preference for RAI1 channel and 70 % of the family has daily preference R4 channel.

We will give the name of Symbolic Object (SO) to a row of non atomic value.

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

3. VISUAL REPRESENTATION

From user requirements in SODAS [NOIRHOMME&00] and ISO-3D projects, we know that users need to visualise a Symbolic Object (SO), a SO inside the reference population, various SOs in principal components space, SO during time and important changes, they also need to point out particular SOs on other graphs (like hierarchies, result of SO classification) and visualise them.

In ISO-3D, to meet these requirements, we have suggested the Temporal Star and the Simple Star graphical representation.

First, we will remember the principles of the Zoom Star [NOIRHOMME&00].

The **Zoom Star** representation is a radial graph where each axis corresponds to a variable .

We allow variables in intervals, multivaluate values, weighted values to be represented. We chose conventions for axes representation (colour, dots). A two dimensional and a three-dimensional representation have been designed.

In the 2D Zoom star, axes are linked according to each variable values (see figure 1 of Simple Star). Most weighted value of a categorical variable are linked as well as extremities of intervals. A surface is drowned by joining extremities of intervals or points of highest weight. This representation does not allow details about distributions associated to weighted categorical variables. However, the user can ask for the complete distribution to be displayed in another window by selecting the axis.

On the 3D Zoom Star representation, distributions corresponding to each weighted values are shown directly on the axis. We provide the user with the opportunity to animate the graphic by turning the picture around a vertical and an horizontal axis in order to make easier the retrieval of pertinent information. On the 3D representation, axes are not linked because the image is continuously changing due to animation feature. The iconic representation is then meaningless in that case.

Let us note that evaluation of the Zoom Star has been made in the past with students [NOIRHOMME& 98] and that validation has been made by many users of SODAS project [Bisdorff99] and by researchers[MENNESSIER&98].

Whereas users had no preliminary experiment with this kind of representation, it appeared that the Star was user friendly and answered user problems.

3.1 Simple Star

The Simple Star will be mostly used to show the detail of a Symbolic Object represented on a preliminary graphic (like Temporal Star or classification representation).

The Simple Star is analogue to the Zoom Star in 2D but the functionality are slightly different.

To be consistent with Temporal Star, some conventions of colour and colour shade are used.

Several stars representing different objects can be superposed (figure 1).

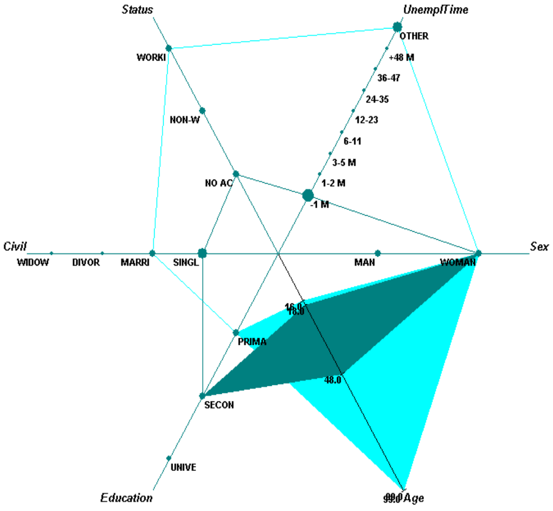


Figure 1. Two Superposed Simple Star.

3.2 Temporal Star

The aim of the Temporal Star is to represent a Symbolic Object at different epochs.

A star in perspective represents the symbolic object at a given epoch (like in the 3D Zoom Star).

The stars at different epochs are thread on one axis representing time (figure 2). This figure can be moved and zoomed. Different colours can be chosen for the different axes. On each axis, representing a categorical variable, the histogram can be shaded, coloured. To emphasise the evolution from one epoch to another, when the axis represents a quantitative variable (mean or intervals), the extremities of the intervals (min, max) or means can be joined. and a transparent veil will be added to the stars (on demand).

It must be also possible to select a particular star on the thread and to display it on the form « Simple Star » (see figure 2).

4. USE OF SOUND IN GRAPHICAL DISPLAY

Most modern graphical displays are highly visually demanding because all information is graphically presented. Our visual sense on the other hand has a rather small area of high focus. A problem arises when a user must concentrate on the visual feedback from one part of the display, so that feedback from another part of the display may be missed as it is outside the area of visual focus. As the amount of information contained by the visual display increases, it may arise that the user become overloaded and the display ineffective. This problem has been extensively documented in recent research, for example by S.A. Brewster who studied auditory enhancements to tool palettes [BREWSTER98a], graphical buttons [BREWSTER&95], etc.

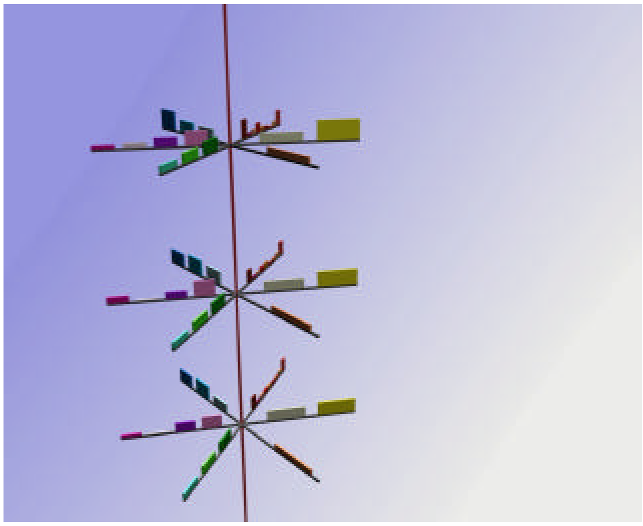


Figure 2. Temporal Star

Another problem arises when representing multi-modal information. In real life, a person is interacting with more then only visual sensations. Most Human-Computer Interfaces represent however only one mode, namely the visual mode. Since it is impossible to convey all the information received by all modes using only visual representation, the user would benefit from a multi-modal interface, because of an increased amount of information and a more natural way of representing it. [ANRIJS99]

How sound can improve the graphical interface?

In [KRAMER94], G.Kramer distinguishes several advantages that result from combining the graphical interface with sound.

The main advantage is that sound can be integrated without interfering with the visual display. Sound is eyes-free. This *nonintrusive enhancement* offered by sound ensures that the user does not become overloaded. It even decreases the overload because sound can replace insufficient or inappropriate visual cues

When representing data, sound offers high dimensionality of up to seven dimensions [CONVERSY&95] and, compared to the visual display, offers *a superior temporal resolution* and *complementary pattern recognition abilities*. The human ear is more sensitive to changes than human vision and perceives the same information in a completely different way, thus providing new and complementary ways to detect occurring trends and relationships in datasets. Moreover, this complementary information leads to *intermodal correlations* where patterns detected by one sensory display are confirmed and verified by the other sensory display.

Another advantage (not mentioned by Kramer) is that short-term memory processes (and therefore remembers) auditory and visual data in a different way. Auditory data has been shown to be remembered longer than visual data.

A sonic interface has one important disadvantage, namely that, unlike the human eye, which can ignore visual stimuli by moving its focus or simply closing its eyelid, a human ear cannot ignore auditory stimuli. In an office environment, this can cause severe problems as users are disturbed by sounds coming from other users.

The most obvious solution is the use of headphones, but this isolates the user from events and communication in the office and is thus undesirable. A method using only one earphone has been proposed and tested with satisfactory results. With this method the right ear has an earphone and captures information coming from the computer, while the left ear is free to capture information from the office.

5. USE OF MUSIC IN GRAPHICAL DISPLAY

The potential of music as an output medium has hardly been examined. However music is the most sophisticated medium of the auditory media. The information is highly organised into complex structures and sub-structures. It can then be used to transmit complex information to a user. As emphasised by James L. Alty [ALTY95], music has some advantages on other media :

« - music is all-pervasive in life. It is very memorable and durable. Most people are reasonably familiar with the language of music in their own culture. Once learned, tunes are difficult to forget ».

« - music involves the simultaneous transmission of complex ideas related over time, within an established semantic framework ».

Let us add two other related arguments:

- Music is often used to transmit information, even if we are not as conscious of that phenomenon as for video channel. We think here about the use of music and sound in movies.

- Western music is organised mathematically with ratios, equal frequency differences between sounds, predefined duration (note value, silence value). It is then not surprising that mathematical structure, and numerical data can be transmitted through music.

Music has a large variety of parameters. Besides characteristics for elementary sounds like pitch, timbre, loudness, duration, reverberation and location, music uses chords (harmony), rhythm and polyphony. This one is very helpful because it can be used to transmit several information simultaneously.

We could think that only musicians are able to perceive musical sound with accuracy.

In [ALTY95], J. Alty describes experiments about perception of tones. Of course, he observes differences between individuals but he can conclude the following :

- Human beings can perceive numerical difference between two tones, usually to about ±1 tone within the octave though the accuracy decreases away from their reference point (large or small intervals are better perceived than middle ones).
- Subjects seem to be able to follow and remember the general pattern of a tune, but the magnitude of variation vary considerably from one to another.
- Rhythm appear as a candidate for improving intelligibility. More rapid presentation of the sequence could help. Not surprisingly, response is more accurate when the notes are organised into a tuneful sequence than when they are sorted.

6. USE OF SOUND IN ISO3D

We have seen the interest of using sound in Human-Computer Interface. The problem is now to decide when choosing graphic or audio. Audio can replace graphic in special cases : when the user is concentrated on a task which does not allow him any distraction by a screen, when the user is blind or is not able to see, for any kind of reason (for example, lack of light).

In ISO 3D, these reasons have not been identified in the user requirements, (but it could happen that some users are partially-sighted). It is thus not appropriate to use audio alone. We have decided to use audio in complement to graphic representation. Graphic representation is the normal way to summarise data. Sound is not usual at all for statisticians. We think that users will be less disturbed if we use sound as a complement to graphic representation.

Let us add that in a previous survey for SODAS project we have asked to statistical users if they were ready to use 3D, colour, sound. The answer was positive for 3D and colour but negative for sound. We have then to take into account a certain negative a priori against sound. Perhaps, that if we ask the same question to younger users, from video game generation, the answer would be different, but at the present moment, with present users, we have to introduce sound carefully. For example, sound cannot be imposed but can be chosen at the beginning of the session.

We have tried, in absence of sound user requirement, to imagine for what kind of information sound could be used. In systems supervision or process control, sound is often used to attract attention on particular problems.

In complex multidimensional data representation, situation is different. The more accurate problem is to represent a large amount of information : large amount of variables of different types, given generally not by single value but by an interval or a distribution. It is then difficult to give all the information at once. In graphic interface, we choose an interactive approach but sound can help to give some kind of information on demand. We have chosen characteristics which seemed hard to be given visually.

6.1 The type of axes

The variable can be numerical, given by an interval of value, categorical, given by a distribution and non applicable, when the axis is represented but is not used for the present object.

A warning can be given on the axes type, when selecting an axis. This warning will be given in the form of a musical earcon. Different sounds have then to be designed for each type and must be very distinct. Rhythm chord and timbre will be important elements to improve the disambiguation between sounds

6.2 Dissimilarity between two symbolic objects

When the symbolic objects are represented on a Temporal Star, linked by a central thread representing time or in different windows, like in the Simple Star way, user can be interested to know in which amount the objects differ from each other.

The first step is to compute a dissimilarity measure between both objects and then to show dissimilarity on the form of a number.

Process can be lighten if dissimilarities are computed automatically, at the beginning of the session, and given to users, in audio way, on request, by a selection click on both objects. A musical sound will be played, with duration proportional to dissimilarity value.

In a first approach, we have divided the range of dissimilarity measure into five equal intervals and to map each of these intervals into a sound with proportional number of chords or notes. For example, if the dissimilarity has a value of 3 (in a scale between 1 to 5), three notes or chords are played successively.

Let us note that sounds for dissimilarity must be very distinct from earcons for axis type.

6.3 The weight of axes

When representing axes in a radial shape, there is no first axis, no last one and axes order is usually arbitrary. When variables are the result of a Principle Component Analysis, they are given with a weight corresponding to the percentage of information (or inertia) explained by each one. This element is important but is usually given in a table, annexed to the graphical representation.

It could be much more helpful using sound. When moving the cursor over an axis, a sound according to the importance of the axis will be played. The visual display will remain unchanged and the user will not have to shift his focus off the symbolic object.

Earcons or musical sounds are of different length/height according to the weight of the axis. A mapping between weight and pitch is tried.

7. ISO3D CHOICE TO GENERATE SOUND

Several decisions had to be taken in order to create sound inside ISO-3D software.

The first step was to make the choice of the type of sonic representation. We have chosen musical sound instead of speech, natural or virtual sound. In preceding paragraphs , we have explained most of the reasons for choosing music but we have to admit that it is also the result of a personal subjective choice.

The second step concerns the way to create musical sound: to copy sound in wave format and transform it into MP3 standard or to design our original sounds.

Copying sound to create musical earcons would be like copy parts of paintings to create icons. It would not be well appropriated, not sufficiently flexible, not easy to tune, not master. We then decided to create our own musical sound.

At the third step, we have to decide who will design the sound. An orchestra ? It is out of our budget ! A musician ? We know some very well.

The first possibility would be to record (in audio or wave format) and then digitalise and compress the result. This solution needs studio record material.

An alternative, and it is the solution that we have chosen, consists to use a synthesiser which transforms directly a large variety of sounds into MIDI format files. This solution allows a large range of timbres, tones, which cannot be obtained with a single instrument. It can also generate polyphony.

The last step concerns hardware/software synthesiser. For our musician, who is pianist, it is much more natural to use keyboard then to tune the sounds on a computer. Moreover, virtual or software synthesiser need still some improvement to have the same quality as hardware synthesiser. The advantage of external production on internal one is that these sounds are easy to integrate in the application and that they are hardware independent, meaning that, no matter the used soundcard is, the sounds produced are the same. The disadvantage is that only few parameters, like balance and loudness, can be modified while using the musical sounds in the application tool.

To design a more adaptable product, we have decided to offer a library of sounds, with a choice by default (like when you choose your bell ring on your portable telephone or the colour of screen).

8. ANIMATION

Animation of an object can help to understand his evolution along time.

We have used this technique to show the evolution of an object represented by a Simple Star. Considering, for example, the data of TV audience of a family recorded by minutes, if we superpose quickly the pictures of Simple Star for each minute, we obtain an animated picture which shows the evolution of family preference all along the day.

9. SOFTWARE DESIGN

A first version of the software has been developed using Java 2.0, Open Inventor, 3D-MasterSuite and Java Media Framework 2.0 (JMF). It runs in a client/server architecture on the web (InfoBus technology).

OpenInventor is a 3D graphic API using OpenGL standard. 3D-MasterSuite extends and includes Open Inventor and high level 3D graphic classes. We have used the Java version so that , when necessary, objects not available in 3D-MasterSuite can be programmed in Java.

The JMF is an application programming interface (API) for incorporating media data such as audio and video into Java applications and applets. It is specifically designed to take advantage of Java platform features and supports several audio and video formats. JavaSound has been incorporated into JMF (which has been lately incorporated into Java 1.3).

The JMF 2.0 API extends the framework by providing support for capturing and storing media data, controlling the type of processing that is performed during playback, and performing custom processing on media data streams. In addition, JMF 2.0 defines a plug-in API that enables advanced developers and technology providers in more easily customising and extending JMF functionality.

To implement the sound, we had to make a link between Open Inventor and JMF. A special sound object has been implemented. Its task is to:

- initialise the soundcard
- to load a sampled file (e.g. WAV) or a MIDI file
- to play that file

10. EVALUATION

Next step is to evaluate the prototype with the users who provided pilot applications. It will be done for November 2000. These applications are TV Audience in RAI (Italy) and portfolio management in ING (The Netherlands).

11. ACKNOWLEDGMENTS

ISO-3D project is an ESPRIT project sponsored by European Community. Partners come from universities, private companies and a bank. It is a two years project and it started in November 1998.

Participants:

Universities : Paris-Dauphine (France), Naples (Italy), Namur (Belgium), Liubljana (Sloveny).

Companies : Matra System & Information, TGS (France), CISIA (France).

Bank : ING (The Netherlands)

Open Inventor environment has been furnished by TGS Company for project duration.

Many thanks to Adolphe Nahimana and Benoît De Greift who worked hardly on software programming, to Koen Anrijs who designed a first prototype with Java Sound and prepared the state of the art.

12. REFERENCES

[1] [ALTY95] Alty, J.(1995). *Can we use Music in Computer- Human Communication?*, HCI 95, UK.

[2] [ANRIJS99] Anrijs, K. (1999) *The use of sound in 3D representations of Symbolic Objects*, Mémoire de Licence en Informatique, Institut d'Informatique, FUNDP.

[3] [BISDORFF00] Bisdorff, R.(2000). *Illustrative Benchmark Examples*, chap. 13 in Bock,H.H., Diday, E., *Analysis of Symbolic Data.Exploratory methods for extracting statistical information from complex data.* , Springer Verlag, Heidelberg, pp 355-385.

[4] [BLY85] Bly, S. (1985). *Communicating with Sound*, Proceedings of CHI'85 Conference on Human Factors in Computing Systems, pp 115-119, ACM.

[5] [BOCK&00] Bock, H.H., Diday E.(eds.) (2000).*Analysis of Symbolic Data.Exploratory methods for extracting statistical information from complex data.* Springer Verlag, Heidelberg, pp 54- 75.

[6] [BREWSTER&95] Brewster, S.A., Wright, P.C., Dix, A.J. & Edwards, A.D.N. (1995). *The sonic enhancement of graphical buttons*. In Nordby, K., Helmersen, P., Gilmore, D. & Arnesen, S. (Eds.), Proceedings of INTERACT'95, Lillehammer, Norway: Chapman & Hall, pp. 43-48.

[7] [BREWSTER98a] Brewster, S.A. (1998). *Using earcons to improve the usability of tool palettes*. In Summary Proceedings of CHI98 (Los Angeles, Ca), ACM Press, Addison-Wesley, pp 297-298.

[8] [CONVERSY&95] Conversy, S. & Beaudouin-Lafon, M. (1995). *Le son dans les applications interactives*, Université de Paris-Sud.

[9] [KIENTZLE98] Kientzle, T., (1998). *A Programmer's Guide to Sound*, Addison-Wesley.

[10] [KRAMER94] Kramer, G. (1994). *An Introduction to Auditory Display, Auditory Display: Sonification, Audification, and Auditory Interfaces*, A proceedings volume of the Santa Fe institute studies in the science of complexity, pp 1-78.

[11] [MENNESSIER&98] Mennessier, M.O., Alvarez, R., Noirhomme, M., Rouard, M. (1998) *Physics and Evolution for LPVs from HIPPARCOS Kinematics*, in IAU191 Symposium, Montpellier.

[12] [MEZRICH&84] Mezrich, J.J., Frysinger, S.P. & Slivjanovski, R. (1984). *Dynamic Representation of Multivariate Time Series Data*, J. Amer. Stat. Assoc. 79,pp 34-40.

[13] [NOIRHOMME&98] Noirhomme-Fraiture, M., Rouard, M. (1998). *Visualisation de données multivariés: évaluation de la représentation en étoile zoom*, in IHM 98, Nantes, pp 121-126.

[14] [NOIRHOMME&00] Noirhomme-Fraiture, M., Rouard, M. (2000).*Visualising and Editing Symbolic Objects*,chap 7 in Bock, H.H., Diday, E. (eds.) *Analysis of Symbolic Data.Exploratory methods for extracting statistical information from complex data.*, Springer Verlag, Heidelberg, pp 125-138.

A self organizing map (SOM) extended model for information discovery in a digital library context

Jean-Charles Lamirel
LORIA,
BP 239
54506 Vandoeuvre Cedex,
FRANCE
33-3-83-59-20-88
lamirel@loria.fr

Jacques Ducloy
INIST,
2, allée du Parc de Brabois
54514 Vandoeuvre Cedex,
FRANCE
33-3-83-50-46-00
Ducloy@inist.fr

Hager Kammoun
LORIA,
BP 239
54506 Vandoeuvre Cedex,
FRANCE
33-3-83-59-20-88
lamirel@loria.fr

ABSTRACT

This paper presents the MicroNOMAD Discovering Tool. Its main characteristic is both to provide an user with emergent analyses of a multimedia database content and with querying and browsing guidelines through the use of an advanced topographic interface model. The model also allows the user to dynamically exploit semantic exchanges between multiple viewpoints on the database. The tool basic principles are firstly described. A tool experimentation which is achieved on the multimedia data-base associated to the BIBAN"Art Nouveau" server is then developed. It clearly demonstrates that the combination of both the topographic structures, the textual and iconographic interaction, and the viewpoint exchanges proposed by the MicroNOMAD core model could play an essential role in several discovering and browsing processes.

Keywords

multimedia, information discovery, classification, neural networks, data visualization, information retrieval

1. INTRODUCTION

Digital Libraries are generally requested to provide access to a large variety of information. As a result, most of DL architectures are issued from Information Retrieval models and are designed to help end-users in retrieving information «he already knows but he has lost a link to». Then, a kernel strategy of many available systems or search engines consists in asking a user to formulate a preliminary query, as an initiate value, and to start an inter-active process of reformulation. From an users point of view there is a risk that he will quickly get lost if he has been too generic or imprecise. We therefore aim at building Information Discovering Systems rather than Information Retrieval Systems. Indeed, we want to give some ways for exploring the knowledge of a corpus. In other words, we would like to answer questions like this: «what is the most important feature in this topic? what's new? what do I not know in this domain? etc.». We further worked on textual «Digital Libraries» in the framework of two projects dealing with biomedical information: MedExplore and WebStress. These experiences have shown that a complex exploration requires the user to handle a fairly big set of various tools. Moreover, these tools are selected in a «non predictable» order, depending on the intermediate results. Thus, the user needs a fast and global analysis of intermediate data. In that context, images, graphics and iconographic resources have appeared to be a very fundamental component in man-machine interface. For instance, when the system delivers to an user a list of titles, the user needs to read the abstracts to determine whether the topic is relevant or not; let us mention that a fast glance is sufficient on an image to achieve this. To build up our own Information Discovering System turning to account this "outstanding explanatory power" of the images, we have experienced a specific approach which led to the MicroNOMAD Discovering Tool. The core model of that tool strongly derives from the multimap topographic model which has been successfully tested on textual data in the framework of the NOMAD IR System [10]. This latter model, which can be itself considered as a extension of the basic Kohonen topographic map model, enables the user to browse through a documentary database by means of an advanced topographic interface. The MicroNOMAD core model added-value is then mainly to develop a synergy between the browsing and discovering capabilities of the NOMADs original multimap model, on the one hand, and the natural capability of the imbedded Kohonen map model to support at the same time concept mapping and image mapping, on the other hand. In a first part we will briefly describe our previous research. In a second part, we will explain the basics of our new Discovering Tool and we will conclude with experiments on this tool.

2. THE BIBAN PROJECT

BIBAN (Bibliographic and Iconographic and Base Art Nouveau) is a research prototype for iconographic Digital Libraries. BIBAN is an application of a generic XML workbench DILIB [3] and has been designed for investigating Digital Libraries containing images and heterogeneous documents in a multilingual context. BIBAN covers the "Art Nouveau" period, a widespread movement for a renewal of the decorative art at the end of 19th century. BIBAN's content includes:

- a set of electronic books in French with their translation in English and German: each book keeps its own style and is implemented as a individual web server (with several HTML pages). One of them is «Nancy and the Art Nouveau Style» [6] which deals with the collections of «Musée de l'Ecole de Nancy».
- an iconographic base: this base contains a set of images. A metadata record which, coded with an elementary XML schema, is made for each image.
- a bibliographic base: this base is a subset (300 references) of BHA (Biblographie de l'Histoire de l'Art) selected by «Art Nouveau» or «Ecole de Nancy». This bibliography is connected with a much larger set (6000 references) randomly extracted from the whole BHA.

All documents are indexed with BHA search entries. Thus, the same description vocabulary is used for images, bibliographic records and pages of electronic books. To this end, each HTML page contains Dublin Core [18] elements.

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

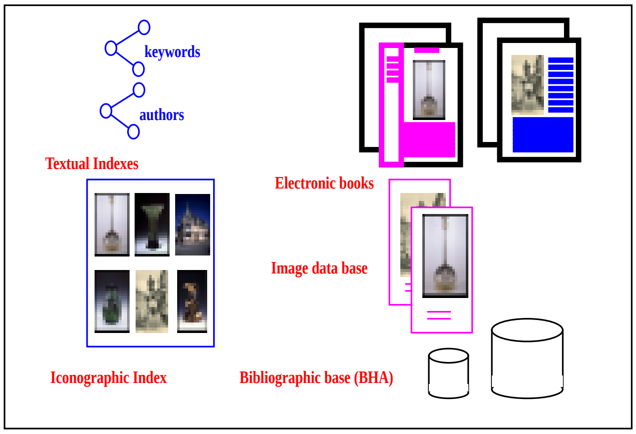


Figure 1: BIBAN context

We have put an early version of the BIBAN server on the INTERNET with a limited advertising, mainly intended for information specialists. BIBAN was put on the Web without any kind of assistance, by people with a poor knowledge of the domain. As a result, one important point was raised: the iconographic tools we have provided in this first experiment have been intensively used during browsing and querying steps. Nevertheless, these tools were mainly based on elementary text to image links or image to image links. They appeared to be useful for widening a query but not sufficient for giving a global view of the main topics of a collection and their relationships. We have then implemented a new set of techniques for producing structured iconographic maps. This point will be developed in the further paragraph.

3. ADAPTIVE MODEL

3.1 Introduction

The implementation of the IR process on iconographic databases considering the specificities of the images has been tackled through time with a lot of different approaches.

On one hand, in some early approaches, as the one proposed in the RIVAGE prototype [2], as well as the one more recently adopted for the first BIBAN prototype, the implementation of the concept of image mosaic or 2D mapping of images has been considered as a convenient way to give the user an overall view of its query results and moreover to help him in its relevance judgments. These approaches rely on the psychological facts that, conversely to textual information, the interpretation and judgment on a relatively large image set could easily be performed in one pass on such a mosaic by an user, without any explicit information on the image content. The early IDIM prototype described by Aigrain and al. [1] has gone one step further on this latter way, proposing a more categorical approach in which the user IR session on a iconographic database is reduced to an user multidirectional browsing through a 3x3 evolutive image mosaic. Conversely to the preceding models, one of the most interesting characteristics of this model is the very exploitation of the 2D structure of the image mosaic for defining different research directions based themselves on different background keywords classification profiles. Nevertheless, the lack of any access to this latter background information often led the user to make arbitrary choice in its browsing.

On the other hand, the use of alternative profiles like visual indexes for image description has been thoughtfully investigated these latter years [5]. Even if these approaches seem to be promising they surely could not cover all the user needs in an image retrieval process. These points have also been noticed by Duffing and al. [4]. Indeed, their original contribution consists in combining an image classification based on visual indexes and another one based on keyword indexes for computing the result of an user query. Unfortunately, their model do not at all rely on an image mosaic approach.

We found an interesting challenge in the trial of both combining the advantages of all the preceding models (i.e. image mosaic mapping, 2D structure exploitation, and multiple classification use) and dealing with advanced discovery capabilities through a federating approach. We therefore choose to derive our approach, which we called MicroNOMAD, from the topographic multimap model of the NOMAD IR System [9]. The role of the NOMAD's topographic multimap model is both to provide an user with emergent and «easy to use» analyses of a documentary database contents and with overall querying and browsing guidelines through an advanced topographic interface. Conversely to a lot of other more classical models, the NOMAD model allows the user to exploit dynamic exchanges between multiple viewpoints (i.e classifications) on the database, those being implemented through Said exchanges could be used in several ways. For instance, they enable an user to highlight semantic correlation between different themes belonging to different viewpoints or to indirectly access to documents which may well be unreachable when considering only one viewpoint. They could also be used by a IR system in an automatic mode for elaborated thematic reasoning tasks [10]. To take benefit of the discovering and browsing properties of the NOMAD multimap model in an iconographic context we have mainly based our adaptation of the original model for the MicroNOMAD approach on a parallel implementation of a thematic mapping and of an image mapping on the same maps.

The basic principles of the multimap model along with these adaptations are presented in the next section.

3.2 Basic maps construction process

The MicroNOMAD basic image classification process is based on the Kohonen topographic map model [8]. This model considers that a data[1] classification can be viewed as a mapping on a 2D neuron grid in which neurons establish predefined neighborhood relation. After the classification process, each neuron of the map will then play the role of a data class representative. The main advantages of the Kohonen map model, as compared to other classification models, are its natural robustness and its very good illustrative power. Indeed, it has been successfully applied for several classification tasks [11] [12] [14] [17]. In our own case, each topographic map is initially built up by unsupervised competitive learning carried out on the whole iconographic database. This learning takes place through the profile vectors extracted from the image descriptions, which describe the characteristics of these images in the viewpoint[2] associated to the map.

For each neuron of a map *M*, the basic competitive learning function has the following global form:

$$W_n^{t+1} = W_n^t + \alpha(t)k(t)(W_{n*}^t - P_n^t)$$

<u>*where*</u>

W_n^t is the external weights profile vector (i.e. the class profile vector) of the neuron n at time t,
P_n^t is the description, considering the viewpoint associated to the map, of the image i chosen as learning sample at time t,
n is the winning neuron at time t, that is the neuron*

[1] For our experiment, the data correspond obviously to the images of the database.

[2] The "viewpoint" notion is an original notion that has been firstly introduced in the NOMAD IR system for playing the role of semantic context of retrieval [10]. In the framework of our image database specific viewpoints have been associated to each specific keywords set of the image description like "Indexer keywords" set, "Title keywords" set or "Author names" set. Other viewpoints could also have been associated to the visual characteristics of the images, if these latter had ever been computed.

which profile has the best match with the i image profile, a(t) a time decreasing function, k(t) a neighborhood adaptation function.

The topological properties associated with the Kohonen maps make it then possible to project the original images (i.e. data) onto a map so that their proximity on the map matches as closely as possible their proximity in the viewpoint associated to said map.

After the preliminary learning phase, each map is organized so as to be legible for the user through analysis of the main components of the neuron profiles.

A first phase of this analysis consists in defining class names that could optimally represent the class contents when the map is displayed to the user. Due to the fact that there is obviously no absolute strategy for achieving that goal (this problem is well known by automatic classification specialists as the "class naming problem") we choose to implement two different kinds of strategies that could be indifferently used during the map consultation phase:

- *The class profile driven strategies:* they consist of attributing to each class a name that represents the combination of the labels of the components having the maximum values in its profile.
  These strategies are well-suited in highlighting for the user the main themes described by the map.
- *The member profiles driven strategies:* they consist of attributing to each class a name that represent the combination of the labels of the components having the maximum values in either the profile of the most representative member of the class or the average member profile computed thanks to all the class member profiles. In this strategies, no name could obviously be attributed to intermediary classes due to the fact that they do not have any associated member.
  These strategies are useful in providing the user with complementary information for the map's themes content interpretation. Indeed, some important information on a theme could be better represented in the theme's member profiles, than in its related class profile[1].

