Journal of Technology Management & Innovation

Received February 14, 2007/ Accepted March 5, 2007

PROBABILISTIC LATENT SEMANTIC ANALYSES (PLSA) IN BIBLIOMETRIC ANALYSIS FOR TECHNOLOGY FORECASTING

Zan.Wang[a,b], Y.C. TSIM[a], W.S. Yeung[a], K.C. Chan[a,*], Jinlan Liu[b]
[a]Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, China
[b] School of Management Tianjin University, Tianjin, PR China. 30072

Abstract

Due to the availability of internet-based abstract services and patent databases, bibliometric analysis has become one of key technology forecasting approaches. Recently, latent semantic analysis (LSA) has been applied to improve the accuracy in document clustering. In this paper, a new LSA method, probabilistic latent semantic analysis (PLSA) which uses probabilistic methods and algebra to search latent space in the corpus is further applied in document clustering. The results show that PLSA is more accurate than LSA and the improved iteration method proposed by authors can simplify the computing process and improve the computing efficiency. Probabilistic Latent Semantic Analyses (PLSA) in Bibliometric Analysis for Technology Forecasting

Keywords: Bibliometric analysis, latent semantic analysis, technology forecasting, probabilistic latent semantic analysis.

1 Introduction

Technological forecasting is the process of predicting the future characteristics and timing of technology. However there are many factors including scientific developments, government policy, organization strategy, organization structure, business opportunity, market needs, and funding availability, which affect the progress and direction of technology, making the forecasting work very difficult. With the advent of information technology, bibliometric analysis which means the analysis of large numbers of scientific documents including research articles and patents, for the purpose of technology forecasting, has attracted much attention (Kostoff, 1997). Bibliometric analysis is a process of summarizing document characteristics into tables for statistical analysis. Swanson and Smalheiser (1997) use an extension of such bibliometric techniques to find indirect links among concepts in documents. Bibliometric techniques, supplemented by visual mapping, have also been applied as an assessment and planning tool

[*] Corresponding author. Dept. of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong. Tel. no.:852 27664981; Fax no.: 85223625267; E-mail address: mfkcchan@inet.polyu.edu.hk

The work described in this paper was supported by the Research Committee of The Hong Kong Polytechnic University (Project no. A-PF99).

to sets of patents and this has yielded good results for competitive intelligence applications (Mogee, 1997).

Bibliometric analysis, when considered as a data mining problem has to address two issues: 1) inter-document similarity, and 2) classification. Morris et al (2002) have recently applied the “similarity functions” method to mathematically express the strength of similarities among documents, together with a classification technique which relies on document mapping, visualization and interactive exploration for their analysis. This technique has been shown to be effective in allowing users to visualize relations among collections of documents from a particular technological field. But users still have to make inferences and draw conclusions about the relations and trends within the technology.

The following flow chart (Figure 1) shows the process of bibliometric analysis

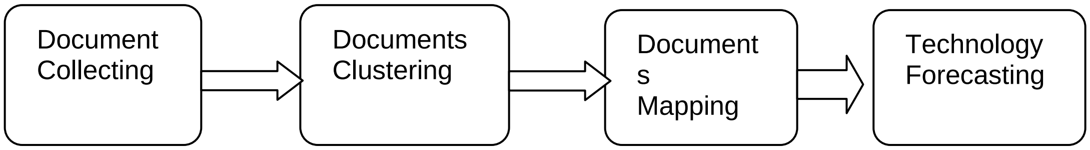


Figure 1 Flow of bibliometric analysis

One of the important phases in bibliometric analysis is to classify documents by finding their links. Since the traditional approach of word-matching can only find the direct link between documents that share the occurrence of same words, latent semantic analysis (LSA) (Dumais, 1995) has been used to find indirect links between documents and to solve the problems that the traditional approach cannot handle such as synonymy and polysemy. Singular Value Decomposition (SVD) as a LSA approach has been applied in technology mining (Porter, 2005) ; the steps are shown in Figure 2.

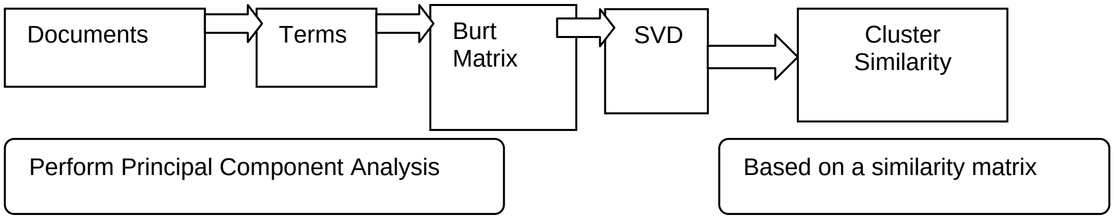

Documents
Terms
Burt Matrix
SVD
Cluster Similarity
Perform Principal Component Analysis
Based on a similarity matrix

Figure 2 Flow of documents clustering

As a new enhancement of LSA, PLSA (probabilistic latent semantic analysis) (Hofmann, 1999) is more efficient than SVD in information retrieval. In this paper, PLSA is further applied to document clustering for technology forecasting.

