COMAD 2008

CUM : An Efficient Framework for Mining Concept Units

By

Mrs. Santhi Thilagam

Department of Computer Engineering

NITK - Surathkal

Outline

- Introduction
- Problem statement
- Design and solution methodology
- Experimental setup and Results
- Conclusion
- References

Introduction

- Data mining
- Web mining

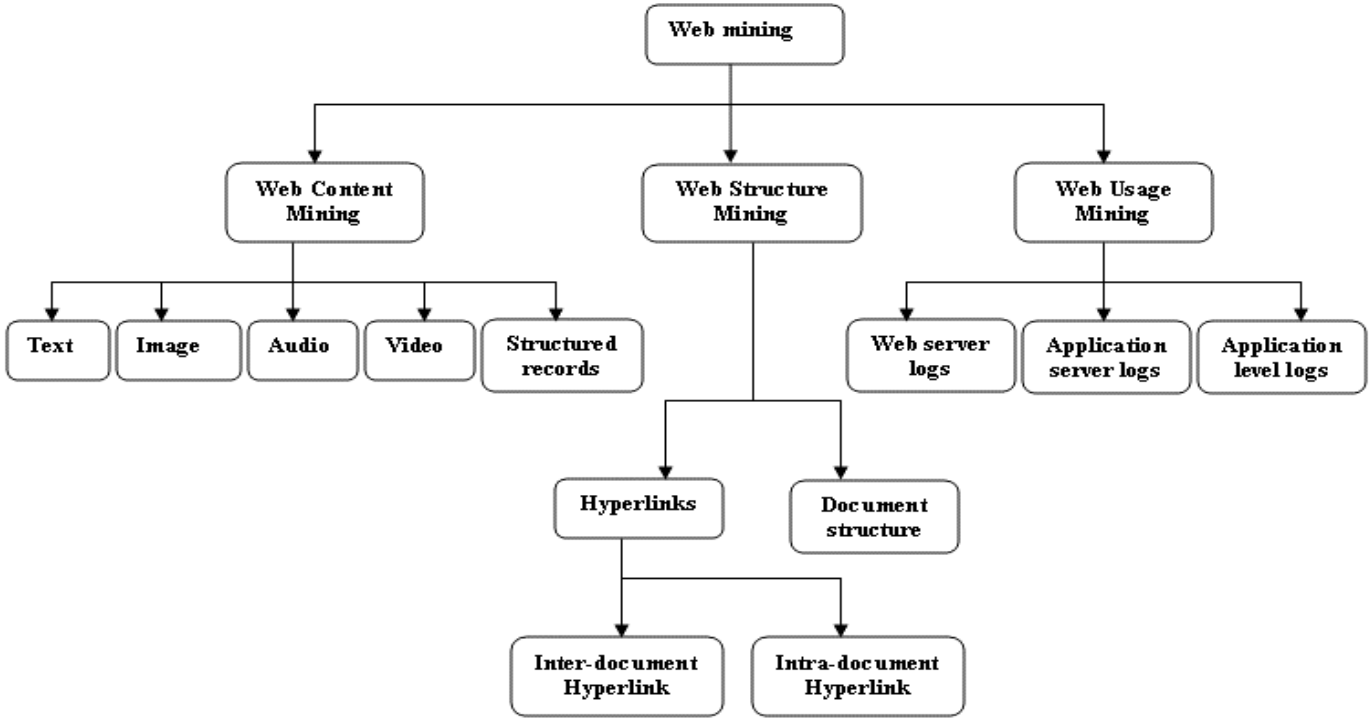

Figure 1: Web mining taxonomy

Introduction

Cont'd..

Web classification

- The process of categorizing a set of objects from the Web into some pre-defined categories.