The second phase of the analysis consist in dividing the map into coherent logical areas or neurons groups. Each area, which can be regarded as a macro-class of synthesis, yields a very reliable information on the relative importance of the different themes described by the map. Main themes are represented as larger areas (i.e. with more neurons) than the marginal themes. This "area effect" could also be considered as a very good illustration of the non linear mapping behavior inherent to the original Kohonen classification method. The area computation is based on the topographic properties of the neuron profiles of a Kohonen map [8]. These properties, that are only valid on a reliable map, guaranteeing both the continuity and the locality of the variations of the map neuron profiles, and indeed the closeness of the computed areas on the map. It has been presented in detail in [10]. The figure 3 represents a partial view of a resulting map in its finalized form. One can see that the "image mosaic" effect is obtained by "illustrating" the map thematic structure by the most representative image of each theme.

3.3 Intermap communication principles

3.3.1 General principles

The communication between Kohonen maps, that has been first introduced in the NOMAD IR model [9], represents a major amelioration of the basic Kohonen model. In MicroNOMAD, this communication is based on the use of the images that have been projected onto the maps as intermediaries neurons or activity transmitters between maps.

The communication process between maps could be divided in two successive steps: original activity setting on source maps (1) and activity transmission to target maps (2). The original activity could be directly set up by the user on the neuron or on the logical areas of a source map through decisions represented by different scalable modalities (full acceptance, moderated acceptance, moderated rejection, full rejection) directly associated to neurons activity levels [10]. This protocol could be interpreted as the users choices to highlight (positively or negatively) different themes representing his centers of interest relatively to the viewpoint associated to the source map. The original activity could also be indirectly set up by the projection of an users query on the neurons of a source map. The effect of this process will then be to highlight the themes that are more or less related to that query. Therefore, the activity of each map neuron is set up to the value of the cosine measure [15] between the neuron profile and the profile vector associated to the query. The activity transmission to target maps is based itself on two elementary steps: a first transmission step from the activated source map to its associated image neurons (down activation) and a second transmission steps from the activated image neurons to the target maps (up reactivation).

The activity A_i^T of a class i of a target map T derived from the activity of a source map S can be computed by the formula:

$$A_i^T = f_{n \in i}(g(A_n)),\ A_n = g(A^S_{j_n})$$

<u>*where*</u>

n represents a neuron associated to a data, j_n its associated class on the source map,
f is a function implementing the semantic correlation computation described hereafter,
g is a bias function.

The activity transmission could be considered as a process of evaluation of the semantic correlations existing between themes of a source viewpoint (source map) and themes belonging to several other viewpoints (target maps). The figure 4 represents the result of such a evaluation on the iconographic database "Art Nouveau" considering three different viewpoints (maps).

3.3.2 Main computation parameters

"Possibilistic"[2] computation of the semantic correlation: in this approach each class inherited of the activity transmitted by its most activated associated data. The *f* function described above can be given as:

$$f = \underset{n \in i}{\mathrm{Max}}(A_n^-) + \underset{n \in i}{\mathrm{Max}}(A_n^+)$$

<u>*where*</u>

A_n^+ represents a positive activity value (positive choice),
and A_n a negative activity value (negative choice).

This approach could help the user to detect weak semantic correlation (weak signals) existing between themes belonging to different viewpoints.

Probabilistic computation of the semantic correlation: in this approach each class inherited of the average activity transmitted by its associated data, either they are activated or not. The *f* function described above can be given as:

$$f = \frac{1}{\|i\|}\sum_{n \in i} A_n$$

<u>*where*</u>

$\|i\|$ represents the number of data associated to the class i.

[1] This phenomenon is due to the fact that the class profiles are drawn from the classification process while the member profiles represent a straightforward information from the original data.

[2] "Possibilistic" is a neologism meaning that our measure is directly related to the measure of possibility defined by the possibility theory.

Conversely to the possibilistic computation, the probabilistic computation give a more reliable measure of the strength of the semantic correlations and may be then used to differentiate between strong and weak matching.

4. EXPERIMENTATION

4.1 Experimental context

We carried out a first experiment with the MicroNOMAD Discovering Tool on the iconographic database "Art Nouveau" managed by the BIBAN server. This database contains approximately 300 images related to the various artistic works of the Art Nouveau school. It covers several domains, such as architecture, painting and sculpture.

The images have associated bibliographic description containing optionally title, indexer keywords and author information. These description are managed by the DILIB workbench in XML format. We choose to use 3 different viewpoints (profiles) in our experiment:

- *The "Indexer keywords" viewpoint.* Its is represented by the keywords set used by the indexer in the keyword description field of the images.
- *The "Title keywords" viewpoint.* Its associated keywords set is build automatically through a basic keywords extraction (use of a stop word list and plural to singular conversion) of image titles. After the keywords extraction a new "Title keywords" field is added to the image description.
- *The "Authors" viewpoint.* It is represented by the set of authors cited in the image descriptions.

The first step of the experiment consists in transforming the image description associated to the chosen viewpoints in profiles vectors. For that step, we also choose to apply a classical Log-Normalization step [19] in order to reduce the influence of the most widespread words of the profiles. The second step is the original classifications building. Its has been implemented through the classical Kohonen SOMPACK algorithm [17]. The results, which consists in three different classifications associated to the three different viewpoints are then "dressed" and converted to XML format thanks to the DILIB tools. For the sake of portability, the core of the MicroNomad Discovering Tool has been developed as a Java application. Its entries are the XML classification files produced in the preceding step and it implements the class naming strategies, the maps division into logical areas, the map on-line generalization and the intermap communication process described above.

From a practical standpoint, the MicroNOMAD interface provides the user with several different querying and browsing capabilities:

- Browsing through the class of the maps in order to access to their main characteristics and to their associated images.
- Producing queries and afterwards reformulation with a classical querying interface, which nevertheless implements an interesting secondary effect consisting of the projection of the queries on the maps. Indeed, this latter effect could significantly help the user to evaluate the query consistency with respect to the database content: a focalized activity on the map will correspond to a thematically consistent query, a widespread or badly matched activity on the map will correspond to a thematically inconsistent query.
- Acting on the classes activities in different ways in order to highlight semantic correspondences between viewpoints, to find connotations of a query, to get complementary information on some images, or to retrieve images similar to the ones of a chosen class but being not indexed by the current viewpoint.
- Collecting image samples in a session memory for all kinds of future operations.
- Using peripheral tools, like variance and projection tools, for the evaluation of the quality of the classifications and for estimating the degree of influence of the different classes on the classifications.
- Activating links with the BIBAN server pages for highlighting the context of the different artistic works associated to the images.

The whole experimental context is synthetically described by the figure 2.

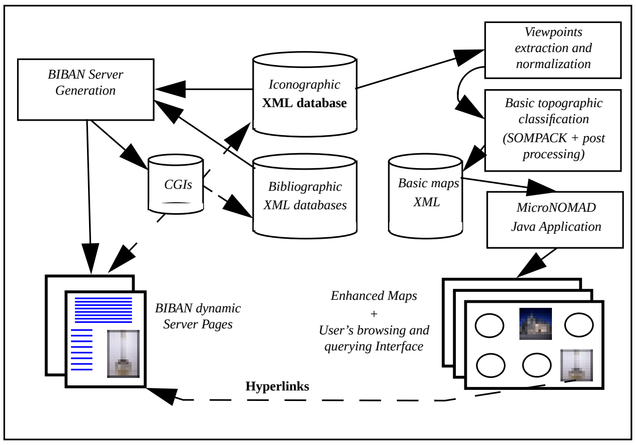

BIBAN Server Generation
Iconographic XML database
Viewpoints extraction and normalization
Basic topographic classification (SOMPACK + post processing)
CGIs
Bibliographic XML databases
Basic maps XML
MicroNOMAD Java Application
BIBAN dynamic Server Pages
Enhanced Maps + User's browsing and querying Interface
Hyperlinks

Figure 2: Experimental context

5. DISCUSSION

The first results which were obtained by our model are very promising. Original multiple viewpoints classification approach have directly produced very interesting results proving one more time the relevance of such an approach which tends to reduce the noise which is inevitably generated in an overall classification approach while increasing the flexibility and the granularity of the analyses[1]. In our experiment we found that a "Title keywords" classification can highlight information that is very complementary to the one highlighted by an "Indexer keywords" classification. For instance, in the context of the Art Nouveau database, we found interesting thematic extrapolation capabilities to the "Title keywords" classification, as well as complementary thematic focalization capabilities to the "Indexer keywords" classification. Indeed, crossings domain, like the common works of the Art Nouveau, are well highlighted by the "Title keywords" classification. As for it, "Indexer keywords" classification isolates very precisely the main artistic domains of the Art Nouveau while focusing on the most investigated sub-domains. Thus, the various naturalist metaphors integrated in the "Art Nouveau" works are very precisely described by this latter classification.

Maps also represent an useful tool for the indexation specialists. They help them in estimating the quality of the indexation of a database. Thanks to the classification method, strong indexation incoherences could be easily highlighted on the map: such incoherences are obvious if themes that specialists judged of equal weight in a domain appear with strongly different surface areas on a map. One example of such an incoherence that has been found by a specialist is the exaggerated representation of the Butterfly ("Papillon") theme, regarding to the Insect ("Insecte") theme, on the map of the figure 3.

After experimentation with several users, the opportunity to have simultaneously images and coherently organized textual information on the same support (map) seems to be definitely of great utility. Classification results interpretation are really made easier by the presence of images, as well as text represent a good

[1] See [LAM 95] for another experimental and theoretical justification on that point.

help in the choice of reliable browsing points in the iconographic database. The model on-line generalization capabilities and its ability to derive the map description context in several ways could also significantly help the user in its database contents interpretation and browsing.

Thanks to the user opinion, the intermap communication process appears to be a very interesting and original feature of the model. It provides the system with a new capability that could be called a dynamic and flexible browsing behavior. Conversely to classical browsing mechanism, like hypertext links, the browsing effect could then be directly tied to the users information and explanation needs (see figure 4). Moreover, the number and the type (i.e. concurrent or complementary) of viewpoints that could be simultaneously used is not limited by the model. For example, one can easily add a new map representing a classification based on "visual indexes" extracted from the images. These last properties could led us to consider our approach as a good basis for building an intelligent multimedia discovering system that could be used for various discovering and analysis tasks, especially for the ones which are strongly tied to image interpretation. Indeed the model is now tested in two important applications:

- Interactive browsing through museum database and intelligent setting up of exhibitions in the framework of the technical collection of the french "Musée de la Villette".
- Management of multiple classifications of butterflies (colour, shape, ...) in the Taiwanese NSC Digital museum of butterflies [7].

6. CONCLUSION

The MicroNOMAD Discovering Tool development represents obviously a important step for providing an Iconographic interface to Digital Library Server with a high level of interactivity. We have said that the first reactions we received in demonstrating it in the BIBAN server context were very encouraging. Nevertheless, we have still a lot of work to do if we want to put such an interface on the Internet or to produce a tool allowing anyone to build this kind of application. The basic browsing and querying capabilities of our Tool seem to be well-suited for over-all browsing and querying tasks, whatever are the users abilities. Nevertheless, a real challenge comes from the relative difficulty for the non specialists of precisely analyzing the classification results that are produced by the Tool (and working on them). As shown in this paper, sophisticated tools give better hypothesis but they are more difficult to validate. Thus, in a «BIBAN like» context, we think that we will have to provide three different ways, depending of the user profile. People who just want to surf in a «tourist approach» will be more confident in a pre-computerized map which will give them a lot of paths to explore. On the opposite, specialists in classification models who want to investigate in a precise strategy could use a dynamic interface, closer to our present prototype. Another problem comes from domain specialists which want to get effective results by a deeper exploitation of both the expressive and the discovering power of the MicroNOMAD Tool but which are not familiar with neural theories and their background behavior: the MicroNOMAD multimap core model will be very useful to them in proposing new assumptions but we will have to connect him with very simple tools enabling non classification specialists to verify these assumptions. For that goal, we planned to interface our model with such a simple validation tool based on gallois lattice and dealing with logical inference [16]. In this way, our programing approach based on XML interfaced components will be very useful in declining various implementations from one model.

As the dimensions of the topographies processed by our system are not "à priori" limited, we are also planning to make use of rather strongly multidimensional topographies in order to represent much better some complex data relations[1]. In order to interpret this relations, the user will then be provided with multiple 2D projections of a same multidimensional topography.

7. REFERENCES

[1] Aigrain P. and Longueville V. A Connection Graph for User Navigation in a Large Image Bank, in Proc. RIAO, Vol. 1, p. 25–44, Barcelona, Spain, avril 1991.

[2] Créhange M. and Halin G. Machine Learning and Vectorial Matching for an Image Retrieval Model, in Proc. SIGIR, p. 99–114, 1990.

[3] Ducloy J. DILIB, une plate-forme XML pour la génération de serveurs WWW et la veille scientifique et technique, in Le Micro Bulletin Thématique, No. 3, april 99.

[4] Duffing G. An alternative image retrieval system based on visual and thematic corpus organisation, in Proc. ICMCS, Florence, Italia, 1999.

[5] El Kwae E. and Kabuka M.R. A robust Framework for Content-Based Analysis Retrieval by Spatial Similarity in Images Database. ACM TOIS, Vol. 17, No. 2, april 1999.

[6] Gnaedig I. & al. Nancy et l'Art Nouveau, http://www.biban.fr/AN96 (restricted use).

[7] Hong J. & al. A digital museum of Taiwanese butterflies, in Proc. ACM/DL 00, San Antonio, Texas, june 2000.

[8] Kohonen T. Self-Organisation and Associative Memory. Springer Verlag, New York, USA, 1984.

[9] Lamirel J.C. and Créhange M. Application of a symbolico-connectionist approach for the design of a highly interactive documentary database interrogation system with on-line learning capabilities, in Proc. ACM-CIKM 94, Gaithersburg, Maryland, november 94.

[10] Lamirel J.C. Application d'une approche symbolico-connexionniste pour la conception d'un système documentaire hautement interactif. PhD, Université de Nancy, France, november 95.

[11] Lin X., Soergel D. and Marchionini G. A Self-Organizing Semantic Map for Information Retrieval, in Proc. SIGIR, Chicago, USA, 1991.

[12] Martinetz T.M. and Schulten K.J. Topology Representing Networks. Neural Networks, 7 (3) : 507-522, 1994.

[13] Michelet B. L'analyse des associations. PhD, Université de Paris 7, Paris, France, october 1988.

[14] Orwig E., Chen H. and Nunamaker Jr J. F. A graphical, Self Organizing Approach to Classifying Electronic Meeting Output.JASIS, 48 (1): 157-170, 1997

[15] Salton G. The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall Inc., Englewood Cliffs, NJ, USA, 1971.

[16] SimonA. and Napoli A. Building Viewpoints in an Object-Based Representation System for Knowledge Discovery in Databases, in Proc. IRI-99.

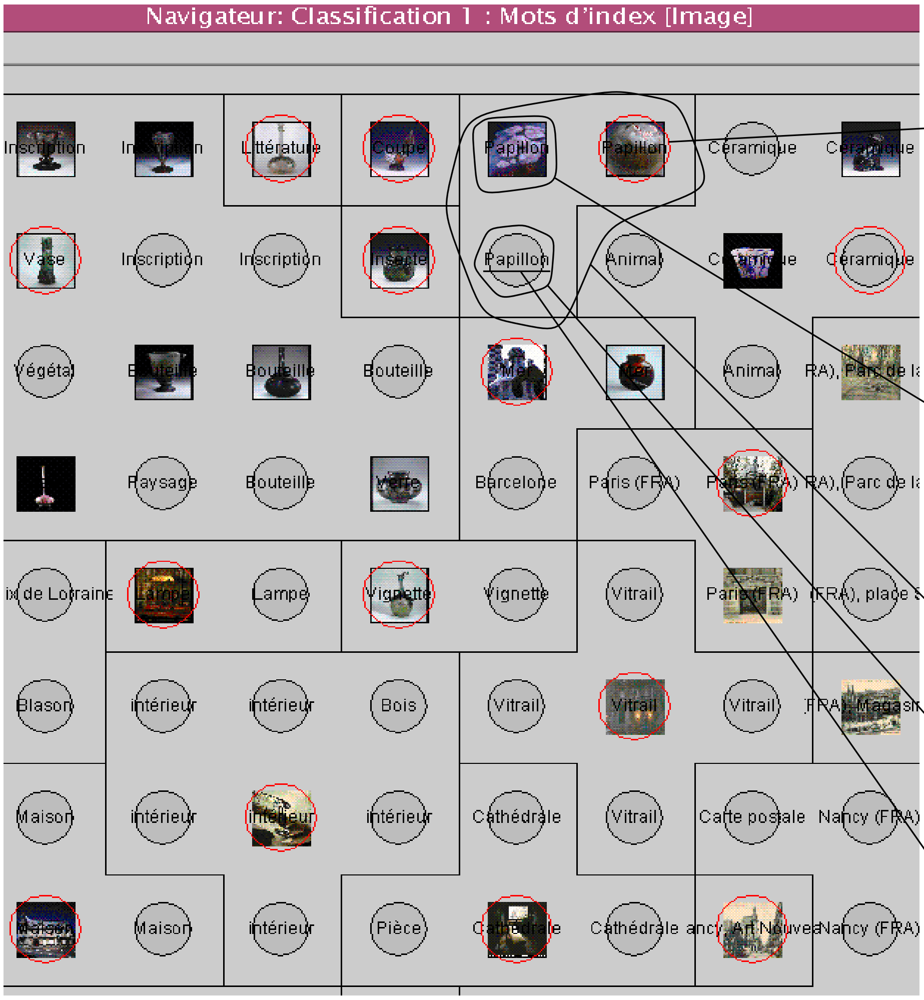
[17] SOM papers, http://www.cis.hut.fi/nnrc/refs/

[18] Weibel S. L., Stuart L. and Miller E. J. Dublin Core Metadata Element Set Reference Page, http://purl.oclc.org/metadata/dublin_core

[19] Wilbur W.J. and Coffee L. The Effectiveness of Document Neighboring in Search Enhancement. Information Processing and Management, 30 (2) : 253-266, 1994.

[1] A straightforward example of such relation is a relation between several authors which is impossible to clearly describe on a 2D map.

8. APPENDIX : MAPS



Navigateur: Classification 1 : Mots d'index [Image]

Inscription	Inscription	Littérature	Coupe	Papillon	Papillon	Céramique	Céramique
Vase	Inscription	Inscription	Insecte	Papillon	Animal	Céramique	Céramique
Végétal	Bouteille	Bouteille	Bouteille	Mer	Mer	Animal	RA), Parc de la
	Paysage	Bouteille	Verre	Barcelone	Paris (FRA)	Paris (FRA)	RA), Parc de la
ix de Lorraine	Lampe	Lampe	Vignette	Vignette	Vitrail	Paris (FRA)	(FRA), place S
Blason	intérieur	intérieur	Bois	Vitrail	Vitrail	Vitrail	(FRA), Magasin
Maison	intérieur	intérieur	intérieur	Cathédrale	Vitrail	Carte postale	Nancy (FRA)
Maison	Maison	intérieur	Pièce	Cathédrale	Cathédrale	ancy, Art Nouvea	Nancy (FRA)

5) *Center of gravity* of the logical area (symbolized by an external thin circle on one specific neuron of the area) = neuron whose profile is the nearest to the average profile of the area.

1) ***Class neuron*** presented jointly with its best associated image or class member. The other images associated to the class neuron can be accessed through the map interface.

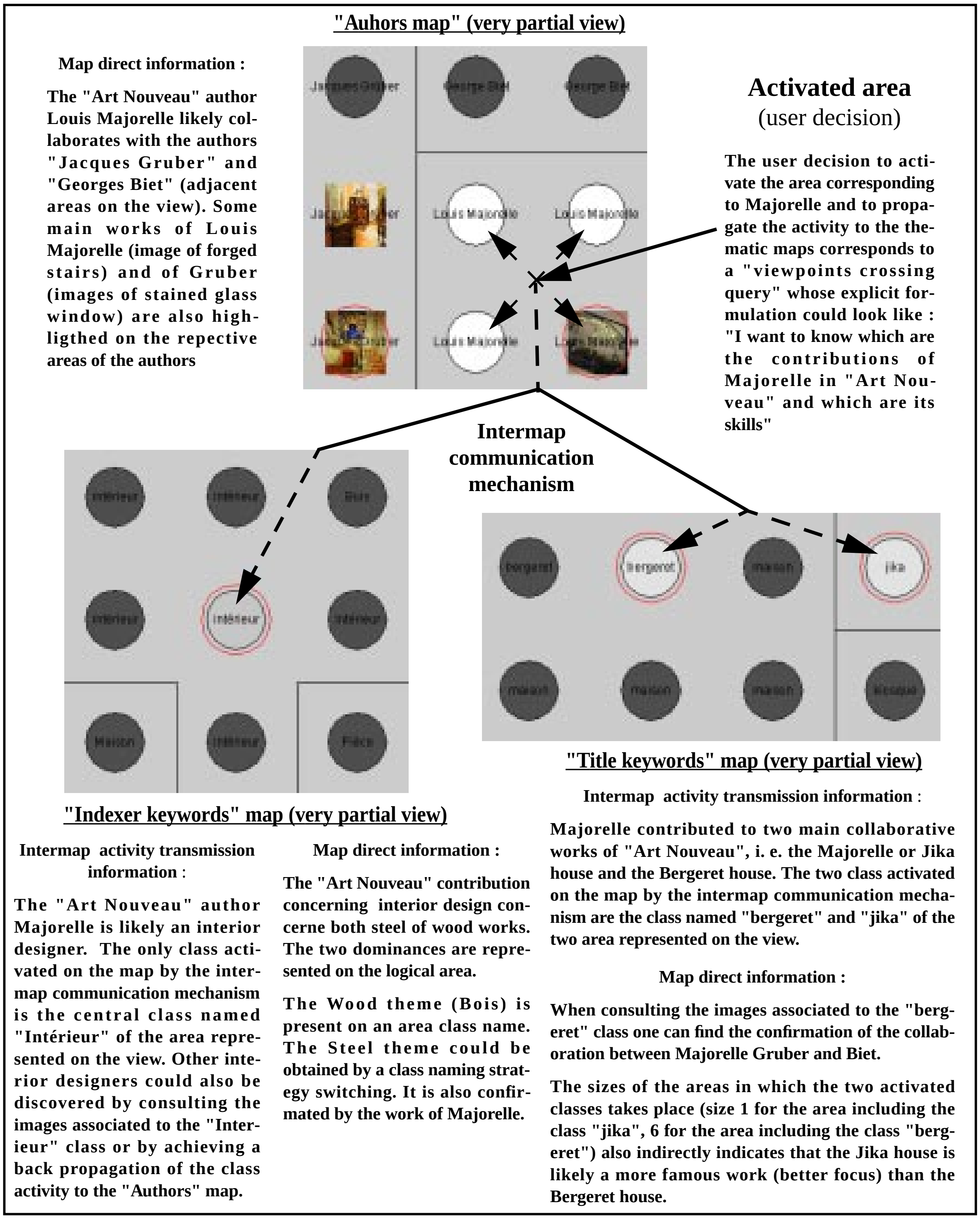
4) ***Logical area*** = group of neurons with the same profile dominances.

2) ***Class neuron*** without members also called intermediary neuron.

3) ***Class label*** or name considering the current naming strategy.

Figure 3: Map example

Partial view of a topographic map of 12 x 12 neurons (i.e. classes). The map is initially organized as a square 2D grid of neurons. The profile of the classes are then generated through an unsupervised competitive learning carried out on the profiles of 300 images of the BIBAN iconographic database considering one specific viewpoint. The viewpoint chosen for the showed map is the "Indexer keywords" viewpoint, with represents the bibliographic description of the images made by indexers (for more details, see Experimentation section).The names of the classes illustrate the themes (considering the chosen viewpoint) that have been highlighted by the learning. After the learning, the neurons related to the same themes have been grouped into coherent areas thanks to the topographic properties of the map. The number of neurons of each area can then be considered as a good indicator of the theme weight in the database. Considering, on the one hand, that themes or areas near one to another represent related notions and, on the other hand, that images, which represents the learning data, have been associated to their nearest classes on the map, the map could be considered both as a analysis tool and as a navigation mosaic with a semantically coherent organization.



"Auhors map" (very partial view)

Map direct information :

The "Art Nouveau" author Louis Majorelle likely collaborates with the authors "Jacques Gruber" and "Georges Biet" (adjacent areas on the view). Some main works of Louis Majorelle (image of forged stairs) and of Gruber (images of stained glass window) are also highligthed on the repective areas of the authors

Activated area

(user decision)

The user decision to activate the area corresponding to Majorelle and to propagate the activity to the thematic maps corresponds to a "viewpoints crossing query" whose explicit formulation could look like : "I want to know which are the contributions of Majorelle in "Art Nouveau" and which are its skills"

Intermap communication mechanism

"Indexer keywords" map (very partial view)

Intermap activity transmission information :

The "Art Nouveau" author Majorelle is likely an interior designer. The only class activated on the map by the intermap communication mechanism is the central class named "Intérieur" of the area represented on the view. Other interior designers could also be discovered by consulting the images associated to the "Interieur" class or by achieving a back propagation of the class activity to the "Authors" map.

Map direct information :

The "Art Nouveau" contribution concerning interior design concerne both steel of wood works. The two dominances are represented on the logical area.

The Wood theme (Bois) is present on an area class name. The Steel theme could be obtained by a class naming strategy switching. It is also confirmated by the work of Majorelle.

"Title keywords" map (very partial view)

Intermap activity transmission information :

Majorelle contributed to two main collaborative works of "Art Nouveau", i. e. the Majorelle or Jika house and the Bergeret house. The two class activated on the map by the intermap communication mechanism are the class named "bergeret" and "jika" of the two area represented on the view.

Map direct information :

When consulting the images associated to the "bergeret" class one can find the confirmation of the collaboration between Majorelle Gruber and Biet.

The sizes of the areas in which the two activated classes takes place (size 1 for the area including the class "jika", 6 for the area including the class "bergeret") also indirectly indicates that the Jika house is likely a more famous work (better focus) than the Bergeret house.

Figure 4: Intermap communication example

When navigating on a single map like an "Authors" map, the user can have a view of an author main works and of its main collaborations. When exploiting the communication between this "Authors" map and different thematic maps (here an "Indexer keywords" map and a "Title keywords" map) he can highlight the author various influence on the main thematic areas of the Art Nouveau, and moreover, its main artistic skills (see Experimentation section for more detailed descriptions of the "Author", "Indexer keywords" and "Title keywords" viewpoints).

On the figure, the maximum activity of a class correspond to white color of a its related neuron, the null activity to a dark grey color. For the sake of readability of these activities on the maps, the "Image display" mode has been switched off on the "Indexer keywords" map. and on the "Title keywords" map.

Learning Feature Weights from User Behavior in Content-Based Image Retrieval

Henning Müller, Wolfgang Müller,
Stéphane Marchand-Maillet, Thierry Pun
Computer Vision Group, University of Geneva
24 Rue du Général Dufour,
CH-1211 Genève 4, Switzerland
henning.mueller@cui.unige.ch

David McG Squire
Computer Science and Software Engineering
Monash University
Melbourne, Australia

ABSTRACT

This article describes an algorithm for obtaining knowledge about the importance of features from analyzing user log files of a content-based image retrieval system (CBIRS). The user log files from the usage of the *Viper* web demonstration system are analyzed over a period of four months. Within this period about 3500 accesses to the system were made with almost 800 multiple image queries. All the actions of the users were logged in a file.

The analysis only includes multiple image queries of the system with positive and/or negative input images, because only multiple image queries contain enough information for the method described. Features frequently present in images marked together positively in the same query step get a higher weighting, whereas features present in one image marked positively and another image marked negatively in the same step get a lower weighting. The *Viper* system offers a very large number of simple features. This allows the creation of flexible feature weightings with high values for important and low values for less important features. These weightings for features can of course differ between collections and as well between users. The results are evaluated with an experiment using the relevance judgments of real users on a database containing 2500 images. The results of the system with learned weights are compared to the system without the learned feature weights.

Keywords

long term learning, log file analysis, content-based image retrieval, web usage analysis, multimedia retrieval

1. INTRODUCTION

Much has been written about Relevance Feedback (RF) in content-based image retrieval (CBIR) [18]. Most feedback methods only takes into account one query step and the knowledge obtained from older query steps of the same session or of other query sessions is forgotten. Often, the feedback is limited to one positive image [16] or several positive feedback images [4]. Only few systems offer both positive and negative feedback as Surfimage [6, 27] and *Viper* [26]. Even these systems often have problems with too much negative feedback as described in [11], although solutions similar to those already used in text retrieval (TR) [17] exist.

Image browsers like *PicHunter* [5] offer the possibility to have feedback over more than one step and thus to really learn from the user interaction in order to find one target image. Using a sequence of queries to discover the user's goal creates another problem whenever the user changes the goal of a query in the querying process. Solutions to this problem referred to as "moving targets" are given in *TrackingViper* [13].

Yet, existing learning algorithms mostly try to find out the goal of a user over one or a few feedback steps. Minka [10] proposes across-session learning for *FourEyes* in *PhotoBook*. In [8], an approach to cluster images marked together positively and divide images marked negatively from the clusters is explained. In the domain of *collaborative filtering* [7], user judgments have been used to propose new items to users based on items being marked together positively by other users. This has been applied to art images of a museum as well [1]. The search for user preferences by giving positive and negative examples for web pages has also been studied [15]. Bayesian networks have been used to find out if an unknown page might fit to the users' profile or not. This supervised learning is out of the scope of this paper as we want to use unsupervised learning techniques to avoid additional work for the user. We also want to learn information for new queries and not just improve one already known query by augmenting important features.

In the domain of electronic commerce, log files resulting from web usage have been analyzed for a long time and the knowledge from this analysis is employed to improve new systems and to adapt them to the users' needs [28]. Part of this research concentrates on analyzing the behavior of users within webpages and the links they use [2]. In the domain of electronic commerce, there are many different concepts to identify users and track their activities, but problems arise with people just trying out pages and making very short

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

visits. Longer visits can be analyzed to facilitate the design of a web page.

The quality of user data gained from the internet might not be the highest. Nevertheless, we can learn from the usage information, and the related analysis in this paper shows that we can get qualitatively and quantitatively better results, even by using potentially poor web user data.

2. THE *VIPER* SYSTEM

The *Viper* system is a CBIRS that is described in more detail in [23, 24]. The system uses many techniques known from TR applications and aims at incorporating them into the domain of CBIR.

2.1 System Architecture

The main difference compared with other systems is the presence of a very large number of more than 85000 possible features. Most images contain between 1000 and 2000 of theses features. The access method to the features is the *inverted file*, which is the most common access method used in TR. Thus, *Viper* allows a fast and efficient access to the large number of features [12].