2 Document clustering and latent semantic analysis

2.1 Document clustering and VSM

2.1.1 Document clustering

Document clustering is one of the most popular computer science techniques used to classify documents and information retrieval (Salton & McGill, 1983; Salton, 1989). It computes the similarity between every pair of documents, and based on this information, it attempts to divide a collection of documents into groups (clusters), such that documents in the same cluster are similar, and documents in different clusters are dissimilar. There are some other methods that have made good documents clustering such as “Bipartite Spectral Graph Partitioning” (Dhillon, 2001)

2.1.2 Vector Space Model (VSM)

Before computing the similarity between documents, a VSM (vector space model) (Salton & McGill, 1983) is used to represent the documents. By extracting features, a document is represented by a term vector, where the magnitude of the vector is the frequency of the term occurring in the document. A text corpus can then be represented by a matrix of term frequency. Porter (1980) has used a stemming algorithm to convert the terms into their origin patterns. As in English, there is a pluralism pattern for nouns, and "-ed" or "-ing" pattern for verbs. If they are not converted, various tenses or forms of the same word will be regarded as different meanings in the documents and this will cause problems. The stemming algorithm converts various forms into the original word by combining them into an item. A similarity matrix can then be computed by the cosine distance of the item in the term frequency matrix. Based on the similarity between documents, the corpus can be divided into clusters.

2.2 Latent Semantic Analysis

2.2.1 Introduction to LSA

The similarity between documents is computed by the frequency with which the term occurs. If the same terms occur many times in two documents, this pair of documents can be regarded as similar. However, one meaning can be expressed by different terms and the same term can represent different meanings. Thus, the similarity of documents which have polysemies or synonymies cannot be efficiently measured by word matching. This is a common problem named polysemy and synonymy in text mining and information retrieval. Latent semantic analysis tries to use statistical analysis and linear algebra method to find a latent semantic space in the corpus. Besides, it can re-represent both documents and terms in a new vector space with smaller number of dimensions, because large number of dimensions can provide many works that make for poor cluster performance (Dumais, 1995).

2.2.2 Singular Value Decomposition (SVD)

It is known that Singular Value Decomposition (SVD) is a reliable tool available for matrix decomposition. It can decompose a matrix as the product of three matrices $X = U\Sigma V^T$ (V^T denotes the transpose of V), where U and V are both column orthonormal. That is, $U^T U = I, VV^T = I$ and Σ is a diagonal matrix of singular values, where those singular values are in descending order as shown in Fig. 3 (where t denotes the number of rows, d denotes the number of columns and m denotes the rank of the matrix).

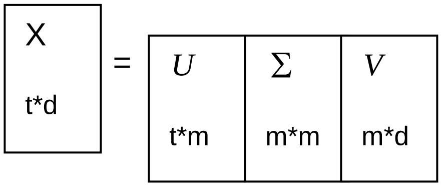


Figure 3 Matrix Decomposition of SVD

After decomposing the matrix, LSA can derive the first k singular values as the principal components, and the right singular values will be regarded as noise and can be deleted. Thus, the rank can be reduced from m to k, and U_k is assumed to be the $t \times k$ matrix by removing right $m-k$ columns, $\sum_k$ be the $k \times k$ matrix by removing the $m-k$ singular values, and V_k be the $d \times k$ matrix by removing $m-k$ columns. So, X can be represented approximately by the following equation:

$$X = U\Sigma V^T \approx U_k \Sigma_k V_k{}^T = X^{'} . \qquad (1)$$

$X^{'}$ minimizes $\left\| X^{'} - X \right\|_F$ over all rank-$kX^{'}$,where $\left\| X^{'} - X \right\|_F$ denotes the Frobenius norm (F-norm).

The LSA method using SVD can be viewed as a technique for the statistical method of "Principal Component Analysis" (PCA) (Kruskal, 1978) used in text mining. PCA gets the most important "components" in a given set of data and replaces the original variants with much fewer principal components to make the problem much easier. In SVD, the first k largest singular value can be viewed as the k most important components of the corpus. It is known that an origin term document matrix is a high-dimensional sparse matrix. After using SVD, the high-dimensional matrix can be replaced by a lower-dimensional one, and the new matrix will typically not be sparse (Deerwester et al, 1990).

This implies that it is possible to compute meaningful association values between pairs of documents, even if the documents do not have any terms in common. The terms having a common meaning, in particular synonyms, can then be roughly mapped to the same direction in the latent space (Brian et al, 1992; Furnas et al, 1988; Ding, 1999).

3 Using probabilistic LSA for documents clustering

3.1 Introduction to PLSA

The goal of LSA is to find latent semantic space in a corpus. And the latent semantic space is made up of some latent topics in the corpus. The different attention on different topics can be regarded as the prior probability as to which different topics will occur. Based on LSA and some probabilistic knowledge such as Bayes rules, the PLSA (Porter, 2005; Story, 1996; Ando, 2000;. Papadimitriou, 1998) method is introduced to find the latent topics and the association of documents and topics, and terms and topics.