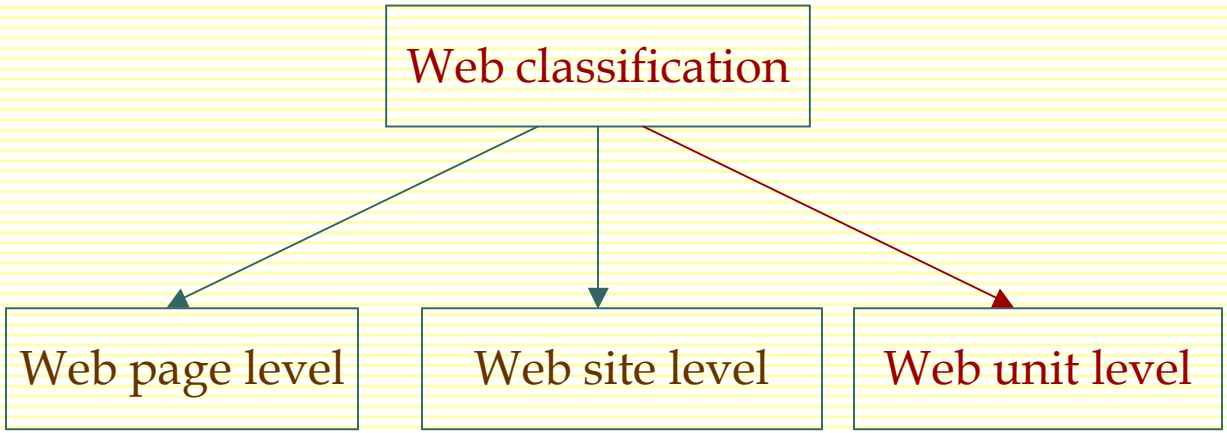

Figure 2 : Web classification

Definitions

Cont'd...

- Web Unit (or) Concept unit
  - Web unit of the domain concept is a web page or a set of web pages from a web site that jointly provides information about a concept instance.
  - A web unit consists of exactly one key page and zero or more support pages.
  - Key page is a page which has links to all the support pages having the supplementary information about the concept.

- **Example:**

http://...path/course/CS100/CS100.html
http://...path/course/CS100/lecture-programs.html
http://...path/course/CS100/instructors.html
http://...path/course/CS100/officehours.html
http://...path/course/CS100/exams/final.html
http://...path/course/CS100/exams/preliminary.html
http://...path/course/CS100/programs/program1.html
http://...path/course/CS100/programs/program2.html

(a) CS100 Web unit

http://...path/user/Johnson/index.html
http://...path/user/Johnson/research.html
http://...path/user/Johnson/publications.html
http://...path/user/Johnson/activities.html
http://...path/user/Johnson/students.html
http://...path/user/Johnson/teaching.html
http://...path/user/Johnson/contact.html

(b) Johnson Web unit

Figure 3: Web unit examples: Course web unit and faculty web unit

- **Web Unit Mining**
  - **Existing Methods**
    - **Base line method**
    - **Base line with fragments**
    - **iterative Web Unit Mining (iWUM)**

Problem statement

Given a collection of web pages and a set of concepts, the web unit mining problem is to construct web units from these web pages and assign them the appropriate concepts (category labels).

- Inputs
  - Training: Labeled web units file and *web Knowledge Base* (webKB) data which contains the actual web pages.
  - Testing: University folder which contains the web pages.
- Output

  Properly constructed and classified web(concept) units.

Design and solution methodology

- Design of Concept Unit Mining (CUM)
  - Training phase
  - Web unit construction phase
  - Web unit classification phase

Design (cont'd…)

