A Simplified Latent Semantic Indexing Approach for Multi-Linguistic Information Retrieval [1]

Liu Yi, Lu Haiming, Lu Zengxiang, Wang Pu
Department of Automation, Tsinghua University
Beijing 100084
P.R.China
yiliu00@mails.tsinghua.edu.cn

Abstract

Latent Semantic Indexing (LSI) approach provides a promising solution to overcome the language barrier between queries and documents, but unfortunately the high dimensions of the training matrix is computationally prohibitive for its key step of Singular Value Decomposition (SVD). Based on the semantic parallelism of the multi-linguistic training corpus we prove in this paper that, theoretically if the training term-by-document matrix can appear in either of two symmetry forms, strong or weak, the dimension of the matrix under decomposition can be reduced to the size of a monolingual matrix. The retrieval accuracy will not deteriorate in such a simplification. And we also discuss what these two forms of symmetry mean in the context of multi-linguistic information retrieval. Although in real world data the term-by-document matrices are not naturally in either symmetry form, we suggest a way to make them appear more symmetric in the strong form by means of word clustering and term weighting. A real data experiment is also given to support our method of simplification.

1 Introduction

Multi-linguistic Information Retrieval (MLIR for short, also "translingual" or "cross-language" IR) enables a query in one language to search document collection in another one or more languages. Many monolingual IR approaches can be extended to multi-linguistic environment and among them Latent Semantic Indexing (LSI for short, Deerwester *et al.*, 1990) has proved effective (Y. Yang *et al.*, 1997, Douglas William Oard, 1996).

The particular technique used in LSI is singular-value decomposition (SVD for short), in which a large term-by-document matrix is decomposed into a set of orthogonal factors from which the original matrix can be approximated by linear combination. However, SVD on the large term-by-document matrix whose size increases with the size of the training corpus brings huge computation costs. This situation becomes even worse when we use LSI for MLIR because the training matrix consisting of various languages is always several times larger. To reduce the cost of SVD in LSI and thus make it feasible in MLIR, we try to exploit the semantic symmetry hidden in the training corpus. We find that theoretically if the term-by-document matrices of multi-linguistic training set have either a weak symmetry form or a strong symmetry form, the SVD step of LSI in multi-linguistic environment can be simplified. Both symmetry forms have clear meanings in the context of MLIR. Further, though we can never reach precisely either of two symmetry forms from real world data, two possible methods are raised to enhance the strong form symmetry of the term-by-document matrices. Our small-scale experiment gives a satisfying result though we only roughly keep the strong symmetry.

In section 2 we will briefly introduce how LSI approach can be extended to multi-linguistic environment. In section 3 we will prove some theorems for the LSI simplification in the two symmetry forms, then discuss symmetry enhancement issues for real world data. Experiments and results will be

[1] Supported by the National Natural Science Foundation of China (No. 60003004)

given in section 4. Finally in section 5 we will draw our conclusions with some problems for future work.

2 LSI for MLIR

LSI is based on the vector space mode (VSM), in which both queries and documents are represented as vectors of term weights

$$\mathbf{q} = (q_1, q_2, \cdots, q_m)^t$$
$$\mathbf{d} = (d_1, d_2, \cdots, d_m)^t$$

where **q** is the query vector, **d** is the document vector, *m* is the number of unique terms (words, phrases or word clusters) in the corpus after stop-word removal and stemming, q_i and d_i are term weights in the query and the document respectively. Terms are usually weighted by term frequency, term frequency inverted document frequency (TFIDF), information gain or other weighting schemes. In monolingual IR, the similarity between a query and a document is defined as

$$sim(\mathbf{q},\mathbf{d}) = \cos(\mathbf{q},\mathbf{d}) = \frac{\sum_{i=1}^{m} q_i d_i}{\sqrt{\sum_{i=1}^{m} q_i^2 \sum_{i=1}^{m} d_i^2}}$$

LSI is a one-step extension of VSM. The claim is that neither terms nor documents are the optimal choice for the orthogonal basis of a semantic space, and a reduced vector space consisting of the most meaningful linear combinations of documents would be a better representative basis for the documents content.