If $D=\{d_1,d_2,......d_n\},W=\{w_1,w_2,......w_m\}$ are assigned as the documents vector and terms vector respectively there must be a latent variable model for co-occurrence data which associates an unobserved class variable $z\in Z=\{z_1,z_2,......z_k\}$ with each observation. A joint probability model over $D\times W$ is defined by:

$$P(d,w)=P(d)P(w\,|\,d),\quad P(w\,|\,d)=\sum_{z\in Z}P(w\,|\,z)P(z\,|\,d) \qquad (2)$$

Based on equation (2), $P(d,w)$ can be represented by the following equation:

$$P(d,w)=\sum_{z\in Z}P(z)P(d\,|\,z)P(w\,|\,z) \qquad (3)$$

Thus, the joint probability of the $D\times W$ model can be obtained. In this model, $P(z),P(d\,|\,z),P(w\,|\,z)$ are parameters. Before clustering documents, the value of parameters must be calculated.

For the corpus, the joint probability of sample S is

$$P(S)=\prod_{w\in W}\ \prod_{d\in D}P(w,d)^{n(w,d)} \qquad (4)$$

where $n(w,d)$ is the frequency of the co-occurrence.

3.2 EM Algorithm

In order to determine P(S), equation (3) is converted to "log" scale:

$$\log P(S)=\sum_{d\in D}\sum_{w\in W}n(w,d)\log P(d,w) =>$$

$$L(\theta)=\sum_{d\in D}\sum_{w\in W}n(w,d)\log[P(z)P(d\,|\,z)P(w\,|\,z)] \qquad (5)$$

To get the maximum likelihood estimation (MLE) of the parameters, the EM algorithm (Dempster et al, 1977) is adopted and the following steps are used. **E-step:**

Initialize the values of the parameters, and then compute the expectation of $L(\theta)$,

$$Q(\theta) = \sum L(\theta) \times P(z \mid d, w) \qquad (6)$$

M-step:

Maximize the function in the E-step.

In the E-step, we need to calculate the posterior probability $P(z \mid d, w)$, based on Bayes rules. The equation below can be obtained:

$$P(z \mid d, w) = \frac{P(z, d, w)}{P(d, w)} = \frac{P(z)P(d \mid z)P(w \mid z)}{\sum_{z^{'} \in Z} P(z^{'})P(d \mid z^{'})P(w \mid z^{'})} \qquad (7)$$

where $d \in D, w \in W, z \in Z$.

In the M-step, the Lagrange multipliers $\lambda_1, \lambda_2, \lambda_3$ are introduced in order to maximize the function $Q(\theta)$ with the following constraints:

$$\sum_{z \in Z} P(z) = 1 \quad \sum_{d \in D} P(d \mid z) = 1 \quad \sum_{w \in W} P(w \mid z) = 1$$

The target function is:

$$\begin{aligned} QQ &= \sum_{z \in Z} (\sum_{d \in D} \sum_{w \in W} n(w, d) \log[P(z)P(d \mid z)P(w \mid z)] * P(z \mid d, w) \\ &- \lambda_1 (\sum_{z \in Z} P(\mathrm{z}) - 1) - \lambda_2 (\sum_{d \in D} P(\mathrm{d}) - 1) - \lambda_3 (\sum_{w \in W} P(w) - 1) \end{aligned} \qquad (8)$$

The derivatives of QQ with θ are then given by

$$\frac{\partial QQ}{\partial P(z)} = \frac{\sum_{d \in D} \sum_{w \in W} n(w, d) * P(z \mid d, w)}{P(z)} - \lambda_1 \qquad (9)$$

$$\frac{\partial QQ}{\partial P(d \mid z)} = \frac{\sum_{d \in D} \sum_{w \in W} n(w, d) * P(z \mid d, w)}{P(d \mid z)} - \lambda_2 \qquad (10)$$

$$\frac{\partial QQ}{\partial P(w \mid z)} = \frac{\sum_{d \in D} \sum_{w \in W} n(w, d) * P(z \mid d, w)}{P(w \mid z)} - \lambda_3 \qquad (11)$$

$$\frac{\partial QQ}{\partial \lambda_1} = \sum_{z \in Z} P(z) - 1 \; , \qquad (12)$$

(and the computing process is same for λ_1, λ_2)

Let $\frac{\partial QQ}{\partial \lambda_1} = 0$ and $\frac{\partial QQ}{\partial P(z)} = 0$,

The following equations can be obtained

$$P(z) = \frac{\sum_{d \in D}\sum_{w \in W} n(w,d) \times P(z \mid d,w)}{\sum_{z' \in Z}\sum_{d \in D}\sum_{w \in W} n(w,d) \times P(z' \mid d,w)} \tag{13}$$

and

$$\sum_{z' \in Z}\sum_{d \in D}\sum_{w \in W} n(w,d) \times P(z' \mid d,w) = \sum_{d \in D}\sum_{w \in W} n(w,d) \times \sum_{z' \in Z} P(z' \mid d,w) = \sum_{d \in D}\sum_{w \in W} n(w,d) \tag{14}$$

So,

$$P(z) = \frac{\sum_{d \in D}\sum_{w \in W} n(w,d) \times P(z \mid d,w)}{\sum_{d \in D}\sum_{w \in W} n(w,d)} \tag{15}$$

