Dimensionality Reduction with Multilingual Resource

YingJu Xia **Hao Yu** **Gang Zou**
Fujitsu Research & Development Center Co.,LTD.
13F Tower A, Ocean International Center, No.56 Dong Si Huan Zhong Rd, Chaoyang District, Beijing, China, 100025
{yjxia,yu,zougang}@cn.fujitsu.com

Abstract

Query and document representation is a key problem for information retrieval and filtering. The vector space model (VSM) has been widely used in this domain. But the VSM suffers from high dimensionality. The vectors built from documents always have high dimensionality and contain too much noise. In this paper, we present a novel method that reduces the dimensionality using multilingual resource. We introduce a new metric called *TC* to measure the term consistency constraints. We deduce a *TC* matrix from the multilingual corpus and then use this matrix together with the term-by-document matrix to do the Latent Semantic Indexing (LSI). By adopting different *TC* threshold, we can truncate the *TC* matrix into small size and thus lower the computational cost of LSI. The experimental results show that this dimensionality reduction method improves the retrieval performance significantly.

1 Introduction

1.1 Basic concepts

The vast amount of electronic information that is available today requires effective techniques for accessing relevant information from it. The methodologies developed in information retrieval aim at devising effective means to extract relevant documents in a collection when a user query is given. In information retrieval and filtering, Query and document representation is a key problem and many techniques have been developed. Among these techniques, the vector space model (VSM) proposed by Salton (1971; 1983) has been widely used. In the VSM, a document is represented by a vector of terms. The cosine of the angle between two document vectors indicates the similarity between the corresponding documents. A smaller angle corresponds to a larger cosine value and indicates higher document similarity. A query, which describes the information need, is encoded as a vector as well. Retrieval of documents that satisfy the information need is achieved by finding the documents most similar to the query, or equivalently, the document vectors closest to the query vector. There are several advantages to this approach beyond its mathematical simplicity. Above all, it is efficient to compute and store the word counts. This is one reason that why VSM is widely used for query and document representation. But this method has problem that the vectors built from documents always have high dimensionality and contain too much noise. The high dimensionality causes high computational and memory requirements while noise in the vectors degrades the system performance.

1.2 Related works

To address these problems, many dimensionality reduction techniques have been applied to query and document representation. Among these techniques, Latent Semantic Indexing (LSI) (Deerwester et al., 1990; Hofmann, 1999; Ding, 2000; Jiang and Littman, 2000; Ando, 2001; Kokiopoulou and Saad, 2004; Lee et al., 2006) is a well-known approach. LSI constructs a smaller document matrix that retains only the most important information from the original by using the Singular Value Decomposition (SVD). Many modifications have been made to this approach (Hofmann, 1999; Ding, 2000; Jiang and Littman, 2000; Kokiopoulou

and Saad, 2004; Sun et al., 2004; Husbands et al., 2005). Among them, IRR (Ando and Lee, 2001) is a subspace-projection method that counteracts tendency to ignore minority-class documents. This is done by repeatedly rescaling vectors to amplify the presence of documents poorly represented in previous iterations.

In concept indexing (CI) (Karypis and Han, 2000) method, the original set of documents is first clustered into *k* similar groups, and then for each group, the centroid vector (i.e., the vector obtained by averaging the documents in the group) is used as one of the *k* axes of the lower dimensional space. The key motivation behind this dimensionality reduction approach is the view that each centroid vector represents a concept present in the collection, and the lower dimensional representation expresses each document as a function of these concepts. George and Han (2000) extend concept indexing in the context of supervised dimensionality reduction. To capture the concept, phrase also has been used as indexing entries (Mao and Chu, 2002).

The LPI method (Isbell and Viola, 1999) tries to discover the local structure and obtains a compact document representation subspace that best detects the essential semantic structure. The LPI uses Locality Preserving Projections (LPP) (Xiaofei He and Partha, 2003) to learn a semantic space for document representation. Xiaofei He et al., (2004) try to get sets of highly-related words, queries and documents are represented by their distance to these sets. These algorithms have successfully reduced the dimensionality and improve the retrieval performance but at the mean time they led to a high computational complexity.