In monolingual IR, let **W** be the term-by-document matrix of the training corpus consisting of only one language. By SVD analysis

$$\mathbf{W} = \mathbf{USV}^t$$

where matrices **U** and **V** are unitary matrixes, which means that $\mathbf{U}^t = \mathbf{U}^{-1}$, $\mathbf{V}^t = \mathbf{V}^{-1}$. And $\mathbf{S} = \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ in which Σ is a diagonal matrix with *r* nonzero diagonal entries. These *r* nonzero values are called *singular values* of the matrix **W**. If we take the first *k* biggest nonzero diagonal entries of **S** and the corresponding columns in matrices **U** and **V**, the bilingual term-by-document matrix **W** can be approximated as

$$\mathbf{W} \approx \mathbf{U}_k \mathbf{S}_k \mathbf{V}_k^t$$

where matrices $\mathbf{U}_k$ and $\mathbf{V}_k$ contain a set of *k* orthogonal singular vectors each (one for the representation of terms and the other for the representation of documents). Matrix $\mathbf{S}_k$ is *k*-diagonal, containing the singular values indicating the importance of the corresponding singular vectors in matrices $\mathbf{U}_k$ and $\mathbf{V}_k$. The monolingual retrieval criterion of LSI approach is defined to be

$$sim(\mathbf{q},\mathbf{d}) = \cos(\mathbf{U}_k^t\mathbf{q}, \mathbf{U}_k^t\mathbf{d})$$

LSI approach can be extended to multi-linguistic environment. Let us take bilingual case for example. We define $\mathbf{A}_{m\times n}$ be a term-by-document matrix for the training documents in the source language (also the language of queries), $\mathbf{B}_{s\times n}$ be a term-by-document matrix for the training documents in the target language. The corresponding columns of **A** and **B** are the matching pairs of documents in the bilingual corpus.

Now, **W** is defined as an $(m+s)\times n$ term-by-document matrix of the entire corpus, representing the bilingual document pairs

$$\mathbf{W} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix}$$

Then we do the SVD analysis and k singular value approximation on the bilingual term-by-document matrix **W** just as in the monolingual case. Let **q** be a query in the source language, **d** be a document in the target language. The retrieval criterion of LSI approach in bilingual environment is (Y. Yang *et al.*, 1997)

$$sim(\mathbf{q},\mathbf{d}) = \cos(\mathbf{U}_k^t \begin{bmatrix} \mathbf{q} \\ \mathbf{0} \end{bmatrix}, \mathbf{U}_k^t \begin{bmatrix} \mathbf{0} \\ \mathbf{d} \end{bmatrix})$$

The bilingual case above can be easily extended to multi-linguistic cases. When the training corpus has more than one target language, the matrix **W** will be composed of all the term-by-document matrices of the languages involved. Suppose that the training corpus consists of l languages, and for each language L_i we choose m_i terms and the same n matching documents. We define $\mathbf{A}_i$ as the term-by-document matrix for the training documents in the i-th language. That the documents of different languages are “matching” means out of n training documents of any language the i-th one has the same content but in different languages. Accordingly we align the corresponding columns representing those matching documents to the same position in $\mathbf{A}_i$. Now the term-by-document matrix **W** representing the entire training corpus is defined as

$$\mathbf{W} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_l \end{bmatrix}$$

The query **q** can be in any one of the l languages, for example L_1. And the document **d** in search can be in any other l-1 languages. After SVD analysis on the matrix **W**, the retrieval criterion is

$$sim(\mathbf{q},\mathbf{d}) = \cos(\mathbf{U}_k^t \begin{bmatrix} \mathbf{q} \\ \mathbf{0} \\ \vdots \\ \vdots \\ \mathbf{0} \end{bmatrix}, \mathbf{U}_k^t \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{d} \\ \vdots \\ \mathbf{0} \end{bmatrix})$$

However, when the size of the matrix **W** increases the SVD becomes more and more time-consuming and even infeasible in practical use. This is one of the major problems of LSI approach and it is intensified in the field of MLIR.