The computing process is similar for $P(w \mid z)$ and $P(d \mid z)$, and all the parameters can be obtained by the following four equations through iterations.

$$P(w \mid z) = \frac{\sum_{d \in D} n(w,d) \times P(z \mid d,w)}{\sum_{d \in D}\sum_{w' \in W} n(w',d) \times P(z \mid d,w')} \tag{16}$$

$$P(d \mid z) = \frac{\sum_{d \in D} n(w,d) \times P(z \mid d,w)}{\sum_{d' \in D}\sum_{w \in W} n(w,d') \times P(z \mid d',w)} \tag{17}$$

$$P(z) = \frac{\sum_{d \in D}\sum_{w \in W} n(w,d) \times P(z \mid d,w)}{\sum_{d \in D}\sum_{w \in W} n(w,d)} \tag{18}$$

$$P(z \mid d,w) = \frac{P(z,d,w)}{P(d,w)} = \frac{P(z) \times P(d \mid z) \times P(w \mid z)}{\sum_{z' \in Z} P(z') \times P(d \mid z') \times P(w \mid z')} \tag{19}$$

When the above result converges, the process of iteration can be terminated.

3.3 Improvement of the process of iteration

The process of iteration mentioned in section 3.2 is not very efficient, because the number of parameters is $K*(M+N)$ .To simplify the process of iteration, a new method of matrix iteration instead of the method mentioned above is proposed.

The parameters are $P(z_k), P(w_i \mid z_k), P(d_j \mid z_k)$, where $k \in \{1,2,......K\}$, $i \in \{1,2,......M\}$, and $j \in \{1,2,......N\}$. The number of all parameters is $K*(M+N)$, in the document-term matrix, where M, N represent the number of terms and documents respectively. They are generally very large, and this makes the iteration process very difficult and the computing process very complex.

The frequency of document-terms is represented by a matrix, and all the probabilities can be represented by the matrix. As in SVD, the matrix is decomposed into three matrices. If in PLSA, the probability can also be represented by a matrix as it is much easier to make an analogy between SVD and PLSA.

For i, j, k , the probability can be represented by:

$$Z = \begin{pmatrix} P(z_1),0,................0 \\ 0, \quad P(z_2) \\ ....\ddots \\ 0,....................P(z_k) \end{pmatrix}, \tag{20}$$

$$W = \begin{pmatrix} P(w_1 \mid z_1), P(w_1 \mid z_2),.......P(w_1 \mid z_K) \\ P(w_2 \mid z_1), P(w_2 \mid z_1),.......P(w_2 \mid z_K) \\ ....\ddots \\ P(w_M \mid z_1), P(w_M \mid z_1),.......P(w_M \mid z_K) \end{pmatrix} \tag{21}$$

$$D = \begin{pmatrix} P(d_1 \mid z_1), P(d_1 \mid z_2),.......P(d_1 \mid z_K) \\ P(d_2 \mid z_1), P(d_2 \mid z_1),.......P(d_2 \mid z_K) \\ ....\ddots \\ P(d_N \mid z_1), P(d_N \mid z_1),.......P(d_N \mid z_K) \end{pmatrix}. \tag{22}$$

It is now possible to conduct the iteration for the three matrices. The relevant equation can be represented by:

$$\begin{aligned} W &= \{W \bullet\times((X \bullet\div(W \times Z \times D^{'})) \times D \times Z)\} \\ &\bullet\div\{E(M) \times \{W \bullet\times((X \bullet\div(W \times Z \times D^{'})) \times D \times Z\}\} \end{aligned} \tag{23}$$

which corresponds to

$$P(w \mid z) = \frac{\sum_{d \in D} n(w,d) \times P(z \mid d,w)}{\sum_{d \in D} \sum_{w^{'} \in W} n(w^{'},d) \times P(z \mid d,w^{'})} \tag{24}$$

and "$\bullet\times$" denotes the product of the item of same i th row and j th column of two matrices. Whereas "$\bullet\div$". $\mathrm{E}(M)$ denotes the matrix $M \times M$ and all the cells are all "1".

$$D = \{D \bullet\times (Z \times W' \times (X \bullet\div (W \times Z \times D')))\} \bullet\div \{E(N) \times \{\{D \bullet\times (Z \times W' \times (X \bullet\div (W \times Z \times D')))\}\} \quad (25)$$

which corresponds to

$$P(d \mid z) = \frac{\sum_{d \in D} n(w,d) \times P(z \mid d, w)}{\sum_{d^{'} \in D} \sum_{w \in W} n(w, d^{'}) \times P(z \mid d^{'}, w)} \quad (26)$$

and $E(n)$ denotes the matrix $N \times N$ and all the cells are all “1”.

$$Z = \{Z \bullet\times (W^{'} \times (X \bullet\div (W \times Z \times D^{'})) \times D)\} / Sum(Sum(X)) \quad (27)$$

which corresponds to

$$P(z) = \frac{\sum_{d \in D} \sum_{w \in W} n(w,d) \times P(z \mid d, w)}{\sum_{d \in D} \sum_{w \in W} n(w,d)} \quad (28)$$

Using the above three equations, the iteration can be terminated when the iteration becomes convergent. Thus, the probability of each parameter can be obtained.