The emphasis of the project is on user interaction. Hence, it embeds several interaction strategies using several steps of positive and negative feedback. Both online and offline learning are employed in the system. *Viper* offers a good flexibility for learning as it has a very large number of features for the creation of feature weights. Especially the extensive use of negative feedback has shown to be very effective [11] and is also very important for the long term learning approach in this paper.

2.2 *Viper* Features

The system used for this study implements four different groups of image features:

- A global color histogram based on the HSV color space which corresponds roughly to the human color vision [22];
- local color blocks at different scales for fixed regions by using the mode color for each of the fixed blocks; the image is successively partitioned into four equally sized blocks and each block is partitioned again four times;
- global texture characteristics are represented by the histograms of the response to gabor filters of different frequencies and directions; gabor filters are known to be a good model for the human perception of edges [9];
- local Gabor filters at different scales and regions by using the same blocks as for the local color features and applying Gabor filters with different directions and frequencies to these blocks.

These features are only low level features, but because of the high number, very complex queries can be constructed with them. Higher level features like image regions may provide better results, but we still suffer from the semantic gap between the semantics the user is looking for and the visual content the system can offer.

2.3 Weighting schemes

We have implemented several weighting schemes known from the TR literature [21]. They are all based on the collection and document frequencies of the features. For the experiments in this paper, we use the *inverse document frequency* weighting, which weights the features in the following way:

$$relevance_j = \frac{1}{N}\sum_{i=1}^{N}(tf_{ij}\cdot R_i)\cdot log^2\left(\frac{1}{cf_i}\right), \quad (1)$$

$$score_{kq} = \sum_j (tf_{kj}\cdot relevance_j), \quad (2)$$

where *tf* is the *term frequency* of a feature, *cf* the *collection frequency* of a feature, j a *feature number*, q corresponds to a query with $i = 1..N$ input images, k is one *result image* and R_i is the *relevance* of an input image i within the range $[-1; 1]$.

We can see in Equation 1 that the final result mainly depends on the collection frequency of a feature. Rare features are weighted high, whereas features very common in the collection are weighted low because they contain less information. The term frequency of a feature in the input images has a has a minor influence. We can see in Equation 2 that besides the relevance factor for a feature, the term frequency of the feature in the resulting image has a small influence on the final score.

3. LEARNING FEATURE WEIGHTS FROM USER BEHAVIOR

Reference [26] points to the web demonstration of the CBIRS *Viper* we used for this study. Every time a user accesses this page and does an action, it is logged with a time stamp. Like this, we can always see what the user did and which problems he might have encountered with the system. This also offers the possibility to make an offline analysis of the data to better suit the information needs of a user. The host name of the user is also saved, but no other private data.

3.1 Analyzing the Log Files

This section gives a general overview of the data we logged into a file. Between September 1999 and January 2000, we had 3500 accesses to the system. About half of the accessors just looked at random or sorted image sets or watched the parameters, but about 1700 accesses actually were queries. This shows that many people visited the page, but a large number of them just played around with the system. This can be confirmed with the fact that only 24 of the 201 hosts which accessed the system had more than 20 actions with the system. About 40 percent of the queries came from different hosts within the University of Geneva. Of the 1700 queries, 786 where multiple image queries. Only multiple image queries contain enough information for the algorithm we want to employ.

In the log files, the query data from 10 different databases is regarded. It is hard to map the importance of features from one database to another database although they use the same set of features. The distribution of the features actually present in the database is very different for every

Table 1: The different functions of the system and their usage statistics in the web demonstration

Chose Database	668 times
Browse Image Names	251 times
Image Queries	1586 times
Random Images	586 times
Change Options	114 times
Clear Judgments	100 times

database. Only the histogram features for color and texture are present in a very large number of images in every database.

From the log files, we could also analyze the problems the user had with our system. Several people did queries without marking any image as relevant. As a result, we inserted a comment telling the user that at least one image needs to be marked. Another problem encountered while analyzing the log files was related to using too much negative feedback. This can as well remove all the important features from the query and lead to bad results. We therefore implemented a modified version of Rocchio's formula [17] for separately weighting positive and negative relevance feedback [11].

3.2 Learning from Log Files

The two rationales for our learning algorithm are:

- Features which occur often in two images marked together positively in the same query step should have a higher weighting than others;
- features which occur often in images marked once positive and once negative in the same query step should have a low weighting.

Based on these principles, we identified all pairs of images marked together. Queries with two input images just have one pair, whereas queries with three images have three and queries with four images have six. Thus, the number of image pairs in a query with n images is:

$$\mathit{number\ of\ image\ pairs} = \frac{n \cdot (n-1)}{2}. \tag{3}$$

The 786 multiple image queries lead to more than 31.000 image pairs marked together. If images are marked together, both negatively, the image pair is discarded, as this does not contain much information. Images can be marked together negatively for different reasons and may not have anything negative in common.

We then analyzed which features the two images of a pair have in common. Positive image pairs lead to a positive mark for the features they have in common and negative image pairs to a negative mark. Negative pairs have in general a smaller number of features in common. On average the image pairs have slightly more than 300 features in common. In total, the 31.000 image pairs lead to 10 million feature marks (6.1 million positive and 3.9 million negative). We separately analyzed the image database of the Télévision Suisse Romande (TSR) because our user experiments are based on this database. For the TSR database, we had 3.800 image pairs and 1.02 million feature pairs (0.47 million positive and 0.55 million negative), which represents about 10% of all the accesses to the system.

Features with a very high collection frequency like the histogram features occur about the same time as positive and negative pairs. Hence, their respective weight should stay very similar as before.

The additional factor we want to calculate should be in the range of [0; 2] to allow poor features to disappear completely and good features to be weighted significantly high. Features which occur only negatively should have a value of zero and features which occur only positively should have a value of 2.

This leads to the following simple formula for the additional factor factor_j:

$$\mathit{factor}_j = 1 + \frac{p_j}{p_j + n_j} - \frac{n_j}{p_j + n_j}, \tag{4}$$

where j is the feature number, p_j then number of positive marks for feature j and n_j the number of negative marks.

The new weighting formula for a feature is basically the same as it was before with only the additional factor from Equation 4 being calculated and included into Equation 1 as can be seen in Equation 5.

$$\mathit{relevance}_j = \mathit{factor}_j \cdot \frac{1}{N} \sum_{i=1}^{N} (\mathit{tf}_{ij} \cdot R_i)\, \mathit{log}^2 \left(\frac{1}{\mathit{cf}_i} \right). \tag{5}$$

Because improvements were lower than expected for the queries with relevance feedback (see Figure 2), we implemented a second factor similar to the factor in Equation 4 for comparison. We think that a complete disappearance of poor features might reduce the possibility to move in feature space, an effect which is stronger visible in feedback queries. Hence, we implemented a factor where the negative value can only reach a minimum of 0.25, whereas the maximum factor can be up to four, when only positive marks occur. If positive and negative marks occur with the same frequency, the factor stays at one. The resulting $\mathit{factor2}_j$ is obtained by rescaling the positive and negative parts of factor_j in a different way as can be seen in Equation 6.

$$\mathit{factor2}_j = \begin{cases} 0.25 + \frac{\mathit{factor}_j}{0.75} & : \mathit{factor}_j < 1 \\ 1 & : \mathit{factor}_j = 1 \\ 1 + (\mathit{factor}_j - 1) \cdot 3 & : \mathit{factor}_j > 1 \end{cases} \tag{6}$$

For the calculation of the weight, $\mathit{factor2}_j$ is used in exactly the same way as factor_j in Equation 5.

The fact that there are slightly more pairs marked together positively than there are negative pairs may lead to a different quantization of positive and negative parts, but does not alter the quality of the results.

4. EXPERIMENTAL RESULTS

To analyze the success of this method, we use a user experiment performed in [12]. This includes a very heterogeneous database of 2500 images from the Télévision Suisse Romande (TSR). 14 queries were presented to 3 users for relevance judgments. The users had to mark all the images in the database they regard as being similar to each of the 14 query images. Interestingly, the result sets for each user differ strongly in size and also in the images being selected. Similar effects were already reported in [25].

To evaluate the performance, we use precision/recall (PR) graphs which are the standard evaluation method in TR [19] and are more and more used in CBIR [24]. The results shown below are the PR graphs averaged over the relevance sets of all users and all queries from the user experiment. To simulate relevance feedback based on the user judgments, we used the algorithm explained in [11]. We feed back all images the user regards as relevant and which are in the first 20 images the system returns for the initial query.

The training data is only taken from the usage of the web demonstration system and does not have any connection with the user experiment we performed.

We see in Figure 1 that the results of the system with the additional factor are up to 10% better than the original graph when all queries of the same database (TSR) are used to calculate the weights. Using all queries of all the different databases still gives an improvement of 7% to 8%, but only in the beginning of the graph. The overall improvement is lower when using the data of all databases.

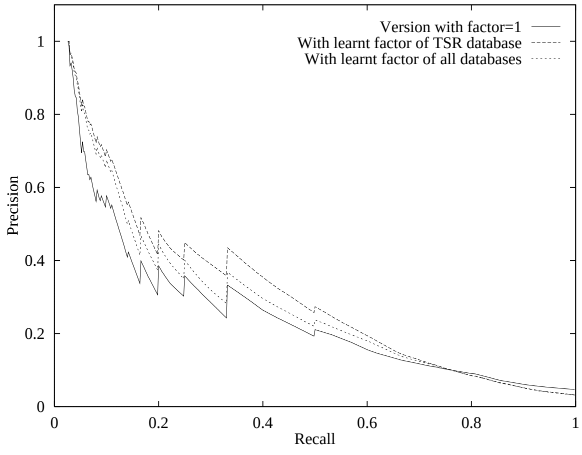


Figure 1: PR-Graph for a system with and without a learned factor (without feedback).

In Figure 2, we can see that the results of the first feedback step are much better (up to 100% in the middle parts) than the results before feedback (compare Figure 1). An improvement in the beginning of the graphs is especially important because this part represents the images the user actually views. The results with the learned factor are significantly better than without the factor, even on this high level. This shows that the gain with the additional $factor_j$ is not just limited to one query step as it favors image pairs already marked together.

When we use the factor learned from all the different databases, the results are about 3% to 5% better than without the learned factor. We think that this improvement was only small because the additional factor can become 0 for bad features which limits the flexibility to move in feature space. As a consequence, we repeated the experiments with a second factor explained in Equation 6.

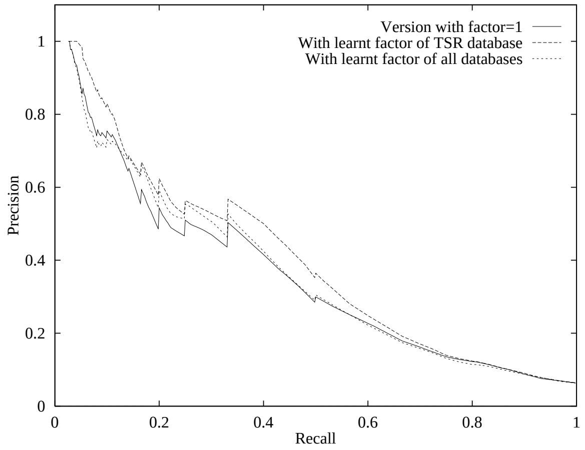


Figure 2: PR-Graph for a system with and without a learned factor (with feedback).

Figure 3 compares the results obtained using the two factors, respectively learned on the queries of the TSR database and learned with all databases. We can see that the results with the second factor are in both cases better for $factor2_j$. The beginning part of the graphs is almost identical, but in the middle parts of the graph the results improve with the second factor.

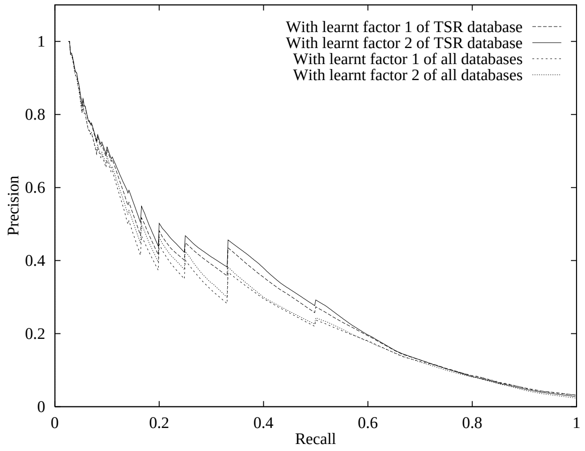


Figure 3: Comparison of the two different weighting factors (without feedback).

Figure 4 shows a comparison of the results obtained using the two factors for the queries with feedback. Here, we can clearly see the improvements of $factor2_j$ compared to $factor_j$ of up to 7%, especially in the middle parts of the graph. This shows that it might be better to let the factor always be above zero to not reduce the mobility in feature space. The small drop off in the beginning of the curve for $factor2_j$

can be explained with one query with only very few features and basically no textures, where the first returned image was non-relevant.

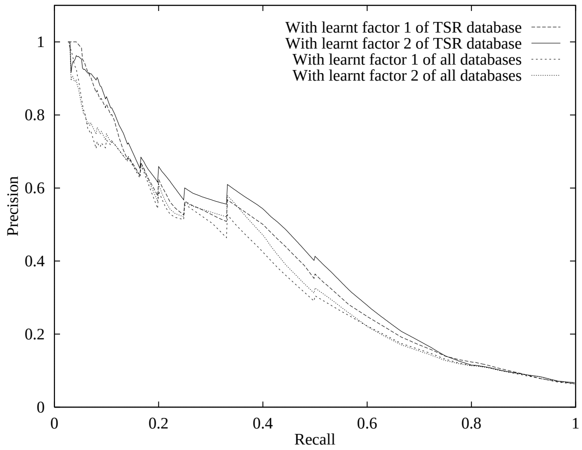


Figure 4: Comparison of the two different weighting factors (with feedback).

We also made some experiments where we tried to learn $factor_j$ based on feedback queries performed on a completely different database. The results were basically the same as without learning. The results using all the feedback from every database show clearly that not much feedback of the same database is necessary to improve the results of a query. Feedback from other databases does not change the results much as the feature space is only very sparsely populated and the databases populate different areas of the feature space.

We see that calculating weights from user log files brings strong improvements, especially, when the factor is learned based on queries of the same database. Learned over all queries and all databases, the improvement was not extremely strong, but clearly visible. Defining a user profile for learning could bring even stronger improvements, especially if the user is often performing similar search tasks. Therefore, we propose to have a hierarchy of factors, corresponding to a user, a domain and a global factor to be learned. To do this, a user identification needs to be inserted into the log file.

5. CONCLUSIONS AND FURTHER WORK

In this paper, an approach is presented on how to learn the importance of features in CBIR from log files containing user behavior of a web demonstration system. The problems of log files on the web is of course that we do not know much about the quality of the user data. Many people may come to a web page to try out the system and to see how it reacts and might even challenge the system with inconsistent data. This means that we can not always learn much from this kind of data. With the proposed approach, artifacts can be minimized as combinations of image pairs with high feature similarity have much more importance than image pairs with low feature similarity. The experiments with the factors we use show that, even with this kind of data, a significant improvement in retrieval quality can be reached. This is mostly true when the feature importance is learned on the same database. In this case, the results are very good.

Much better results will be possible once the data is obtained from serious users and even better if the study is restricted to a certain domain or a certain user. Like this specific user profiles or group profiles can be learned. We propose a hierarchy of learned feature weightings on a user, domain and global level.

Besides the learning of a feature weight for future queries we can evaluate the usefulness of features. This can also be used for the creation of new features. New features can be extracted for the old images and can directly be evaluated by using this method with the old log files.

More work needs to be done on finding an optimal factor to calculate a feature weight. We only proposed a very simple factor without any optimization. Another promising approach is to not only analyze pairs of images marked together, but directly evaluate multiple image queries by looking at all the images marked in a query. Features contained in $n > 2$ images marked together in the same query step should for example get a much higher weighting than features only contained in two images.

6. REFERENCES

[1] Active web museum. `http://abyss.eurecom.fr:1111/ AWM/login.html`, 2000.

[2] B. Berendt and M. Spiliopoulou. Analysis of navigation behaviour in web sites integrating multiple information systems. *VLDB Journal: Special Issue on Databases and the Web - to appear*, 2000.

[3] *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'99)*, Fort Collins, Colorado, USA, June 22 1999.

[4] Compass web page. `http://compass.itc.it`, 2000.

[5] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. Target testing and the `PicHunter` Bayesian multimedia retrieval system. In *Advances in Digital Libraries (ADL'96)*, pages 66–75, Library of Congress, Washington, D. C., May 13–15 1996.

[6] Surfimage webdemo. `http://www-rocq.inria.fr/ cgi-bin/imedia/surfimage.cgi`, 1999.

[7] A. Kohrs and B. Merialdo. Clustering for collaborative filtering applications. In *Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation*, Vienna, Austria, February 1999. IOS Press.

[8] C. S. Lee, W.-Y. Ma, and H. Zhang. Information Embedding Based on User's Relevance Feedback for Image Retrieval. In Panchanathan et al. [14]. (SPIE Symposium on Voice, Video and Data Communications).

[9] W. Y. Ma, Y. Deng, and B. S. Manjunath. Tools for texture- and color-based search of images. In B. E. Rogowitz and T. N. Pappas, editors, *Human Vision*

and Electronic Imaging II, volume 3016 of *SPIE Proceedings*, pages 496–507, San Jose, CA, February 1997.

[10] T. Minka. An image database browser that learns from user interaction. Master's thesis, MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139, 1996.

[11] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Strategies for positive and negative relevance feedback in image retrieval. In *Proceedings of the 15th International Conference on Pattern Recognition (ICPR 2000)*, Barcelona, Spain, September 2000. IEEE.

[12] H. Müller, D. M. Squire, W. Müller, and T. Pun. Efficient access methods for content-based image retrieval with inverted files. In Panchanathan et al. [14]. (SPIE Symposium on Voice, Video and Data Communications).

[13] W. Müller, D. M. Squire, H. Müller, and T. Pun. Hunting moving targets: an extension to Bayesian methods in multimedia databases. In Panchanathan et al. [14]. (SPIE Symposium on Voice, Video and Data Communications).

[14] S. Panchanathan, S.-F. Chang, and C.-C. J. Kuo, editors. *Multimedia Storage and Archiving Systems IV (VV02)*, volume 3846 of *SPIE Proceedings*, Boston, Massachusetts, USA, September 20–22 1999. (SPIE Symposium on Voice, Video and Data Communications).

[15] M. Pazzani and D. Billsus. Learning and revising user profiles: The identification of interesting web sites. *Journal on Machine Learning*, 27:313–331, 1997.

[16] QBIC$^{\mathrm{TM}}$ – IBM's Query By Image Content. `http://wwwqbic.almaden.ibm.com/~qbic/`, 1998.

[17] J. J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System, Experiments in Automatic Document Processing* [20], pages 313–323.

[18] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool in interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, September 1998. (Special Issue on Segmentation, Description, and Retrieval of Video Content).

[19] G. Salton. Evaluation parameters. In *The SMART Retrieval System, Experiments in Automatic Document Processing* [20], pages 55–112.

[20] G. Salton. *The SMART Retrieval System, Experiments in Automatic Document Processing.* Prentice Hall, Englewood Cliffs, New Jersey, USA, 1971.

[21] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[22] J. R. Smith and S.-F. Chang. VisualSEEk: a fully automated content-based image query system. In *The Fourth ACM International Multimedia Conference and Exhibition*, Boston, MA, USA, November 1996.

[23] D. M. Squire, H. Müller, and W. Müller. Improving response time by search pruning in a content-based image retrieval system, using inverted file techniques. In CBAIVL99 [3], pages 45–49.

[24] D. M. Squire, W. Müller, H. Müller, and J. Raki. Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. In *The 11th Scandinavian Conference on Image Analysis (SCIA'99)*, pages 143–149, Kangerlussuaq, Greenland, June 7–11 1999.

[25] D. M. Squire and T. Pun. Assessing agreement between human and machine clusterings of image databases. *Pattern Recognition*, 31(12):1905–1919, 1998.

[26] Viper webdemo. web page: `http://viper.unige.ch/`, 1999.

[27] A. Winter and C. Nastar. Differential feature distribution maps for image segmentation and region queries in image databases. In CBAIVL99 [3], pages 9–17.

[28] K.-L. Wu, P. S. Yu, and A. Ballman. Speedtracer: A web usage mining and analysis tool. *IBM Systems Journal on Internet Computing*, 37(1), 1998.

When image indexing meets knowledge discovery

Chabane Djeraba
IRIN, Ecole Polythechnique de l’Université de Nantes,
2 rue de la Houssinière, BP 92208 - 44322 Nantes Cedex 3, France
E-mail : djeraba@irin.univ-nantes.fr

ABSTRACT

In our paper, we deal with the challenge of extending automatically the classic image indexing by visual relationship features. The visual relationship features are discovered automatically from images. They contribute to make more efficient the content-based indexing. More particularly, we develop an advanced content-based indexing articulated around the following notions : - classic indexing, - clustering algorithm, - visual feature book and relationship qualification.

Keywords

Image, Indexing, retrieval, content, similarity, knowledge discovery, relations.

1. INTRODUCTION

In large image databases, finding images that contain semantic content, such as flowers during autumn or goals during football plays, is not simple. To do so, images should be well annotated by experts when inserted in the database. So, the quality of retrievals depends on the quality of the manual annotations. This solution characterizes classic information retrieval systems initiated by [Moo 51], and developed by [Sal 68], [Rij 79], and others. However, manual annotations tend to be incomplete and inconsistent, and they do not allow visual content-based image indexing and retrieval. Visual information systems, also known by content-based indexing and retrieval systems, such as in [Dje 00], [Jai 98] and others, overcome some of these shortcomings. The index is, generally, created automatically, and the final users have the possibility to formulate content-based queries. In spite of these appreciable advantages, the automatic indexing, which is the most important advantage of visual information systems, support weak semantic description, and therefore weak semantic queries. So finding images that contain flowers during autumn remains a very difficult query.

Content-based image indexing associated to knowledge discovery may be seen as a new way of thinking and regarding retrieval of multimedia information and it opens up to a lot of new applications which have not been possible, previously. For image archives the new possibilities given by content-based image indexing and knowledge discovery lies in the ability to perform "advanced queries-by-example”, meaning that we can present an image of an object, pattern, texture, etc., and fetch the images in the database that most resemble the example of the query. For image databases the new possibilities lie in the ability to access efficiently and directly selected images of the database.

Our paper deals with the following challenge : how do we build automatically the semantic content of images, based on basic content descriptions ? We believe that discovering hidden relations among basic features contributes to extract semantic descriptions useful to make the content-based image retrieval more efficient. In our case, the relationship discovery are held into two important steps : symbolic clustering based on the new concept of visual feature book and relevant relationships discovery.

The originality of our work concerns the following points :

- the definition of a new algorithm of global/local clustering and classification, based on : - visual quantization, powerful image descriptors and - suitable similarity measures,
- the creation of an efficient feature (texture, color) book which is the most representative of database image features,
- the power qualification of the relationship among visual features. They are composed of conditional probability and implication intensity measures,
- the extension of the classic indexing by relevant relationships that are automatically discovered.

The implementation of these notions together in the same framework constitute our advanced content-based indexing which is the scope of the paper.

We organize the paper as follow : in section 2, we describe the classic and advanced content based indexing and retrieval. We answer to the following questions : how images are searched in image database. We will not focus on speed data structures necessary to support the index, however, we will focus on the knowledge necessary to advanced content-based retrieval. In section 3, we present how the content of images are extracted and represented, how descriptors of images may be used to discover

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

relations between descriptors, and how the discovered relations are useful to content-based image retrieval. In section 4, we describe some experiment results.

2. YOU SAID CONTENT-BASED IMAGE INDEXING AND RETRIEVAL ?

The content-based image indexing and retrieval architecture is composed of three important components : extraction, representation and retrieval. Extraction and representation components constitute the heart of the architecture, together, they constitute the indexing component. The extraction component extract, automatically or semi-automatically, regions in images and compute features such as color, texture and shape of these regions. The whole image may constitute itself a region. The extracted contents are represented as or transformed into suitable models and data structures, and then stored in a persistent index.

The retrieval component constitute the eyes of the architecture. It searches images by selecting target images or content properties such as color, sketched shape, texture of image regions, or combinations of these. The retrieval process computes distances between source (example) and target features, and sorts the most similar images.

The central question is : how to extract and represent the content in order to make the retrieval process efficient ? Before answering this question, we will start by presenting the classic approach, and we will compare the benefits of the knowledge discovery to image indexing and retrieval efficiency.

2.1 Classic indexing

Indexing responds to how the content should be extracted and represented to allow efficient and effective search and access ?

Sequential searching of images with simple similarity computations is quite appropriate in a small database. However, the larger the database is, the slower the sequential approach is. So efficiency will not be respected. Classic access structures such as B-trees [Bay 72], K-D trees [And 85], point quadtrees [Fin 74] and R-trees [Gut 84] have advantages and disadvantages. Point quadtrees are simple to implement. However, there is a complexity of both insertion and search. Furthermore, deletion in point quadtrees is complex because finding a candidate replacement node for the node being deleted is generally difficult. Finally, the range retrieval in point quadtrees is time consuming. It takes $O(2\sqrt{n})$, where n is the number of image references in the tree. K-D-trees are very simple to implement. However, the search and insertion complexity in k-d-tree is high. In MX-quadtrees, range retrieval is very efficient, and the insertion, deletion and search take time proportional to O(n). We assume that the image (ex. map) is split up into a grid of size (2^n x 2^n) cells. R-trees have been preferred over k-d trees and point quadtrees, because they store a large number of rectangles in each node. So, they are suitable for disk accesses by reducing the height of the tree, this leading to fewer disk accesses. The disadvantage of R-trees is that, in certain cases, instead of following one path in the search process, multiple paths may be followed, because bounding rectangles associated with different nodes may be overlapped. Multiple paths means more disk accesses that might be compared to disk accesses of the other quadtrees.

These representations are physical access structures, they deal with applications that require massive amounts of storage and disk accesses. So they concern low level representation of the access structures. These access structures are necessary, but not enough to access effectively image materials. They need to be completed by high level representations (logical representation) that organize efficiently the descriptors of images, independently of their physical representations.

2.2 Advanced indexing

To obtain efficient access data structures, we should combine physical and logical representations of high-dimensional features. In our context, to effective up the content-based retrieval, we consider semantic representations that include image class hierarchy (images of flowers, panorama, etc.) characterized by knowledge and access speed data structures (K-D-trees). The K-D trees are implemented for high-dimensional features, at least eleven-dimension color and texture attributes, and voluminous classes. However sequential search is used for low-dimensional features and less voluminous classes. The K-D-trees are implemented at eleven-dimension because the color is represented by one dimension and the texture is represented by ten dimensions (ten couple of coefficients).

For example, when the user asks for images that contain waterfalls (figure 1), the system matches the user examples with the knowledge in the form of rules. In certain case, the image may belong to several classes, because the distance between the gravity center of the examples and the knowledge of the image classes are near together. In all cases, the retrieval process focuses its matches in the sub-classes of the current ones. In the sub-class, it triggers the same match process. When the leaf class is reached, the physical data structure is used to find the best images. When the number of the images in a class is low (ex. less than 100), than the search process is limited to sequential order.

In the example presented bellow, the first images returned contain waterfall, and the other images contain flowers. The whole images are visually similar to the example images. This example illustrates the « advanced query by examples » that is based on combination of visual features (texture and color) and knowledge. «advanced query by examples» specifies a query that means «find images that are similar to those specified». The query may be composed of several images. Several images accurate the quality of retrieval. For example, Several images of a «waterfall» accurate the description of the waterfall. This property makes possible the refinement of retrieval based on the feed backs (results of previous queries).

In the retrieval task (figure 2), features (colors, textures) of the query specification are matched with the knowledge associated to classes (ex. natural, people, industries, etc.). The suited classes are « Natural », then the matching process focus the search on the sub-classes of Natural : « Flowers », « Mountain », « Water », « Snow », etc. The knowledge associated to flowers and waterfalls are verified, so the matching process focuses the search on the « Flower » and « Water » classes. « Flowers » and « Water » classes are leaves, so the matching process compares the features of the examples with features of the image database to determine which images are similar to the example features. The matching task is based on computing the distance between target and source image regions. When mixing several features, such as colors and textures, the resulting distance is equal to the Sum taking into account the ponderation values of the considered features. The resulting images are sorted, the shortest distance corresponds to the most similar images.

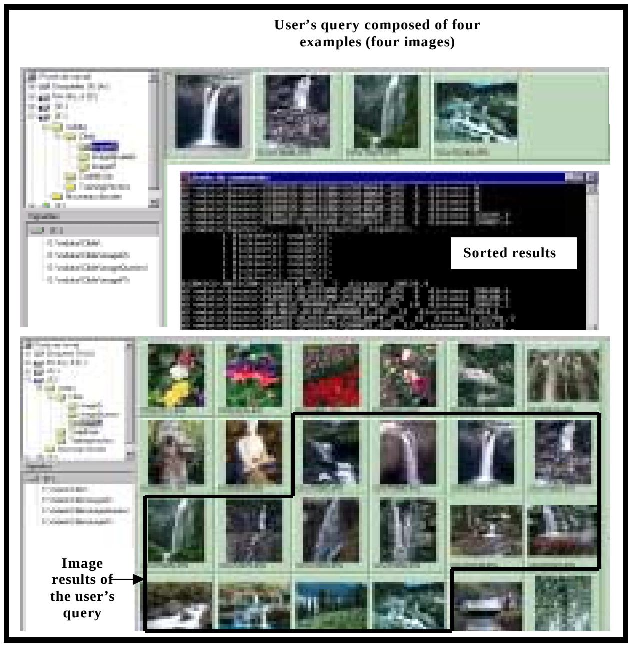

User's query composed of four examples (four images)
Sorted results
Image results of the user's query

Figure 1 : «find images that contain waterfalls».

An important advantage of the advanced indexing is the efficiency of the content-based retrieval. When the user gives examples of image to formulate his query, and asks "find images similar to the examples", the system will not match the source image with all the images in the database. It will match the source image features with only the target image features of suited classes. If the knowledge associated to a class is globally verified, then the considered class is the suited one. Then, the system will focus the search on the sub-classes of the current one. In the target classes that contain few instances, the search is limited to sequential accesses. Another advantage is the richness of descriptions contained in the results of queries since the system presents both similar images and their classes.

2.3 New architecture

The advanced approach for content-based image indexing needs an advanced architecture. The advanced architecture extends the classic architecture by knowledge in the form of simple rules. Simple rules that characterize each semantic class (flowers, natural, mountain, etc.) are automatically extracted. The classic indexing is base exclusively on low level representations of images and physical access structures, without any knowledge and logical representations of the content. The rules describe relationships between visual features (colors and textures of images). Each set of rules associated to a class summarizes image contents of the class. Rules contribute in the discrimination of each class, so they represent knowledge shared by the classes. When images are inserted in the database, it is classified "automatically" in the class hierarchy. At the end of the classification process, the image is inserted in a specific class. In this case, the distance between the image and the knowledge associated to the class is the shortest one, compared to the distance between the image and the other classes. Otherwise, the instantiation relationship between the image and the class, will not be considered.

