Incremental Document Clustering for Web Page Classification

By Wai-chiu Wong and Ada W. Fu
The Chinese University of Hong Kong

Presented by: Zhigang Shi

The problem this paper tries to solve

- How to automatically organize the massive amount of Web pages retrieved from the internet by a search engine

The main contributions of this paper:

- Proposed a feature extraction mechanism which is more suitable for Web page classification
- Introduced DC-tree to make the clustering process incremental and less sensitive to the document insertion order
- Experiments of applying the proposed algorithm to real data

What is document classification

- A process to help the information retrieval systems organize the vast amount of documents
- It can help to make the retrieved results easier to browse.

Traditional document classification

- Done manually, need large amount of human effort
- Available automatic text classification algorithms not suitable for web page classification

How clustering techniques can help to make document classification process automatic

- It can find clusters directly from the given data, without relying on any pre-determined information such as training examples provided by domain experts.

Existing clustering algorithms cannot be applied to Web document classification

- Most of them (e.g. CLARANS, BIRCH) require the supply of the number of clusters to work on. Unfortunately, the number of clusters of the document set is usually unknown to user
- Document databases are now facing high rate of update

Traditional feature extraction methods

- Select the n highest-weighted terms as the features
- The term weighting scheme is commonly based on the term frequency (TF)

Statistical analysis in the Web domain

	No. of distinct words	Average word freq.
Mean	174.75	1.68
Median	116.00	1.55
Standard Dev.	225.82	0.71
Maximum	4351	17.59
Minimum	1	1.0

Proposed new feature extracting method

1. Randomly select a subset of documents with size m from the corpus.
2. Extract the set of words that appear at least once in the documents. Remove stop words and combine the words with the same root by using the stemming technique.
3. Count the document frequency of the words which are extracted in Step 2.

Proposed new feature extracting method (Continued)

4. Set *lower* = k and *upper* = k
5. Select all words with document frequency in the range from lower to upper
6. Check if the coverage of these words is larger than the pre-defined threshold. If so, stop. Otherwise, set *lower* = *lower* –1 and *upper* = *upper* + 1 and go to step 5

Document representation

A document D_i is represented as: $D_i=(W_i, ID_i)$

Where Id_i is the document identifier and W_i is the feature vector of the document: $W_i = (w_{i1}, \ldots, w_{in})$ Here w_{ij} is the weight of the j-th feature. It is equal to 1 if D_i contains the j-th feature, otherwise, it is equal to 0.

Document cluster (DC Definition)

Given N documents in a cluster: $\{D_1, D_2, D_N\}$, the Document Cluster entry of a node is defined as a triple: DC= (N, ID, W), where N is the number of documents in the cluster, ID is the identifiers of the documents in the cluster: ID = $\{ID_1, \ldots, ID_N\}$ and W is the feature vector of the document cluster: W = $(w_1, \ldots, w_n)$ where wj = $\sum w_{ij}$

DC-tree

A DC-tree is a tree with four parameters: branching factor (B), two similarity thresholds (S1, S2, where $0 \leq S1, S2 \leq 1$) and the minimum number of children of a node (M).

Example of a DC-tree

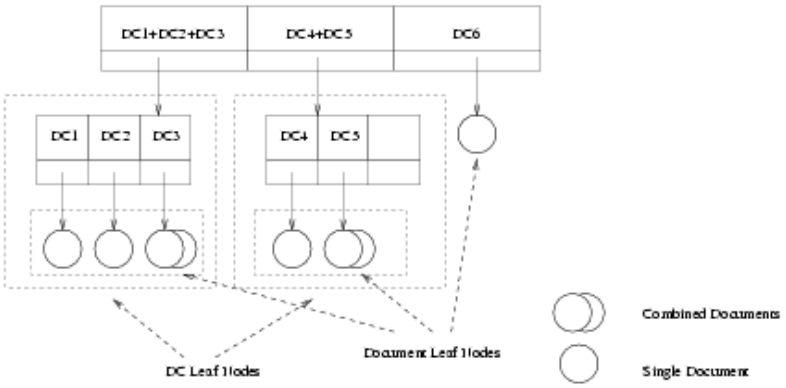

Fig. 1. Example of a DC-tree

Experiments

- Two sets of web pages from Yahoo search engine are selected
- On set contains non- correlated document set using ten sub-topics such as author, wine, film etc
- Another set contains correlated topic Web document set using another ten sub-topics such as Java, Perl, Python, etc.
- 20000 Web pages for each document set

Results

Cluster Topic	Precision	
	DC-tree	B^+-tree
author	0.61	0.59
internet game	0.73	0.55
wine	0.63	0.65
credit bank	0.66	0.57
soccer	0.73	0.47
astronomy organization	0.88	0.77
psychology department	0.85	0.63
job opportunity	0.91	0.72
film	0.72	0.73
disease organization	0.64	0.65
Average	**0.736**	**0.633**

(a)

Cluster Topic	Precision	
	DC-tree	B^+-tree
Cobol	0.64	0.54
Fortran	0.72	0.72
Java	0.73	0.62
JavaScript	0.64	0.67
Lisp	0.73	0.63
Pascal	0.81	0.67
Perl	0.78	0.72
Python	0.82	0.55
Smalltalk	0.61	0.57
VisualBasic	0.77	0.65
Average	**0.725**	**0.634**

(b)

Table 2. Precision of clusters (a) non-correlated topics (b) correlated topics

Conclusion

- The paper introduced DC-tree for Web document clustering
- Proposed a feature extracting method for document clustering in the Web domain
- Show by experiment that the features of the DC-tree give good results in terms of accuracy and efficiency

Thank you

Questions?