3.4 Documents clustering by PLSA

For document clustering, the matrix $(P(d_j \mid z_k))_{N \times K}$, for each row of the matrix is considered. Based on the rule of MLE, document j belongs to the latent topic class k when $P(d_j \mid z_k)$ is the maximum for all k. And the documents that belong to the same latent topic class could be grouped into a cluster.

4 Analysis of the Experiments

4.1 Basic analysis

In the experiments, about 3500 documents were extracted from the SCI database (http://portal.isiknowledge.com) with the keyword “data mining”. After data cleaning, there are 2549 documents. Using VSM, about 3000 keywords are extracted from the abstract of the documents to form a term document matrix. In this matrix, the columns denote the documents and the rows denote the keywords. Fig. 4 summarizes the changes of paper numbers over the past 14 years. It shows clearly that after 1996, there is an obvious increase in the number of papers on data mining. The increase becomes even more rapid after 2000. Because the data was collected before Nov. 2005, there is a slight decrease in the number of papers from 2004 to 2005.

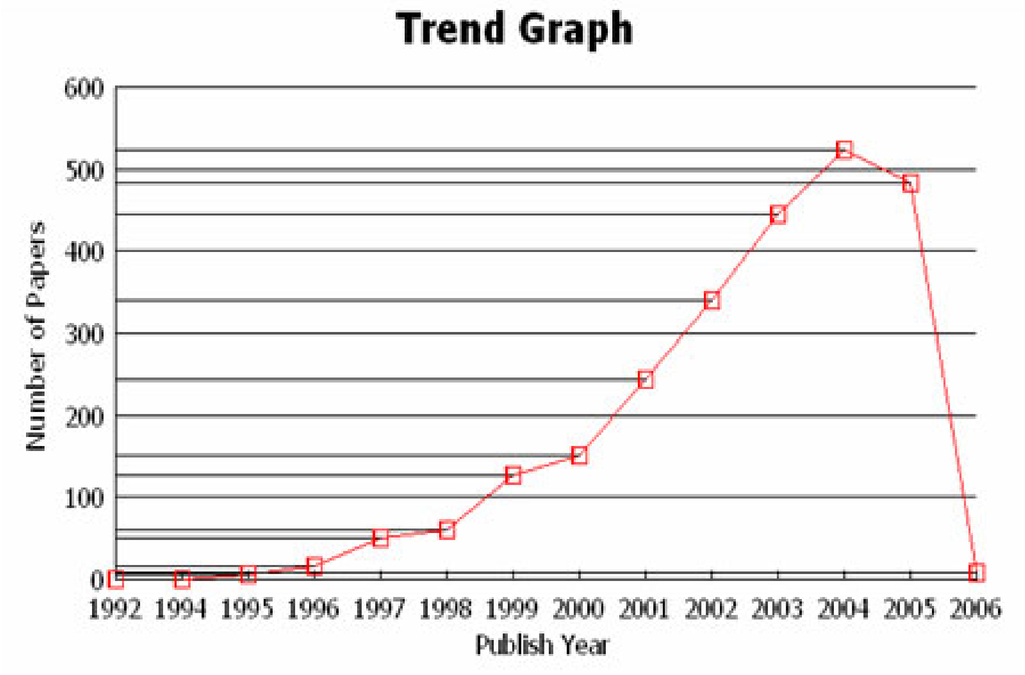


Figure 4 Trend of the number of papers

The numbers of papers in different countries/cities are also examined, and their distribution is shown in Fig. 5. It shows that USA produced a significantly larger number of papers than other countries/cities, in the area of data mining.

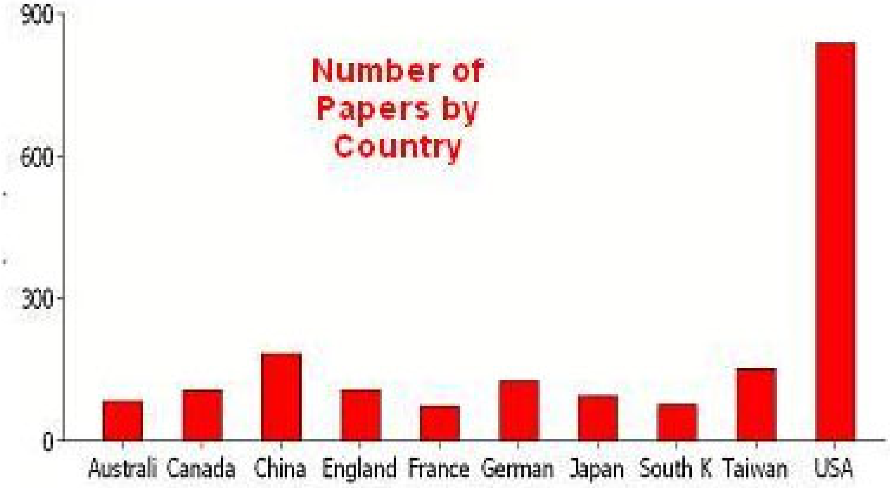
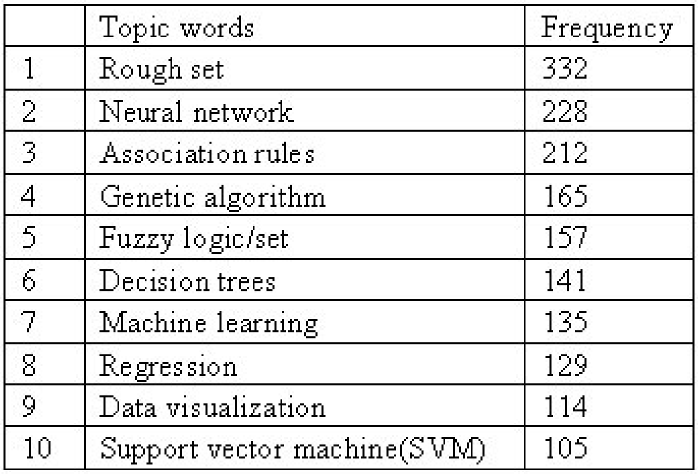


