An Assessment on Character-based Chinese News Filtering Using Latent Semantic Indexing

Shih-Hung Wu, Pey-Ching Yang, Von-Wun Soo
Department of Computer Science
National Tsing Hua University
Hsin-Chu 30043 Taiwan R.O.C.

e-mail: *dr828307@cs.nthu.edu.tw, mr854359@cs.nthu.edu.tw, soo@cs.nthu.edu.tw*

Abstract

In this paper, we assessed the Latent Semantic Indexing (LSI) approach for Chinese information filtering. The assessment was for Chinese news filtering agents that used a character-based and hierarchical filtering scheme. The traditional vector space model was employed as information filtering model, and each document was converted into a vector of weights of terms. Instead of using words as terms in IR denominating tradition, the terms were referred to Chinese characters. LSI captured the semantic relationship between the documents and Chinese characters. We used the Singular-value Decomposition(SVD) technique to compress the terms space into a lower dimension which achieves latent association between document and terms. We showed by experiments that the recall and precision results of Chinese news filtering by character-based approach incorporating the LSI technique into the information filtering system were satisfactory.

1. Introduction

The rapid growth of Internet precipitates the need of the Network Information Retrieval System. Most of the famous systems that assist people in locating information on the Internet such as Lycos, Infoseek, Alta Vista, WebWatcher[Armstrong 95] are designed for English in-

formation retrieval. To our knowledge, only the Csmart[Chien96] and GAIS [http://gais.cs.ccu.edu.tw/] systems are designed for Chinese information retrieval. However, information filtering is conceptually slightly different from information retrieval, we have to modify the techniques of information retrieval into information filtering. In this paper, we assessed the LSI technique for a hierarchical Chinese information filtering scheme. In particular we assess the SVD approach for Chinese news filtering, which to our knowledge has never been investigated for Chinese language.

Usenet news is one of the rich information resources on Internet, filtering out useful news among thousands of available news is a crucial problem [Lang95]. Imagine a client user who needs a software agent to automatically recommend interesting news in Chinese from the Internet. Since the news is updated every day, the traditional technique for information retrieval to retrieve news with a fixed set of database would not work. Also the task that a news filtering agent faces is to select relevant news according to the user's interest or preference from a huge amount of dynamically growing news. Belkin and Croft [Belkin 92] pointed out that one major difference between information retrieval and filtering is that: The queries in information retrieval typically represent user's short-term interests, while the user profiles in information filtering tend to represent user's long-term interests. To model user's long term interest, a user profile plays an important role in information retrieval [Mayeng 90] and filtering. Profiles can be represented in many ways and at different psychological and abstraction levels. A collection of documents in a user's personal digital library may approximate the user profile. The information filtering is a document-find-document style of information retrieval. A document that is similar to the documents in the user's personal digital library is regarded as relevant.

We adopted the vector space model [Yan 94] in our design of Chinese news filtering agents. In this model, each document is represented as a vector of weights of terms. We form each user profile by merging document vectors of the same interest category. The similarity of the incoming document vectors with the profile vectors can be computed by the cosine angles between the two vectors to determine if a document is to be filtered out.

In Chinese, there is no word delimiters to indicate the word boundaries, therefore word segmentation is a difficult task to deal with. Many proper nouns or unknown words could not be found in a word dictionary with a large vocabulary [Chien 95]. The size of Chinese character vocabulary is about 13000 among which about 5,000 characters are the most commonly used characters. However, the number of Chinese words in a document collection set can be easily up to 1,000,000. To represent the personal profile in terms of words will face the difficulty of word segmentation in Chinese [Chien 96]. We will show by experiments that without word segmentation, character-based filtering incorporated LSI can be a satisfactory information filtering method.

Filtering method incorporated LSI will possibly select relevant documents whose contents have no exactly matched keywords. This is quite different from traditional technique such as the Boolean models. The probabilistic model, Bayesian belief network Model [Turtle 91] [Ribeiro 96] shared the similar feature. The Boolean models exactly matched document's terms with the combination of the search terms specified in the query. The probabilistic models estimated the degree of relevance between documents and user query by considering the appearance frequency of certain terms in the document and the user query, together with the information about term distribution in the document collection.