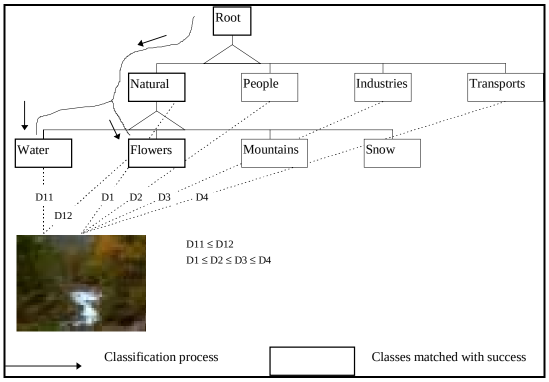


Figure 2 : Example of image insertion into the class hierarchy

This architecture avoids efficient retrievals and browsing through classes. For example, the user may ask "find images similar to the source image but only in People classes" or "find me all images that illustrate the bird class with such colors and such shapes".

3. DISCOVERY HIDDEN RELATIONS

Based on image content description, the knowledge are discovered. The discovered knowledge characterizes visual properties shared by images of the same semantic classes (Birds, Animals, Aerospace, Cliffs, etc.).
The discovery is held into two steps : symbolic clustering and relationship discovering and validation.

1- symbolic clustering

2- relationship discovery and validation

In the first step, numerical descriptions of images are transformed into symbolic form. The similar features are clustered together in the same symbolic features. Clustering simplifies, significantly, the extraction process. For example, in the figure presented bellow (figure 3), the image is composed of region1 and region2. Region1 is characterized by light red color, and region2 by water color and water texture.

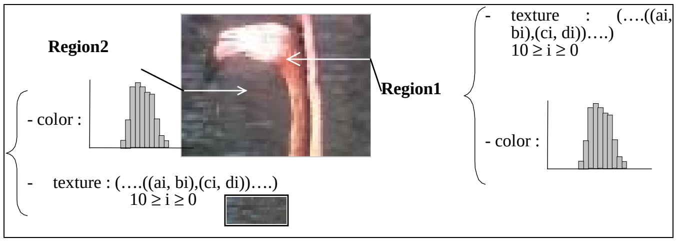

Region2
Region1
- color :
- texture : (….((ai, bi),(ci, di))….)
10 ≥ i ≥ 0
- texture : (….((ai, bi),(ci, di))….)
10 ≥ i ≥ 0
- color :

Figure 3 : Original representation of the image. Numeric representation of image B8169

Light red color is not described by a simple string, but by a color histogram. Even if the region colors of different images of the same class, as presented in figure 4, are similar (i.e. light red), the histograms (numerical representation of color) associated with them are not generally identical.

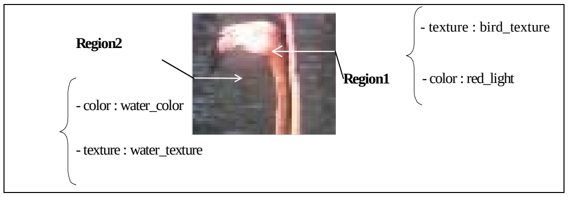


Figure 4 : Symbolic description of image B8169

```
/* Declaration of composition relations
between images and regions. */
  is_composed_of(imageB8169,      [region1,
region2]).
/* Region features declaration. A region is
usually described by texture and color */
/* text attributes. */
  features(region1,                [texture,
bird_texture], [color, red_light]]).
  features(region2,               [[texture,
water_texture], [color, water_color]]).
/* Image features declaration. An image is
usually described by the texture, color. */
features(imageB8169, [[text, text1]]).
```

In the second step, the knowledge discovery engine automatically determines common features between the considered images in rule form. These rules are relationships in the form of `Premise => Conclusion` with a certain accuracy. These rules are called statistical as they accept counter-examples.

```
(texture,  water_texture)  =>  (color,
water_color) (CP 100%, II 96.08%)
(texture, waterfall_texture) => (color,
white_color) (CP 100%, II 87.43%)
(texture,  texture_bird)  =>  (color,
red_light) (CP 100%, II 40.45%)
```

Before presenting the algorithm of discovering, we will present how the image content (color, texture) are represented and extracted automatically. More details about image descriptors have been presented in [Dje 00].

3.1 Image descriptors

3.1.1 Color

The color is the first descriptor of image content. The color feature is extracted automatically from an image or a region. In the first step of the extraction process, based on a physical format, the region or image color is extracted and represented in the RGB model. Based on the RGB model, the color is transformed into HSV model, characterized by three means H, S and V. The HSV model is more suited than the RGB model, in which certain ambiguities appear between colors (ex. Yellow and Green).

In the object-oriented modeling, we define a class of colors called HSV. HSV class includes color histogram and methods (ex. distance measures). The color of a region is represented by a histogram of 256 colors. Each element of the histogram represents the number of pixels that have the suited color (see figures 5, 6). So, comparing the colors of two regions is equivalent to compute the distance between the histogram of the target and the source regions. Before submitting the query, the user may choice the distance, by default quadratic distance is activated.

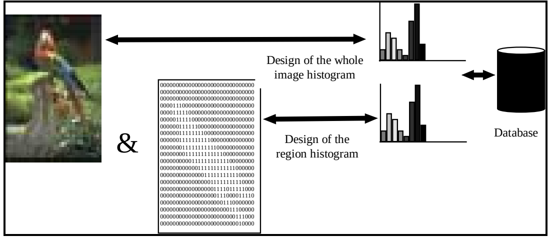

Design of the whole image histogram
&
Design of the region histogram
Database

Figure 5 : Extraction of colors.

In figure 5, the color is represented by a histogram. One histogram represents the color of the whole image, and other histograms represent image region colors. An image region is designed by a binary mask. For example, the binary mask designs the image region that characterizes the bird. The binary mask is equal to 1 inside the region, and 0 outside the region. The histogram of colors are calculated on the basis of the binary mask and the photo.

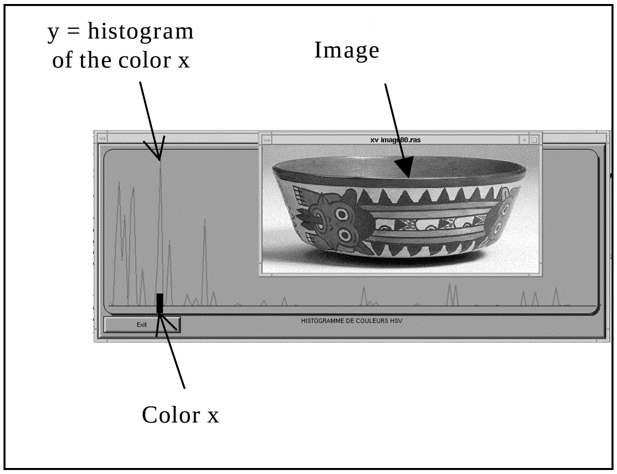


Figure 6 : Color histogram.

In figure 6, the graphic representation of the image color histogram is displayed. For example, y is the histogram of the color x ⇔ y = number of pixels that have the color x.

3.1.2 Texture

The texture is an important aspect of human visual perception, and it is the second important feature extracted automatically from image regions.

When two patterns differ only in scale, the magnified one is coarser. The variance measures the dispersion of the difference of gray-level with a certain distance. The contrast measures the vividness of the texture and is a function of the gray-level difference histogram. The directionality measures the « peakedness » of the distribution of gradient directions in the image. For example the region may have a favored direction. It is not a powerful texture representation, but may be interesting for retrieval process when mixing it with color features.

The approach, considered, implements a powerful texture representation. Thus, we use a mathematical model which is one of the best : Fourier model [Zah 72]. Fourier model has very interesting advantages : - the texture can be reconstructed from the descriptors. – it has a mathematical description rather than a heuristic one. - And finally, the model supports the robustness of description to translation, rotation and scale transformations. An important contribution of our representation is our extension of Fourier model to texture description. This extension considers the matching process. In this extension, we consider texture(t) composed of two functions : x(t) and y(t).

So texture(t) =(x(t), y(t)). x(t) represents the different level of gray of x, and y(t) represents the different level of gray of y. t indicates the different indices of the signal texture. t = 0, N-1. N is the period of the function, and N = number of x values and y values = length of the normalized image. So, we have two suites of coefficients S(a_n, b_n) and S(c_n, d_n) that represents Fourier coefficients of x(t) and y(t) respectively.

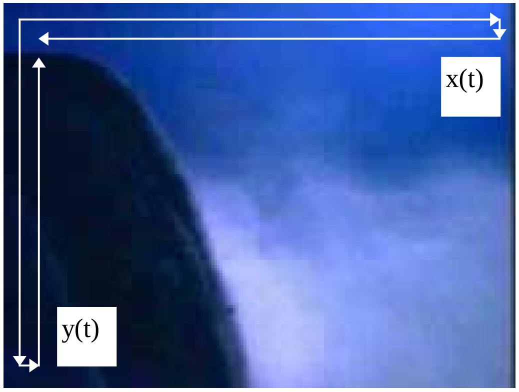


Figure 7 : x(t), y(t)

$$x(t) = a_0 + \sum_{k=1,N} a_n \cos(2\pi kt/N) + b_n \sin(2\pi kt/N)$$

$$y(t) = b_0 + \sum_{k=1,N} c_n \cos(2\pi kt/N) + d_n \sin(2\pi kt/N)$$

and

$$a_n = 2/N \sum_{k=1,N} x(t) \cos(2\pi kt/N)$$

$$b_n = 2/N \sum_{k=1,N} x(t) \sin(2\pi kt/N)$$

$$c_n = 2/N \sum_{k=1,N} x(t) \cos(2\pi kt/N)$$

$$d_n = 2/N \sum_{k=1,N} x(t) \sin(2\pi kt/N)$$

Figure 8 : Fourier Coefficients formulas

We consider only eleven coefficients of Fourier that select the lowest frequencies of the sub-band k ∈ [0-10]. In this extension, we modify the similarity measures (Euclidean distance) in order to consider the coefficients of the two signals x(t) and y(t), as we will see in the following section.

3.2 Symbolic clustering algorithm

The clustering of numeric features in symbolic form raises several problems. The first problem is that a feature may belong to one or several symbol(s). The problem is the same for texture and color features. The second problem is a consequence of the first one. After the symbol creation, we can obtain two different symbols that may be either composed of the same numerical features (equal symbols), or composed of several symbols that differ on only one feature. If we obtain two different symbols composed of the same features, the system keeps only one symbol among symbols composed of the same features. If we obtain several symbols that differ on only one numerical feature, then, it is more difficult to resolve. The problem is the same for the other features. The third problem is that the system generates a symbolic feature base bigger than the numeric feature base since the system computes for one fact containing numeric values, several facts containing symbolic values. The figure presents a part of a symbolic feature and illustrates the possibility of feature fact explosion.

To resolve these problems, we implemented a technique that clusters numerical representation of color, texture, by using data quantization of colors and textures, we use also the term of feature book creation. The color and texture clustering algorithms are similar, the difference is situated in the distance used.

3.2.1 Principle of the algorithm

The algorithm is a classification approach based on the following observation. The scalar quantification of Lloyd developped in 1957 is valide for our vectors (color histogram, fourier coefficients), four rate distribution and for a large variety of distortion criteria. It generalizes the algorithm by modifying the feature book iteratively. This generalization is known by k-means [Lin 80]. The objective of the algorithm is to create a feature book, based on automatic classifications themselves based on a learning set. The learning set is composed of feature vectors of unknown probability density. Two steps should be distinguished :

- A first step of classification that clusters each vector of the learning set around the initial feature book that is the most similar. The objective is to create the most representative partition of the vector space.

- A second step of optimization that permits the correct adaptation in a class of the feature book vector. The gravity center of the class created in the previous step is computed.

The algorithm is reiterated in the new feature book in order to obtain a new partition. The algorithm converges to stable position by evolving at each iteration the distortion criteria. Each application of the iteration of the algorithm should reduce the mean distortion. The choice of the initial feature book will influence the local minimum that the algorithm will achieve, the global minimum corresponds to the initial feature book. The creation of the initial feature book is inspired of the splitting technique [Gra 84].

The splitting method decomposes a feature book Y_k into two different feature books $Y_{k-\varepsilon}$ and $Y_{k+\varepsilon}$, where ε is a random vector of weak energy, and its distortion depends of the distortion of the splited vector. The algorithm is then applied to the new feature book in order to optimize the reproduction vectors.

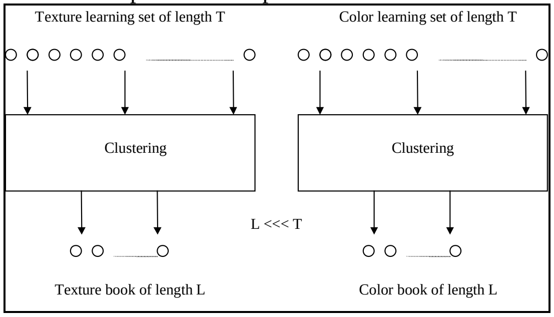


Figure 9 : Clustering and reduction algorithm. In our experiments T = 30.000 and L = 256

3.2.2 *Distances*

The system clusters similar colors together in a symbolic form by using a suitable distance. In our case, for the color, we implement the quadratic distance which is one of the most accurate distances.

$$D_Q(H,I)=\sqrt{(H-I).A(H-I)^T} \text{ or } D_Q^2(H,I)=\sum_{p=1}^{n}\sum_{q=1}^{n} a_{pq}\left(h_{c_p}-i_{c_p}\right)\left(h_{c_q}-i_{c_q}\right)$$

$$\text{With } A\text{: the similarity matrix } (n\times n), A=\left[a_{pq}\right], a_{pq}\text{: weight of the similarity between the } p \text{ and } q \text{ bins}$$

Figure 10 : Quadratic_distance definition.

This distance takes into account the color similarity between the histogram bins by using the symmetrical similarity matrix *A*. The matrix weights may be normalized to obtain $0 \le a_{pq} \le 1$. So, the matrix diagonal is equal to 1, since any color is identical with itself (a_{pp}=1). A coefficient a_{pq} close to 0, represents a dissimilarity between *p* and *q* bins. For example, in QBIC, the quadratic distance between two color histograms, is used with a similarity matrix *A* whose elements are defined by [Haf 95]: $a_{ij} = (1 - d_{ij}/d_{max})$, with $d_{max} = max_{ij}(d_{ij})$, d_{ij} being Euclidean distance between the color i and j in any color space. The two distributions *H* and *I*, may also be normalized in order that $0 \le h_{c_p}, i_{c_p} \le 1$ $and \sum_p h_{c_p} = 1 = \sum_p i_{c_p}$.

$$D_{L2}(H,I)=\sqrt{\sum_{l=1}^{n}(h_{c_l}-i_{c_l})^2}$$

Figure 11 : L2-distance or Euclidean distance definition.

This distance makes it possible to obtain satisfactory results since it appreciates color similarity correctly. However, its major drawback is that it is time-consuming compared to the other distances. Euclidean distance results from the quadratic distance where *A* matrix is the identity matrix (no correlation between the histogram bins).

In our example, the light red color zones in the different images are grouped together in the symbolic form red_light as they are similar. Water color in not clustered in red_light, because the distance between them is not short enough. However, it is clustered in the symbolic form water_color shared with other images. In the same way and based on appropriate distances, the system clusters respectively similar shapes, similar textures together in a symbolic form.

For the texture, we implement an adaptation of the Euclidean distance to Fourier coefficients, we call it « texture_Fourier_distance ». So, the matching distance between the Fourier descriptors of the texture *t'* of an image image' and the Fourier descriptors of the texture *t* of an image « image », is triggered by computing the distance between *t* and *t'*, namely: $d(t,t') = \sqrt{(\sum_{n=1,N} (|T'_n - K.|T_n|)^2)}$, N=10, for t and t' textures, we have a positive constant K, and for any n ≠ 0, $|T'_n| = K*|T_n|$, where $Z_n = \sqrt{(|X_n|^2 + |Y_n|^2)} = \sqrt{(a_n^2 + b_n^2 + c_n^2 + d_n^2)}$
That is to say, the textures are identical near to one geometric transformation. The translation, scale and rotation have no effect on the module of Fourier coefficients. $K = 1/N*(\sum_{n=1,N}(|T'_n|/|T_n|))$ is an estimation of *K* which minimizes the error on the *N* (e.g. 11) first coefficients of Fourier.

3.2.3 *Algorithm*

Based on the learning set of length equal to T, the algorithm finds a feature book of colors and textures of length equal to L, that are the most representative colors and textures of image databases.

Global Clustering

```
FeatureBook $Y_f$ = SymbolicClustering (visual feature = VisualFeature, learning set = LearningSet, $Y_0$, T, L)
{
if the VisualFeature = color then LearningSet = {$H_1$, $H_2$, $H_3$, ..., $H_T$}, a set of T histograms.
If VisualFeature = texture then LearningSet = {S1, S2, ...., $S_T$}, a set of T sequence of Fourier coefficients. $Y_0$ is the initial feature book with distortion $D_0$ and cardinal equal to L.
Pre-conditions : L << T
Invariant : s ≤ S=L/2
1- Initialization : $D_0$ = Distortion ($Y_0$) ; $E_0$ = Entropy($Y_0$) ; s = 0 ; s = number of splitting activated. $Class_0$ = {$Class_{0,k}$ ; k = 1, ..., L}
```

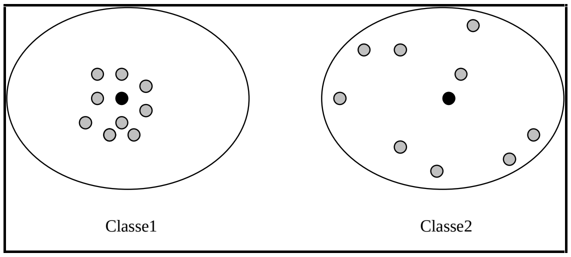


Figure 12 : Distorsion(Classe1) < Distorsion(Classe2)

```
    While (s < S)
    {
2 - s = s + 1
```

3 - Splitting of the VisualFeature of the feature book Y_{s-1} that support the highest apparition probability p_i. p_i corresponds to the $class_{s,i}$ that has the maximum number of instances. The VisualFeature of the feature book corresponds to the gravity center of the $class_{s,i}$. $(Y_{s-1,i'}, Y_{s-1,i''}) = splitting(Y_{s-1,i})$.

4 - Deletion of the VisualFeature of the feature book Y_{s-1} that support the lowest apparition probability p_j. p_j corresponds to the $class_{s,j}$ that has the minimum number of of instances. The VisualFeature of the feature book corresponds to the gravity center of the $class_{s,j}$.

Each splitting is followed by a deletion, so the cardinal of the feature book remains constant (equal to L).

5 - A local clustering with the parameter E_1 is executed on the class $class_{s,i}$ on the local feature book composed of $Y_{s-1,i'}$, $Y_{s-1,i''}$ and E_1 the stop criteria of the algorithm.

Y_s = Clustering(visual feature = VisualFeature, feature book = $(Y_{s-1,i'}, Y_{s-1,i''})$, E_1, learning set = $class_{s,i}$).

6 - A global clustering is executed on the global feature book composed of Y_s with the parameter E_2. E_2 is the stop criteria of the algorithm.

Y_s = Clustering(visual feature = VisualFeature, feature book = Y_s, E_2, learning set = $class_s$) ;

D_s = Distortion (Y_s) ;

E_s = Entropy(Y_s).

$D_s < D_0$: the distortion is reduced and $H_s > H_0$: the entropy is augmented}}

Ideally, the stop criteria of the algorithm should depend of the distortion D_s, however, the distortion D_s depends of the number of splitting.

Local clustering

FeatureBook Y_f = Clustering(visual feature = VF, learning set = LS, Y_0, Y_f, T, L, E)

{

Y_0 is the initial feature book with distortion D_0 and length equal to L. LS is the learning set with a length is equal to L. E is the stop criteria.

Pre-conditions : L << T

1 - Initialization : D_0 = Distortion (Y_0) ; s = 0 ; s = number of splitting activated.

Do

{

2 - Based on the feature book $Y_s = \{Y_{s,k}\ k=1,..,L\}$ and the learning set LS; we extract the partition $Class_s = \{Class_{s,k} ; k = 1, ..., L\}$, in which distance(x, y) is minimal. So :

$x_t \in Class_{i,k}$ when $distance(x_t, y_k) \le distance(x_t, y_j)\ \forall j \neq k$.

$D_s = 1/T \sum_{t=1,T} \min_Y distance(x_t, y), y \in Y_s$

if VF = texture then distance = texture_fourier_distance, presented bellow.

if VF = color then distance = quadratic_distance, presented bellow.

3 - Creating the optimal catalogue $Y_{s+1} = \{centroid(Class_{s,k})\ k=1,..,L\}$; $centroid(Class_{s,k})$ corresponds the gravity center of the class $Class_{s,k}$. $centroid(Class_{s,k}) = (1/|Class_{s,k}|) * \sum x_t / t : x_t \in Class_{s,k}$. $|Class_{s,k}|$ is the number of instances in $Class_{s,k}$.

4 - s = s + 1

} Until $(D_{s-1} - D_s)/D_s < E$}

The distortion D_s is a positive and decreasing function. Each iteration of the algorithm reduce the distortion. So, $D_{s-1} \ge D_s$.

The experimental results showed that the distortion values decrease quickly compared to splitting evolution. After the quick decreasing, the distortion values decrease very slowly. Conversely, The entropy increase quickly compared to splitting evolution, and then, it increases very slowly.

3.3 Relationship discovery and validation

Based on the feature book, the discovery engine is triggered to discover the shared knowledge in the form of rules, and this constitutes the second step the general algorithm.

Accuracy is very important in order to estimate the quality of the rules induced. The user should indicate the threshold above which rules discovered will be kept (relevant rules). In fact, the weak rules are rules that are not representative of the shared knowledge. In order to estimate the accuracy of rules, we implement two statistical measures : conditional probability and implication intensity. The conditional probability formula of the rule a => b makes it possible to answer the following question: ''what are the chances of proposition b being true when proposition a is true ? The definition of this measure is $P(b/a) = Card(A \cap B)/Card(A)$

More intuitively, conditional probability allows us to estimate the accuracy of a rule, considering the number of counter-examples. For example, let us consider p_1 (a => b) and p_2 (b => a) conditional probabilities are respectively 100% and 5.6%. So, the rule b =>a has a lot of counter-examples. In *E* (universe set), there are lots of objects that belong to B, but not to A. Conversely, the rule a => b has no counter-example. So, objects that respect proposition a, respect also proposition b.

Conditional probability allows the system to determine the discriminating characteristics of considered images. Furthermore, we completed it by the intensity of implication [Gra 82]. For example, implication intensity requires a certain number of examples or counter-examples. When the doubt area is reached, the intensity value increases or decreases rapidly contrary to the conditional probability that is linear. In fact, implication intensity simulates human behavior better than other statistical measures and particularly conditional probability. Moreover, implication intensity increases with the considered population sample representativity. The considered sample must be large enough in order to draw relevant conclusions. Finally, implication intensity takes into consideration the sizes of sets and consequently their influence. For example, conditional probability of a => b is P_1 (100%) and implication intensity of a =>b is φ_1 (23%) values are very different because conditional probability does not take into consideration the fact that proposition b is verified by lots of objects. On the contrary, implication intensity considers that it is not surprising that an object of *A* verifies proposition b because proposition b is verified by many objects of the considered sample.

Let A,B and E sets respectively be the sets of instances that verify proposition a, the set of instances that verify proposition b, and the set of all instances or the universe set. From a theoretical point of view, implication intensity measures the degree of statistical astonishment of size $A \cap \overline{B}$ (this set contains objects that verify proposition a and that do not verify proposition b) considering the sizes of *A*, *B* and E sets, and assuming there is no a priori link between A and B. The cardinals or the sizes of A and B subsets of E are determined by the objects of the database belonging to A and B.

The knowledge discovery engine returns the rules in the form of Premise => Conclusion whose intensity and conditional probability are greater than or equal to a certain threshold. For the moment, this threshold is defined manually (ex. 90 %). Samples of extracted rules by the prototype are (texture, water_texture) => (color, water_color), (texture, waterfall) => (color, white) with respective conditional probability values of 100% and 100%, and implication intensity values of 96.08% and 87.08 %.

3.4 Some comments

The set of induced rules corresponds to knowledge shared by classes. This knowledge is helpful for user's comprehension of the class. Extracted rules are validated when the conditional probability and the rule intensity are greater than a special value (i.e. 90% for conditional probability and 80% for implication intensity). For example, (texture, bird_texture) => (color, red_light) has 100% conditional probability and 40.4598% implication intensity. Since the rule intensity is less than 80%, the system will not store it. We explain this weak measure of rule intensity by the fact that there are few examples that respect this rule.

In our example, the searched class is characterized by a set of rules such as rule 1. So, if we have the ''water_texture'' texture in an image of the class, then the region color inside the image is red_light with 100% conditional probability and 96,08 % rule intensity. So, during image database creation, the classification of an image in a class is possible if the class rules, previously extracted and validated, are globally respected. At least 50 % of rules are respected. If not, we will not consider the instantiation relationship between the image and the class.

x => y has 15.3846% conditional probability and 61.79% implication intensity, that is to say that the conditional probability value is less than 90%. So, the system did not store this rule. We explain this weak measure of conditional probability by the fact that there are a lot of counter-examples of the considered rule.

(texture, waterfall) => (color, white) is a good rule because the conditional probability value is 100% and the implication intensity is 81.79%. This rule means that when we have a texture that includes water, then we would have a white region color.

In the retrieval task, when the user specifies an image (called source image) as the basis of his query, and asks ''find images similar to the source image'', the system will not match the source image with all the images of the database. It will match the source image features with all the target images of the appropriate classes. These classes contain rules globally respected by the source image.

For example, if we have a source image that contains a ''texture_waterfall'', but it does not globally verify the rules associated with this concept, we can deduce the weakness of the relationship between the source image and the class. The system matches the source image with classes through their rules stored in the database.

4. EXPERIMENTAL RESULTS AND CONCLUSION

We have conducted extensive experiments of varied data sets to measure the performance of the advanced content-based query.

The recall and precision graphic for our system are computed as follows. References («query») of images are selected from a test collection. A sub-set of images is selected per class (waterfalls, fires, panorama, etc.). For each image, a knowledge content-based query is formulated. For an image reference, we associate a knowledge content-based query that includes visual features (color, texture, color + texture). We also associate a classic content-based query that uses classic indexing (there is no knowledge integration).

To demonstrate the efficiency of the knowledge content-based queries, the results of the advanced content-based queries are compared with the results of queries that do not use classic content-based queries. Since it is not possible to retrieve all relevant images, our experiment evaluates only the first ranked images.

Judging on the results, it is obvious that the use of knowledge leads to improvements in both precision and recall over majority queries tested. The average improvements of advanced content-based queries over classic content-based queries are 23% for precision and 17 % for recall. Precision and recall are better for concept-based queries (queries that mix visual features and textual descriptions with different degrees of importance) than for queries that use only visual features such as color or shapes or textures or textual descriptions, but not both.

Acknowledgement

Many thanks to Henri Briand for his encouragement.

REFERENCES

[And 85] Andrew W. Appel, An Efficient Program for Many-Body Simulation, SIAM Journal of Statistical and Scientific Computing, 6(1), January 1985.

[Bay 72] Bayer, R., E. McCreight. Organization and Maintenance of Large Ordered Indexes. Acta, 1972, Informatica 1(3), 173-189.

[Dje 00] Djeraba C., Bouet M., Henri B., Khenchaf A. « Visual and Textual content based indexing and retrieval », to appear in International Journal on Digital Libraries, Springer-Verlag 2000.

[Fay 96] Fayyad U. M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R., «Advances in Knowledge Discovery and Data Mining», AAAI Press, MIT Press, 1996.

[Gut 84] Antonin Guttman: R-Trees: A Dynamic Index Structure for Spatial Searching. SIGMOD Conference 1984, pages 47-57, 1984.

[Gra 82] Gras Régis, THE EISCAT CORRELATOR, EISCAT technical note, Kiiruna 1982, EISCAT Report 82/34, 1982.

[Gra 84] Gray R. M. « Vector Quantization », IEEE ASSP Mag., pages 4-29, April 1984.

[Gup 97] Amarnath Gupta, Ramesh Jain «Visual Information Retrieval», A communication of the ACM, May 1997/Vol. 40, N°5.

[Haf 95] Hafner J., al. «Efficient Color Histogram Indexing for Quadratic Distance Functions». In IEEE Transaction on Pattern analysis and Machine Intelligence, July 1995.

[Jai 98] Ramesh Jain: Content-based Multimedia Information Management. ICDE 1998: 252-253

[Lin 80] Linde Y., Buzo A., Gray R. M. « An algorithm for Vector Quantizer Design », IEEE Trans. On Comm., Vol. COM-28, N° 1, pages 84-95, January, 1980.

[Moo 51] Moores C. N. «Datacoding applied to mechanical organization of knowledge» AM. Doc. 2 (1951), 20-32.

[Rag 89] Raghavan, V., Jung, G., and Bollman, P., “A Critical Investigation of Recall and Precision as Measures”, ACM Transactions on Information Systems 7(3), page 205-229, 1989.

[Rap 74] Raphael A. Finkel, Jon Louis Bentley: Quad Trees: A Data Structure for Retrieval on Composite Keys. Acta Informatica 4: 1-9, 1974.

[Rij 79] C. J. Keith van Rijsbergen «Information retrieval», Second edition, London: Butterworths, 1979

[Sal 68] Salton Gerard «Automatic Information Organization and Retrieval», McGraw Hill Book Co, New York, 1968, Chapter 4.

[Zah 72] C. T. Zahn, R. Z. Roskies, « Fourier descriptors for plane closed curves », IEEE Trans. On Computers, 1972.