Figure 5 Distribution by country

4.2 Clustering analysis

In order to forecast technology, a very detailed analysis should be carried out. For example, it is necessary to know which areas/topics in the field are hot and which problems are most important, by document clustering. By using PLSA, Table 1 is generated, showing the top ten areas/topics in the field.



	Topic words	Frequency
1	Rough set	332
2	Neural network	228
3	Association rules	212
4	Genetic algorithm	165
5	Fuzzy logic/set	157
6	Decision trees	141
7	Machine learning	135
8	Regression	129
9	Data visualization	114
10	Support vector machine(SVM)	105

Table 1 Topic frequency

In order to find the trend, the number of papers on each area/topic in the period from 1998 to 2005 is determined and shown in Fig. 6. It shows that the increase in the number of papers involving "rough set" is most frequent, and so it is the hottest topic in data mining.

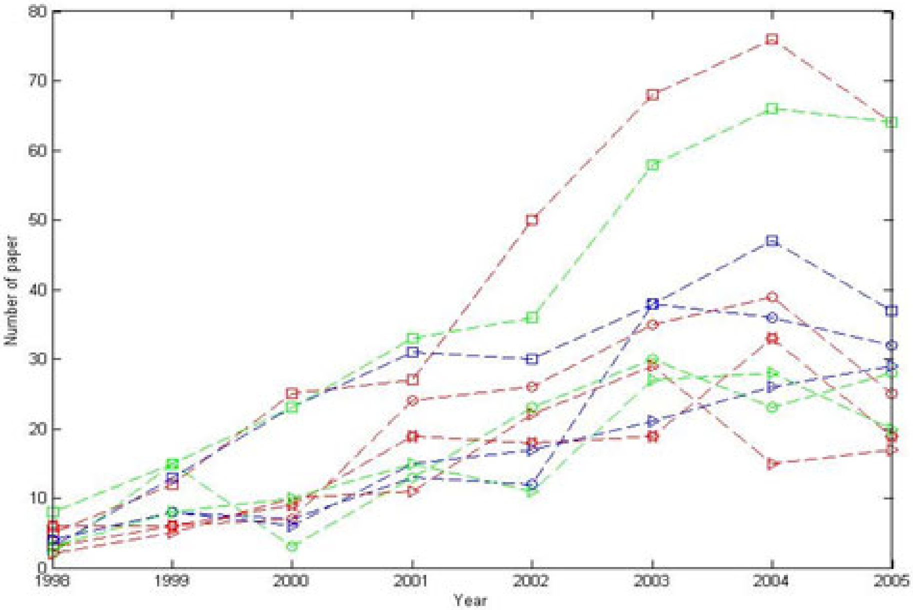


Figure 6 Trend of top ten topics during 1998-2005

ISSN: 0718-2724. (http://www.jotmi.org)

4.3 Comparison between PLSA and LSA

4.3.1 Difference between LSA and PLSA

Both LSA and PLSA can find the latent semantic space in a given corpus. But there is a big difference between the two methods. Firstly, SVD is based on principal component analysis (PCA) and matrix decomposition. The reduced matrix is the $F-norm$ approximation of the term-frequency matrix, while PLSA relies on the likelihood function and wants to get the maximization conditional probability of the model. It introduces a prior probability of latent class. The prior probability for a class is the probability of seeing this class in the data for a randomly chosen record, ignoring all attribute values. Mathematically, this is the number of records with a class label divided by the total number of records. Using EM algorithm, a local maximum of likelihood function can then be obtained. Secondly, LSA does not define a properly normalized probability distribution and $X^{'}$ may even contain negative entries while in PLSA, the matrix of the co-occurrence table is a well-defined probability distribution and the factors have a clear probabilistic meaning (Porter, 2005). Lastly, the similarity between a pair of documents can be obtained from the reduced matrix of SVD. Some clustering algorithms must be adopted in order to cluster documents. But it is possible to get the probability of which document occurs in a latent topic class by PLSA. Also a document can be grouped into a latent class of which the probability is the maximum of all classes.