1.3 Our method

In this study, we propose a novel method that reduces the dimensionality using multilingual resource. We first introduce a new metric called *TC* to measure the term consistency constraints. We use this metric to deduce a *TC* matrix from the multilingual corpus. Then we combine this matrix to the term-by-document matrix and do the Latent Semantic Indexing. By adopting different *TC* threshold, we can truncate the *TC* matrix into small size and thus lower the computational cost of LSI.

The remainder of this paper is organized as follows. Section 2 describes the dimensionality reduction method using multilingual resource. Section 3 shows the experimental results to evaluate the dimensionality reduction method. Finally, we provide conclusions and remarks of future work in Section 4.

2 Dimensionality reduction using multilingual resource

2.1 Motivation

As mentioned above, the queries and documents are represented by vectors of terms. The weight of each term indicates its contribution to the vectors. Many weighting schemes have been proposed. The simplest form is to use the term-frequency (TF) as the term weight. In this condition, a document can be represented as a vector $\vec{d} = (tf_1, tf_2, ..., tf_n)$, where tf_i is the frequency of the *i*th term in the document. A widely used refinement to this model is to weight each term based on its inverse document frequency (IDF) in the documents collection. This is commonly done by multiplying the frequency of each term i by $\log(N / df_i)$, where N is the total number of documents in the collection, and df_i is the number of documents that contain the *i*th term. This leads to the TF-IDF representation of the documents. Although the TF-IDF weighting scheme has many variants (Buckley, 1985; Berry et al., 1999; Robertson et al., 1999), the idea is the same one that uses the statistical information such as TF and IDF to calculate the term weight of vectors.

This kind of statistical information is independence with languages. For example, in one language, say L^a, we have a vocabulary $V^a = \{w_1^a, w_2^a, ..., w_n^a\}$ and a documents collection $D^a = \{d_1^a, d_2^a, ..., d_m^a\}$. If this documents collection has a parallel corpus in language L^b, say, $D^b = \{d_1^b, d_2^b, ..., d_m^b\}$ and a vocabulary $V^b = \{w_1^b, w_2^b, ..., w_n^b\}$. When we put a query $Q_k^a = \{q_{k1}^a, q_{k2}^a, ..., q_{kl}^a\}$ ($q_{ki}^a \in V^a$) into an information retrieval system. The information retrieval system will converts the query Q_k^a and the documents in the collection D^a into vectors. By calculating the similarity between query Q_k^a and each document d_i^a, the system selects the documents whose similarity is higher than a threshold as the results R_k^a. If we translate the query Q_k^a into language L^b and get query Q_k^b, when putting the Q_k^b into the same information retrieval system, we get the retrieval results R_k^b. Since the Q_k^a and Q_k^b contain the same content and only expressed in different languages. We expect that R_k^a

and R_k^b will contain the same content. If this assumption holds, the vocabulary which is used to build queries and documents vectors should have high representative ability. Since the weight of each term in the vector is calculated by the statistical information such as TF and IDF. If the vocabulary V^a and V^b have high representative ability, their statistical information will be consistent as well. This is the main motivation of our dimensionality reduction method.

2.2 Dimensionality reduction method

The most straightforward way to measure the word's representability in multilingual resource is to calculate the TF and IDF of each word in different languages. But this method has one problem that the TF-IDF scheme is dedicated for each single document, the same word will have different weight in different documents. It is impractical to impose the consistency constraint to every document. Even we can do that, this method still has the drawback that it is very difficult to port to another documents collection. To address this problem, we consider the whole documents collection as one single document. In this condition, the IDF will be a fixed number.

We introduce a new metric to measure the term consistency called *TC*. Figure 1 and Figure 2 illustrate the basic idea. In these figures, the curve L_a shows the word logarithmic frequency in the documents collection of language L^a, the curve L_b shows the corresponding translation's logarithmic frequency in the documents collection of language L^b. TC_i and TC_j are the term consistency of w_i and w_j respectively.