Semantic Indexing and Temporal Rule Discovery for Time-series Satellite Images

Rie Honda
Kochi University
Akebono-cyo 2-5-1
Kochi, JAPAN 780-8520
honda@is.kochi-u.ac.jp

Hirokazu Takimito
Kochi University
Akebono-cyo 2-5-1
Kochi, JAPAN 780-8520
takimoto@is.kochi-u.ac.jp

Osamu Konishi
Kochi University
Akebono-cyo 2-5-1
Kochi, JAPAN 780-8520
konishi@is.kochi-u.ac.jp

ABSTRACT

Feature extraction and knowledge discovery from a large amount of image data such as remote sensing images have become highly required recent years. In this study, we present a framework for data mining from a set of time-series images including moving objects using clustering by self-organizing mapping(SOM) and extraction of time-dependent association rules. We applied this method to weather satellite cloud images taken by GMS-5 and evaluated its usefulness. The images are classified automatically by two-stage SOM. The results were examined and the cluster addresses were described in regard to season and prominent features such as typhoons or high-pressure masses. Sequential images are then transformed into a data series expressed by cluster addresses and time of occurrence, from which time-dependent association rules (simple serial rules) are extracted using a method for finding frequently co-occurring term-pairs from text. Semantic indexed data and extracted rules are stored in the database, which allows high-level queries by entering SQL through user interface, and thus supports knowledge discovery for domain-experts. We believe that this approach can be widely useful and applicable to knowledge discovery from an enormous amount of multimedia data, which includes unknown sequential patterns.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*data mining, image database, scientific databases*
; H.3.3 [**Information Systems**]: Information Search and Retrieval—*clustering*; J.2 [**Computer Applications**]: Physical Science and Engineering—*earth and atomspheric science*

General Terms

DESIGN, EXPERIMENTATION, PERFORMANCE

Keywords

Satellite image database, clustering, self-organizing feature map, time dependent association rule, R-tree, content-based image retrieval, SQL query

1. INTRODUCTION

A huge amount of data has been stored in databases in the areas of business or science. Data mining or knowledge discovery from database(KDD) is a method for extracting unknown information such as rules and patterns from a large-scale database. The well-known data mining methods include decision tree, association rules[3], classification, clustering, and time-series analysis[1][2].

The process of the data mining is composed of the following six parts: (1) acquisition of input data, (2) selection of input data, (3) preprocessing, (4) transformation, (5) extraction of patterns, rules, etc., and (6) interpretation and evaluation of the results.

There are two main areas of in the data mining: one focused on business data and one focused on scientific data.

One of well-known cases of scientific data mining is the Sky Image Cataloging and Analysis tool (SKICAT) developed for the second Palomar Observatory Sky Survey[6]. They extracted astronomical body candidates from enormous raw images and classified them using a decision tree. In this process the researchers discovered both the classification rules and the novel bodies. Smyth et al.[8] and Burl et al.[7] have also reported a discovery system for venusian volcanoes based on synthetic aperture radar images taken by the spacecraft Magellan, which are very effective as recognition guides.

Image data such as satellite images and medical images often amount to several Tera bytes, thus manual and detailed analysis of these data becomes impractical[5]. Therefore an automated(or semi-automated) procedure

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

to extract knowledge from these data should be included in the data mining from the image database.

In our recent studies[14][15], we have applied data mining methods such as clustering and association rules to a large number of the satellite weather images over the Japanese islands taken by Japanese stationary satellite GMS-5. These weather images are accumulated everyday and form a large amount of raw database.

Metrological events are considered to be chaotic phenomena in that an object such as a mass of cloud changes its position and form frequently. Furthermore they are time-sequential data such as video images.

Features of our studies applied to the weather images are summarized as follows:

(1) The application of data mining method to image classification and retrieval.

(2) Feature description from time-series data.

(3) Implementation of the result of classification as the user retrieval interface.

(4) Construction of the whole system as a domain-expert supporting system.

We describe an overview of the system in Section 2. A clustering algorithm for time-sequential images and its experimental results are described in Section 3. Section 4 describes the algorithm of extraction of time-dependent association rules and its experimental results. Section 5 describes details of the construction of the database by using R-tree and the results of its implementation. Section 6 provides a conclusion.

2. SYSTEM OVERVIEW

We constructed a weather image database that gathers the sequential changes of cloud images and the domain-expert analysis support system for these images. We characterize the system's images using clustering method (Section 3) and describe the image changes in terms of the sequential cluster numbers. Then we derive the time-dependent association rules from the sequential data (Section 4) and index them. The flow of this system is shown in Figure 1 and described as follows:

step 1 Clustering using a self-organizing map.

step 2 Generation of time-sequential data from a series of cluster addresses.

step 3 Extraction of time-dependent rules from the time-sequential data.

step 4 Indexing of rules and a series of cluster addresses by using R-tree, and construction of the database.

step 5 Searching for time-sequential variation patterns and browsing for the retrieved data in the form of animation.

The above-described process enables us to characterize enormous amount of images acquired at a certain time interval semi-automatically, and to retrieve the images by using the extracted rules. For example, this process enables queries like "search for frequent events that occur between one typhoon and the next typhoon", or "search for a weather change such that a typhoon occurs within 10 days after a front and high pressure mass developed within the time interval of 5 days".

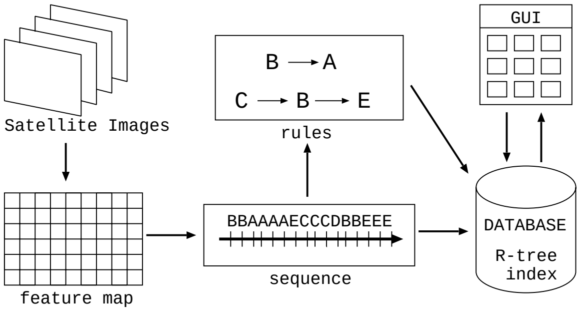


Figure 1: Overview of the system.

3. TIME-SEQUENTIAL DATA DESCRIPTION BY USING CLUSTERING

3.1 Data set description

Satellite weather images, taken by GMS-5 and received at the Institute of Industrial Science, Tokyo University, are archived at the Kochi University weather page (http://weather.is.kochi-u.ac.jp). The images used in this study are infrared band(IR3: moisture band, wavelength of 6.5-7 μm) images taken Japanese islands, which are of 640-pixels in width and 480-pixels in height. Each image is taken every hour, and about 9000 images are archived every year. Figure 2 shows an example of the image sequence.

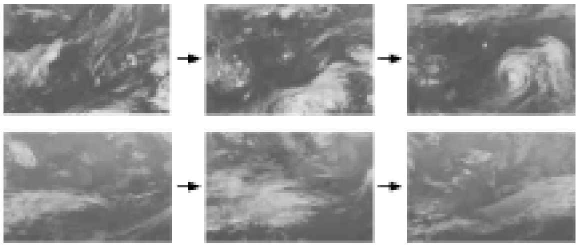


Figure 2: Example of weather image(GMS-5 IR3 band) sequence.

We considered that conventional image processing methods might be unable to detect moving objects such as the cloud masses that change their position and form during the time sequence. Thus we used the following SOM-based method for the automatic clustering of images by using the raster-like scanned image intensity vectors as the inputs.

3.2 Clustering and Kohonen's self-organizing map

Similarity analysis from full-text databases or image databases use sophisticated retrieval methods based on the indexing of space or the indexing of feature spaces. Clustering based on similarity is one of the extensions of these methods. When standard feature patterns are not given for the object data set, distance criteria in the feature spaces are used to divide the object set into the subset. This provides the rough structure to the given non-structured information.

Kohonen's self-organizing map (SOM)[9] is a paradigm which was suggested in 1990. The SOM is a two layer network that organizes a feature map by discovering feature relations based on input patterns through iterative non-supervised learning.

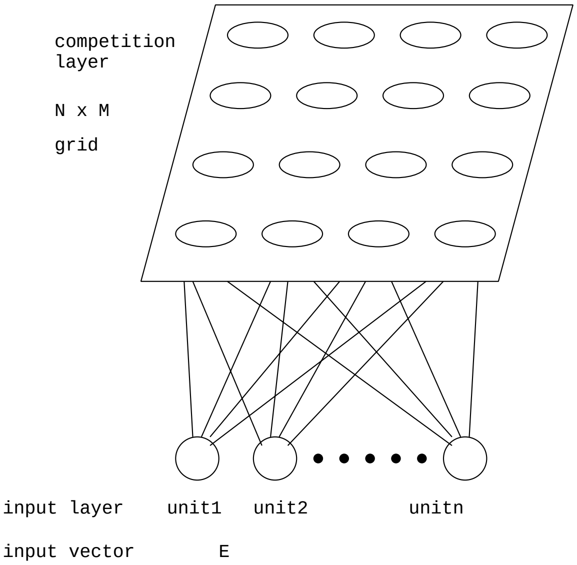


Figure 3: Basic structure of Kohonen's self-organizing map

Figure 3 presents basic schematic structure of Kohonen's self-organizing map. The network, a combination of the input layer and the competition layer, is trained through non-supervised learning. Each unit of the input layer has a vector whose components correspond to the input pattern elements.

The algorithm of the SOM is described as follows:

step 1 Let the input pattern vector $E \in R^n$ as,

$$E = [e_1, e_2, e_3, \cdots, e_n] \qquad (1)$$

step 2 Assume the weight of union from the input vector the to a unit i as

$$U_i = [u_{i1}, u_{i2}, u_{i3}, \cdots, u_{in}]. \qquad (2)$$

Initial values of u_{ij} are given randomly.

step 3 E is compared with all U_i, and the best matching node which has the smallest Euclidean distance $|E - U_i|$ is determined and signified by the subscript c,

$$c = \text{argmin}_i |E - U_i|. \qquad (3)$$

step 4 Weight vectors of the best matching node c and its neighbors,N_c, are adjusted to increase the similarity as follows,

$$u_{ij}^{new} = u_{ij}^{old} + \Delta u_{ij} \qquad (4)$$

where

$$\Delta u_{ij} = \begin{cases} \alpha(e_j - u_{ij}) & (i \in N_c) \\ 0 & (i \notin N_c) \end{cases} \qquad (5)$$

$$\alpha_t = \alpha_0 \left(1 - \frac{t}{T}\right) \qquad (6)$$

The α_t is the learning rate at the time of t iterations, α_0 is the initial leaning rate, and T is the total number of iterations.

step 5 The learning rate and the size of neighbor decreases as the learning proceeds.

The input signals E are classified into the activated (nearest) unit U_c of the input layer and projected onto the competition grids. The distance on the competition grids reflects the similarity between the patterns. After the training is completed, the obtained competition grids. i.e., the feature map, represents a natural relationship between the patterns of input signals entered into the network.

3.3 Clustering by two-stage SOM

Figure 4 represents the problem of clustering of weather images. Two images in Figure 4(a) are considered to have features similar to those of typhoon and a front, although their forms and positions are changed. When we take the input vectors simply as the raster-like scanned intensity vectors, these images are classified into the different groups based on the spatial variations of intensity. We considered that this difficulty is avoided by dividing the images into blocks as shown in Figure 4(b).

The procedure adopted here, named two-stage SOM, is described as follows:

stage 1 **Clustering of pattern cells**

step 1 All Images are divided into N×M blocks.

step 2 SOM generates the feature map, taking the each block's raster-like scanned intensity vectors as the input vectors.

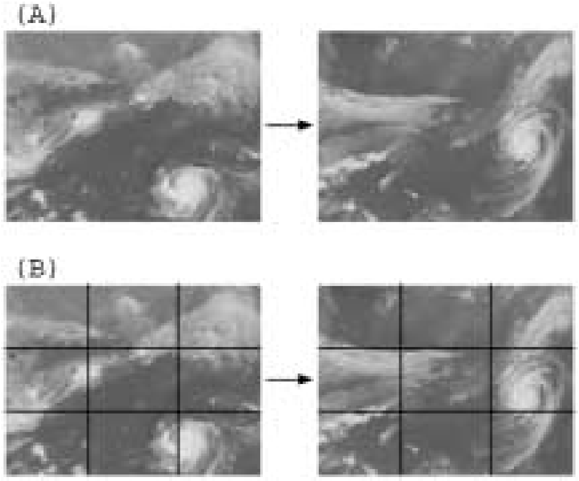


Figure 4: Problem for clustering of weather images.

step 3 The SOM map cluster address is used to describe blocks of the original images. We refer to this characterized blocks as the pattern cells.

stage 2 **Clustering of the images by using frequency histograms of pattern cells.**

step 1 Each image is represented as the frequency histogram of the pattern cells.

step 2 The feature map of SOM is generated by taking the frequency histogram of each image's pattern cell as the input.

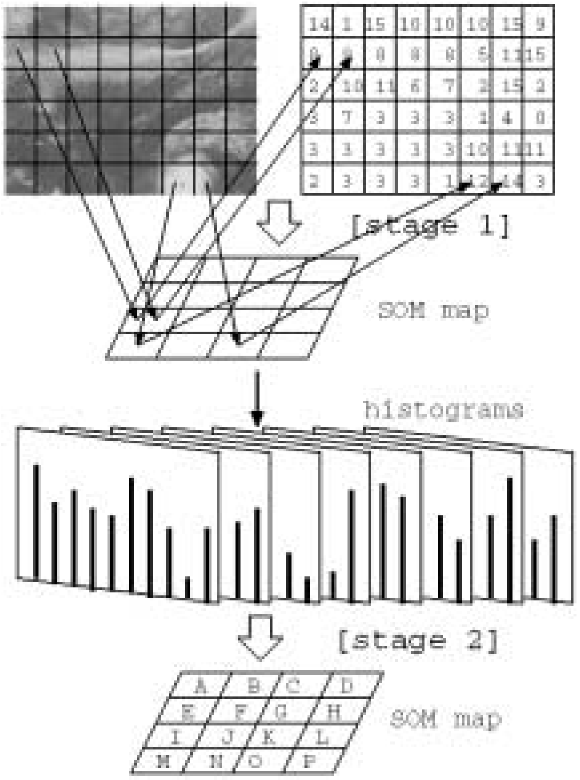
Extraction of frequency histogram of pattern cells in step 1 of stage 2 reduces the spatial information of blocks included in the images. Thus this process enables to classify time-series images which have similar objects at different positions as the same type of images.

Figure 5 schematically shows the above-described process of the two-stage SOM. The images that have similar objects are clustered into similar cells on the second stage feature map. Note that the difference in seasons is not distinguished at this point.

3.4 Result of experiments on clustering

In our experiments, we sampled GMS-5 IR3 images with 8 hour time intervals obtained between 1997 and 1998, and composed two data set for 1997 and 1998 which include 1044 and 966 images, respectively. We defined number of blocks for each image to be 12 × 16. The sizes of feature maps of both first stage SOM and second stage SOM are defined to be 4 × 4. Learning processes are iterated 8000-10000 times.

The results of the experiment show that images with similar features are classified into similar cells. To evaluate the accuracy of clustering quantitatively, we defined the following parameters,



14	1	15	10	10	10	15	9
0	0	0	0	0	5	11	15
2	10	11	6	7	2	15	2
3	7	3	3	3	1	4	0
3	1	3	3	3	10	11	11
2	3	3	3	1	12	14	3

Figure 5: Clustering of weather images by SOM.

$$Precision = B/(B+C), \tag{7}$$

$$Recall = B/(A+B), \tag{8}$$

where A is the number of the nonrelevant images that are classified into the cells, B is the number of the relevant images that are classified into the valid cell, and C is the number of the relevant images that are classified into the invalid cell.

Table 1 show the precision values for 1997 and 1998 to be 86.0% and 86.7%, respectively, and that the values of recall are 84.6% and 86.7%, respectively. These values indicate that the clustering of weather images by two-stage SOM can successfully learn the features of images and can classify them with a high degree of accuracy.

Table 1: Accuracies of clustering

year	Recall	Precision
1997	86.0%(876/1022)	84.6%(876/1044)
1998	86.7%(838/945)	86.7%(838/966)

Furthermore, we describe the semantic representation of clusters by specifying the season in which the clusters are observed, based on the frequency of each cluster every month, and by describing the representative object such as front or typhoon by means of visual observation of images in the cluster from a domain-expert like view. Table 2 shows the semantical descriptions of 1997 and 1998. The distribution of similar clusters for 1997 is different from 1998 since we performed the SOM leaning for these datasets independently. However, most of the groups are observed in both maps, thus the obtained result is meaningful even in the view of the domain-expert knowledge.

The obtained map is considered to be dependent on the block size of the original images and size of SOM map. Hierarchical division of each block in the original image by using standard deviation of intensity will be a solution to the determination of block sizes. The algorithm of Growing Hierarchical SOM[16], which is capable of growing both in terms of map size as well as the three-dimensional tree structure, will be effective for the adaptation of map size.

4. SEQUENTIAL ANALYSIS AND EXTRACTION OF TIME-DEPENDENT ASSOCIATION RULES

4.1 Association rules

Association rules are one of the key concepts of data mining[4]. An item i is defined to be a minimum element for extraction of rules. We define the set of items I and transaction database D as

$$I = [i_1, i_2, \cdots, i_m], D = [T_1, T_2, \cdots, T_n], (T_i \subseteq I), \quad (9)$$

where T_i is an element of the transaction database. A combination of k items is referred to as the item set with the length of k.

Then association rule is represented as

$$X \Rightarrow Y(X, Y \subset I, X \cap Y = \phi). \quad (10)$$

Evaluating parameters of the association rule $X \Rightarrow Y$, support and onfidence, are defined by

$$support(X \Rightarrow Y) = \frac{N(T_i \mid T_i \supseteq X \cup Y)}{N(D)}, \quad (11)$$

$$confidence(X \Rightarrow Y) = \frac{N(T_i \mid T_i \supseteq X \cup Y)}{N(T_i \mid T_i \supseteq X)}, \quad (12)$$

where N is the number of transactions in each condition. These parameters reflect the processing time and effectiveness of the rule.

Rule extraction is defined to find all rules that have larger confidence and support than the minimum threshold defined by users. The following process describes the extraction of association rules.

1. The item set that has larger support than the threshold is selected (referred to as the large item set).
2. The rules that have larger confidence than the threshold are selected from the large item set.

Table 2: Semantical description of each cluster. Cluster address is represented by the character of A, B, C, $\cdots$, P for the raster-like cells scanned from the upper left corner to the lower right corner.

1997

cluster address	season	prominent characteristics
A	spring summer	front, typhoon
B,C	spring autumn	high pressure in the west and low pressure in the east
D,H	spring autumn	band-like high-pressure
E	autumn	migratory anticyclone
F	spring autumn	front
G	autumn winter	linear clouds
I	summer	Pacific high pressure, front
J	spring summer	rainy season's front, typhoon
K,L	winter	winter type, whirl-like cloud
M	summer	Pacific high pressure, typhoon
N	spring summer	high pressure, typhoon
O	winter	cold front
P	spring autumn	migratory anticyclone

1998

cluster address	season	prominent characteristics
A,F,O	spring summer	front, typhoon
B	spring autumn	front, migratory anticyclone
C	summer	Pacific high-pressure
D	autumn	migratory anticyclone
E	spring autumn	band like high-pressure
G	spring summer	Pacific high pressure, front
H	spring summer	rainy season's front
I,K,N	winter	winter type, linear clouds(high pressure in the west and low pressure in the east)
J	summer	Pacific high pressure, front
L,M	winter	cold front
P	autumn winter	linear clouds

4.2 Time-series pattern analysis

Time-sequential data analysis is the method used to extract unknown patterns from time-sequential information, is related to the association rules, and is remarkable in the area of data mining. Episode rule[10][11] are known as one of those methods.

Episodes are defined as the event pairs in a certain time window. Events in time sequence are represented by (e, t), where e is the class of the event and t is its occurrence time. In Figure 4, an event sequence given by a string are represented by (E,31)(F,34)(A,35)(B,37)(C,38) $\cdots$ (D,49), where A, B, C are the event classes, and the number is the time of occurrence.

Figure 6 represents simple examples of episode rules such as those regarding serial episodes as "event B occurs after event A", parallel episodes such as "both events E and F occurs", or a combination of serial episodes and parallel episodes such as "event C occurs after event E and F".

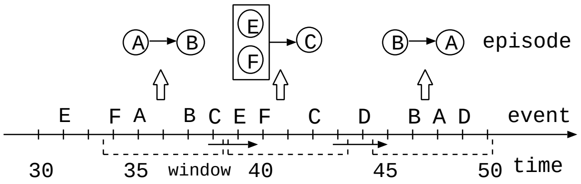


Figure 6: Example of event sequence and episode.

In order to define how closely these events occur, Mannila et al.[10] considered the time window that is shifted in an orderly manner in the sequence. Candidates of episodes are extracted as the co-occurring events in the time window. And combinations of events that have larger frequencies than the threshold frequency are determined to be episodes. A more flexible method that uses the minimal occurrence interval has also been suggested in [11].

4.3 Time-dependent association rule

In this study we present time-dependent association rules which modify the episode rules using the concept of cohesion, and represent local association rules such as "weather pattern B occurs after weather pattern A".

First we generate the sequential data of a weather pattern using cluster addresses as $(A, 1), (A, 2), (C, 3), \cdots$. We define the event as continuously occurring clusters. The event e_i in the sequence is then represented by

$$e_i =< C_i, S_{if}, T_{is}, T_{ie} > (i = 1, \cdot, \cdot, n), \qquad (13)$$

where C_i is the cluster addresses, S_{if} is the continuity, T_{is} is the starting time, and T_{ie} is the ending time. The sequence S is then represented by

$$S =< e_1, e_2, \cdots, e_n >, \qquad (14)$$

where n is the total number of the events in the sequence. Figure 7 shows a representation of event sequence in the case of $S_{if} \geq 2$.

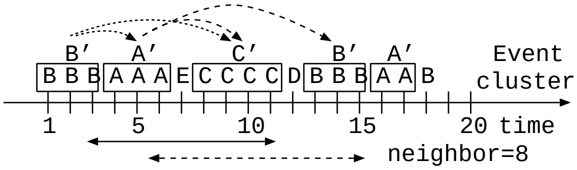


Figure 7: Example of description of cluster sequence, event sequence, and extraction of time-dependent association rules.

We extract the event pairs that occur closely in the sequence by introducing the neighborhood distance. The pattern change E is then represented by

$$E = \langle [e_i, e_j], neighbor \rangle (i = 1, \cdots, n-1, j = 1, \cdots, n), \qquad (15)$$

where $[e_i, e_j]$ represents a combination of the two events of e_i and e_j which satisfies $i < j$, and *neighbor* is the neighborhood distance.

Although *neighbor* is an idea similar with a time window in episode rules[10][11], we use this concept as the time interval necessary to extract only serial episodes such as $A \Rightarrow B$. We exclude parallel episode rules and combination of serial/parallel episode rules which are included in [10][11].

Furthermore we use the method of co-occurring term-pair [13] to evaluate the set of combinations of events which occurs closely and frequently in the local time window, and to extract them. The cohesion of the event e_i and e_j in a local time window is represented by

$$cohesion(e_i, e_j) = \frac{E_f(e_i, e_j)}{\sqrt{[f(e_i) \times f(e_j)]}}, \qquad (16)$$

where $f(e_i)$ and $f(e_j)$ are the frequencies of e_i and e_j, respectively, and $E_f(e_i, e_j)$ is the frequency of the co-occurrence of both e_i and e_j. The time-dependent association rules are extracted when the event pair has larger cohesion than the threshold.

The procedure of extraction of time-dependent association rule is shown schematically in Figure 8, and is described in the following:

step 1 The frequency of each event is determined (Fig. 8(1)).

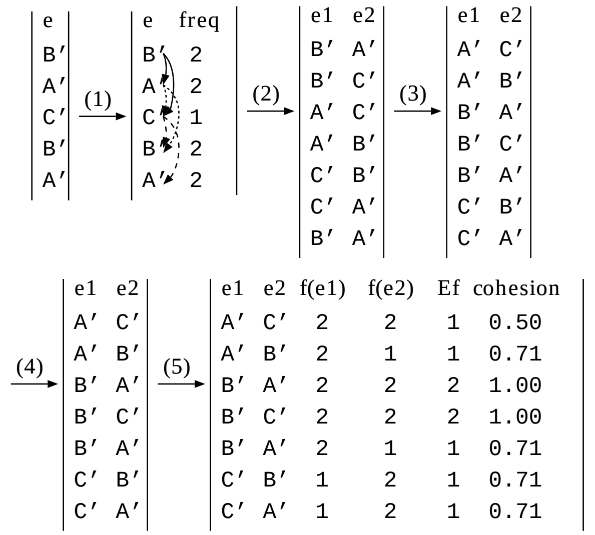
step 2 A combinational set of event pairs are determined as the candidates of rules, assuming the neighborhood distance (Figure 8(2)).

step 3 Event pairs are sorted lexicographically in regard to the first event (Fig. 8(3)).

step 4 Event pairs are sorted lexicographically in regard to the following event (Fig. 8(4)).

step 5 The candidates' frequency of co-occurrence and cohesion are calculated(Fig. 8(5)).

step 6 The event pairs that have larger cohesions than the threshold are extracted.



e
B′
A′
C′
B′
A′

e	freq
B′	2
A′	2
C′	1
B′	2
A′	2

e1	e2
B′	A′
B′	C′
A′	C′
A′	B′
C′	B′
C′	A′
B′	A′

e1	e2
A′	C′
A′	B′
B′	A′
B′	C′
B′	A′
C′	B′
C′	A′

e1	e2
A′	C′
A′	B′
B′	A′
B′	C′
B′	A′
C′	B′
C′	A′

e1	e2	f(e1)	f(e2)	Ef	cohesion
A′	C′	2	2	1	0.50
A′	B′	2	1	1	0.71
B′	A′	2	2	2	1.00
B′	C′	2	2	2	1.00
B′	A′	2	1	1	0.71
C′	B′	1	2	1	0.71
C′	A′	1	2	1	0.71

Figure 8: Procedure of a extraction of time-dependent association rule, where $e1$ and $e2$ are the first event and the following event, respectively, $f(e1)$ and $f(e2)$ are the frequencies of $e1$ and $e2$, respectively, E_f is the frequency of co-occurrence, and *cohesion* is the strength of cohesion between $e1$ and $e2$. The neighborhood distance is taken to be 8 in this case.

Strongly correlated event pairs in *neighbor* have large *cohesion* even if each event occurs less frequently. Inversely, weakly correlated event pairs have small *cohesion* even if each event occurs very frequently.

4.4 Result of experiments regarding time-dependent association rules

We performed the experiment by applying the above-described time-dependent association rule to the result of the clustering described in 3.4. Here we take the threshold of cohesion as 0.4, and *neighbor* ranging from 10 to 50. Since we sampled data every 8 hours, the virtual length of *neighbor* is between 3.3 days and 16.7 days.

Table 3 shows the relationship between *neighbor* and the number of extracted rules. Although the assessment of the context of the extracted rules is ongoing, the result suggests the similar numbers of rules are extracted from the different year's data set, which indicates that our present method is useful and robust.

Table 3: Relationship between *neighbor* and number of rules.

neighbor	10	20	30	40	50
number of rules(1997)	17	63	116	165	207
number of rules(1998)	7	50	98	166	218

5. CONSTRUCTION OF WEATHER IMAGE DATABASE

We constructed a weather image database which retrieves the above-described characteristics of weather images, visualizes time-dependent variation pattern, and supports the analysis and scientific discovery by domain-experts.

First we indexed the sequential time data by using events and time-dependent association rules, and constructed a weather satellite image database which contains index information regarding patterns such as time variations in weather.

5.1 Definition of attributes

We stored weather patterns extracted in the experiments in the following three tables: "series", "date_id", and "e_series" that represent contexts of time-dependent rules, the relationship between the observation date and image ID, and the contents of time-dependent rule candidates (those obtained in step 5 in 4.3), respectively.

Table 4: List of three table "series", "data_id", "e_series"

(a) "series" that represents time-dependent rule in which l_term is the cluster number of left term, r_term is the right term, location is the reference to the R-tree data(rectangular), and first and last are the image ID of the l_term starting point and r_term ending point, respectively .)

l_term	r_term	cohesion	location	first	last
int	int	float	box	int	int

(b) "date_id" that indicates the relationship between the observation date *date* and image ID *id*.

id	date
int	int

(c) "e_series" that indicates the candidate of time-dependent event rules, where *term* is the cluster number, *first* and *last* are the image ID of the starting point and ending point of *term*, respectively.

term	first	last
int	int	int

5.1.1 *Indexing by using R-tree*

We indexed the image IDs at the starting and ending point of the obtained pattern using R-tree. As shown in Figure 9, spring encloses March to May, and summer encloses June to August. Taking note of this relation, we index the enclosure relation between seasons and months, and index the starting and the ending times of variation patterns. This allows each variation pattern to contain an index which includes the enclosure relations by month or season as keys.

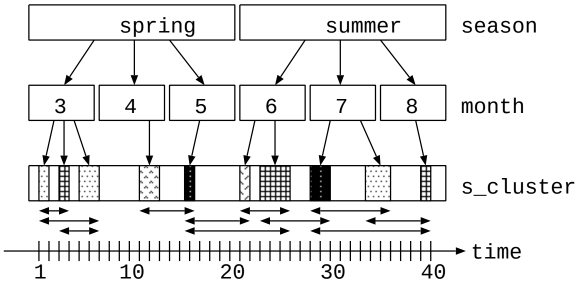


Figure 9: Indexing by using R-tree, remarking at the continuing sequence.

5.2 Query by SQL

Rule storage in the database enables the retrieval of the various queries by using SQL statements. We show examples of the queries and corresponding SQL statement[1] in the following:

"Search for typhoons that occurred within 20 days after July 16th, 1997."

```
select first, last, t1.date, t2.date
from series, date_id t1, date_id t2
where t1.date = 97071617 and t1.id = first
and t2.id = last and(r_term = 0 or r_term =
9 or r_term = 12 or r_term = 13) and location
@ '((570, 0),
(630,15))'::box
```

"Search for weather changes between one typhoon and the another."

```
select first, last, t1.date, t2.date
from series, date_id t1, date_id t2
where(l_term = 0 or l_term = 9 or l_term =
12 or l_term = 13) and(r_term = 0 or r_term
= 9 or r_term = 12 or r_term = 13) and t1.id
= first and t2.id = last
```

or

```
select t1.first, t2.last, t1.date, t2.date
from e_series t1, e_series t2, date_id t1,
date_id t2
where(t1.term = 0 or t1.term = 9 or t1.term
= 12 or t1.term = 13) and(t2.term = 0 or t2.term
= 9 or t2.term = 12 or t2.term = 13) and t1.id
= t1.first and t2.id = t2.last and t1.first
< t2.first order by t1.first
```

"Search for weather patterns in which typhoon occurs within 10 days after the development of front and typhoon during 5 days."

```
select t1.first, t2.last, t1.date, t2.date
from series t1, e_series t2, date_id t1,
date_id t2
where (t1.l_term = 0 or t1.l_term = 5 or
t1.l_term = 8 or t1.l_term = 9 or t1.l_term
= 14)
and (t1.r_term = 1 or t1.r_term = 2 or t1.r_term
= 3
or t1.r_term = 4 or t1.r_term = 7 or t1.r_term
= 8
or t1.r_term = 12 or t1.r_term = 13 or t1.r_term
= 15)
and t1.first >=(t1.last - 15) and t1.id = first
and
t2.id = last and(t2.term = 0 or t2.term = 9
or
t2.term = 12 or t2.term = 13) and
t1.first >=(t2.last - 30) and t1.last <= t2.last
```

5.3 Result of implementation

Figure 10 shows the browse page[2] of the system which retrieves weather images using R-tree index. Entering the SQL in the upper frame performs retrievals. This example shows the results of query: "Is there any weather pattern in which a typhoon occurred in 10 days after the development of front and typhoon during 5 days". Seven periods are retrieved and listed in the lower left frame as the result, and the weather variation in these periods is shown as an animation in the lower right frame.