4.3.2 Analysis of Experiments in LSA and PLSA

The main difference between PLSA and LSA has been discussed in the previous section. In order to compare the results of PLSA and LSA, document clustering is also carried out by LSA using SVD. As SVD can only obtain the correlation matrix of a semantically meaningful projection of documents and terms, it is necessary to use other clustering approaches to cluster the documents; the job is difficult when the number of documents is large. Only a total number of 200 documents are therefore selected from the corpus for comparison. To evaluate the efficiency of the two methods, precision and recall are the basic measures used in clustering the documents. The precision and recall for each cluster are defined as follows (http://www.hsl.creighton.edu/hsl/Sea-rching/Recall-Precision.html):

A: The number of relevant documents retrieved
B: The number of relevant documents not retrieved
C: The number of irrelevant documents retrieved

$precision = \frac{A}{A+C}$, $recall = \frac{A}{A+B}$, for the total precision and recall, there can be,

$P_{all} = \frac{\sum_{i=1}^{M} p_i}{M}$, $R_{all} = \frac{\sum_{j=1}^{M} r_j}{M}$, where p_i denotes the ith precision, r_j denotes the jth recall and M denotes the number of clusters.

Fig. 7 shows the precision achieved by the two methods.

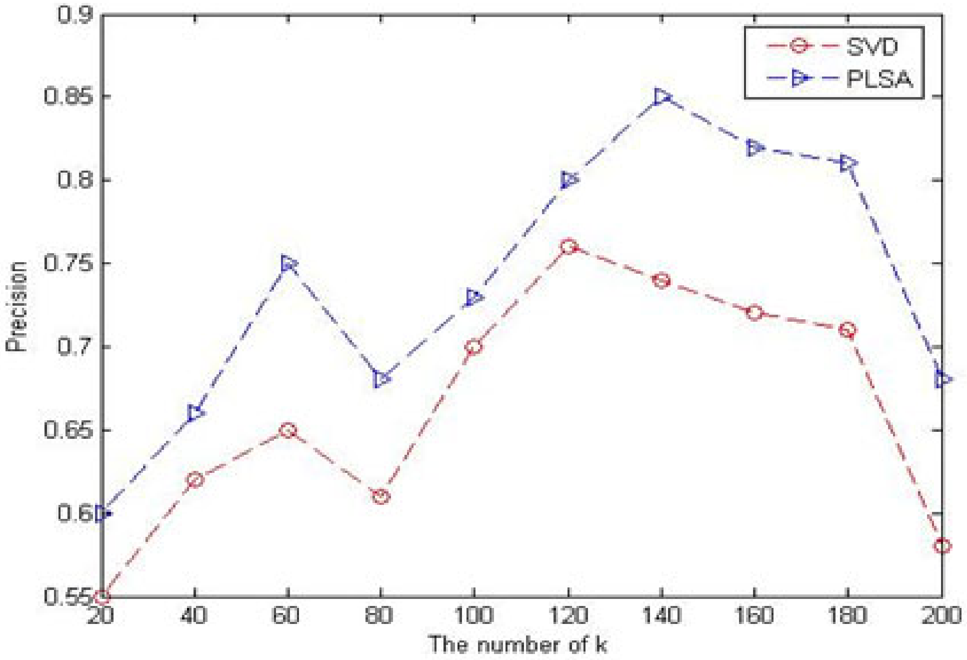


Figure 7 A comparison between PLSA and SVD

It is found that the precision achieved by PLSA is obviously higher than that of SVD for the 200 documents. PLSA is shown to be able to get the final result without applying other methods, but SVD needs other clustering algorithms.

5 Related work

Bibliometric analysis, especially document clustering has attracted more and more attention. Before latent semantic analysis, there was another method to find the indirect links between documents. S. Morris constructed a similarity function that generated links that yield meaningful clusters. The similarity functions are based on citations (Morris et al, 2002; Ronald et al, 2001). There are a direct citation and three types of indirect relations to derive the inter-document similarity values. After getting the similarity functions, they formulate these values into a similarity matrix and then use a clustering method such as Self-organized Mapping (SOM) (Morris et al, 2001) to divide these documents into groups. Unfortunately, this method is based only on citations and there may be few links between a pair of documents where one cites the other. On the other hand, there may be no citations between those correlated documents. So, semantic analysis, especially latent semantic analysis can give a more convincing warranty as to which documents are correlative.

In latent semantic analysis, there are some variation of LSA besides SVD and PLSA, for example, Riemann LSA (Jiang & Berry, 1998) is another variation of LSA. Some combine LSA and SOM methods to cluster documents. Lin and Chiang (2005), using a geometric structure in combinatorial topology, proposed a method to find a simplicial complex structure in latent semantic space. And there are many kinds of explanation for LSA including probabilistic analysis, multidimensional scaling analysis, etc.