Figure 1 shows the *TC* in normal condition that the average word frequency in language *a* is proximate to that of language *b*. In this case, the *TC* is defined as below:

$$TC(w_i^b) = \min(\log(f_i^a)/\log(f_i^b), \log(f_i^b)/\log(f_i^a)) \quad (1)$$

Here f_i^a is the frequency of w_i^a in language *a*. f_i^b is the frequency of the w_i^a's translation in language *b*. In multilingual case, the $TC(w_i)$ will be defined as below:

$$TC(w_i) = \min(TC(w_i^b), \ldots TC(w_i^n)) \quad (2)$$

In the case that the average word frequency in language *a* is different with that of language *b*, we will first calculate the moving average as shown in the Figure 2. After that, we use the moving average to calculate the *TC* of w_i as below:

$$TC(w_i^b) = \min((\log(f_i^a)+H)/\log(f_i^b), \log(f_i^b)/(\log(f_i^a)+H)) \quad (3)$$

Here *H* is distance between the moving average and the original one.

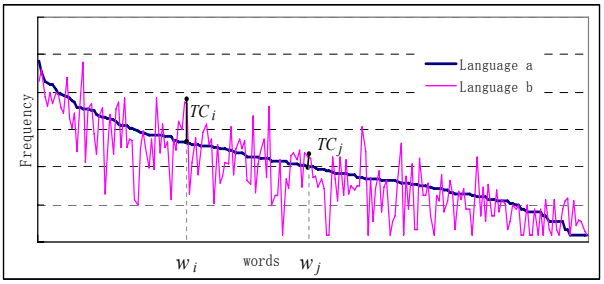

Figure 1. *TC* in normal condition

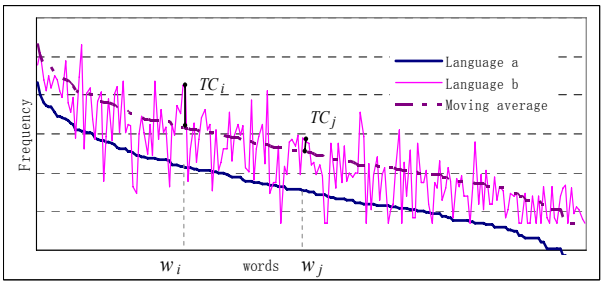

Figure 2. *TC* in shift condition

Once we get the *TC* of every word in language *a*, we present it in a diagonal matrix $T_{t\times t}=diag(TC_1, TC_2, \ldots, TC_t)$, $TC_1 \geqslant TC_2 \geqslant \ldots \geqslant TC_t$.

When applying the *TC* matrix $T_{t\times t}$ in information retrieval, we combine $T_{t\times t}$ into the term-by-document matrix $A_{t\times d}$. Where $A_{t\times d} = [a_{ij}]$ and the a_{ij} is the weight of term *i* in document *j*. We get a new matrix $B_{t\times d}=T_{t\times t}A_{t\times d}$. Then following the classical LSI, we replace $B_{t\times d}$ by a low-rank approximation derived from its truncated Singular Value Decomposition (SVD):

$$B_{t\times d} = U_{t\times n}\Sigma_{n\times n}V_{d\times n}^T$$

Here $UU^T = I$, $VV^T = I$, $\Sigma = diag(\sigma_1, \sigma_2, ..., \sigma_n)$ $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r \geq \sigma_{r+1} = ... = \sigma_n = 0$.

The main problem of LSI is that it usually led to a high computational complexity since the matrix $B_{t\times d}$ usually in 10^3-10^5 dimensional space. To lower the computational cost, we truncate the *TC* matrix $T_{t\times t}$ according to different *TC* threshold and get a new matrix $\hat{T}_{t\times t} = diag(TC_1, TC_2, ..., TC_t)$, $TC_1 \geq TC_2 \geq ... \geq TC_r \geq TC_{r+1} = ... = TC_t = 0$. Then we get $\hat{B}_{r\times d}=\hat{T}_{r\times r}A_{r\times d}$. Since *r* is small than *t*, the

computational cost on the matrix $\hat{B}_{r\times d}$ will lower than $B_{t\times d}$. Note that the matrix $\hat{B}_{r\times d}$ is deduced from the *TC* matrix $T_{t\times t}$ which is sorted by word representative ability. It will contain less noise and outperform the original matrix $A_{t\times d}$. The experimental results have shown the effective of this method.