3 Simplified LSI approach

LSI for MLIR is based on parallel corpus training. In parallel corpus, each documents in one language has its counterparts in all other languages. This kind of parallel structure makes the corpus documents semantically symmetric. If the symmetry can be embodied in text representation, we could use it to simplify computation with mathematical methods. That is the basic idea of our simplified LSI approach.

In this section we will show that in two particular circumstances the dimension of the matrix under decomposition in MLIR can be reduced to the size of a monolingual matrix. In either of the two circumstances, the term-by-document matrix of each language has some kind of symmetry, which we call weak and strong symmetry form respectively. Theoretically we prove that our simplification will not deteriorate the accuracy of information retrieval. For practical use, we will also discuss how to enhance the symmetry of real world data to approximate the conditions required in theoretical conclusions.

3.1 Simplified LSI for the Weak Symmetry form

The simplification of LSI under the weak symmetry form is based on the following theorem.

Theorem 1

If two matrices $\mathbf{B}\in \square^{mxn}, \mathbf{C}\in \square^{sxn}$ $(m,s\geq n)$ *satisfy that* $\mathbf{B}^t\mathbf{B}=\mathbf{C}^t\mathbf{C}$, *then*

(1) $\mathbf{B}$ *and* $\mathbf{C}$ *have the same singular values*

(2) Let $\sigma_1,\sigma_2,\cdots\sigma_r$ *be the singular values of* $\mathbf{B}$ *and* $\mathbf{C}$. *If all these singular values are different, i.e.* $\sigma_1>\sigma_2>\cdots>\sigma_r>0$, *and we define* $\mathbf{\Sigma}=\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_r)$, *then there must exist SVD*

$$\mathbf{B}=\tilde{\mathbf{P}}\tilde{\mathbf{Q}}\tilde{\mathbf{R}}^t \textit{ and } \mathbf{C}=\tilde{\mathbf{E}}\tilde{\mathbf{F}}\tilde{\mathbf{G}}^t$$

satisfying that

I. $\tilde{\mathbf{Q}}=\begin{bmatrix}\mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\in \square^{mxn}, \tilde{\mathbf{F}}=\begin{bmatrix}\mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\in \square^{sxn};$

II. If we define the first r columns of $\tilde{\mathbf{R}}$ *and* $\tilde{\mathbf{G}}$ *as* $\tilde{\mathbf{R}}_1$ *and* $\tilde{\mathbf{G}}_1$ *respectively, then* $\tilde{\mathbf{R}}_1=\tilde{\mathbf{G}}_1$.

Before proving Theorem 1 we first give a lemma with proof.

Lemma

Define unit matrix

$$\mathbf{M}(m,k)=\mathrm{diag}(\sigma_1,\sigma_2,\cdots\sigma_m)$$
$$\sigma_i=-1,\quad i=k$$
$$\sigma_i=1,\quad i=\text{others}$$

$\mathbf{B}=\mathbf{PQR}^t$ *is an arbitrary SVD on a matrix* $\mathbf{B}\in \square^{mxn}$. *If we define*

$$\tilde{\mathbf{P}}=\mathbf{PM}(m,k),\ \tilde{\mathbf{R}}=\mathbf{RM}(n,k)$$
$$k\leq\min(m,n)$$

then $\mathbf{B}=\tilde{\mathbf{P}}\mathbf{Q}\tilde{\mathbf{R}}^t$ *is also a SVD on the matrix* $\mathbf{B}$.