6 Conclusion and future work

In this paper, a new document clustering algorithm is successfully applied to group documents for technology forecasting. The probabilistic latent semantic analysis algorithm (PLSA) which uses probabilistic method and algebra to find latent semantic space in a corpus is applied, and its results are compared with those produced by the existing LSA algorithm using SVD. The theoretical and experimental findings show that the PLSA is more accurate. In order to improve the efficiency of PLSA, a new iteration technique which uses matrix iteration instead of equation iteration is also proposed in this paper. Although the new algorithm is more accurate, large computing time is still an issue needing further improvement

References

[1] Ando, R.K., 2000, 'Latent Semantic Space: Iterative Scaling Improves Precision of Inter-document Similarity Measurement', *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval,* p.216-223, July 24-28, 2000, Athens, Greece.

[2] Brian, T., Bartell, G W. and Cottrell R. K. B., 1992, 'Latent Semantic Indexing is an Optimal Special Case of Multidimensional Scaling', *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*.

[3] Deerwester, S, Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harslman, R., 1990, 'Indexing by latent semantic analysis'. *Journal of the American Society for Information Science*, 41(6):391-407.

[4] Dempster, A., Laird, N., and Rubin, D. 1977. 'Maximum likelihood from incomplete data via the EM algorithm'. *Journal of the Royal Statistical Society*, Series B, 39(1):1–38.

[5] Dhillon, I.S., 2001, 'Co-clustering documents and words using Bipartite Spectral Graph Partitioning'. *Knowledge Discovery and Data Mining.*

[6] Ding, C.H., 1999, 'A similarity-based probability model for Latent Semantic Indexing'. In *Proceedings of S1-GIR'99*, pages 1:58-65.

[7] Dumais, S.T. 1995. 'Using LSI for information filtering', The Third Text REtrieval Conference(TREC3). *National Institute of Standards and Technology Special Publication*s 500-215, pp. 219-230.

[8] Furnas, G.W., Deerwester, S., Dumais, S.T., 1988, 'Information retrieval using a singular value decomposition model of latent semantic structure', *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval* **,** pages: 465 - 480

[9] Golub, G. and Loan, C.V., *'Matrix Computations*'. The Jason Hopkins University Press, Baltimore, Maryland, second edition edition.

[10] Hofmann, T., 1999, 'Probabilistic latent semantic indexing'. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM Press.

[11] Jiang, E.P., and Berry, M.W., 1998, 'Information filtering using the Riemannian SVD (R-SVD)'. In *Proceedings of IRREGULAR'98*, pages 386-395.

[12] Kostoff, R.N., 1997. '*The Handbook of Research Impact Assessment*'. Seventh Edition, Summer 1997.

[13] Kruskal, J.B., 1978 , 'Factor Analysis and Principal Components'. *International Encyclopedia of Statistics*. New York: Free Press.

[14] Lin, T.Y. and Chiang, I.J., 2005, 'A simplicial complex, a hypergraph structure in the latent semantic space of document clustering', *International Journal of approximate reasoning.* 40(2005) 55-80.

[15] Mogee, M. E., 1997. 'Patents and technology intelligence' . In W. B. Ashton, & R. A. Klavans (Eds.), *Keeping abreast of science and technology, technical intelligence for business*. Columbus, OH: Battelle Press..

[16] Morris, S., De Yong, C., Wu, Z., Salman, S. and Yemenu, D., 2002, 'DIVA: a visualization system for exploring documents database for technology forecasting ', *Computers and industrial engineering.* 43(2002) 841-862.

[17] Morris, S.A., Wu, Z., Yen, G. 2001. 'A SOM mapping technique for visualizing documents in a database'. *Proceedings of the IEEE International Joint Conference on Neural Networks*, Washington DC, USA, July 14-19

[18] Papadimitriou, C.H., Tamaki, H., Raghavan, P., and Vempala, S., 1998, 'Latent Semantic Indexing: A Probabilistic Analysis'. *PODC'98*.

[19] Porter, A., 2005, '*Tech Mining'*, pp 137- 169. New Jersey, Wiley.

[20] Porter, M.F.,1980, '*An algorithm for suffix stripping'* Program, 14 no. 3, pp 130-137, July 1980.

[21] Ronald. N., Kostoff, J., del Río, A., Humenik, J.A., García, E.O., and Ramírez, A.M. 2001, 'Citation mining: integrating text mining and bibliometric for research user profiling'. *Journal of the American Society for Information Science and Technology,* Volume 52 Issue 13 page 1148 - 1156.

[22] Salton, G. and McGill, M., 1983, 'Introduction to Modern Information Retrieval'. McGraw-Hill.

[23] Salton, G., 1989, ‘Automatic text processing: the transformation, analysis, and retrieval of information by computer’. Reading, MA, USAL Addison-Wesley.

[24] Shima, K., Todoriki, M., and Suzuki, A., 2004, ‘SVM-based feature selection of latent semantic features’, *Pattern Recognition Letters* 25 (2004) 1051–1057.

[25] Story, R.E., 1996, ‘An explanation of the effectiveness of Latent Semantic Indexing by means of a Bayesian regression model.’ *Information Processing & Management*, 32(3):329--344.

[26] Swanson, D.R. and Smalheiser, N.R., 1997, ‘An interactive system for finding complementary Literatures: a stimulus to scientific discovery’, *Artificial Intelligence*, 91, 183-203.

[27] http://www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html