Since individual terms and keywords are not adequate discriminators of the semantic content of the documents and queries, the performance of the conventional retrieval models often suffers from either missing relevant documents which are not indexed by the keywords specified in the query, or retrieving irrelevant documents which are indexed by unintended sense of the keywords in the query. Therefore, there has been great interest in text retrieval research that is based on semantics matching instead of strictly keyword matching.

Latent Semantic Indexing(LSI) using Singular-value Decomposition (SVD) is an approach to overcoming this deficiency of exact keyword matching techniques. We use truncated SVD to capture the semantic structure of word usage among certain documents, and hope this relation can be applied to other documents. Using the singular values matrix from the truncated SVD, a high-dimensional vector space representing term-document matrix is mapped to a lower dimension matrix that reflects the major concept factors in the certain

documents, while ignoring the less important factors. Terms occur in similar documents will be near in the reduced vector space. Documents may satisfy a user's query when they share terms that are closer in the reduced space. Since the reduced vector spaces are more robust indicators of the semantic meaning than individual words, the performance may be better than that of the original space.

Several papers report the use of the LSI method. Conference uses the LSI method to assign submitted manuscripts to the reviewers of the *Hypertex'91* conference based on the interests of each reviewer, a set of relevant manuscripts was sent to the reviewer[Dumais 92]. The automated assignment method achieved better matching between the reviewers and their interests than the assignment by the human experts. [Syu96] presented the technique of incorporating Latent Semantic Indexing into a neural network model for text retrieval. The performance, in terms of precision and recall, was comparable to text retrieval models.

The remainder of this paper is as follows. Section 2 provides an overview of the Latent Semantic Indexing method as applied to information retrieval, and how to use truncated SVD as a LSI approach. Section 3 briefly reviews our information filtering scheme. Section 4 reports the experimental results comparing the LSI-based model and Section 5 is the discussion and conclusion.

2. Latent Semantic Indexing method applied to information retrieval

Latent Semantic Indexing(LSI) is an extension of the vector space retrieval method. We assumed that there is some unknown "Latent" association in the pattern of terms or keywords used among documents [Dumais 92], and tried to estimate this latent association. Singular-Value Decomposition (SVD) is a technique about eigenvector decomposition and factor analysis used in statistics[Cullum 85], and Latent Semantic Indexing(LSI) using SVD is one approach to modeling the latent semantic relationships between the documents and the index terms. This approach performs singular-value decomposition on a term-by-document matrix, generating a reduced space with lower dimension. The similarity between two documents is calculated according to the index terms used in each of the documents occur in other documents. Using the LSI representation, documents satisfy a user query when they share terms of

similar semantic meaning in the reduced vector space. The dimension of the resulting vector space is much smaller than the number of exact index terms used in a document collection(e.g. from several thousands to 100 or 300[Dumais 94]), a filtering model using LSI can benefit from requiring less time and memory.

2.1 Singular-Value Decomposition(SVD) and truncated SVD

SVD is a reliable tool for matrix factorization. For any matrix A, A^TA has nonnegative eigenvalues. The nonnegative square roots of the eigenvalues of A^TA are called the singular values of A, and the number of the non-zero singular values are equal to the rank of A, *rank(A)*. Assume that A is an *m by n* matrix and *rank(A)* = r, the singular -value decomposition of A is defined as

$$A = U W V^T,$$

where the size of U is *m by m*, the size of V is *n by n* and the size of W is *m by n*. Both the U and V^T, are orthogonal matrices, i.e., $UU^T = I_m$, and $VV^T = I_n$; W is a diagonal matrix consists of the singular values of A: σ_1, σ_2, ..., σ_r. And the σ_j's are the singular values of A, $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r \geq 0$, and $\sigma_j = 0$ for $j \geq r+1$.