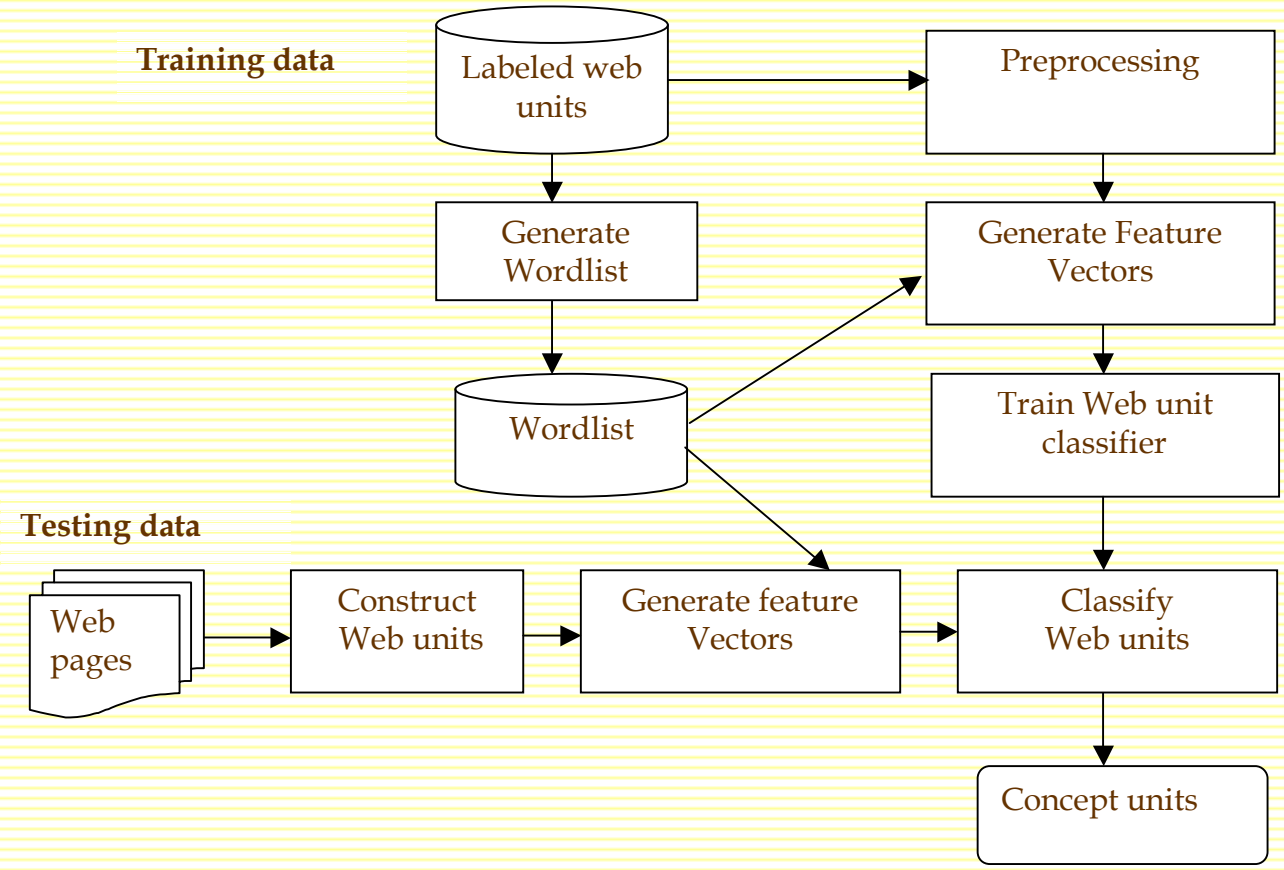

Figure 4: Design of Concept Unit Mining method

Solution methodology

1.Generate word List

- Preprocess the training web pages to get the words
- word document count (d) is calculated
- Inverse document frequency (*idf*) is calculated as follows

Let N is the number of training documents

$$idf_i = \log(\frac{N}{d_i})$$

- List of words and the corresponding inverse document frequency are created

Solution methodology

Cont'd...

2.Preprocessing

- Get the feature content from the web page
- Tokenize the content and transformed to lower case
- Remove stop words
- Apply the stemming
- Calculate the term frequency of each word

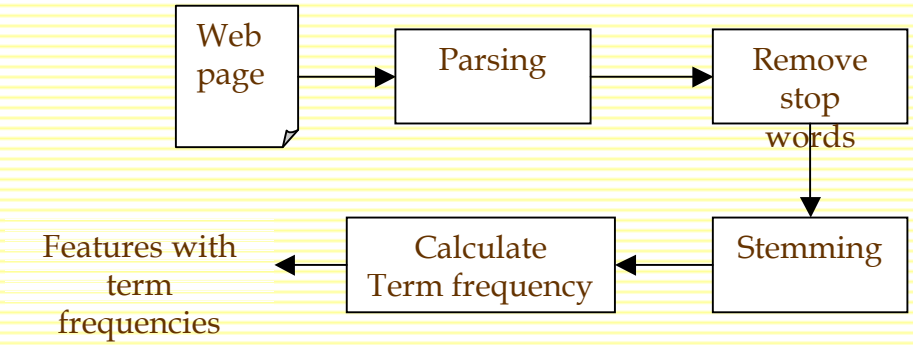

Figure 5: Preprocessing of web page

Solution methodology

Cont'd...