Proof of Lemma

According to the definition of SVD, $\mathbf{P}$ and $\mathbf{Q}$ are both orthogonal matrices. And the unit matrix $\mathbf{M}(n,k)$ is an symmetric orthogonal matrix. So $\tilde{\mathbf{P}}$ is also an orthogonal matrix. Hence

$$\begin{aligned}\tilde{\mathbf{P}}\mathbf{Q}\tilde{\mathbf{R}}^t &= \mathbf{PM}(m,k)\mathbf{QM}(n,k)^t\mathbf{R}^t\\ &= \mathbf{PM}(m,k)\mathbf{QM}(n,k)\mathbf{R}^t\\ &= \mathbf{P}[\mathbf{M}(m,k)\mathbf{QM}(n,k)]\mathbf{R}^t\end{aligned}$$

For any matrix $\mathbf{C}\in \square^{mxn}$, $\mathbf{M}(m,k)\mathbf{C}$ equals to multiplying the k-th row of matrix $\mathbf{C}$ by -1, and $\mathbf{CM}(m,k)$ equals to multiplying the k-th column of matrix $\mathbf{C}$ by -1. According to the definition of SVD, the matrix $\mathbf{Q}$ is composed of a diagonal matrix $\mathbf{\Sigma}$ and three zero matrices, i.e.

$$\mathbf{Q}=\begin{bmatrix}\mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$$

So $\mathbf{M}(m,k)\mathbf{QM}(n,k)=\mathbf{Q}$

Hence $\mathbf{B}=\tilde{\mathbf{P}}\mathbf{Q}\tilde{\mathbf{R}}^t$ holds true, i.e. $\mathbf{B}=\tilde{\mathbf{P}}\mathbf{Q}\tilde{\mathbf{R}}^t$ is also a SVD on the matrix $\mathbf{B}$.

Then we prove Theorem 1.

Proof of Theorem 1

(1) According to the definition of SVD, the singular values of matrix **B** are the square roots of the positive eigenvalues of matrix $\mathbf{B}^t\mathbf{B}$, and the singular values of matrix **C** are the square roots of the positive eigenvalues of matrix $\mathbf{C}^t\mathbf{C}$. Because of $\mathbf{B}^t\mathbf{B} = \mathbf{C}^t\mathbf{C}$, **B** and **C** have the same singular values.

(2) Let $\mathbf{B} = \mathbf{PQR}^t$ and $\mathbf{C} = \mathbf{EFG}^t$ be two arbitrary SVDs of matrices **B** and **C**. Here $\mathbf{P}, \mathbf{R}, \mathbf{E}$ and $\mathbf{G}$ are all orthogonal matrices.

According to (1) and the definition of SVD, we have $\mathbf{Q} = \mathbf{F} = \begin{bmatrix} \mathbf{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$. Let us define

$\mathbf{S} = \mathbf{Q}^t\mathbf{Q} = \mathbf{F}^t\mathbf{F} = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_r^2, 0, \cdots, 0)$.

Because of $\mathbf{B}^t\mathbf{B} = \mathbf{C}^t\mathbf{C}$, i.e.

$$(\mathbf{PQR}^t)^t\mathbf{PQR}^t = (\mathbf{EFG}^t)\mathbf{EFG}^t$$

$$\mathbf{RQ}^t\mathbf{QR}^t = \mathbf{GF}^t\mathbf{FG}^t$$

$$\mathbf{RSR}^t = \mathbf{GSG}^t$$

$$\mathbf{SR}^t\mathbf{G} = \mathbf{R}^t\mathbf{GS}$$

Let $\mathbf{L} = \mathbf{R}^t\mathbf{G} = (l_{i,j})_{n\times n}$, and **L** is also an orthogonal matrix. So

$$\mathbf{SL} = \mathbf{LS}$$

i.e.