To apply SVD as a LSI tool, term-by-document matrix A must be constructed. Using SVD to generating optimal approximation of the document representation specified by the matrix A. Since the singular values in matrix W are ordered from largest to smallest, the first k largest may be kept and the remaining smaller ones are set to zero. As a result, the representations of the matrices U, V, and W can be reduced by reform a new diagonal matrix W_k by removing column and rows which are zeros from W; reform a matrix U_k by removing the *(k+1)*st to the mth columns from U; and reform a matrix V_k by removing the *(k+1)*st to the nth rows from V. The product of the resulting matrices is a matrix A_k which is an approximation of the matrix A, and *rank*(A_k)=k.

$$A_k = U_k W_k V_k^T,$$

The LSI method using SVD can be viewed as a technique for deriving a set of non-correlated indexing of factors(i.e. the singular values), which represent different concepts in

the usage of words in the documents collection. The documents and queries are then represented by vectors of factor values, instead of the individual index terms. Using the *k*-largest factors may captures the most important latent semantic relation between documents and index terms, and avoids unintended sense in word usage.

2.2 Document and query representations

Since the term-by-document matrix *A* has been reduce to a lower dimension matrix, the vector which represent the query and all the new-add document must be mapped to the same lower dimension. Using the singular-value decomposition, a term-by-document matrix *A* is mapped into a reduced *k by n* matrix represented by $W_k V_k^T$, which relates *k* factors to n documents. A query *q*, originally of dimension size *m*, can be mapped into a size *k* vector *q'*

$$q' = (q^T U_k W_k^{-1})^T,$$

The similarity between two documents then is computed using this shorter vector representation.

3. The character-based Chinese news filtering Scheme

3.1 The character-based vector representation of documents and personal profiles

A Chinese character is the basic processing unit and is used equivalently as the concept of a "term" in IR denominating tradition, we use terms to refer to Chinese characters in the context of the paper. In our approach, no stemming and stop word lists or a thesaurus is used. We represent the weight of a term in a given document by adopting Salton's well-tested *TFIDF* formula in IR, the term frequency (tf) multiplied by the inverse document frequency (idf) [Salton 89] [Salton 91]. Namely, the weight of a term t in a given document d, namely w(t,d), is represented as

$$w(t, d) = tf_{t,d} * \log(N/df_t)$$

where documents number N is the total number in a collection of documents, term frequency $tf_{t,d}$ is the number of appearance of term t in document d, and the document frequency df_t is the number of documents which content term t in the collection.

A document D can be represented as a vector V with elements $v_1, v_2, \ldots, v_n$, where n is equal to the size of character vocabulary, and vi is the weight of term i in the document. All vector are normalized, for convenience and by convention. We can calculate the similarity between two documents Di and Dj by the cosine of the angle between their vector representations:

$$\text{Similarity}(Di, Dj) = \frac{Vi * Vj}{\|Vi\| \|Vj\|}$$

where V_i is the vector representation of Di, * represents the inner product between two vectors; $\|Vi\|$ represents the norm of a vector Vi. Based on the formula, two documents with same character set will have the highest similarity between them because the inner product of the two document vectors would be one, while two documents without any character in common will have the lowest similarity zero.

We merge the document vectors in the same interest group (either grouped by the user or by a classification/clustering agent) into a higher level profile vector by their vector sum. The profile vector is also normalized.

3.2 The tasks of a news filtering agent

A Chinese character-based news filtering agent will carry with it a set of weights in terms of inverse document frequencies as discussed in section 2.2.1 for a vocabulary of terms (characters), a profile vector that represents a certain interest category and a similarity threshold associated with the profile vector. For each news document in the news server, the filtering agent will convert it first to a document vector and then the similarity between the document vector and the profile vector is computed according to the method discussed in section 2.1. If the similarity of the document with the profile is lower than the threshold, it is filtered out.

3.3 The hierarchical information filtering scheme

The hierarchical information filtering scheme reduces agent's total task. By composition of profile vectors, we reduce the number of vectors that each agent must compare with document

vectors on the web. All the lower-level profile vectors are combined to form higher-level profile vectors. We may assume that the final highest level profile vector can represent an overall interest of the user. The intelligent news filtering agent can then carry this profile vector in search for relevant documents on the web.