For one word w_i^a in language L^a, there are always several translations in language L^b, say (w_{i1}^b, w_{i2}^b,..., w_{ik}^b). To handle this one-to-many phenomenon, we calculate the co-occurrence of w_i^a and each translation and select the highest one as the translation of w_i^a.

3 Experiments

We adopt a VSM based IR system to evaluate the dimensionality reduction method presented in Section 2. The term weight in the term-by-document matrix is calculated by the TF-IDF weighting scheme.

3.1 Training and test corpora

The training corpus comes from Chinese Linguistic Data Consortium (http://www.chineseldc.org/, abbreviate as CLDC). Its code number is "2004-863-009". This parallel corpus contains parallel texts in Chinese, English and Japanese. It is aligned to sentence level. The sentence alignment is manually verified and the sampling examination shows the accuracy reaches 99.9%.

The experiments are conducted on two test corpora. The first one is the information retrieval test corpus gotten from CLDC ("2003-863-006"). It is a Chinese IR corpus and contains 20 topics for test. Each topic has key words and description and narrative. The second one is the Reuters 2001 data (http://about.reuters.com/researchandstandards/corpus/). This corpus is a collection of about 810,000 Reuters English news stories from August 20, 1996 to August 19, 1997. It was used by the TREC-10 Filtering Tracks (Robertson and Soboroff, 2002). In TREC-10, 84 Reuters categories were used to simulate user profiles.

The evaluate measure is a version of van Rijsbergen(1979)'s *F* measure with β=1(we denote it as *F1*).

3.2 Experimental results

The table1 and table2 show the experimental results conducted on Chinese and English test Corpus respectively. In these tables, we compare our method with basic LSI and LPI (Xiaofei et.al, 2004). In the table1, the 'C-E' means the *TC* matrix gotten from Chinese-English training collection (deduced from the trilingual training corpus). The 'C-J' means that the *TC* matrix gotten from Chinese-Japanese training collection, and so force the 'C-E-J'. All the *TC* matrices have been normalized to range from 0 to 1. The threshold θ is used to truncate the *TC* matrix into small size. Bigger θ corresponds to smaller truncated *TC* matrix. Note that here θ is discrete since for some θ, the size of truncated matrix is very similar. For example, when θ = 0.85 and θ = 0.9, the size of truncated *TC* matrices are the same one.

LSI: 0.3785, LPI: 0.405			
θ	C-E	C-J	C-E-J
0.3	0.404	0.4014	0.4124
0.4	0.4098	0.406	0.4185
0.45	0.4159	0.4185	0.4226
0.5	0.4204	0.4124	0.4105
0.55	0.4061	0.4027	0.3997
0.6	0.3913	0.3992	0.396
0.8	0.3856	0.3867	0.3842
0.85	0.3744	0.3754	0.3768

Table1.*F1* measure of Chinese test corpus

LSI: 0.3416, LPI: 0.3556			
θ	E-C	E-J	E-C-J
0.3	0.356	0.3478	0.3578
0.4	0.3578	0.3596	0.3702
0.45	0.3698	0.3651	0.3734
0.5	0.3636	0.3575	0.363
0.55	0.3523	0.3564	0.3477
0.6	0.3422	0.3448	0.3458
0.8	0.3406	0.3397	0.3378
0.85	0.3304	0.3261	0.3278

Table2. *F1* measure of English test corpus

From the experimental results, we can see that our method make great enhancement to the basic LSI method. And our method also outperforms the LPI method in both test corpora. Comparing the performance on different training collection, we can find that the difference is subtle. In Chinese test corpus, the *TC* matrix gotten from C-E-J training collection get the best performance (*F1*=0.4226) at θ=0.45 while the C-E test collection get 0.4204

at θ=0.5 and the C-J test collection get 0.4185 at θ=0.45. For the English test corpus, the trilingual training collection also gets the best performance. But the difference between bilingual and trilingual training collection is also subtle (E-C-J: *F1*=0.3734, E-C: *F1*=0.3698, E-J: *F1*=0.3651). In the English test corpus, all the training collection get the best performance at θ=0.45.