$$\left[\begin{array}{ccc:c} \sigma_1^2 & & & \\ & \ddots & & \mathbf{0} \\ & & \sigma_r^2 & \\ \hdashline & \mathbf{0} & & \mathbf{0} \end{array}\right]\left[\begin{array}{ccc:cc} l_{1,1} & \cdots & l_{1,r} & \cdots & l_{1,n} \\ \vdots & & \vdots & & \vdots \\ l_{r,1} & \cdots & l_{r,r} & \cdots & l_{r,n} \\ \hdashline \vdots & & \vdots & & \vdots \\ l_{n,1} & \cdots & l_{n,r} & \cdots & l_{n,n} \end{array}\right] = \left[\begin{array}{ccc:cc} l_{1,1} & \cdots & l_{1,r} & \cdots & l_{1,n} \\ \vdots & & \vdots & & \vdots \\ l_{r,1} & \cdots & l_{r,r} & \cdots & l_{r,n} \\ \hdashline \vdots & & \vdots & & \vdots \\ l_{n,1} & \cdots & l_{n,r} & \cdots & l_{n,n} \end{array}\right]\left[\begin{array}{ccc:c} \sigma_1^2 & & & \\ & \ddots & & \mathbf{0} \\ & & \sigma_r^2 & \\ \hdashline & \mathbf{0} & & \mathbf{0} \end{array}\right]$$

$$\left[\begin{array}{ccc:cc} \sigma_1^2 l_{1,1} & \cdots & \sigma_1^2 l_{1,r} & \cdots & \sigma_1^2 l_{1,n} \\ \vdots & & \vdots & & \vdots \\ \sigma_r^2 l_{r,1} & \cdots & \sigma_r^2 l_{r,r} & \cdots & \sigma_r^2 l_{r,n} \\ \hdashline & \mathbf{0} & & & \mathbf{0} \end{array}\right] = \left[\begin{array}{ccc:c} \sigma_1^2 l_{1,1} & \cdots & \sigma_r^2 l_{r,1} & \\ \vdots & & \vdots & \mathbf{0} \\ \sigma_1^2 l_{1,r} & \cdots & \sigma_r^2 l_{r,r} & \\ \hdashline \vdots & \cdots & \vdots & \\ \sigma_1^2 l_{n,1} & \cdots & \sigma_r^2 l_{n,r} & \mathbf{0} \end{array}\right]$$

Compare the corresponding entries of both sides

$$\sigma_i^2 l_{i,j} = \sigma_j^2 l_{i,j} \qquad 1 \le i, j \le r$$

$$\sigma_i^2 l_{i,j} = 0 \qquad 1 \le i \le r,\ r+1 \le j \le n$$

$$\sigma_j^2 l_{i,j} = 0 \qquad r+1 \le i \le n,\ 1 \le j \le r$$

Considering that $\sigma_1 > \sigma_2 > \cdots > \sigma_r > 0$, we get

$$l_{i,j} = 0 \qquad 1 \le i \le r \text{ or } 1 \le j \le r \text{ and } i \ne j$$

So $\mathbf{L}$ can be represented in the form of $\begin{bmatrix} \mathbf{L}_1 & \\ & \mathbf{L}_2 \end{bmatrix}$, and $\mathbf{L}_1 = \text{diag}(l_{1,1}, l_{2,2}, \cdots, l_{r,r})$.

$\mathbf{L}$ is an orthogonal matrix, i.e. $\mathbf{L}^t\mathbf{L} = (\mathbf{R}^t\mathbf{G})^t\mathbf{R}^t\mathbf{G} = \mathbf{G}^t\mathbf{R}\mathbf{R}^t\mathbf{G} = \mathbf{I}$, or

$$\left[\begin{array}{ccc|c} l_{1,1} & & & \\ & \ddots & & \mathbf{0} \\ & & l_{r,r} & \\ \hline & \mathbf{0} & & \mathbf{L}_2^t \end{array}\right] \left[\begin{array}{ccc|c} l_{1,1} & & & \\ & \ddots & & \mathbf{0} \\ & & l_{r,r} & \\ \hline & \mathbf{0} & & \mathbf{L}_2 \end{array}\right] = \left[\begin{array}{ccc|c} l_{1,1}^2 & & & \\ & \ddots & & \mathbf{0} \\ & & lr_{,r}^{\,2} & \\ \hline & \mathbf{0} & & \mathbf{L}_2^t\mathbf{L}_2 \end{array}\right] = \mathbf{I}$$

hence $l_{i,i} = \pm 1, \quad 1 \le i \le r$.