4. Experimentation

4.1 Data collection and document vectorization

We gathered three sets of articles from on-line China Times [http://www.chinatimes.com.tw/] for 2 consecutive weeks from Mar.2nd 1997 to Mar.15th 1997. There were 671 articles in the first week, 669 in the second week. These articles were written by professional reporters and we collected all the articles from all the nine categories that China Times provided. The categories are: *Entertainment, Sports, Economy, Focus, International, Mainland, Social, Taiwan and Editorial.*

Table 1. The number of documents in the document collection sets.

Category \ Set	1st week	nd week
Economy	86	95
Editorial	14	14
Entertainment	80	82
Focus	80	79
International	70	63
Mainland	63	58
Social	67	73
Sports	90	87
Taiwan	114	111
Total	671	669

Table 1. shows the number of documents in each of the categories. The length of each article is about 500-2000 Chinese characters. In order to test the usage of words in the document collection sets is stable or not, we use the 671 articles in the first week as the training set to compute the document frequency df_t for each term t and the 669 articles in the second week as the testing set for the filtering experiment.

The articles were first transformed into normalized document vectors as discussed in section 2.1, all English characters and Arabic numerals were ignored. The similarity between two documents is then equal to the inner product of two vectors. To mimic a user's interests, we choose news articles from three categories (*Entertainment, Sports, Economy*) on the same day (Mar. 2^{nd}) to form the initial user's profile. The user profile is treated as a set of documents and are transformed into normalized document vectors. In composition of the user profile vector we treated the importance of all articles equally.

4.2 Experiments on information filtering with SVD

To evaluate the effectiveness of news filtering based on character-based method for Chinese news document, the tf-idf weighting and vector space model were adopted in our experiment on the nine news categories, and we tried different k using truncated SVD. As discussed above, several arbitrary articles of each category in the training set were selected and merged into one *query* document. The query document was then transformed into a normalized vector named as a query vector or a profile vector. By comparing the query (profile) and document vectors in the test set, we retrieved the most similar documents in the test set and measured the precision against different recall values as plotted in **Figure 1-3**.

From **Figure 1-3**, we observed that different k number in SVD had different performance. The performance was worse either when the k number was small, e.g., 2, or when the k number was large, e.g., 100. The experiments show the suitable value for k is 10 for our document collection sets. [Dumais 94] suggested that the probable k number is from 100 to 300 in English document. Our experiments disagree with this. The great reduction of vector dimension will save a lot of memory space and time consuming on further utilization of document vectors. Before this, we perform more experiments to justify our observation.

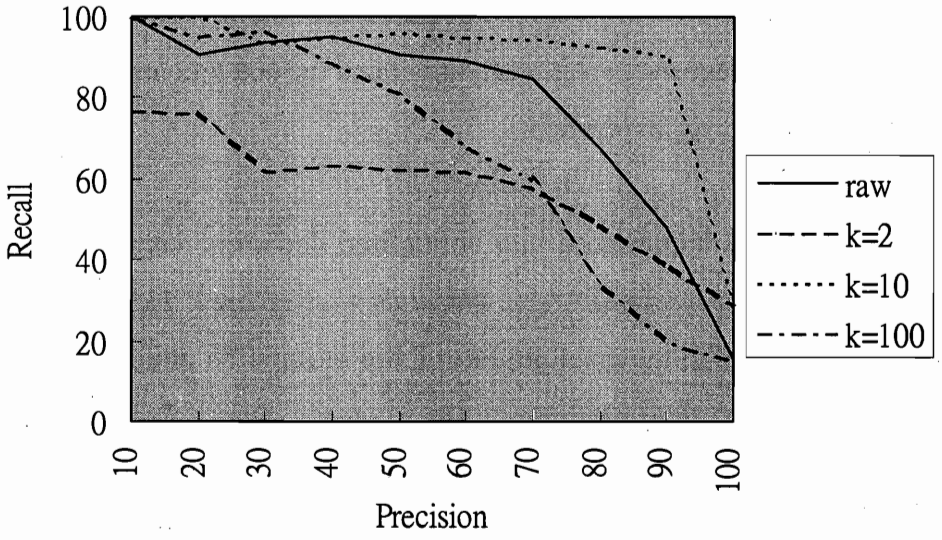

Figure 1. Recall-precision curve. Four different processing methods (raw vector form and SVD with k=2, 10, 100) on economy category.