3.Generate feature vectors

- Feature vector is a vector containing feature ids and corresponding weights of a web page.
- Positive training examples for a concept C_j
  - key pages of the web units labeled with concept C_j
- Negative training examples
  - support pages of web units in C_j
  - key pages and support pages of the web units that do not belong to C_j
- The weight of a term is calculated using term frequencies (*tf*) of each term of a page and *idf* values as follows

$$w_i = tf_i * idf_i$$

Solution methodology

Cont'd…

4.Support vector machine

- SVM has been proved to be very effective in dealing with high-dimensional feature spaces.
- Widely used in
  - pattern recognition areas such as face detection
  - isolated handwriting digit recognition
  - gene classification
  - text categorization ,etc …
- The basic idea
  - find an optimal hyperplane to separate two classes with the largest margin from pre-classified data.

Solution methodology

Cont'd...

6.Construction of web units

- Build web directory
- generate web units
  - process the directory from bottom up manner
  - web folder will be examined only after all its child web folders have been examined or it has no child web folder.
  - if the folder is well connected
    - construct a candidate unit with pages in the folder
  - if the folder is not well connected
    - construct a candidate unit with pages in the subfolder

Solution methodology

Cont'd...

7.Classification of web units

- The constructed web units are classified based mainly on the key page.
- For each constructed web unit the feature vector is generated from the key page.
- This vector is given to the SVM classifier which will classify based on the previously constructed models.

Web Unit Evaluation

- *u* is defined as constructed unit , *u'* is defined as the perfect web unit, *u.k* is the key page, *u.s* is the set of support pages of the unit. The following is the contingency table for web unit u_i .

Web unit evaluation		Perfect web unit u_i'		
		$u_i'.k$	$u_i'.s$	NU
Constructed web unit u_i	$u_i.k$	TK_i	SK_i	–
	$u_i.s$	KS_i	TS_i	FS_i
	NU	NK_i	NS_i	–

$$TK_i = \{u_i.k\} \cap \{u_i'.k\} \qquad NK_i = \{u_i'.k\} - \{u_i.k\} - u_i.s$$

$$TS_i = u_i.s \cap u_i'.s \qquad NS_i = u_i'.s - \{u_i.k\} - u_i.s$$

$$FS_i = u_i.s - (u_i'.s \cup \{u_i'.k\})$$

Cont'd…

- Precision and recall values of a web unit u_i are defined as follows.

$$Pr_{u_i} = \frac{\alpha \cdot |TK_i| + (1-\alpha) \cdot |TS_i|}{\alpha + (1-\alpha) \cdot (|KS_i| + |TS_i| + |FS_i|)}$$

$$Re_{u_i} = \frac{\alpha \cdot |TK_i| + (1-\alpha) \cdot |TS_i|}{\alpha + (1-\alpha) \cdot (|SK_i| + |TS_i| + |NS_i|)}$$

- Precision and recall values of the Concept C_j are defined as follows.

$$PrC_j = \frac{\sum_{u_i \in C_j} Pr_{u_i}}{M}$$

$$RecC_j = \frac{\sum_{u_i \in C_j} Re_{u_i}}{N}$$

Experimental Setup and Results

- Experimental set up
  - Standard stop word list is used.
  - WebKB data set is used for experiments.
  - UnitSet data which is labeled used for training.
- UnitSet Web unit distribution

Concept University	student		course		faculty		project	
	u	p	u	p	u	p	u	p
Cornell	128	301	42	219	34	60	20	78
Texas	148	370	38	95	46	104	20	115
Washington	126	495	74	360	31	71	21	129
Wisconsin	156	416	82	413	42	83	25	90

u-web unit , p-web page

Results

Table 1: Web unit mining results (α = 1)

Concept	CUM Pr	CUM Re	CUM F1
student	0.757	0.558	0.642
faculty	0.815	0.689	0.747
course	0.732	0.938	0.822
Project	0.824	0.545	0.656
MacroAve	0.782	0.682	0.729
MicroAve	0.772	0.798	0.785

Cont'd…

Table 2: Web unit mining results (α = 0.5)

Concept	CUM Pr	CUM Re	CUM F1
student	0.746	0.535	0.623
faculty	0.797	0.733	0.764
course	0.747	0.955	0.838
Project	0.806	0.555	0.657
MacroAve	0.774	0.695	0.732
MicroAve	0.767	0.834	0.799