The problem of this method is that the accuracy of clustering and the semantical description of clusters changes the retrieval results significantly. Interactive processing interface, such as adjustment of the sample data or assumed parameters with metrological experts who are potential users, are required to solve this problem.

6. CONCLUSION

We applied clustering and time-dependent association rules to a large-scale content-based image database of weather satellite images. Each image is divided into $N \times M$ blocks and automatically classified by two-stage SOM. We also extracted unknown rules from time- sequential data expressed by a sequence of cluster addresses by using time-dependent association rules. Furthermore, we developed a knowledge discovery support system for domain experts, which retrieves image sequences using extracted events and association rules.

[1] Here cluster addressees are represented by numbers ranging from 0 to 15 instead of characters A-P in Table 2.

[2] http://zeus.is.kochi-u.ac.jp/˜takimoto/java/servlets/ index6e.
html

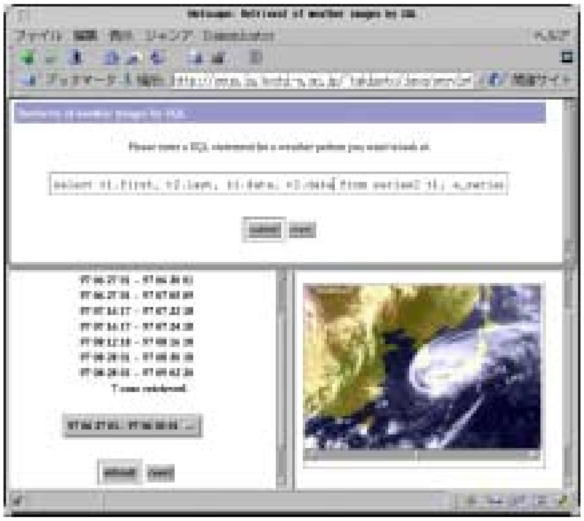


Figure 10: Example of the result of retrieval result from sequential image data.

From the perspective that high-level queries make the analysis easier, we stored the extracted rules in the database to admit sophisticated queries described by SQL. The retrieval responses to various queries shows the usefulness of this approach.

The framework presented in this study, clustering ⇒ transformation into time-sequential data ⇒ extraction of time-dependent association rules, is considered to be useful in managing enormous multimedia datasets which include sequential patterns such as video information and audio information.

7. ACKNOWLEDGEMENTS

The authors are grateful to Prof. T. Kikuchi for offering us well-prepared GMS-5 images. This research is partly supported by grants-in-aid for intensive research(A)(1)(Project 10143102) from the Ministry of Education, Science, and Culture of Japan.

8. REFERENCES

[1] J. Fomg,(Edt.). Data mining, data warehousing & client/server databases. In *Proceedings 8th International Database Workshop*, Springer, 1997.

[2] A. F. Alex and H. L. Simon. *Mining very large databases with parallel processing.* Kluwer Academic Publishers, 1998.

[3] R. Agrawal, T. Imelinski and A. Swani. Mining in association rules between sets of items in large database. In *Proc. ACM SIGMOD International Conference*, pages 207–216, 1993.

[4] R. Agrawal and R. Srikant. Fast Algorithms for mining association rules. In *Proceedings of 20th International Conference on VLDB*, pages 487–499, 1994.

[5] O. R. Zaiane, J. Han, Z. N. Li, J. Y. Chiang and S. Chee. Multimedia-miner : a system prototype for multimedia data mining. At *Proceedings ACM-SIGMOD Conference on Management of Data*, system demo, 1998.

[6] U. M. Fayyad, S. G. Djorgovski and N. Weir. Automatic the analysis and cataloging of sky surveys. *Advances in Knowledge Discovery and Data Mining*, pages 471–493 , AAAI Press/MIT Press, 1996.

[7] M. C. Burl, L. Asker, P. Smyth, U. M. Fayyad, P. Perona, L. Crumpler and J. Aubele. Learning to recognize volcanoes on venus. *machine learning*, 30(2/3):165–195, February, 1998.

[8] P. Smyth, M. C. Burl and U. M. Fayyad. Modeling subjective uncertainty in image annotation. In *Advances in Knowledge Discovery and Data Mining*, pages 517–539. AAAI Press/MIT Press, 1996.

[9] T. Kohonen. *Self-organizing maps.* Springer, 1995.

[10] H. Mannila, H. Tovinen and A. I. Verkano. Discovering frequent episodes in sequences. In *First International Conference on Knowledge Discovery and Data Mining(KDD'95)*, pages 210–215 , AAAI Press, 1995.

[11] H. Mannila, H. Tovinen. Discovering generalized episodes using minimal occurrences. In *Proceeding of the Second International Conference on Knowledge Discovery and Data Mining(KDD'96)*, pages 146–151, AAAI Press, 1996.

[12] T. N. Raymond, J. Han. Efficient and effective clustering methods for spatial data mining. In *Proceeding of 20th VLDB Conference*, Santiago, Chile, 1994.

[13] O. Konishi. A statistically build knowledge based terminology construction system (in Japanese). *Transaction of Information Processing Society of Japan*!$30(2):179–189, 1989.

[14] K. Katayama and O. Konishi. Construction satellite image databases for supporting knowledge discovery(in Japanese). *Transaction of Information Processing Society of Japan*, 40(SIG5(TOD2)):69–78, 1999.

[15] K. Katayama. and O. Konishi. Discovering co-occurencing patterns in event sequences (in Japanese). *DEWS'99*, 1999.

[16] D. Merkl and A. Rauber. Uncovering the hierarchical structure of text archives by using unsupervised neural network with adaptive architecture. In *PAKDD 2000*, pages 384–395, 2000.

Data Mining from Functional Brain Images

Mitsuru Kakimoto
Corporate Research &
Development Center,
Toshiba Corporation
1, Komukai Toshiba-cho,
Saiwai-ku, Kawasaki
212-8582 Japan
mitsuru.kakimoto@
toshiba.co.jp

Chie Morita
Corporate Research &
Development Center,
Toshiba Corporation
1, Komukai Toshiba-cho,
Saiwai-ku, Kawasaki
212-8582 Japan
chie.morita@
toshiba.co.jp

Hiroshi Tsukimoto
Corporate Research &
Development Center,
Toshiba Corporation
1, Komukai Toshiba-cho,
Saiwai-ku, Kawasaki
212-8582 Japan
hiroshi.tsukimoto@
toshiba.co.jp

ABSTRACT

Recent advances in functional brain imaging enable identification of active areas of a brain performing a certain function. Induction of logical formulas describing relations between brain areas and brain functions from functional brain images is a category of data mining. It is difficult, however, to apply conventional mining techniques to functional brain images due to several reasons, such as the difficulty of reducing images to symbolic data, possible existence of correlations between adjacent pixels in a image and the limited number of samples available from a single subject. Tsukimoto and Morita presented an algorithm for data mining from functional brain images and showed that the algorithm works well for artificial data. The algorithm consists of two steps. The first step is nonparametric regression. The second step is rule extraction from the linear formula obtained by the nonparametric regression. The authors have applied the algorithm to real f-MRI images. This paper reports that the algorithm works well for real f-MRI data and has led to the discovery of certain rules for a finger tapping action and a speech-related action.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Concept learning; J.3 [**Life And Medical Sciences**]: Medical information systems

Keywords

Knowledge discovery, Functional brain images, Nonparametric regression, Rule extraction, Human brain mapping

1. INTRODUCTION

Conventional data mining techniques deal with symbolic and/or numerical data contained in tables in which the independence of rows is tacitly assumed. This simple structure makes the data easy to mine. As demand for knowledge discovery from real world data grows, however, methods for deriving knowledge from structured data (including time series, images and data embedded in graphical structures) are strongly desired.

Knowledge discovery from functional brain images is a candidate field for the application of such methods. As a result of the ongoing development of non-invasive measurement of brain function, detailed functional brain images can be obtained, from which the relations between brain areas and brain functions can be understood. These relations, however, could be complicated since several brain areas might be responsible for a brain function. Some of them are connected in series, and others are connected in parallel. Brain areas connected in series are described by "AND" and brain areas connected in parallel are described by "OR". Therefore, the relations between brain areas and brain functions are described by rules.

It is of crucial importance for researchers involved in mapping brain functions to find such rules from functional brain images. Although several statistical methods, such as the Statistical Parametric Map[2] and Independent Component Analysis[3], are being used in human brain mapping, these methods can only present some principal areas for the function and cannot discover rules. Furthermore, several factors prevent conventional data mining techniques from being applied to functional brain imaging. First, observed images consist of real values and it is not easy to reduce them to simple symbolic data such as "active" / "inactive". Second, it is expected that strong correlations exist between adjacent pixels in an image. Therefore, a mining scheme should take the structure of the image into account so as to improve its quality. Third, in a usual function brain imaging experiment, the number of samples obtained from a single subject is limited. This scarcity of samples makes it hard to obtain accurate rules.

Tsukimoto and Morita have presented a new algorithm capable of extracting rules from such structured data, which is now called the Logical Regression Analysis(LRA). They confirmed that the LRA works well for artificial functional brain image data[11]. The LRA consists of two steps. In the first step, a linear formula describing the relation between an image and a brain function is derived using regression

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

analysis. In order to obtain a linear formula from relatively few samples, nonparametric regression is used. The subsequent step extracts rules from the linear formula obtained in the previous step.

This paper reports the application of the LRA to real f-MRI data obtained in the experiments of finger tapping and speech actions. Section 2 briefly outlines a scheme of the data mining from functional brain images. Section 3 gives exact formulation of nonparametric regression used in the analysis. Section 4 explains the rule extraction algorithm from a linear formula. Section 5 shows experimental results obtained by applying the LRA to real f-MRI images and discusses its meaning in brain science. Conclusions are presented in Section 6.

2. DATA MINING FROM FUNCTIONAL BRAIN IMAGES

Fig.1 is a schematic illustration of a 2-dimensional functional brain image with a circle representing a contour of a brain. In Fig.1, the image is divided into 6×6(=36) pixels. In an experiment using f-MRI, subjects are told to do some task and rest for a while repeatedly. If a pixel includes or intersects brain areas responsible for that task, activation of the area by the task results in enhanced value of the pixel. Detecting difference of the value of a pixel between the image taken while the subject is doing the task and one while he/she is resting thus makes it possible to identify areas responsible for the task.

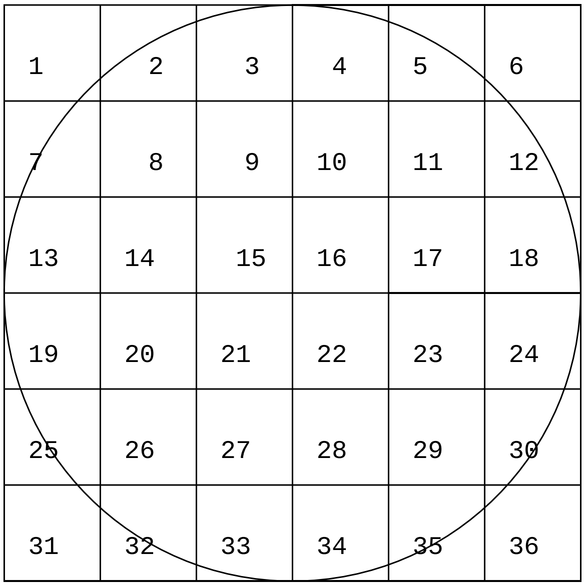


Figure 1: Brain image

Although values of pixels in a real f-MRI image are continuous, for simplification, we assume here that the values are binary, that is, on(active) or off(inactive). Also, we assume that there are seven samples. Table 1 shows the data. In Table 1, 'on' and 'off' mean that the pixel is active and inactive, respectively. Y in class shows that a subject is doing a certain task and N shows that he/she is resting.

Table 1: Data

sample	1	2	·	36	class
S1	on	off	·	off	Y
S2	on	on	·	off	N
S3	off	off	·	on	N
S4	off	on	·	on	Y
S5	on	off	·	off	N
S6	off	on	·	on	N
S7	off	off	·	on	Y

If an activity of a pixel is strongly correlated with the class value, the pixel is considered to be a part of an area responsible for the function. On the contrary, if a pixel is always inactive, it is considered to be a part of inhibitory area. Otherwise, i.e. a pixel's activity does not have any correlation with the class value, it is regarded as irrelevant to the function. Combining pixel values of responsible areas and negation of pixel values in inhibitory areas produces a logical formula that describes a rule governing the brain function. It is thus clear that rule extraction from functional brain images is formulated as a typical supervised inductive learning.

What makes the rule extraction difficult, however, is that variation of a pixel value correlated to brain function is so subtle that there is no clear-cut way to reduce observed numerical pixel values to simple 'on'/'off' symbols. Combining regression analysis and rule extraction, the LRA evaluates quantitative significance of each pixel respecting the brain function, which makes fast and rigorous rule extraction possible.

3. NONPARAMETRIC REGRESSION

As explained earlier, the LRA uses regression in its first step. In functional brain image analysis, each image has more than a thousand pixels, which mean that there are more than a thousand independent variables. The number of samples obtained in a single experiment is around one hundred, which is significantly small compared with a number of independent variables. It is therefore impossible to use conventional linear regression which requires a larger number of samples than the number of independent variables.

Another problem inherent in image analysis is that strong correlation is expected between adjacent pixels. This also applies in the analysis of functional brain images where contribution of each pixel to brain function cannot be regarded as truly independent.

As shown below, these two problems are resolved simultaneously in a framework of nonparametric regression. The next subsection explains the conventional 1-dimensional nonparametric regression[1]. Extension to 2-dimensions, which is used in the analysis of functional brain images, is described in the later subsection.

3.1 1-dimensional nonparametric regression

Nonparametric regression is defined as follows: Let y stand for a dependent variable and $t_j (j = 1, .., m)$ stand for independent variables. Then, the regression formula is as follows:

$$\hat{y} = \sum a_j t_j + e(j = 1, .., m),$$

where a_j are real numbers and e is a zero-mean random variable. When there are n measured values of y,

$$\hat{y}_i = \sum a_j t_{ij} + e_i(i = 1, .., n).$$

In usual linear regression, the coefficients a_j are defined so that residual error(i.e., difference between measured value of y and calculated value by the formula) is minimized, whereas in nonparametric regression, continuity of coefficients is also taken into account[Miwa, private communication]. Therefore, the evaluation value is now described as follows:

$$1/n \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{m} (a_{j+1} - a_j)^2$$

where $\hat{y}$ is an estimated value. The second term in the above formula is the difference of first order between the adjacent coefficients, that is, the continuity of the coefficients. λ is the coefficient of continuity. Consideration of two extreme cases facilitates understanding of the characteristics of the above evaluation value. When $\lambda \rightarrow 0$, the evaluation value consists of only the first term, that is, error, which means the usual regression. In this case, the effective number of coefficients is exactly the same as the actual number of coefficients. On the other hand, if λ is infinitely large, the evaluation value consists of only the second term, that is, continuity, which means that the error is ignored and a_j is a constant. This is equivalent to the case where there is a single coefficient. The effective number of coefficients is thus controlled by the value of λ. By determining the value of λ adaptively, a nonparametric regression scheme can handle the situation in which the number of samples available is smaller than the number of coefficients.

We determined the value of λ using the *leave-one-out method* cross validation[6], whose formulation can be described as follows. Let $\mathbf{X}$ stand for $n \times m$ matrix. Let t_{ij} be an element of $\mathbf{X}$. Let $\mathbf{y}$ stand for a vector consisting of y_i. $m \times m$ matrix $\mathbf{C}$ is as follows:

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & & & \cdots \end{pmatrix}$$

Off diagonal elements of $\mathbf{C}$ not written explicitly are exactly 0. Cross validation function CV is as follows:

$$CV = n\tilde{\mathbf{y}}^{\mathbf{t}}\tilde{\mathbf{y}}$$

$$\tilde{\mathbf{y}} = \mathbf{Diag}(\mathbf{I} - \mathbf{A})^{-1}(\mathbf{I} - \mathbf{A})\mathbf{y}$$

$$\mathbf{A} = \mathbf{X}(\mathbf{X}^{\mathbf{t}}\mathbf{X} + (\mathbf{n} - \mathbf{1})\lambda\mathbf{C})^{-1}\mathbf{X}^{\mathbf{t}},$$

where $\mathbf{DiagA}$ is a diagonal matrix whose diagonal components are $\mathbf{A}$'s diagonal components. Then the coefficients $\hat{\mathbf{a}} = (a_1, \cdots, a_m)^t$ are obtained by

$$\hat{\mathbf{a}} = (\mathbf{X}^{\mathbf{t}}\mathbf{X} + \mathbf{n}\lambda_{\mathbf{o}}\mathbf{C})^{-1}\mathbf{X}^{\mathbf{t}}\mathbf{y},$$

where λ_o is the λ that minimizes the cross validation function CV.

3.2 2-dimensional nonparametric regression

The nonparametric regression scheme explained in the previous subsection is extended to the 2-dimensional case and applied to f-MRI data. In 2-dimensional data, there are four adjacent measured values [1], whereas in 1-dimensional data, there are only two. Hence, the evaluation value for the continuity of coefficients a_j is modified so that continuity with adjacent four pixels is taken into account. For example, pixel 8 in Fig.1 has four adjacent pixels (2, 7, 9 and 14), and the evaluation value is as follows:

$$(a_8 - a_2)^2 + (a_8 - a_7)^2 + (a_9 - a_8)^2 + (a_{14} - a_8)^2.$$

4. RULE EXTRACTION

4.1 Rule extraction in the discrete domain

In this subsection, a method for rule extraction in the discrete domain is explained. The main idea is to find a Boolean function which is nearest to a given linear formula in the Boolean function space.

Let (f_i) be the values of a linear formula. Let $(g_i)(g_i = 0$ or $1)$ be the values of Boolean functions. The basic method is as follows:

$$g_i = \begin{cases} 1(f_i \geq 0.5), \\ 0(f_i < 0.5). \end{cases}$$

This method minimizes Euclidean distance.

Generally, let $g(x_1, ..., x_n)$ stand for a Boolean function, and let $g_i(i = 1, ..., 2^n)$ stand for values of a Boolean function and then the Boolean function is represented by the following formula:

$$g(x_1, ..., x_n) = \sum_{i=1}^{2^n} g_i a_i,$$

where $\sum$ is disjunction, and a_i is the atom corresponding to g_i, that is,

$$a_i = \prod_{j=1}^{n} e(x_j) \ (i = 1, ..., 2^n),$$

where

$$e(x_j) = \begin{cases} \overline{x_j}(e_j = 0), \\ x_j(e_j = 1), \end{cases}$$

where $\prod$ stands for conjunction, $\overline{x}$ stands for the negation of x, and e_j is the substitution for x_j, that is, $e_j = 0$ or 1. The above formula can be easily verified.

Fig. 2 shows a case of two variables, x and y. Crosses stand for the values of a linear formula and circles stand for the values of a Boolean function. Values on the horizontal axis, $00, 01, 10$ and 11, stand for the domains. For example, 00 stands for $x = 0, y = 0$.

In this case, the values of the Boolean function $g(x, y)$ are as follows:

$$g(0, 0) = 1, g(0, 1) = 1, g(1, 0) = 0, g(1, 1) = 0.$$

[1] One might think of a neighbor consisting of eight surrounding pixels. We did not pursue this possibility.

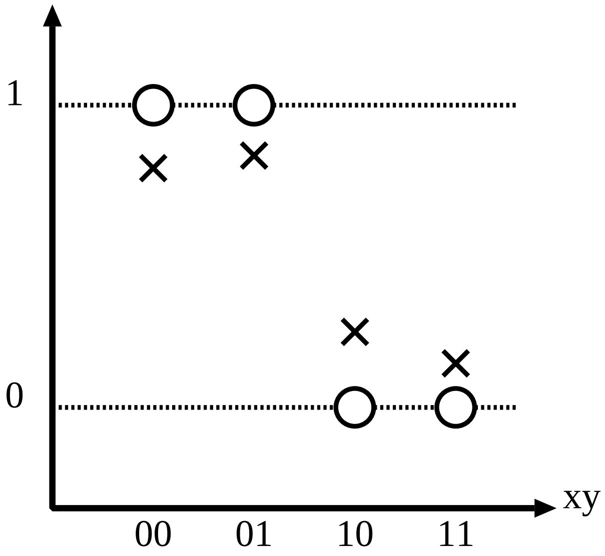


Figure 2: Approximation

Therefore, in the case of Fig. 2, the Boolean function is as follows:

$$
\begin{aligned}
& g(x,y) \\
= \; & g(0,0)\bar{x}\bar{y} + g(0,1)\bar{x}y + g(1,0)x\bar{y} + g(1,1)xy \\
= \; & 1\bar{x}\bar{y} + 1\bar{x}y + 0x\bar{y} + 0xy \\
= \; & \bar{x}\bar{y} + \bar{x}y \\
= \; & \bar{x}.
\end{aligned}
$$

4.2 The fast polynomial time algorithm

A naive implementation of the rule extraction above requires computational time which grows exponentially with the number of independent variables. Tsukimoto presented the polynomial time algorithm[9], [10], the outline of which is now described.

Let a linear formula be as follows:

$$f = p_1x_1 + \ldots + p_nx_n + p_{n+1},$$

The Boolean function which approximates f is obtained by the following steps.

1. Check if

   $$x_{i_1}\cdot\cdot x_{i_k}\overline{x}_{i_{k+1}}\cdot\cdot\overline{x}_{i_l}$$

   exists in the Boolean function after the approximation by the following formula:

   $$p_{n+1} + \sum_{i_1}^{i_k} p_j + \sum_{1\le j\le n, j\ne i_1,..,i_l, p_j\le 0} p_j \ge 0.5$$

   .

2. Connect the terms existing after the approximation by logical disjunction to make a DNF formula.

3. Execute the above procedures up to a certain (usually two or three) order.

4.3 Extension to the continuous domain

Continuous domains can be normalized to [0,1] domains by some normalization method. So we assume here that the values lie in [0,1] domains without loss of generality. First, we have to present a system of qualitative expressions corresponding to Boolean functions, in the [0,1] domain. The expression system is generated by direct proportion, reverse proportion, conjunction and disjunction. The direct proportion is $y = x$. The inverse proportion is $y = 1 - x$, which is a little different from the conventional one ($y = -x$), because $y = 1 - x$ is the natural extension of the negation in Boolean functions. The conjunction and disjunction will be also obtained by a natural extension. The functions generated by direct proportion, reverse proportion, conjunction and disjunction are called continuous Boolean functions, because they satisfy the axioms of Boolean algebra. For details, refer to [8]. In the domain [0,1], linear formulas are approximated by continuous Boolean functions. The method for deriving such an expression is exactly the same as the one in the discrete domain[7], [12].

5. EXPERIMENTS

Two f-MRI experiments are conducted, each with a single subject. Data obtained consist of 32 series of cross-sectional brain images, i.e. slices(Fig. 3). Nonparametric regression is performed on each series of images and average residual error of the regression is calculated. Small error value of a slice is deemed to be an evidence of the existence of signals correlated to the task on the slice. Rule extraction is performed on these slices to identify brain areas related to the task. Detailed results of each experiment are given in the following subsections.

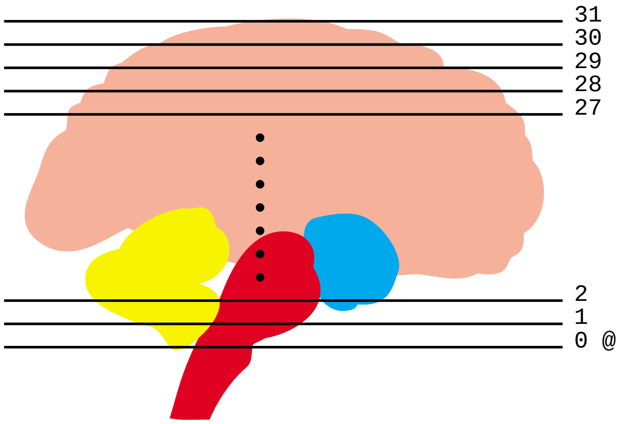


Figure 3: Slices

5.1 Finger tapping

In a finger tapping task experiment, the subject is asked to tap four fingers with thumb using his right hand.

The experimental conditions are summarized as follows:

magnetic field intensity:	1.0 Tesla
number of pixels:	64×64
number of slices:	32
subject:	male(36 years old)
number of task samples:	30
number of rest samples:	33

Table 2 shows the error of nonparametric regression for each slice. Note that slices are numbered from bottom to top, i.e., slice 0 is lower part of the brain and slice 31 is located at the top of the brain.

Table 2: Error (Finger tapping)

slice	err.	slice	err.	slice	err.	slice	err.
0	0.73	8	1.00	16	0.11	24	0.96
1	0.02	9	0.83	17	0.96	25	0.10
2	0.38	10	0.45	18	0.40	26	0.54
3	0.58	11	1.00	19	0.71	27	0.53
4	0.09	12	0.10	20	0.93	28	0.71
5	0.01	13	0.90	21	0.09	29	0.58
6	0.62	14	0.75	22	0.39	30	0.71
7	0.37	15	0.09	23	0.47	31	0.89

Slices with small errors (0.1 or less) are listed as follows:

slice	5	1	4	21	15	25	12
error	0.01	0.02	0.09	0.09	0.09	0.10	0.10

The errors of No.8 and No.11 slices are nearly 1.0, which means that these slices have no relations to finger tapping.

Figure 4 - Figure 10 are graphical representations of the rules for slices of No. 1, 4, 5, 12, 15, 21 and 25, respectively. In the figures, white areas show activation areas related to finger tapping and dark gray areas show inhibition areas. Note that these figures illustrate cross sections seen from below, i.e., left side of figures correspond to the right side of subject's head and vice versa. Note also that upper side and lower side of figures correspond to the front and back of subject's head, respectively.

Interpretations of these rules in terms of brain physiology are given as follows:
Movements of the non-dominant hand usually induce neural activity in motor and sensory areas in both hemispheres, and cause higher activity in the dominant hemisphere (contralateral to the non-dominant hand). In this case, the subject was left-handed and he moved his right hand (non-dominant). Higher activity was observed in the right (dominant) motor and sensory areas than in those on the left as shown in Figs. 7 - 9.

- Fig.6, 4, 5(slice No. 5, 1, 4)
  Higher activity was observed in the right cerebellum. The result agrees with the fact that the neural activity in the cerebellum ipsilateral to the moving hand is higher than in the cerebellum contralateral to the moving hand.

- Fig.7(slice No.12)
  Activity in the right (dominant) motor-sensory area.

- Fig. 8(slice No.15)
  Neural activity was clearly observed in motor, sensory, and supplementary motor areas in the right (dominant) hemisphere, and diffusive activity in the left motor-sensory area.

- Fig.9(slice No.21)
  Activity in the right (dominant) premotor area related to motor programming and pattern generation.

- Fig.10(slice No.25)
  Activity in the right (dominant) premotor and supplementary areas.

5.2 Shiritori

The next experimental task is *shiritori*, which is a well-known Japanese word game for two or more players. Each player utters a word that starts with the same syllable as the last syllable of a word uttered by a previous player. An example is shown below.

to$\underline{\text{ki}}$ → $\underline{\text{ki}}$mo$\underline{\text{no}}$ →
(time) (wear)
$\underline{\text{no}}$mimo$\underline{\text{no}}$ → $\underline{\text{no}}$nki → ...
(drink) (optimism)

This process is repeated until someone fails to come up with a word. In our *shiritori* experiment, the subject is presented a single Japanese character at the beginning of each task period, then he begins playing *shiritori* by himself starting with the character. While doing the task, he does not actually utter words but speaks silently.

The experimental conditions are as follows:

magnetic field intensity:	1.0 Tesla
number of pixels:	64×64
number of slices:	32
subject:	male(45 years old)
number of task samples:	40
number of rest samples:	40

Table 3 shows the errors of nonparametric regression. The result of nonparametric regression is worse than that of finger tapping. In *shiritori*, no slice has an error of 0.1 or less and the least error is 0.11. That means that *shirotori* is complicated and therefore is related to several areas such as speech area, vision area, auditory area, motor area, and so on.

Slices with small errors (0.2 or less) are as follows:

slice	7	13	6	16	3
error	0.11	0.15	0.15	0.17	0.19

There are strong correlations between *shiritori* and slices No.7, 13, 6, 16 and 3, since errors in the slices are small.

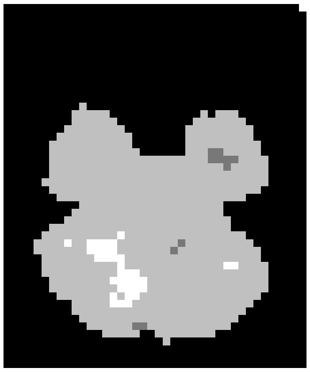


Figure 4: Rule (Finger 1)

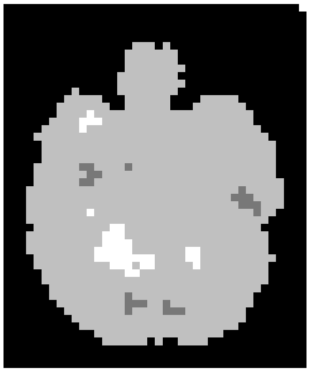


Figure 5: Rule (Finger 4)

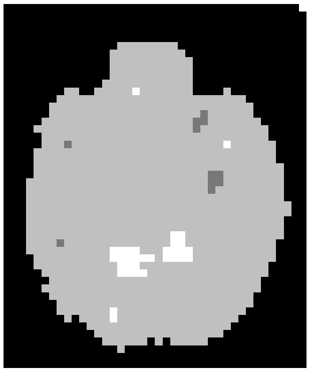


Figure 6: Rule (Finger 5)

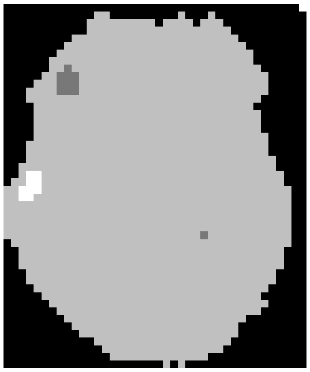


Figure 7: Rule (Finger 12)

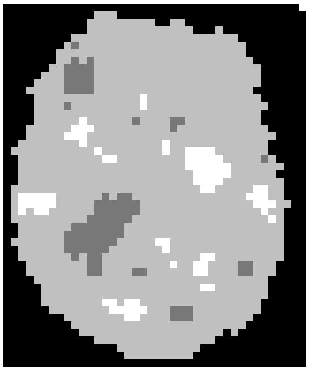


Figure 8: Rule (Finger 15)

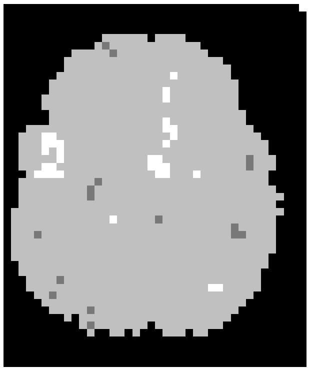


Figure 9: Rule (Finger 21)

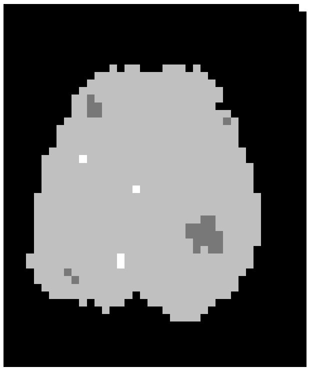


Figure 10: Rule (Finger 25)

Table 3: Error (Shiritori)

slice	err.	slice	err.	slice	err.	slice	err.
0	0.92	8	0.88	16	0.17	24	0.21
1	0.89	9	0.85	17	0.62	25	0.87
2	0.87	10	0.67	18	0.44	26	0.86
3	0.19	11	0.61	19	0.49	27	0.31
4	0.80	12	0.84	20	0.71	28	0.58
5	0.66	13	0.15	21	0.70	29	0.68
6	0.15	14	0.76	22	0.84	30	0.91
7	0.11	15	0.71	23	0.85	31	0.38

Figure 11 - Figure 15 show rules for slices No. 3, 6, 7, 13 and 16, respectively.