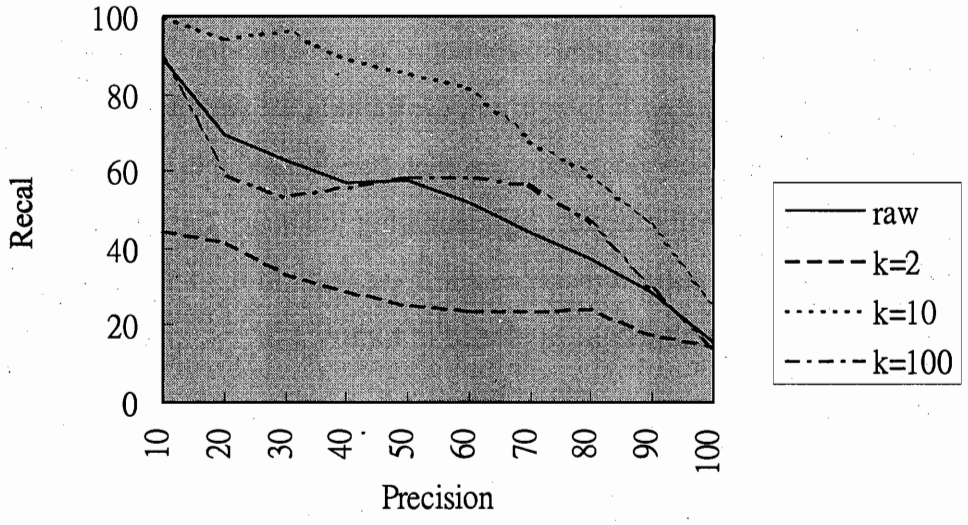

Figure 2. Recall-precision curve. Four different processing methods (raw vector form and SVD with k=2, 10, 100) on entertainment category.

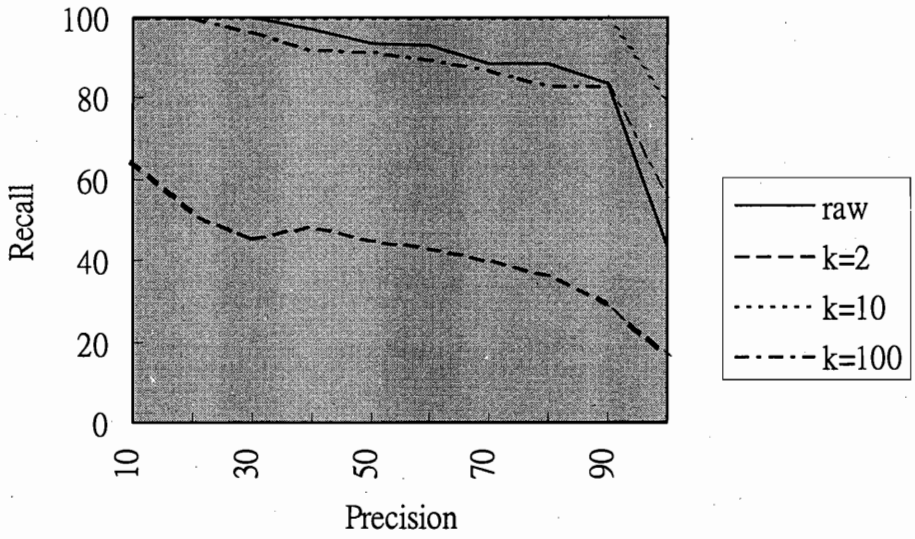

Figure 3. Recall-precision curve. Four different processing methods (raw vector form and SVD with k=2, 10, 100) on sports category.