Cont'd…

Table 3: Web unit mining results ($\alpha = 1/|u|$)

Concept	iWUM			CUM		
	Pr	Re	F1	Pr	Re	F1
student	0.902	0.868	0.883	0.743	0.541	0.626
faculty	0.908	0.733	0.802	0.797	0.750	0.773
course	0.772	0.608	0.656	0.748	0.970	0.845
Project	0.332	0.187	0.239	0.806	0.566	0.665
MacroAve	0.729	0.608	0.656	0.774	0.707	0.739
MicroAve	0.868	0.744	0.801	0.766	0.854	0.808

Cont'd…

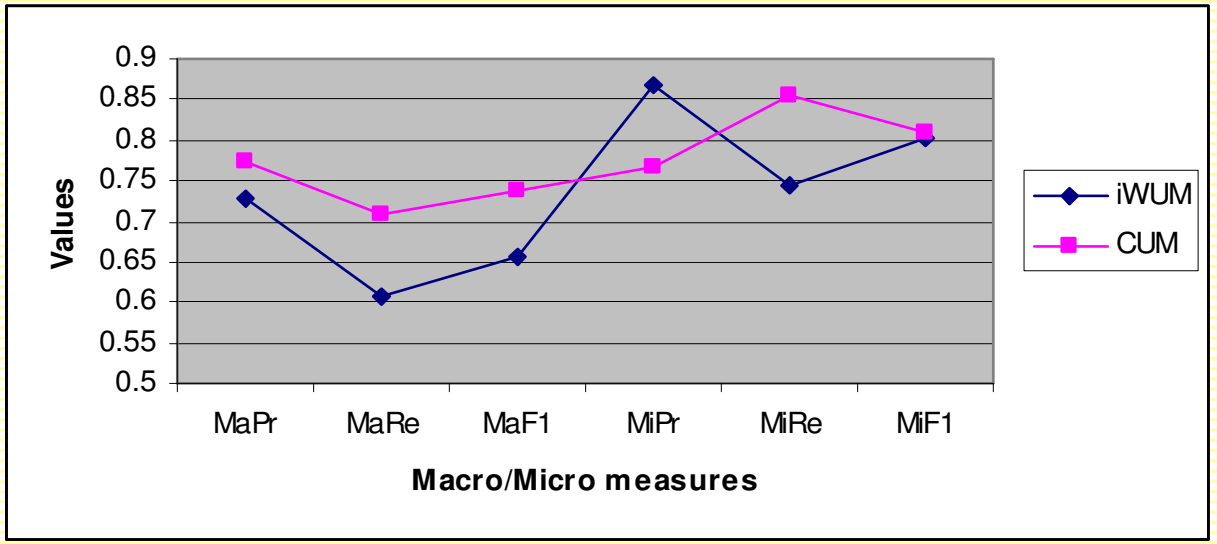

Figure 9: Macro/Micro-averaged results of the two methods

Conclusion and future work

- Day by day the amount of data on the web is increasing. Providing the complete information and reducing the search space is very crucial. Web unit mining is very much useful for this purpose.
- In this work, a new way of mining the concept units is explored and evaluated.
- The proposed approach for Web unit mining shows that a large portion of *incomplete* web units are removed and web unit mining performance is thus improved.
- CUM is not an iterative process thus reduces the time required to mine the concept units.
- Future work can focus on utilizing mined web units to enhance web information retrieval and exploring ways for modeling web units and develop web unit-based ranking strategies.

References

[1] Chen, M.., Han, J., and Yu, P. S. 1996. Data mining: an overview from a database perspective. *IEEE Trans. On Knowledge And Data Engineering 8*, 866(883).

[2] Han, J. and Kamber, M. (2001). *Data Mining Concepts and Techniques*. Series in Datamanagement Systems. Morgan Kaufmann.

[3] Tom M. Mitchell. Machine Learning. The McGraw-Hill Companies, 1997.

[4] Ed Greengrass, Information retrieval: A survey, DOD Technical Report TR-R52- 008-001, 2001.

[5] A.-H. Tan. Text mining: The state of the art and the challenges. In Proc of the Pacic Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, pages 65-70, 1999.

[6] da Costa, M.G., Jr.; Zhiguo Gong,"Web structure mining: an introduction", IEEE International Conference on Information Acquisition, pp. 6, July 2005.