We define the first r columns of $\mathbf{R}$ as matrix $\mathbf{R}_1$, and the rest as matrix $\mathbf{R}_2$; similarly we define the first r columns of $\mathbf{G}$ as $\mathbf{G}_1$, and the rest as $\mathbf{G}_2$, i.e.

$$\mathbf{R} = [\mathbf{R}_1 \quad \mathbf{R}_2], \mathbf{G} = [\mathbf{G}_1 \quad \mathbf{G}_2]$$

$$\mathbf{R}^t\mathbf{G} = \mathbf{L}$$

$$\mathbf{G} = \mathbf{R}\mathbf{L}$$

$$[\mathbf{G}_1 \quad \mathbf{G}_2] = [\mathbf{R}_1 \quad \mathbf{R}_2]\begin{bmatrix} \mathbf{L}_1 & \\ & \mathbf{L}_2 \end{bmatrix} = [\mathbf{R}_1\mathbf{L}_1 \quad \mathbf{R}_2\mathbf{L}_2]$$

So

$$\mathbf{G}_1 = \mathbf{R}_1\mathbf{L}_1$$

The corresponding column of $\mathbf{R}_1$ and $\mathbf{G}_1$ are either equal or opposite, because $\mathbf{L}_1$ is an diagonal matrix with either 1 or -1 as its diagonal entries. Supposing that $l_{k,k} = -1$, according to the *Lemma*, when the k-th columns of $\mathbf{R}$ and $\mathbf{P}$ are multiplied by a factor -1 we get a new form of SVD on matrix $\mathbf{B}$. Repeat such transformation for each -1 entries of $\mathbf{L}_1$ and we will reach a SVD in the following form

$$\mathbf{B} = \tilde{\mathbf{P}}\mathbf{Q}\tilde{\mathbf{R}}^t$$

with the matrix $\tilde{\mathbf{R}}_1$, the first r column of $\tilde{\mathbf{R}}$, satisfying $\mathbf{G}_1 = \tilde{\mathbf{R}}_1$.

Finally, let $\mathbf{E} = \tilde{\mathbf{E}}$, $\mathbf{F} = \tilde{\mathbf{F}}$, $\mathbf{G} = \tilde{\mathbf{G}}$, $\mathbf{Q} = \tilde{\mathbf{Q}}$, and we find SVDs

$$\mathbf{B} = \tilde{\mathbf{P}}\tilde{\mathbf{Q}}\tilde{\mathbf{R}}^t \quad \text{and} \quad \mathbf{C} = \tilde{\mathbf{E}}\tilde{\mathbf{F}}\tilde{\mathbf{G}}^t$$

with $\tilde{\mathbf{R}}_1$ and $\tilde{\mathbf{G}}_1$, the first r columns of $\tilde{\mathbf{R}}$ and $\tilde{\mathbf{G}}$ respectively, satisfying $\tilde{\mathbf{R}}_1 = \tilde{\mathbf{G}}_1$.

The above theorem gives a possible way of reducing SVD computation in a particular case. In the context of multi-linguistic IR the prerequisite condition can be formally written as

Weak Symmetry Form

$\mathbf{A}_1^t\mathbf{A}_1 = \mathbf{A}_2^t\mathbf{A}_2 = \cdots = \mathbf{A}_l^t\mathbf{A}_l$ ($\mathbf{A}_i$ is defined as the term-by-document matrix for the training documents in the i-th language).