Some of the rules presented are interpreted as follows:

- Fig. 13(slice No.7)
  Activation of the left prefrontal area.
  The left prefrontal area was activated related to working memory required for mental word generation .

- Fig. 11(slice No.3)
  Activation of the right cerebellum.
  This showed that the cerebellum was related to some cognitive functions in addition to motor functioning. The cerebellum predominantly connects with the contralateral frontal cortex, and the right cerebellum was activated associated with the left prefrontal area.

Physiological meanings of other rules observed during the task are to be studied in future.

6. CONCLUSIONS

We have applied the Logical Regression Analysis to real f-MRI images obtained by experiments of finger tapping and speech actions, i.e., *shiritori* tasks. It is confirmed that the nonparametric regression extended for functional brain image analysis, which consists of the first step of the LRA, can successfully identify slices of a brain relevant to the tasks. Rule extractions, the second step of the LRA, performed on these relevant slices induced rules which are reasonably interpreted in terms of brain physiology.

7. ACKNOWLEDGMENTS

We would like to thank Dr. T. Kobayakawa of the National Institute of Bioscience Human-Technology for instruction and planning of f-MRI experiments. This research is partly supported by Grant-in-Aid for Scientific Research on Priority Areas "Discovery Science" from the Ministry of Education, Science and Culture, Japan(grant 10143106).

8. ADDITIONAL AUTHORS

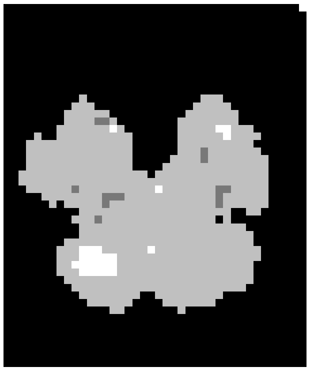


Figure 11: rule (shiritori 3)

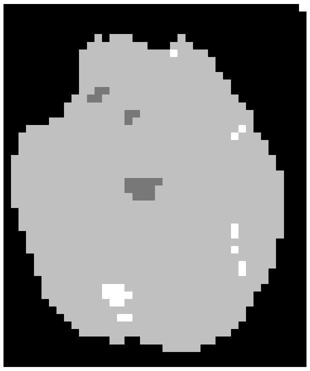


Figure 12: rule (shiritori 6)

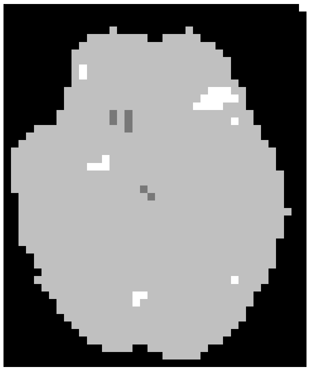


Figure 13: rule (shiritori 7)

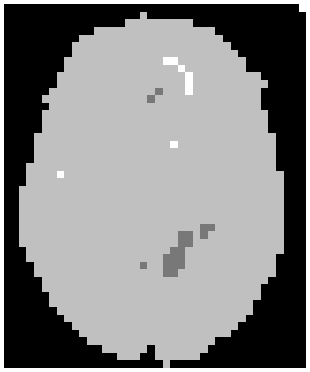


Figure 14: rule (shiritori 13)

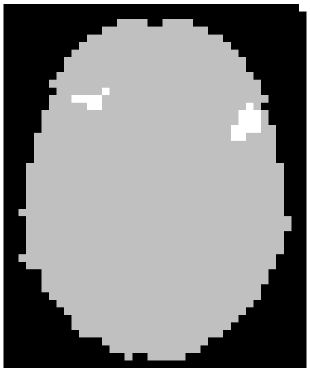


Figure 15: rule (shiritori 16)

Additional authors: Yoshiaki Kikuchi (Tokyo Metropolitan University of Health Sciences, Higashi-ogu 7-2-10, Arakawa-ku, Tokyo,116-8551 Japan
email: `ykikuchi@post.metro-hs.ac.jp`)

9. REFERENCES

[1] Eubank, R.L.: Spline Smoothing and Nonparametric Regression, Marcel Dekker, New York, 1988.

[2] Friston et al.: SPM course notes, 1997. `http://www.fil.ion.bpmf.ac.uk/spm`

[3] McKeown, M., Makeig, S., Brown, G., Jung, T.-P., Kindermann, S., Lee, T.-W., and Sejnowski, T.J. : Spatially independent activity patterns in functional magnetic resonance imaging data during the stroop color-naming task. *Proceedings of the National Academy of Sciences, 95*, pp.803-810, February 1998. `http://www.cnl.salk.edu/~tewon/ica_cnl.html`

[4] Morita,C. and Tsukimoto,H.: Knowledge discovery from numerical data, *Knowledge-based Systems*, Vol.10, No.7, pp.413-419, 1998.

[5] Posner, M.I., Raichle,M.E.: Images of Mind, W H Freeman & Co, 1997.

[6] Stone, M.: Cross-validatory choice and assessment of statistical prediction (with discussion), *Journal of the Royal Statistical Society*, Series B, 36, pp.111-147, 1974.

[7] Tsukimoto, H.: The discovery of logical propositions in numerical data. *AAAI'94 Workshop on Knowledge Discovery in Databases*, pp.205-216, 1994.

[8] Tsukimoto, H.: On continuously valued logical functions satisfying all axioms of classical logic. *Systems and Computers in Japan*, Vol.25, No.12. pp.33-41, 1994.

[9] Tsukimoto,H. and Morita,C.: Efficient algorithms for inductive learning-An application of multi-linear functions to inductive learning, *Machine Intelligence 14*, pp.427-449, Oxford University Press, 1995.

[10] Tsukimoto,H., Morita,C., Shimogori,N.: An Inductive Learning Algorithm Based on Regression Analysis, *Systems and Computers in Japan*, Vol.28, No.3. pp.62-70, 1997.

[11] Tsukimoto,H. and Morita,C.: The Discovery of Rules from Brain Images, *Discovery Science, Proceedings of the First International Conference DS'98*, pp.198-209, 1998.

[12] Tsukimoto,H: Extracting Rules from Trained Neural Networks, *IEEE Transactions on Neural Networks*, Vol.11, No.2, pp.377-389, 2000.

Mining Cinematic Knowledge: Work in Progress

[An Extended Abstract]

Duminda Wijesekera[*]
Department of Information and Software Engineering
George Mason University, MS 4A4,
Fairfax, VA 22101, U.S.A.
duminda@ise.gmu.edu

Daniel Barbará[†]
Department of Information and Software Engineering
George Mason University, MS 4A4,
Fairfax, VA 22101, U.S.A.
dbarbara@ise.gmu.edu

ABSTRACT

This paper presents the blueprint of an on going effort underway at George Mason University to create a movie mining system that uses already existing content detection technology. The emphasis of this project is to examine the suitability of existing concepts in data mining to multimedia, where the semantic content is time sensitive and constructed by fusing data obtained from component streams. Methods used for this purpose have a non zero probability of being incorrect. We discussed the issues involved in mining knowledge from or about movies and propose some solutions.

Categories and Subject Descriptors

data mining [**multimedia**]: cinematic knowledge

Keywords

data mining,multimedia,cinematics

1. INTRODUCTION

Multimedia and data mining are two young and flourishing fields, with their own application domains: multimedia concentrating on video conferencing, VoD services, databases, content based retrieval and data mining concentrating on detecting *interesting* patterns from basket data, event scripts, web-based information etc.

In detecting *interesting patterns* contained in multimedia data, one of the basic problems that needs to be addressed is the issue of extracting semantic information from audio, images, and video - and this has proven to be a challenging problem with limited success in specific applications. Notice that this problem does not arise in conventional data types. Secondarily, even when it is possible to extract and track basic features such as faces and gestures from surveillance video, it is still difficult to translate those gestures and facial movements to semantic information. There are many problems involved in doing this First, with the current state of the art in image and audio, it is not easy to identify *semantic information* easily. Secondly, higher *semantic* content may be contained in more than one stream of data, and consequently, some mechanisms to *fuse* these streams in order to extract the common semantics are needed. The third problem is the suitability of applying existing data mining concepts and algorithms (that have been successful with traditional textual data) to complex, fused multimedia data with different *quality of service* characteristics - which can be answered only when there is a measure of success in addressing the first two issues. Fortunately, as evidenced from work in image and audio analysis [12, 27, 31, 38, 11, 29, 32, 26, 13] [8, 39, 4, 16, 3, 7] there are tools to begin mining knowledge contained in multimedia data.

The objective of this extended abstract is to present a project that is underway at George Mason University where such audio and video analysis techniques are being used to mine *interesting knowledge* from multimedia data contained in movies. There are many reasons for selecting cinema. First, except for the early *silent movies*, cinema involves more than one media stream, and consequently is a good place to begin mining knowledge from data *fused* from multiple media streams. Secondly, movies on compact disks and tapes are relatively easy to access. In the third place, modulo some differences of opinion, movies have some *structure*, such as scenes and shots etc, and there are some standard cinematic styles. Further, at least the structural boundaries of the former can be detected with existing technology [39, 4, 31] (subject to a small amount of uncertainty.) Fourthly, mined knowledge can be verified against human audiences, thereby gaining a measure of validation of the correctness and appropriateness of mined knowledge. Considering all these facts, Cinema serves as a good test case for mining for knowledge contained in more than one media type, and consequently to look for appropriateness of existing mining algorithms and concepts for multimedia.

The rest of the paper is organized as follows. Section 2 de-

[*]Partly supported by an Internal Grant from George Mason University

[†]Partially supported by NSF grant IIS-9732113

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

scribes relevant background in the area of multimedia data mining. Section 3 describes our work in progress, including our objectives and plans to achieve them. Section 4 describes the prototype testbed that is being constructed. Section 5 contains concluding observations.

2. BACKGROUND

This section provides a brief summary of techniques and tools that are relevant to our work. A longer summary of work in multimedia data mining appears in [36].

The *Automatic Movie Content Analysis (MoCA)*[12] project at the University of Mannheim has developed or improved a lot of techniques to segment video and audio. They have also developed other algorithms such as for face detection, genre recognition of movies, recognizing text and music in TV commercials etc. Most of these techniques can be used to mine interesting patterns hidden in movies, or portions thereof.

One of the main issues that have to be recognized in mining for structure in movies is recognizing their boundaries. There are many methods to recognized scene cuts using the video [39, 4], audio [7, 38, 9] or both [31]. With the use of commercial tools such as the speech-to-text translator Dragon [11] and text analyzers such as WordNet [3], it is possible to enhance the detection of structure in movies.

The work at IBM Almaden Research Center covering *Query by Image Content (QBIC)*[13] and *QueVideo* [8] projects have made significant advances in recognition, indexing and content based retrieval of video [33], audio [34] and images. Similarly, the *Informedia* Digital Libraries project at Carnegie Mellon University [7] has made significant advances in indexing and analysis of digital media with respect to content understanding, indexing and creation of digital libraries. The *MultimediaMiner* [40, 41, 23] project and its predecessor projects [28, 24] at Simon Fraser University have constructed many image understanding, indexing and mining techniques in digital media. On related projects, there have been considerable advances in mining knowledge from spatial [20] and geographic [21] databases.

3. WORK IN PROGRESS

The objective of our *cinema miner* is to build a *framework* consisting of concepts, their implementation mechanisms, relevant algorithms and a test-bed to mine *interesting knowledge* contained in movies. We concentrate in mining higher level knowledge from streams using already available detection, identification and querying capabilities for underlying audio and video, with the expectation that our concepts, mechanisms and algorithms will be sufficiently generic and independent of the underlying media specific detection and identification mechanisms and thus will accommodate new advances in these areas.

Data contained in, or about a movie come from several sources. Firstly, there is some meta data, such as names of the crew including actors, director(s) etc, where the movie was filmed and other pertinent business data, criticisms, ratings, popularity measures, advertisements etc. These can be obtained in textual form. Secondly, the *story* may be available in some textual form too. Thirdly, there are audio and video tracks, which is the final outcome prepared for the audience. For our mining efforts, we assume that meta data and the movie is available, but plan to use the story in its textual form only as a means of validating our results.

The type of knowledge we are interested in mining from movies can be classified as follows:

Compositional Structure: This includes the number and sequencing of the movie into scenes, shots segments etc, and will be described shortly.

Interesting Events: Specific events that indicate emotions, mood, serenity, violence level etc., For example, crying may indicate sadness (or happiness), clapping or laughter may indicate happiness and bomb blasts or gun shots may indicate street violence.

Event Patterns: Such as *Violence is followed by sad scenes* with some probability.

Clustering Movies: Finding clusters of movies such as the one's with a *sad ending.*

Relationships Between Movies and Meta Data: These are relationships such as *Movies with sad endings produced by Studio X earns more than 20 million in their first year.*

We now explain each item in detail, and how our planned prototype is to mine them.

3.1 Compositional Structure

The structure of movies can be described using a cinematic hierarchy [10] such as scenes where each scene is divided in to a sequence of shots and each shot is divided into a sequence of frames, as given in Fig. 1. We use already developed techniques [39, 4, 26, 29, 31] to detect video and audio scene-cuts, and to develop a preliminary classification of movies as a collection of successive scenes. Further, the time stamp of the video track can be used to index into the audio track to obtain the corresponding audio transcript that *describes* the scene. By using commercially available audio to text translators [11], we can obtain the textual content of the spoken conversation. By analyzing the textual descriptions, we plan to construct cinematic hierarchies of movies.

Following audio-video technology can be used for the purposes of creating cinematic structure.

Detecting Scene Cuts: Using abrupt changes in chrominance, luminance and other visual parameters between successive video frames it is possible to detect scene cuts with a certain probability [39, 4].

Detecting Cinematic Artifacts: Using similar technology it is also possible to detect cinematic artifacts such as camera pans, tilts, zooming, etc within a scene [39, 31].

Detecting Shots within a Scene: By using cinematic artifacts occurring within a scene, it is possible to divide

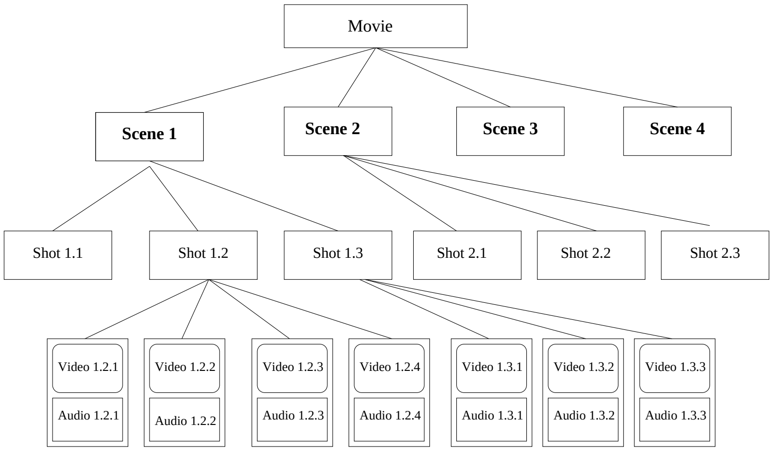


Figure 1: Movie Hierarchy

a scene into a sequence of *shots*. For example, one scene could be a *zooming in* action of 50 frames followed by a *zooming out* of 35 frames [31].

Similarly, there are techniques to segment the audio track consisting of four basic steps [31].

Separation of Music, Voice, Silence and Noise: By using frequency spectrum, relative loudness and other parameters of traditional sound analysis, it is possible to filter the sound track into music, voice. By using similar techniques, it is even possible to segment the music stream into different *beats*. These are currently available as either commercial products or research prototypes [35, 26, 29, 31].

Translating Voice into Text: There currently exist voice to text translators to translate the voice stream into the textual format. These are commercially available [11].

Understanding Text: By *understanding phrases* in texts [3], it is possible to categorize a text stream corresponding to a sequence of sentences by extracting a number of key words.

Recognizing Special Sounds: It has been shown that *gun shots*, *cries*, *clapping* and *bomb blasts* can be identified with some degree of certainty. Also there are research prototypes that can retrieve sounds that are *similar* to ones spoken to a microphone. (The analog of *query by image content*[13], called *query by humming*[14]).

We plan to *merge* knowledge gained by video segmentation, audio segmentation and textual understanding to characterize each movie consisting of a four level hierarchy as given in Fig. 1. At each level of the hierarchy above that of frames, we collect parameters, such as the sequence of camera artifacts, key phrases, special sounds (such as laughter, cries, gun shots, bomb blasts etc) and standard graphics parameters such as average RGB, luminance, chrominance etc. In addition, textual meta data about movies such as the director, rating, popularity, year of production, amount of money spent, studio used, country of production type of movie (i.e. thriller, children's movies, biography, romance, war, ethnic conflicts, Medieval story etc) actors, awards won etc are also available. By using these statistics we plan to mine for the following type of knowledge:

Frequent Episodes: What are the most frequent episodes in movies, or in a movie? Following standard terminology in temporal data mining, an episode consists of a sequence of *events*, where we consider an event to be any item in the movie hierarchy, such as frame, shot, scene or act. Notice that in order to apply known algorithms for mining for frequent episodes we need to enhance them to accommodate the following differences:

Basic Events: Identification of basic events has the following two added complexities.

Compound Nature: In the work of Manilla, Toivonen [25] and others, identification of basic events do not pose any problems: i.e. the associated textual predicate indicates if the event was present or not. In our case, the lowest level of events are complex objects consisting of at least one component of audio and video.

Probabilistic Nature: As stated in [25], the identification of basic objects is non probabilistic: i.e. either they were present or not. In our case, there is a non-perfect probability associated with them, as audio and video detection methods does not guarantee an absolute decision, but comes with an associated probability.

In order to address our mining requirements, we are re-addressing the basic formulation of *episode mining* to account for basic mine-able events consisting of components being identified with non-perfect probabilities. Our long term goal is to extend this work to a case where one or more of these components are *missing* and the assumption of perfect temporal synchrony is no longer valid due to network based media delivery.

Hierarchy: There is a hierarchy of events such as frames, shots, scenes, acts that belong to different levels of granularity. Consequently, using patterns of events at a lower granularity being contained in larger episodes with coarse grain granularity need to be investigated. This has to be done using the knowledge gained at the lower level to help the mining at a higher level. This is similar to the issue faced in conventional episode mining, but requires some enhancements to account for probabilities associated with object identity.

Mining for Trends Within Movies: We plan to investigate trends such as *Do all movies begin happily and end sadly?* or vice versa. In mining for such knowledge we need to characterize *happy beginnings* or *sad endings.* In detecting such trends, we can use (for the lack of better word) a *happiness* index such as the mixture (or percentage mixtures of) laughter or crying within the scene, key words or phrases associated with the scene matching those that have been pre-characterized as indicating happiness, and the existence of special cinematic effects or average RGB colors, and beats in music). Notice that this is equivalent to classifying movies, according to certain criteria. E.g., a movie can be deemed *happy* or not; *violent* or *non-violent.* Once these concepts are associated with a series of parameters (as discussed before for *happiness* index), one can use classification techniques [22, 30, 5, 6, 18, 17] to achieve this task. Notice also that a movie can belong to more than one class: e.g., a movie can be *happy*, and *adult-oriented.* There are enough movies to verify the accuracy of potential predictions such as *A bomb blast is followed by two minutes of sad scenes*, or at a higher level *In action movies, an act of violence is followed by two acts more (perhaps retaliatory or punitive) of violence*, and the validity of such claims are to be tested by user studies.

Association Rules Mining: We plan to develop techniques to mine for association rules [1, 2, 15] at every level of the hierarchy. For instance, one can find that certain kinds of objects in a frame representative of a scene are often associated with other kinds of objects in the frame that represents the next scene. At the higher level, one can discover association rules among high-level concepts and features such as directors, amount of money spent, type of movie, *trend* (as mined by our algorithm) period, money earned, studio of production, and so on. Notice that applying standard association rules techniques to multimedia not only uncovers knowledge, but also helps in the object recognition problem, by potentially enhancing the probability assigned to the object. For instance, if one knows that objects type A and B occur frequently together (within a window of δ frames), and it is discovered in the movie that is being mined currently that an object, believed to be of type A and and object believed to be of type B are occurring frequently together, the probability of being of type A, attached to the first object and the probability of being of type B, attached to the second can be strengthened (via a heuristic procedure).

We also plan to extend our association rule mining for *quantitative knowledge*, which are customarily called quantitative association rule. An example would be *A bomb blast lasting T minutes is followed by two acts of mourning*, but *a gunshot is followed by crying in two consecutive scenes of the same act.*

Clustering Movie Trends and Categorization: We plan to cluster movie trends using our hierarchical information. Using the information obtained from mining trends and association rules, one can form a vector of features that describe the movie. Then, clustering algorithms [19] can be applied to group movies into similar classes.

4. PROTOTYPE CINEMA MINER

Fig. 2 shows the structure of our prototype. As stated in Section 1, it consists of using already available audio and video analysis techniques to separate and identify components and datum that are used in the mining process. The hierarchy constructor, text analyzer and movie miner are the components that will receive our concentrated attention in this project. Each of these units can be considered a module in a prototype we plan to build in stages as a result of this research agenda.

Textual Analysis of Audio: As shown in Fig. 2, this module is fed voice extracted from the audio stream with appropriate time stamps. It is then fed to text off-the-shelf text analysis module, where the audio will be translated to a parse tree of text. The parse tree can be used to analyze the *story* to some extent. We also propose to use selective phrases and words that are indicative of noteworthy or interesting events.

Cinema Miner: This module collects all the information in the form of a feature vector, including those detected by the video and audio analysis components, and mine for *knowledge* out of them. It is divided into trend prediction, event sequence mining, association rules and clustering sections, as described earlier.

User Validation of Mined Data: We plan to carry out user surveys to figure out the validity of mined knowledge from movies, in the spirit of perception studies related to multimedia usability [37].

5. CONCLUSIONS

We have presented a blueprint for mining information in and about movies that has become feasible by combining existing technology in image, audio and text analysis and understanding. As stated in the introduction, even under the assumptions of perfect time wise synchrony, mining common patterns existing in a collection of fused media streams present a challenge that goes even beyond the difficult problems of image, text and audio understanding. Even so the more important issues is to find out if common concepts used in mostly text based mining such as clustering, event patterns and commonly occurring pairs would be sufficient to describe common patterns and hidden knowledge available in cinema.

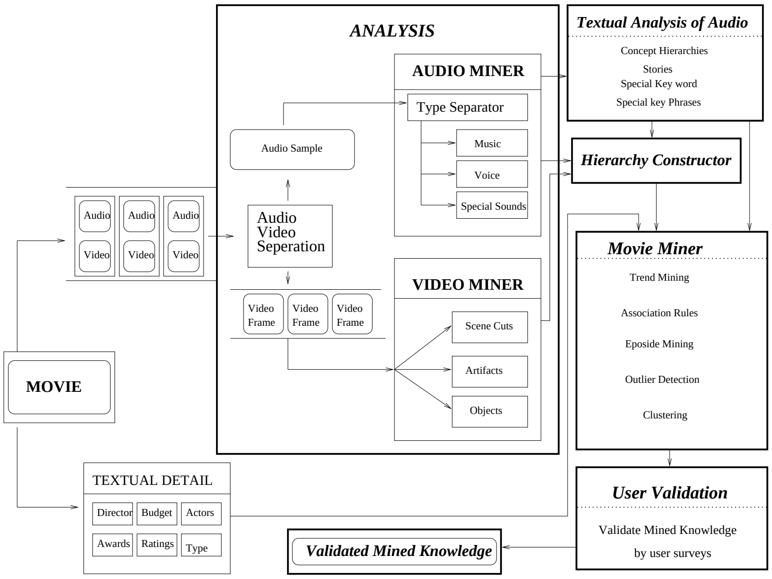


Figure 2: Prototype Cinema Miner

6. REFERENCES

[1] R. Agrawal, T. Imielinski, and A.Swami. Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, may 1993.

[2] R. Agrawal, H. Mannila, R.Srikant, H. Toivonen, and A.Inkeri. Fast Discovery of Association Rules. In U.Fayyad, G. Shapiro, P. Smyth, and R. Uthurusamy, editors, *In Advances in Knowledge Discovery and Data Mining*. AAAI press, 1996.

[3] R. Beckwith, F. C., D. Gross, K. Miller, G. A. Miller, and R. Tengi. Five papers on wordnet: Special issue of the journal of lexicography. *Journal of Lexicography*, 3(4), 1990.

[4] J. S. Boreczky and L. A. Rowe. A comparison of video shot boundary detection techniques. *Journal of Electronic Imaging*, pages 122–128, September 1996.

[5] L. Breiman, J. Friedman, R. A. Olsen, and C. Stone. *Classification and Regression Trees*. Chapman and Hall, 1984.

[6] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 1998.

[7] M. Christel, S. Stevens, and H. Watlar. Informedia digital library. *Communications of the ACM*, 34(9):57–58, April 1994.

[8] The cuevideo project. http://www.almaden.ibm.com/cs/cuevideo/index.html.

[9] A. Czyzewski. Mining knowledge in noisy audio data. In *Proceedings of the Second International Conference of Knowledge Discovery and Data Mining (KDD'96)*, pages 220–225, Portland, Oregon, 1996. AAAI Press.

[10] G. Davenport, T. A. Smith, and N. Pincever. Cinematic primitives for multimedia. *IEEE Computer Graphics and Applications*, July 1991.

[11] Audiomine by dragon systems inc. Available at `http://dragonsys.com`.

[12] W. Effelsberg. Automatic movie content analysis. http://www.informatik.uni-mannheim.de/informatik/pi4/projects/MoCA, 1994.

[13] M. Flickner, H. Sawhney, and W. Niblack. Query by image and video content: The qbic system. *IEEE Computer*, 28:23–32, 9 1995.

[14] A. Ghais, J. Logan, D. Chamberline, and B. C. Smith. Query by humming: Musical information retrieval in an audio database. In *Proceedings of the Third ACM International Conference on Multimedia*, pages 231–236. ACM Press, 1995.

[15] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proceedings of the 1995 Intl. Conf. on Very Large Data Bases (VLDB'95)*, pages 420–431, September 1995.

[16] I. Haritaoglu, D. Harwood, and L. S. Davis. w^4: Who? when? where? what?: A real time system for detecting and tracking people. *FGR'98*, 1998. Available at `http://umiacs.umd.edu/users/lsd/vsam/Pubs.html`.

[17] T. Joachims. A probabilistic analysis of the rocchio algorithms with tfidf for text categorization. In *In Proceedings of the Intl. Conference on Machine Learning*, 1997.

[18] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In

Proceedings of the 10th European Conference on Machine Learning, 1998.

[19] D. Keim and A. Hinneburg. Clustering Techniques for Large Data Sets - From the Past to the Future. Tutorial session in ACM SIGKDD International Conference On Knowledge Discovery and Data Mining. San Diego, California, June 1999.

[20] K. Koperski, J. Adikary, and J. Han. Spatial data mining: Progress and challenges. In *SIGMOD Workshop on Research Issues on Data mining and Knowledge Discovery ((DMKD'96)*, pages 27–32, Montréal, Canada, 1996.

[21] K. Koperski, J. Han, and Adhikary. Mining knowledge in geographical data. *Communications of the ACM*, 1998.

[22] L. Kubat, I. Bratko, and R. Michalski. *Machine Learning and Data Mining, Methods and Applications.* John Wiley and Sons, 1998.

[23] Z.-N. Li, O. R. Zaiane, and Z. Tauber. Illumination invariance and object model in content-based image and video retrieval. *Journal of Visual Communication and Image Representation*, 1998.

[24] Z.-N. Li, O. R. Zaiane, and B. Yan. C-bird: Content-based image retrieval in digital libraries using chromaticity and recognition kernal. In *International Workshop on Storage and Retrieval Issues in Image and Multimedia Databases, in conjunction with the 9th International Conference on Database and Expert Systems (DEXA'98)*, Vienna, Austria, 1998.

[25] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovering generalized episodes using minimal occurrences. In *Proceedings of the second International Conference Knowledge Discovery and Data Mining*, pages 146–151. AAAI Press, 1996.

[26] K. Minami, A. Akutsu, and H. Hamada. Video handling with music and speech detection. *IEEE Multimedia*, pages 17–25, july-September 1998.

[27] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura. Video handling with music and speech detection. *IEEE Multimedia*, pages 17–25, July-September 1999.

[28] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the International Conference of Very Large Databases (VLDB'94)*, pages 144–155, Santiago, Chile, 1994.

[29] S. Pfeiffer, S. Fischer, and W. Effelsberg. Automatic audio content analysis. In *Proc of ACM Multimedia*, pages 21–30, ACM Press, New York, 1996.

[30] R. Quinlan. *C4.5 - Programs for Machine Learning.* Morgan Kauffman, 1993.

[31] C. Saraceno and R. Leonardi. Audio as a support to scene change detection and characterization of video sequences. In *Proceedings of the ICASSP*, IEEE Computer Society Press, 1997.

[32] S. Satoh, Y. Nakamura, and T. Kanade. Nameit: Naming and detecting faces in news media. *IEEE Multimedia*, pages 22–35, January-March 1999.

[33] S. Sirinivasan, D. Ponceleon, and D. Amir, A. Petkovic. What is that video anyway?: In search of better browsing. In *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, pages 388–392. IEEE, IEEE Press, June 1999.

[34] S. Srinivasan, D. Ponceleon, and D. Petkovic. Towards robust features for classifying audio in the cuevideo system. In *ACM Multimedia 99*. ACM, ACM Press, November 1999.

[35] Y. Tonomura, A. Akutsu, Y. Taniguchi, and G. Suzuki. Structured video computing. *IEEE Multimedia*, pages 34–43, Fall 1994.

[36] D. Wijesekera and D. Barbara. *Mining Multimedia Datasets*, chapter F7, Handbook of Data Mining. Oxford University Press, 2000.

