Web Page Categorization and Feature Selection Using Association Rule and Principal Component Clustering

Jerome Moore, Eui-Hong (Sam) Han, Daniel Boley, Maria Gini, Robert Gross, Kyle Hastings, George Karypis, Vipin Kumar, and Bamshad Mobasher

Professor:
Dr. Osmar ZAIANE

Student:
Alexandru COMAN

Why Do We Need It?

- retrieve, filter and categorize documents available on the World Wide Web;
- extract salient features of related web documents to automatically formulate queries and search for other similar documents on the web;
- organize bookmark files;
- construct a user profile.

What Tools We Have?

Search engines:

- rely on large indexes of documents located on the web;
- determine the URLs of those documents satisfying a user's query.

Traditional clustering algorithms:

- use *a priori* knowledge of document structure to define a distance or similarity among documents;
- use probabilistic techniques.

What Is the Problem?

- World Wide Web is a vast resource of information and services that continues to grow rapidly;
- often queries return inconsistent search result with documents meeting the search criteria but of no interest;
- dimensionality of the feature space is high relative to the size of the document space.

What Tools Should We Build?

Intelligent software agents:

• extract semantic features from the words or structure of an HTML document;

• use these features to classify and categorize the documents.

Advantage:

• don't need *a priori* knowledge;

• categorization process is unsupervised.

What Are the Problems of Traditional Clustering Methods?

They use a selected set of words (features) from different documents as the dimensions. Each document is represented by a feature vector -> a point in this multi-dimensional space.

Problems:

1. It is not trivial to define a distance measure;

2. The number of words could be very large.

Proposed Methods

1. Association Rule Hypergraph Partitioning Algorithm

2. Principal Component Analysis Partitioning Algorithm

Advantages:

• efficiently handle very high dimensional spaces;

• don't depend on pre-specified distance functions;

• capable of automatically discovering document similarities or associations.

Association Rule Hypergraph Partitioning Algorithm (ARHP)

Some notions:

1. **Hypergraph** - a graph whose hyperedges connect two or more vertices;

2. **Hyperedge** - a connection between two or more vertices of a hypergraph;

3. **Confidence level** - conditional probability that a feature occurs in a document or group of documents given that it occurs in the remaining documents in that hyperedge.

Association Rule Hypergraph Partitioning Algorithm (ARHP) (cont.)

Steps:

• each document correspond to an item, each possible feature correspond to a transaction;

• an association rule discovery algorithm is used to find sets of documents with many features in common (frequent item set) satisfying a user specified minimum support criteria;

• a hypergraph is formed with documents as vertices and frequent item sets as hyperedges;

Association Rule Hypergraph Partitioning Algorithm (ARHP) (cont.)

Steps (cont.):

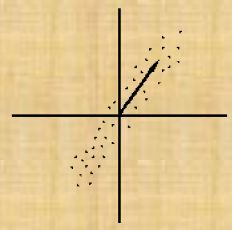
• each hyperedge will have a weight - average confidence of all the association rules involving the related documents of the hyperedge;

• a hypergraph partitioning algorithm is used to partition the hypergraph such that the weight that are cut by the partitioning is minimized.

Principal Component Analysis (PCA) Partitioning Algorithm

Some notions:

1. **Principal Component Analysis** - is a technique to find the directions in which a cloud of data points is stretched most.



2. **Scatter value** - average distance from the documents in a cluster to the mean.

Principal Component Analysis (PCA) Partitioning Algorithm (cont.)

Steps:

• each document is represented by a feature vector of word frequencies scaled to unit length;

• the space of documents is cut with a hyperplane passing through the overall arithmetic mean of the documents and normal to the principal direction;

• each group is further split in the same manner ->tree structure;

• the leaves are document clusters, each with a computed mean and principal direction;

• scatter value is used to determine the next cluster to split.

Experimental Setup

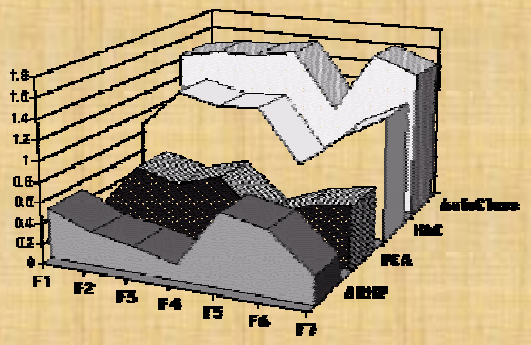
- 98 web pages in four broad categories are used: business and finance, electronic communication and networking, labor, and manufacturing;
- every page is labeled -> facilitates entropy calculation;
- seven experiments were created;
- four algorithms were used:
  - Association Rule Hypergraph Partitioning (ARHP);
  - Principal Component Analysis Partitioning (PCA);
  - Hierarchical Agglomeration Clustering (HAC);
  - Bayesian Classification Method (AutoClass).

Experimental Setup (cont.)

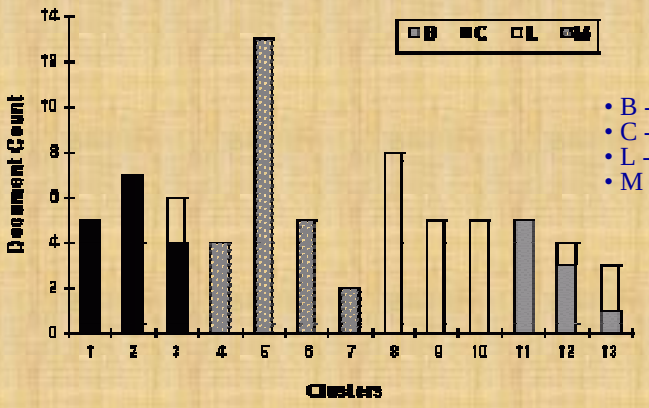
Experiment	Selection Criteria	Dataset Size
F1	All words	98 x 5623
F2	Quantile Filtering	98 x 619
F3	Top 20+ words	98 x 1239
F4	Top 5+ words plus emphasized words	98 x 1432
F5	Frequent itemsets	98 x 399
F6	All words with text frequency > 1	98 x 2641
F7	Top 20+ with text frequency >1	98 x 1004

Evaluation of Results

Entropy will measure the goodness of the cluster.



Evaluation of Results (cont.)



- B - Business
- C - Communication
- L - Labour
- M - Manufacturing

ARHP Algorithm and F4 feature selection method

Conclusions

1. Clustering methods based on partitioning work best because:

- do not depend in a choice of a distance function;
- do not require calculation of the mean of the clusters;
- are not sensitive to the dimensionality of the data set.

2. Careful selection of a small number of representative features from each document is important.

Related Work

Other web agents use:

- semantic information embedded in link structures;
- pattern recognition methods and word clustering;
- boolean feature vector to represent a HTML page;
- single well-defined profile to find similar web documents for a user;
- Kohonen Self-Organised Feature Map.