4.3 Information filtering experiments based on different k values

To justify our observation, we tried more different k value, and calculate 11-point average precision against different k values as plotted in **Figure 4.** As in experiment 1, several arbitrary articles of each of the three categories (Sports, Economy, Entertain) in the training set were selected and merged into one *query* document. The query document was then transformed into a normalized vector named as a query vector or a profile vector. By comparing the query (profile) and document vectors in the test set, we retrieved the most similar documents in the test set, and measured performance by the 11-point average precision (average over different recall values from 0% to 100%, 10% each step).

From Figure 4, we observed that the performance reaches its maximum when k is about10 for each of the three testing profiles. The experimental result is consistent with the result in experiment 1, but quite differently from what [Dumais 94] suggested. We conjecture that the probable *k* number is different for Chinese and English and for different document

collection sets. To prove the conjecture, more experiments are needed.

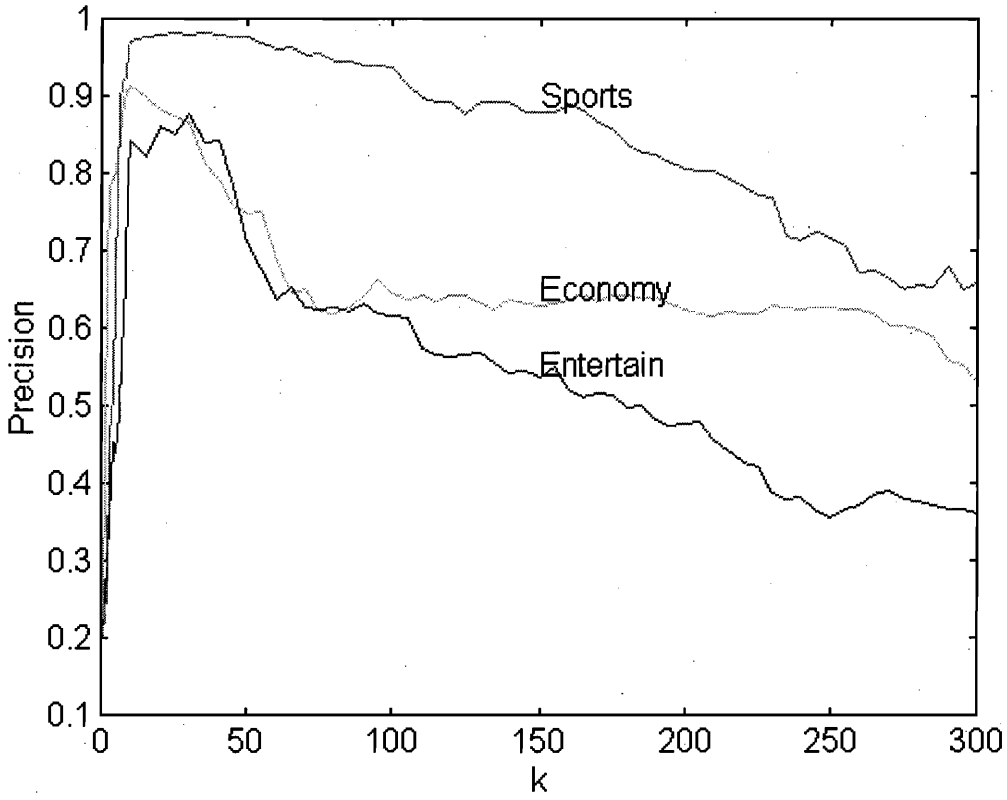

Figure 4. Performance(11-point average precision) varies against different *k* values.

5. Discussions and Conclusion

In the experimental results, we found that the recall-and-precision are surprisingly satisfactory in the character-based document-find-document style of information filtering for Chinese news filtering. The SVD technique can be used to reduce the need of storage space of the term-by-document matrix and the processing time on further utilization. The difference on performance for different choice of k when using truncated SVD value is quit interesting.

Without the word segmentation, neither stemming and stop word lists nor a thesaurus is used, the well performance of Chinese character-based information filtering is an interesting

finding. This finding experience has been also shared by Dr. Lee-Feng Chien in personal communication. This finding suggested that the semantic meaning of a Chinese news article can be implied by the character set. Articles with similar character sets tend to have similar meaning. Even though in Chinese different orders of the same set of characters may have different meaning, and the same word may have ambiguities in part of speech, character-based filtering seem to provide more information.