As mentioned before, the bigger θ means the smaller size of the truncated *TC* matrix. While small size of the truncated *TC* matrix means low computational cost and high system speed. This is one of the advantages of our method over the traditional LSI method. We conducted some experiments to test the system speed on different threshold θ. We use the number of documents per second (docs/s) to denote this kind of system speed. The experiment is conducted on the personal computer with a Pentium (R) 4 processor @2.8GHz, 256 KB cache and 512 MB memory. Table 3 shows the experimental results that the θ vs. system speed and Figure 3 illustrates the *F1* measure vs. the system speed.

Baseline(LSI): 566.5 docs/s			
θ	C-E	C-J	C-E-J
0.3	1039.3	1034.4	1355.0
0.4	1148.4	1188.9	1372.5
0.45	1290.5	1246.9	1391.3
0.5	1323.9	1323.3	1469.6
0.55	1393.3	1392.6	1563.8
0.6	1413.3	1508.8	1590.1
0.8	1513.1	1555.6	1660.5
0.85	1641.1	1778.2	1773.5

Table 3. θ vs. system speed

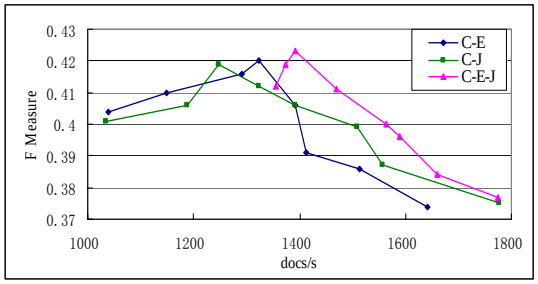

Figure 3. *F1* measure vs. system speed

4 Conclusions

In this paper, we present a novel method that reduces the dimensionality using multilingual resource. We deduce a *TC* matrix from the multilingual corpus and then truncate it to small size according to different *TC* threshold. Then we use the truncated matrix together with the term-by-document matrix to do the LSI analysis. Since the truncated *TC* matrix is sorted by word representative ability. It will contain less noise than the original term-by-document matrix. The experimental results have shown the effectiveness of this method.

In the future, we will try to find the optimal truncate threshold θ automatically. And since it is more difficult to get the parallel corpora than comparable corpora, we will explore using comparable corpora to do the dimensionality reduction.

Acknowledgement

This research was carried out through financial support provided under the NEDO International Joint Research Grant Program (NEDO Grant).

References

Ando R. K., “Latent Semantic Space: Iterative Scaling improves precision of inter-document similarity measurement”, in Proc. of the 23th International ACM SIGIR, Athens, Greece, 2000.

Ando R. K., and Lee L., “Iterative Residual Rescaling: An Analysis and Generalization of LSI”, in Proc. of the 24th International ACM SIGIR, New Orleans, LA, 2001.

Arampatzis A., Beney J., Koster C.H.A., and T.P. van der Weide. KUN on the TREC9 Filtering Track: Incrementality, decay, and theshold optimization for adaptive filtering systems. The ninth Text Retrieval Conference, November 9-12, 2000 Gaithersburg, MD,

Avi Arampatzis and Andre van Hameren The Score-Distributional Threshold Optimization for Adaptive Binary Classification Tasks , SIGIR’01, September 9-12,2001, New Orleans, Louisiana,USA. 285-293

Berry M., Drmac Z., and Jessup E.. Matrices, vector spaces, and information retrieval. SIAM Review, 41(2):pp335-362, 1999.

Bingham E. and Mannila H., “Random Projection in dimensionality reduction: applications to image and text data”, Proc. Of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 245-250,2001.