[37] D. Wijesekera, J. Srivastava, A. Nerode, and M. Foresti. Experimental evaluation of loss perception on continuous media. *Multimedia Systems*, 7:486–499, October 1999.

[38] E. Wold, T. Blum, and J. Wheaton. Content-based classification, search and retrieval of audio. *IEEE Computer*, 28:27–36, 9 1996.

[39] R. Zabin, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying production effects. *ACM Multimedia Systems*, 7:119–128, March 1999.

[40] O. R. Zaiane, J. Han, Z.-N. Li, S. H. Chee, and C. J. Y. Multimediaminer: A system prototype for multimedia data mining. In *Proceedings of the 1998 ACM-SIGMOD Conference on Management Data (system Demo)*, volume 38, Seattle, Washington, 1998.

[41] O. R. Zaiane, J. Han, Z.-N. Li, and J. Hou. Mining multimedia data. In *Proceedings of the CASCON'98: Meeting of Minds*, pages 27–32, Toronto, Canada, 1998.

Variations on Multimedia Data Mining

Simeon J. Simoff
Faculty of Information Technology
University of Technology, Sydney
Broadway, NSW 2009, Australia
+61 2 9514 1838

simeon@socs.uts.edu.au

ABSTRACT

Is multimedia data mining just a new combination of buzz-words or is it a new interdisciplinary field which not only incorporates methods and techniques from the relevant disciplines, but is also capable to produce new methodologies and influence related interdisciplinary fields. Rather than making an overview of existing methods and techniques, or presenting a particular technique, this paper aims to present some facets of multimedia data mining in the context of its potential to influence some relatively new interdisciplinary domains.

Keywords

multimedia, digital media, data mining, knowledge discovery, knowledge representation, case-based reasoning, computer-supported collaborative work.

1. INTRODUCTION

Multimedia and digital media, and data mining are perhaps among the top ten most overused terms in the last decade. The field of multimedia and digital media is at the intersection of several major fields, including computing, telecommunications, desktop publishing, digital arts, the television/movie/game/broadcasting industry, audio-video electronics. The advent of Internet and low-cost digital audio/video sensors accelerated the development of distributed multimedia systems and on-line multimedia communication. The list of their application spans from distance-learning, digital libraries, and home entertainment to fine arts, fundamental and applied science and research. As a result there is some multiplicity of definitions and fluctuations in terminology [2]. In this paper digital (multi)media denotes computer-mediated and controlled integration of numeric, text, graphics and other geometry representations (CAD drawings, 3D models, virtual universes), images, animation, sound, video and any other type of information medium which can be represented, stored, processed and transmitted over the network in digital form.

Another result of the rapid progress in these fields is the number of challenges for computer systems research and development, including:

- enlarged data sets with variety of formats and structures;
- variety of models for integration of media elements and components;
- demands on computational efficiency of media analysis and retrieval algorithms;
- knowledge representation schemes;
- visualisation metaphors.

Multimedia (or digital media) representations comprise a collection of domain descriptions in "native" for the domain format. Figure 1 illustrates that variety in terms of the degree to which representation is structured and to what degree the specification of this representation complies with some formal models. The term "hypermedia" is added to stress the presence of links in the digital media under consideration. of knowledge representation and specification.

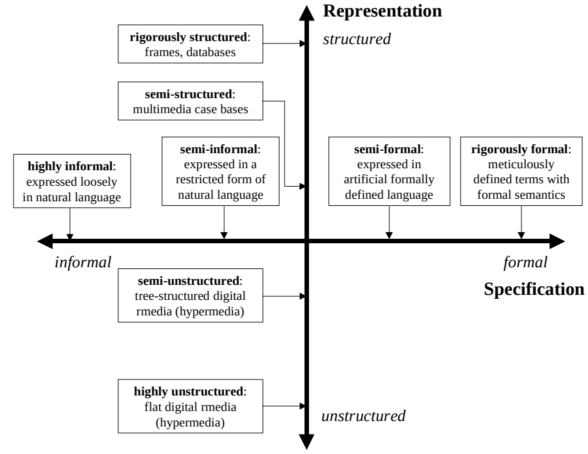

Representation
structured
rigorously structured: frames, databases
semi-structured: multimedia case bases
highly informal: expressed loosely in natural language
semi-informal: expressed in a restricted form of natural language
semi-formal: expressed in artificial formally defined language
rigorously formal: meticulously defined terms with formal semantics
informal
formal
Specification
semi-unstructured: tree-structured digital rmedia (hypermedia)
highly unstructured: flat digital rmedia (hypermedia)
unstructured

Figure 1. Digital media and domain description (adapted from [8]).

The specifics of the domain, where multimedia is used may influence the conceptual model that defines the representation of the multimedia content. The knowledge about the model can be invaluable in the development of multimedia data mining schemes, providing some initial assumptions and structure insights and assisting in the attribute identification.

Researchers in the database community are viewing multimedia data mining basically as an extension of the knowledge discovery

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

in databases, for example, as "the mining of high-level multimedia information and knowledge from large multimedia databases," a "subfield of data mining that deals with the extraction of implicit knowledge, multimedia data relationships, or other patterns not explicitly stored in multimedia databases [14]. Consequently, the framework of KDD applied in multimedia database mining provides similar knowledge representation schemes - association, classification, characterisation and other types rule patterns. This approach, more extensively described in [15], is consistent with the overall KDD methodology.

In the information system approach, the methods of multimedia data mining are better known as multimedia information analysis and retrieval. The research in the field includes a collection of works in content-based image and video search, fusion of pictorial and other media and efficient storage organisation for multimedia data [3].

Is multimedia data mining just a new combination of buzz-words or is it a new interdisciplinary field which not only incorporates methods and techniques from the relevant disciplines, but is capable to produce new methodologies and influence related interdisciplinary fields? The aim of this paper is to present some facets of multimedia data mining in the context of case-based reasoning and computer-supported collaborative work environments.

2. MULTIMEDIA DATA MINING IN HYPERMEDIA CASE BASES - AN EXAMPLE OF EXTENDING AI SCHEMES

Methods developed in multimedia data mining can have significant impact on related fields from data analysis and artificial intelligence. The potential is illustrated on the use of multimedia data mining approach in hypermedia case bases for automating case-based reasoning with unstructured case data.

2.1 Knowledge representation

Case models based on hypermedia representations are becoming popular alternatives to the strict format of object-oriented and attribute-value representations. What exactly is denoted by a case and how it is represented are major structural issues in CBR. When in financial and business applications cases are usually well-structured object-oriented or relational attribute-value representations, in interdisciplinary domains like design, digital media production and visual reasoning, cases are represented in more informal way (see Figure 1). Among the reasons for such diversity, perhaps the major one is the limited expressive power of formal representations. Hypermedia case representation is suitable for domains where it is difficult to fit domain knowledge into structured knowledge representation schemes (see Figure 1). Hypermedia case models offer richer semantics which may be considered both as alternative and extension to the strict format of object-oriented and attribute-value representations. The hypermedia representations comprise a collection of case descriptions represented as text in free or table format and other multimedia data, such as CAD drawings, images, video, sound, etc. Another characteristic of hypermedia is the use of links, where the links can connect information within a case, between different cases, or links to data that lies outside the case library.

Usually, the representation of the cases in the library is organised according to some conceptual model of the domain. This approach towards complexity is based on the following assumptions:

- a case is a hierarchy of concepts, or “subcases”;
- a case is represented by different views.

This supports case-based reasoning paradigm because subdividing a case in this way allows reasoning to focus only on the relevant parts of that case. By processing only some of the knowledge associated with a case, reasoning can become more efficient. The development of a case-base that has a hierarchical structure usually requires defining a typical decomposition of domain experience. Figure 2 presents an example of domain decomposition - the decomposition of the building design domain according to a structural engineering view of the building. Figure 3 illustrates the use of the ontology in Figure 2 for organising the access to case elements (subcases).

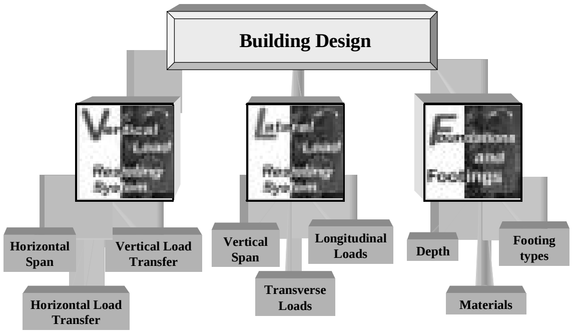


Figure 2. Case as a hierarchy of concepts.

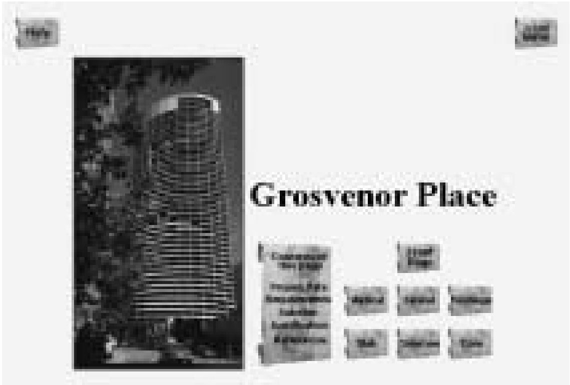


Figure 3. Example of hypermedia case organisation.

The use of different views of a domain case recognises that the experience in a domain can be understood from different perspectives. An example of this approach is presented in [12]. In this approach, a single, complex design project is represented as multiple cases. The use of multimedia can make it easier to understand complex systems - icons, images, sketches, etc. can highlight and illustrate corresponding text or tabular information.

Figure 4 illustrates a typical case page, which includes text description, images, CAD drawings and videos. In general, page layout and format are not restricted. Different people develop the actual page content of the different cases over time. The developments in web and multimedia technologies may influence the variety of elements, included in these loosely structured case descriptions.

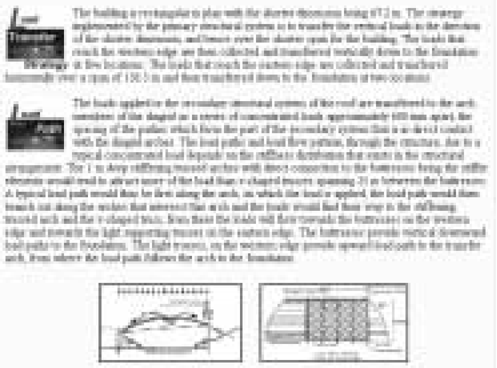
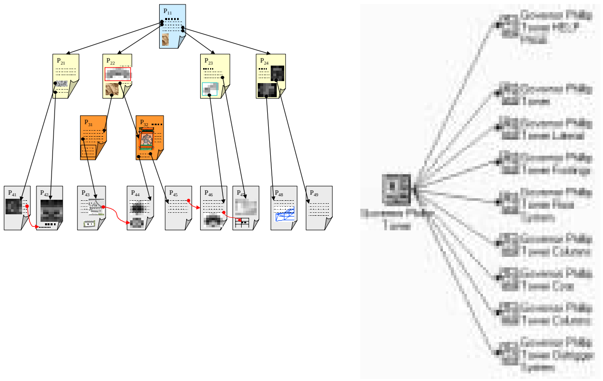


Figure 4. An example of a case page.

Figure 5 illustrates the idea of shaping the case structure according to a conceptual hierarchy. The semantics of the links depends on the relations between concepts that constitute the representation (Figure 5a). The tree-like structure (Figure 5b) may include some links between pages within a same level. Such links usually appear at a lower level in the hierarchy, where a concept may be related to concepts that belong to different branches in the hierarchy (Figure 5a).


P_{11}
P_{21}
P_{22}
P_{23}
P_{24}
Governor Phillip Tower
Governor Phillip Tower HECF Plant
Governor Phillip Tower
Governor Phillip Tower Lateral
Governor Phillip Tower Footings
Governor Phillip Tower Floor System
Governor Phillip Tower Columns
Governor Phillip Tower Core
Governor Phillip Tower Columns
Governor Phillip Tower Outrigger System

a. b.

Figure 5. Hierarchy of domain concepts, reflected in the case representation.

The use of hypermedia to develop case base systems, however, does not solve the problems in building automated reasoning algorithms, which benefit directly from the information in the library. Hypermedia representation as it is supports human reasoning rather than automated reasoning. When this is a reasonable compromise for educational purposes, there is not much use of this approach in the research and industrial case base systems – the system remains simply a structured hypermedia handbook. A common solution is to build an additional structured representation layer, a vector of case attributes. As a rule, the methodology for building such additional layers employs some knowledge engineering techniques for identifying the attributes that represent the domain case and involves domain expert(s). The attribute-value representation of the case is linked to the "entry" page of the corresponding multimedia case representation (an example of entry case page is shown in Figure 3), as shown in Figure 6. The reasoning algorithms operate over the values of the attributes, without utilising the advantages of the multimedia information available in the case.

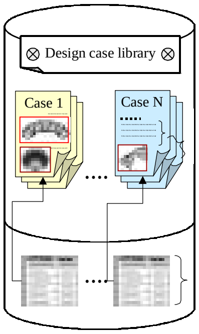


Figure 6. Hypermedia case base with additional attribute-value representation

2.2 A framework for multimedia data mining in hypermedia case libraries

Discovering implicit knowledge in hypermedia case bases is substantially different to data mining in databases. The data organisation units in database data mining are the data tables, in particular their columns or rows. Inside the hypermedia case base the organisational unit is the case or subcase, which merely consists of one or more multimedia pages. The pages comprise a variety of data formats. Knowledge discovery then, in our use of the term, involves finding patterns in primarily unstructured data. More formally the knowledge discovery process in multimedia case representations can be viewed as *machine learning where a case library replaces the training set*. The approach is illustrated in Figure 7.

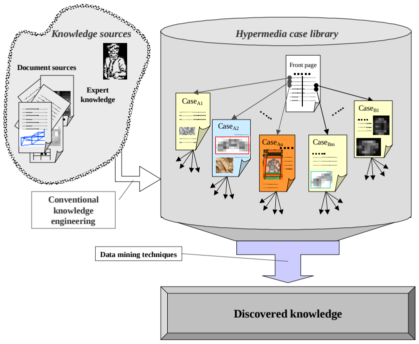


Figure 7. Knowledge discovery in hypermedia case libraries.

As illustrated in Figure 8, the information model of the case library constitutes the initial basis for the multimedia data mining [10]. During the data segmentation multimedia data are divided into logical interconnected segments. For example, in a hypermedia case library each segment can include one or more pages. Within particular media type a segment may have different meaning. For example, within a text a segment could be a paragraph, a sentence. The actual mining and analysis procedures are expected to reveal some relations in the segments and between the segments at different levels. For instance, the text analysis can identify relations between concepts presented in the text and the CAD drawings presented on that page (or vice versa, find that the actual CAD drawings are not connected with the content of the text description). The analysis within text segments can identify word concordances that denote complex terms, not explicitly defined in the case base. Extracted patterns are incorporated and linked under the framework of the information model. As a result there can be additional attributes, change in links, revision of identified attributes, changes in some attribute values. Some paragraphs, images or other media segments could become insignificant. Consequently, the information model should be able to accommodate changes in the structure and media content.

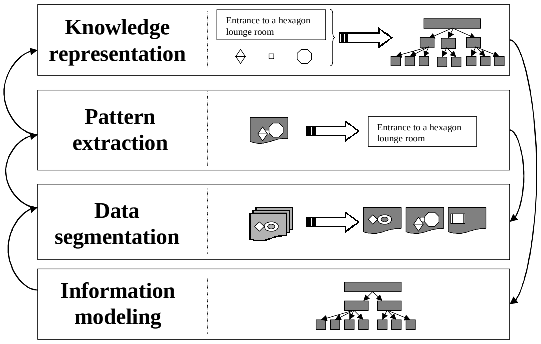


Figure 8. A model of multimedia data mining

The model is expanded in Figure 9. In case-based reasoning systems the specific interest can be focused in finding patterns in the cases that can assist with indexing and adapting cases as a way of improving the retrieval of related previous experience and indication when an adaptation lies outside some reasonable constraints, based on the experience in the case base. Patterns in the form dynamic thematic paths[1] can assist with the navigation in retrieved cases. The framework combines two consecutive complementary strategies - data- and hypothesis-driven exploration, discussed in more details in [12].

The above described data mining schema can be integrated in the learning loop of the cased-based reasoning. The overall enhanced model of case-based reasoning with knowledge discovery back-end, as shown in Figure 10, illustrates the dynamics of case manipulation, analysis, mining, knowledge formulation and case update. The visual "symmetry" in Figure 10 reflects in some sense the mutual benefit from the amalgamation of these computing approaches. On the one hand, multimedia data mining has the potential to improve the case-based system. On the other hand, the case-based paradigm provides mechanism for incorporation and management of discovered knowledge. On the indexing side potential advantages in the extended case-based reasoning model include:

- generation of term indexing schemes, based on the words used in the text representation [11];
- generation of term indexing schemes, based on relating terms to regularities discovered in other media types;
- generation of alternative indexing schemes (for example, a graph structure indexing scheme), based on structural patterns discovered in graphics, image, audio and video media in the case
- association of multiple indexing schemes to one case library;
- dynamic generation of the above listed indexing schemes.

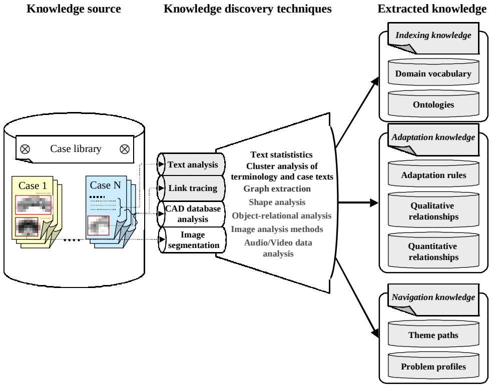


Figure 9. The process for multimedia data mining in case libraries

On the retrieval side the major advantage comes from the terminological flexibility in formulating queries due to the ontology-guided semantic transformation of the initial query and its match against the ontology-based indexing scheme [7].

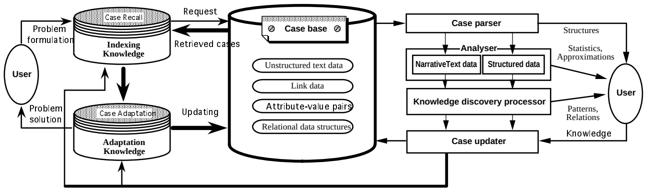


Figure 10. Enhanced case-based reasoning model based on the symbiosis between MDM and case-based reasoning.

The advantage for the adaptation in the enhanced model is the possibility to modify the new case description based on the information discovered from the multimedia analysis of retrieved cases in the context of the requirements.

3. MULTIMEDIA DATA MINING IN CSCW[2] ENVIRONMENTS

There are numerous approaches and techniques for setting up a computer-mediated environment for collaborative work [8]. The most common approach is to extend the desktop environment to include tools for meeting and sharing files. This approach takes the individual work environment and adds tools for communicating with others. An alternative approach is to create a virtual world environment in which the collaborators meet, work, and organise their projects. This approach differs conceptually because it creates a sense of place that is unique to the project, sort of a shared office space. A variation on this approach is to create a virtual world that is the model of the product or system being designed or developed.

The major feature of this kind of collaborative environment is the development of the project within the collaborative, multi-user

[1] Thematic path is a set of multimedia pages, each of which is part of the case library, that are relevant to the explanation and illustration of particular concept and have to be visited in particular sequence.

[2] Computer Supported Collaborative Work

environment. Project participants can work alone or collaboratively building the model and discussing the product as they view the model. There is only one representation of the model so there isn’t a problem with simultaneous changes to different versions. There is a continuum of the process – a person does not shift environments when designing alone or collaboratively, and there is a continuum of the workspace during the design session - all working information about the product is accessed and shared through the same environment. An example of a project scenario in such environment based on Active Worlds, Inc. virtual world support is shown in Figure 11.

Such environment is a repository of multimedia data. Multimedia data mining in such environment can be used to enhance the functionality of computer support to project participants. The data includes 3D geometry of the product, data about allocation and behavior of participants in such environments, web multimedia data used in project documentation, presentation of ideas in collaborative sessions, communication transcripts, audio and video records. The example, from on-line analysis of bulletin board records, illustrates the potential for multimedia data mining and support in this field.

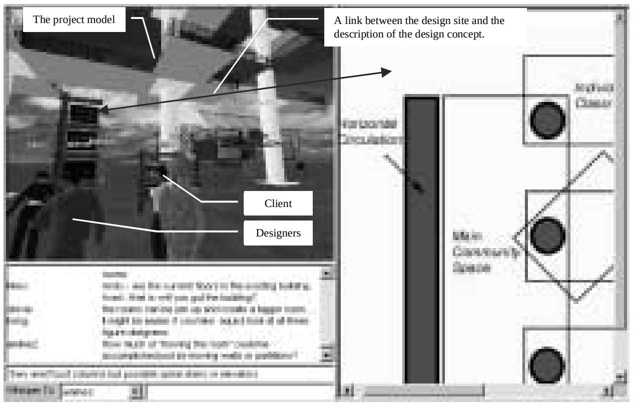

The project model
A link between the design site and the description of the design concept.
Client
Designers

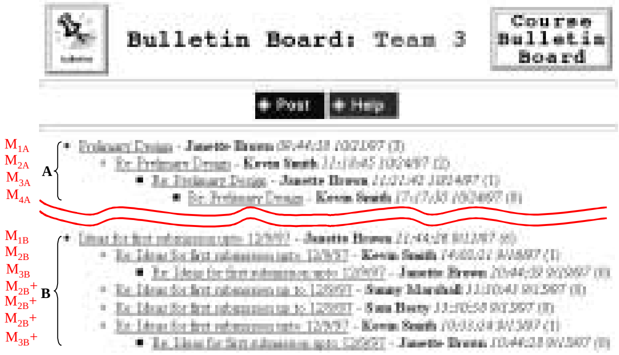
Figure 11. An example of a collaborative project in 3D/2D virtual world

Transcripts from online sessions, audio and video files can be used in the CSCW research or in education to form part of the student’s assessment by including the amount and content of the student's participation. Text-based virtual worlds provide in explicit form a descriptive record of all activities inside the world. 3D virtual worlds provide transcripts from synchronous communications. Personal contribution to a collaborative session can be evaluated using text analysis of seminar transcripts [13] and multimedia analysis of related web pages.

Multimedia bulletin boards preserve the threads and the content of each message. Thus, the analysis of these data sets can be used to evaluate team collaboration. Below is an example of using a visualisation technique, which can provide quick feedback for monitoring collaborative projects.

Figure 12 presents a fragment from a team bulletin board. The messages on the board are grouped in threads.

A threefold split of the thread structure of e-mail messages in discussion archives in order to explore the interactive threads was proposed in [1, 9]. It included (i) reference-depth: how many references were found in a sequence before this message; (ii) reference-width: how many references were found, which referred to this message; and (iii) reference-height: how many references were found in a sequence after this message. The threefold split was extended in [4] to include the time variable explicitly. This model, expressed graphically as tree, allows the comparison of the structure of discussion threads both in a static mode (for example, their length and width at corresponding levels) and in a dynamic mode (for example, detecting moments of time when one thread dominates another in multi-thread discussions).



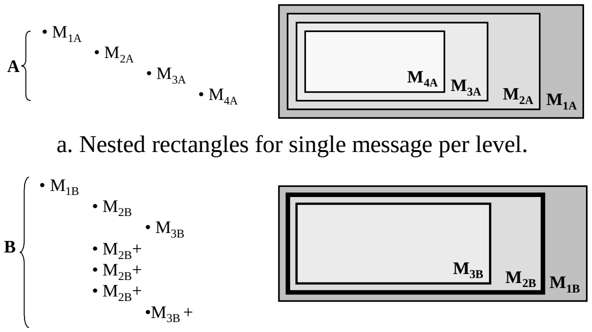
- Preliminary Design - Janette Brown [illegible]
    - Re: Preliminary Design - Kevin Smith [illegible]
        - Re: Preliminary Design - Janette Brown [illegible]
            - Re: Preliminary Design - Kevin Smith [illegible]

- Ideas for first submission upto 12/9/97 - Janette Brown [illegible]
    - Re: Ideas for first submission upto 12/9/97 - Kevin Smith [illegible]
        - Re: Ideas for first submission upto 12/9/97 - Janette Brown [illegible]
    - Re: Ideas for first submission up to 12/9/97 - Sonny Marshall [illegible]
    - Re: Ideas for first submission up to 12/9/97 - Sam Berry [illegible]
    - Re: Ideas for first submission upto 12/9/97 - Kevin Smith [illegible]
        - Re: Ideas for first submission upto 12/9/97 - Janette Brown [illegible]

Figure 12. Fragments from an asynchronous communication in a virtual world bulletin board.

Visualisation techniques based on this model are modified versions of the nested set visualisation of tree structures [5]. Figure 13 shows an example of such visualisation applied to threads "A" and "B" from Figure 12. Each first message in a level is represented by a corresponding rectangle, labeled in this example to illustrate the message correspondence. Thus, there are four nested rectangles in Figure 13a. When messages are at the same level the thickness of the line is estimated based on the content-analysis of the message, including the text, included graphics and images. Each of the relevant messages on the same level is represented as additional 0.5 pt to the baseline thickness. In Figure 13b the base line thickness is 1 pt, thus rectangle "M_{2B}" has thickness 2.5 pt.

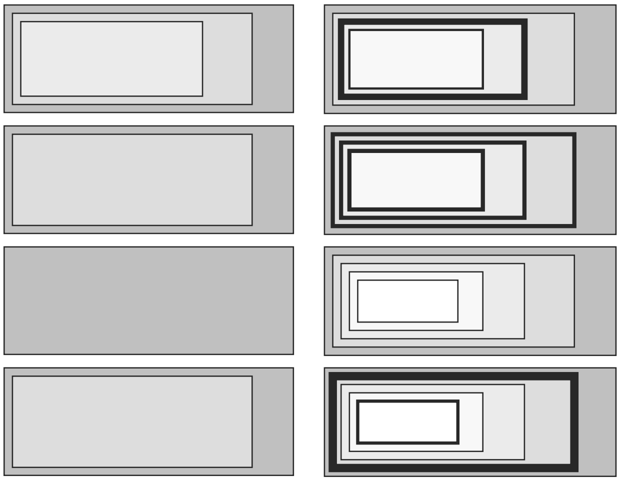
Figure 14 illustrates the application of the technique for monitoring collaborative design teams. Collaboration on a shared design and development task can be considered at different levels of abstraction and "degrees" of task sharing. Two extreme approaches to sharing design tasks during collaboration are identified in [6]: single task collaboration and multiple task collaboration. During single task collaboration the resultant design (or project development) is a product of a continued attempt to construct and maintain a shared conception of the design task. In other words each of the participants has his/her own view over the whole design problem and the shared conception is developed during intensive discussions. An example, of the visual pattern of such type of collaboration is presented in Figure 14b. It is characterised with relatively large amount of nested rectangles, usually indicating also several messages in respond to particular message. During multiple task collaboration the design problem is divided among the participants so that each person is responsible for a particular portion of the design. Thus, multiple task collaborative design does not necessarily require the creation of a single shared design conception, thus messages are usually related to the project management. Isolated messages and short threads dominate this collaboration style, as illustrated in Figure 14a.



a. Nested rectangles for single message per level.

b. Nested rectangles when there are multiple messages on some levels.

Figure 13. Visualisation of discussion threads.



a. collaboration connected more with coordinating project tasks and submissions

b. intensive collaboration for creating a joint understanding of the problem

Figure 14. Patterns of collaboration.

Such visualisation techniques, combined with multimedia analysis of (i) video sequences of communications, (ii) elements of the 3D models in the scenery, and (iii) 2D representation of the project media, are expected to be part of the next generation CSCW environments.

4. EPILOGUE

So is multimedia data mining just a new combination of buzz-words or is it an exciting new area, capable to produce new methodologies and influence related interdisciplinary fields. The answer is left to the reader.

5. ACKNOWLEDGMENTS

Author would like to thank the Australian Research Council for the support of the related projects.

6. REFERENCES

[1] Berthold, M. R., F. Sudweeks, S. Newton and R. Coyne "Clustering on the Net: Applying an Autoassociative Neural Network to Computer-Mediated Discussions," *Journal of Computer Mediated Communication*, 2 (4), http://www.ascusc.org/jcmc/vol2/issue4/bert-hold.html (1997).

[2] Fluckiger, F. Understanding Networked Multimedia: Applications and Technology, Prentice Hall, London (1995).

[3] Ip, H. H. S. and A. W. M. Smeulders, eds., *Multimedia Information Analysis and Retrieval*, Springer, Heidelberg, (1998).

[4] Jones, S. ed., *Doing Internet Research*, Sage Publications, Thousand Oaks, CA, 29-55 (1999).

[5] Knuth, D E., *The art of computer programming, Vol 1: Fundamental algorithms*, Addison-Wesley, Reading, MA, 311-312 (1973).

[6] Maher, M. L., S. J. Simoff and A. Cicognani, "Potentials and limitations of Virtual Design Studio," *Interactive Construction On-line*, January, a1 (1997).

[7] Maher, M. L. and S. J. Simoff, "Knowledge discovery from multimedia case libraries", in Smith, I., ed., *Artificial Intelligence in Structural Engineering*, Springer, Berlin, 197-213 (1998).

[8] Maher, M. L., S. J. Simoff and A. Cicognani, *Understanding Virtual Design Studios*, Springer, Heidelberg (2000).

[9] Sudweeks, F., M. McLaughlin and S. Rafaeli, eds, *Network and Netplay: Virtual Groups on the Internet*, AAAI/MIT Press, Menlo Park, CA, 191-220 (1998).

[10] Simoff, S. J. and M. L. Maher, "Ontology-based multimedia data mining for design information retrieval", *Proceedings of the ACSE Computing Congress*, Cambridge, MA, 310 - 320 (1998).

[11] Simoff, S. J. and M. L. Maher, "Deriving ontology from design cases", *International Journal of Design Computing*, 1, http://www.arch.usyd.edu.au/kcdc/journal/vol1 (1998).

[12] Simoff, S. J. and M. L. Maher, "Knowledge Discovery in Hypermedia Case Libraries - A Methodological Framework," *Proceedings of the Knowledge Acquisition Workshop, 12th Australian Joint Conference on Artificial Intelligence, AI'99*, 213-225 (1999).

[13] Simoff, S. J. and Maher, M. L. "Analysing Participation in Collaborative Design Environments," *Design Studies*, 21, 119-144 (2000).

[14] Zaïane, O. R., J. Han, Z.-N. Li, S. H. Chee, S. H. and J. Y. Chiang. "MultiMediaMiner: A system Prototype for MultiMedia Data Mining, *Proceedings of ACM SIGMOD International Conference on Management of Data*, 581 - 583 (1998).

[15] Zaïane, O. R., J. Han, Z.-N. Li, J. Hou, "Mining Multimedia Data" *Proceedings CASCON'98: Meeting of Minds*, Toronto, Canada, 83-96 (1998).