Character-based information filtering scheme makes a lot of sense in the sense that no dictionary of a large size of words is available and the word segmentation task in Chinese is difficult. Only about the weights and counts of most commonly used characters in the documents collection set are needed to design an intelligent news filtering agents. A truncated SVD approach with yield better performance and save more computation time for the filtering agents. The effect of SVD method is: reduce the size of the term-by-document matrix, and sort the significance of dimensions for the matrix. This should be the reason why a choice of a suitable k will give a better performance. The first k dimensions are necessary and sufficient for discriminating the categories. If we view stop words as noise, the larger the k value, the more the noise will be considered. On the other hand, the smal k may be insufficient for the discrimination among the categories.

To represent user profile and perform news filtering hierarchically not only has the merit of saving computation cost but also has the potential to perform the information filtering task in distributed and parallel manner. The efficiency will be promoted even further if each profile vector runs independently on a distributed system. This could be achieved because of the independence property among profile and document vectors, i.e., they don't interfere each other while executing similarity calculations.

In the future, the relevance feedback from the user can be used to improve performance by adjusting several system parameters. It can be used to adjust the thresholds at each stage or to adjust the weights of combining lower level profile vectors into higher level ones. We are looking into such machine learning techniques as neural networks [Pannu 95] [Syu 96] along this direction.

Acknowledgment This work is financially supported by Institute for Information Industry and National Science Council of Taiwan, Republic of China under the grant No. NSC86-2213-E-007-53.

References

Armstrong, R. and D. Freitag, T. Joachims, and T. Mitchell, "WebWatcher: A learning apprentice for the world wide web," 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, Stanford, March 1995.

Belkin, N.J. and Croft, W.B., "Information filtering and information retrieval: two sides of the same coin?," Comm. ACM 35, 12(Dec.), pp. 29-38.

Chien, L.F., "Fast and quasi-natural language search for gigabytes of Chinese texts," ACM SIGIR 95, 1995.

Chien, L.F., "An intelligent Chinese information retrieval system for the Internet," Proceedings of the ROCLING IX, 1996.

Cullum, J.K. and R.A. Willoughby, "Lanczos Algorithms for Large Symmetric Eigen value Computations - Vol. 1, Theory(Ch 5: Real Rectangular Matrices)," Brikhauser, Boston, 1985.

Dumais, S.T. and J. Nielsen, "Automating the Assignment of Submitted Manuscripts to Reviewers," Proc. Of the 15th International Conference on Research and Development in Information Retrieval, pp. 233-244, 1992.

Dumais, S.T., "Latent Semantic Indexing and TREC-2," The Second Text Retrieval Conference(TREC-2), NIST Special Publication 500-215, pp. 105-115, 1994.

Lang, K., "Newsweeder: learning to filter Netnews," Proceedings of the Twelfth International Conference on Machine Learning, 1995.

Mayeng, S. H. and R. R. Korfhage, "Integration of user profiles: models and experiments in information retrieval. Information Processing and Management," Vol. 26, No. 6, 1990.

Ribeiro, B.A.N. and R. Muntz, "A Belief Network Model for IR," In Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.253-269, 1996.

Salton, G., "Automatic Text Processing," Addison Wesley, Reading, Massachusetts, 1989.

Salton, G., "Developments in automatic text retrieval," Science 253, 1991.

Pannu, A. S. and K. Sycara, "A learning personal agent for text filtering and notification," Proceedings of the International Conference of Knowledge Based Systems (KBCS 96), Dec. 1996.

Syu, I., S. D. Lang, and N. Deo, "Incorporating latent semantic indexing into a neural network model for information retrieval," Proceedings of the Fifth International Conference on Information and Knowledge Management, Nov. 1996.

Turtle, H. and W. B. Croft, "Evaluation of an Inference Network based Retrieval Model," ACM Transactions on Information Systems, Vol. 9, No. 3, July 1991.

Yan, T. W. and H. Garcia-Molina, "Index structures for information filtering under the vector space model," Technical Report STAN CS-TR-93-1494, Nov. 1993.