Buckley C.. Implementation of the SMART information retrieval system. Technical Report TR85-686, Department of Computer Science, Cornell University,

Ithaca, NY 14853, May 1985. Source code available at ftp://ftp.cs.cornell.edu/pub/smart.

C.H. Lee, H.C. Yang, and S.M. Ma, "A Novel Multi-Language Text Categorization System Using Latent Semantic Indexing", The First International Conference on Innovative Computing, Information and Control (ICICIC-06), Beijing, China, 2006.

C.J. van Rijsbergen. Information Retrieval, chapter 7. Butterworths, 2 edition, 1979.

Deerwester S. C., Dumais S. T., Landauer T. K., Furnas G. W., and harshman R. A., "Indexing by Latent Semantic Analysis", Journal of the American Society of Information Science, 41(6):391-407, 1990.

Ding C. H.. A probabilistic model for dimensionality reduction in information retrieval and filtering. In Proc. of 1st SIAM Computational Information Retrieval Workshop, October 2000.

George Karypis, Eui-Hong (Sam) Han, Fast supervise d dimensionality reduction algorithm with applications to document categorization & retrieval ,Proceedings of the ninth international conference on Information and knowledge management, November 2000

Hofmann T., "Probabilistic Latent Semantic Indexing", in Proc. of the 22th International ACM SIGIR, Berkeley, California, 1999.

Husbands, P., Simon, H., and Ding, C. Term norm distribution and its effects on latent semantic indexing, Information Processing and Management: an International Journal, v.41 n.4, p.777-787, July 2005

Isbell C. L. and Viola P., "Restructuring Sparse High Dimensional Data for Effective Retrieval", Advances in Neural Information Systems, 1999.

Jiang F. and Littman M.L., Approximate dimension equalization in vector-based information retrieval. Proc. 17th Int'l Conf. Machine Learning, 2000.

Karypis G. and Han E.H.. Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval & categorization. Technical Report TR-00-016, Department of Computer Science, University of USA

Kokiopoulou E., Saad Y., Polynomial filtering in latent semantic indexing for information retrieval ,Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR '04,July 2004

Mao W. and Chu W.W.. Free-text medical document retrieval via phrase-based vector space model. In Proceedings of AMIA Annual Symp 2002.

Minnesota, Minneapolis, 2000. Available on the WWW at URL http://www.cs.umn.edu/~karypis.

Robertson SE, Walker S, Beaulieu M,Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track- Proceedings of the seventh Text Retrieval Conference, TREC-7, pp. 253-264 ,1999

Robertson, S., & Soboroff, I., The TREC-10 Filtering track final report. Proceeding of the Tenth Text REtrieval Conference (TREC-10) pp. 26-37. National Institute of Standards and Technology, special publication 500-250., 2002

Salton, G, the SMART Retrieval System – Experiments in Automatic Document Processing. Prentice-Hall, Englewood. Cliffs, New Jersey,1971.

Salton, G., Dynamic Information and Library processing. Prentice-Hall, Englewood Cliffs, New Jersey,1983.

Salton, G and McGill. M.J., Introduction to Modern Information retrieval. McGraw Hill, New York,1983.

Sun, J.T. , Chen , Z. , Zeng , H.J. , Lu, Y.C. , Shi, C.Y. and Ma, W.Y. ,"Supervised Latent Semantic Indexing for Document Categorization" , In Proceedings of the Fourth IEEE International Conference on Data Mining 2004

Xiaofei He and Partha Niyogi, "Locality Preserving Projections",in Advances in Neural Information Processing Systems 16, Vancouver, Canada, 2003.

Xiaofei He, Deng Cai, Haifeng Liu, Wei-Ying Ma, Locality preserving indexing for document representation, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR '04, July 2004

Zhai C., Jansen P., Roma N., Stoica E., and Evans D.A.. Optimization in CLARIT adaptive filtering. In proceeding of the Eight Text Retrieval Conference 1999, 253-258.

Zhang Y., and Callan J.. Maximum likelihood Estimation for Filtering Thresholds. SIGIR'01, September 9-12,2001, New Orleans, Louisiana,USA. 294-302