[7] R. Kosala and H. Blockeel, Web mining research: A survey, SIGKDD Explorations, 2(1), pages 1 - 15, 2000.

[8] Jaideep srivastava, prasanna Desikan, vipin kumar,"Web mining accomplishment and future directions", In National science foundation workshop 2001, pp3-6.

[9] A. Sun and E.-P. Lim, "Web unit mining: finding and classifying sub graphs of web pages", in *Proceedings of the twelfth international conference on Information and knowledge management*, 2003.

[10] S. Chakrabarti, B. Dom, and P. Indyk, "*Enhanced hypertext categorization using hyperlinks", in Proceedings of the ACM SIGMOD international conference on Management of Data, 1998, pp.307–318.*

Cont'd…

[11] S. T. Dumais and H. Chen, "Hierarchical classification of Web content", In *Proc. of ACM SIGIR*, pages 256–263, Athens, Greece, 2000.

[12] A. Z. Broder, R. Krauthgamer, and M. Mitzenmacher, "Improved classification via connectivity information", In *Proc. of 11th ACM-SIAM Sym. on Discrete Algo.*, pages 576–585, 2000.

[13] L. Getoor, E. Segal, B. Taskar, and D. Koller. "Probabilistic models of text and link structure for hypertext classification", In *Proc. of Intl Joint Conf. on Artificial Intelligence Workshop on Text Learning: beyond Supervision*, Seattle, WA, 2001.

[14] M. Craven and S. Slattery. *"Relational learning with statistical predicate invention: Better models for hypertext," Journal of Machine Learning, vol. 43, no. 1-2, pp. 97–119, 2001.*

[15] J. Furnkranz. *"Hyperlink ensembles: A case study in hypertext classification", Journal of Information Fusion, vol. 1, pp. 299–312, 2001.*

[16] H.J. Oh, S. H. Myaeng, and M.H. Lee. *"A practical hypertext categorization method using links and incrementally available class information", in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000, pp. 264–271.*

[17] A. Sun, E.P. Lim, and W.K. Ng. *"Web classification using support vector machine," in Proceedings of the 4th international workshop on Web information and data management, 2002, pp. 96–99.*

[18] Y. Yang and X. Liu. *"A re-examination of text categorization methods", in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 42–49.*

[19] L. Terveen, W. Hill, and B. Amento. *"Constructing, organizing, and visualizing collections of topically related web resources", ACM Transactions on Computer-Human Interaction, vol. 6, no. 1, pp. 67–94, 1999.*

[20] M. Ester, H.P. Kriegel, and M. Schubert. *"Web site mining: a new way to spot competitors, customers and suppliers in the World Wide Web", in Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, pp. 249–258.*

Cont'd…

[21] Y. Tian, T. Huang, W. Gao, J. Cheng, and P. Kang. *"Two-phase web site classification based on hidden markov tree models", in proceedings of IEEE/WIC Web Intelligence, October 13-17, 2003.*

[22] N. Eiron and K. S. McCurley. *"Untangling compound documents on the web", In Hypertext, 2003, pp. 85–94.*

[23] M.F Porter .An algorithm for suffix stripping, program, 14(3):130-137, 1980.

[24] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2): 955-974, 1998.

[25] Smola, A. and Schölkopf, B. A tutorial on support vector regression. Technical Report NC2-TR-1998-030 NC2-TR-1998-030, NeuroCOLT 2, 1998.Available from http://www.neurocolt.com/.

[26] Bennett K. and Bredensteiner E.,"Duality and Geometry in SVMs", In P. Langley editor, Proc. Of 17th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, 65-72, 200.

[27] Crisp D. and Burges C.," A geometric interpretation of v-svm classifiers", Advances in Neural Information Processing Systems, 12 ed. S.A. Solla, T.K Leen and K.R. Muller, MIT Press, 2000.

[28] Shuonan Dong,"Support Vector Machines Applied to Handwritten Numerals Recognition.", machine learning, 2005.

[29] A. Sun and E.-P. Lim. Hierarchical text classification and evaluation. In Proc. of the 1st IEEE Int. Conf. on Data Mining, pages 521--528, California, USA, Nov 2001.

[30] WebKb data from http://http://www.cs.cmu.edu//afs/cs.cmu.edu/project/theo-20/www/data/, feb 2007

Thank You