If the term-by-document matrices are in the symmetry form above, according to Theorem 1 there exist SVDs

$$\mathbf{A}_i = \mathbf{U}_i\mathbf{S}_i\mathbf{V}_i \approx \mathbf{U}_{ik}\mathbf{S}_{ik}\mathbf{V}_{ik}$$

$$k \le \min(\text{rank}(\mathbf{A}_i)), \quad i = 1, 2, \cdots, l$$

satisfying that

$$\mathbf{S}_{ik} = \mathbf{S}_{jk}, \quad \mathbf{V}_{ik} = \mathbf{V}_{jk} \qquad 1 \le i, j \le l$$

Note that the condition in Theorem 1 that all singular values of $\mathbf{A}_i$ are different is always satisfied because SVD is calculated numerically in practice.

Hence

$$\mathbf{U}_{ik}^t \mathbf{A}_{ik} = \mathbf{S}_{ik} \mathbf{V}_{ik}^t = \mathbf{S}_{jk} \mathbf{V}_{jk}^t = \mathbf{U}_{jk}^t \mathbf{A}_{jk} \qquad 1 \le i, j \le l$$

which means that in two different LSI-generated reduced vector spaces for two different languages, the matching documents have the same vector representations. Therefore if the query **q** is in the language L_l and the document **d** is in the language L_d, the criterion of LSI for weak symmetry form can be simplified as

$$sim(\mathbf{q}, \mathbf{d}) = \cos(\mathbf{U}_{1k}^t \mathbf{q}, \mathbf{U}_{dk}^t \mathbf{d})$$

3.2 Simplified LSI for the Strong Symmetry Form

Hongxing Zou, Dianjun Wang *et al.* have reached some useful conclusions on the SVD for unitary symmetric matrix (Hongxing Zou *et al.*, 2000 & 2002), which we call strong symmetry form in this paper. Their conclusions are recapitulated below as Theorem 2.

Theorem 2

We define

$$\mathbf{W} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_l \end{bmatrix}.$$

where $\mathbf{A}_i \in \square^{m \times n}$, $i = 1, 2, \cdots, l$ *satisfying* $\mathbf{A}_2 = \mathbf{P}_1 \mathbf{A}_1, \mathbf{A}_3 = \mathbf{P}_2 \mathbf{A}_1, \cdots, \mathbf{A}_l = \mathbf{P}_{l-1} \mathbf{A}_1$, *and* $\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_{l-1}$ *are all permutation matrices.* $\mathbf{A}_1 = \mathbf{U}_1 \mathbf{S}_1 \mathbf{V}_1^t$ is an *SVD on* $\mathbf{A}_1$, *where* $\mathbf{S}_1 = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \square^{m \times n}$ *and* $\mathbf{\Sigma}_1 = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_r)$. *Then there must exist SVD* $\mathbf{W} = \mathbf{U}\mathbf{S}\mathbf{V}^t$, *satisfying*

I. $\mathbf{S} = \begin{bmatrix} \mathbf{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \square^{lm \times n}$, $\mathbf{\Sigma} = \mathrm{diag}(\sqrt{l}\sigma_1, \sqrt{l}\sigma_2, \cdots, \sqrt{l}\sigma_r)$

II. $\mathbf{U} = \begin{bmatrix} \frac{1}{\sqrt{l}} \mathbf{U}_1 & \frac{1}{\sqrt{l}} \mathbf{P}_1 \mathbf{U}_1 & \cdots & \frac{1}{\sqrt{l}} \mathbf{P}_{l-1} \mathbf{U}_1 \end{bmatrix}$

III. $\mathbf{V} = \mathbf{V}_1$

It is not difficult to find that Theorem 2 is a special case of Theorem 1. That is why we call their corresponding symmetry form "weak" and "strong" respectively. For this reason we do not prove Theorem 2 here and you can find a wonderful proof for that in Hongxing Zou *et al.*, 2000 & 2002. Accordingly, Theorem 2 gives basis of LSI simplification for the following strong symmetry form.

Strong Symmetry Form

$\mathbf{A}_2 = \mathbf{P}_1 \mathbf{A}_1$, $\mathbf{A}_3 = \mathbf{P}_2 \mathbf{A}_1$, $\cdots$, $\mathbf{A}_l = \mathbf{P}_{l-1} \mathbf{A}_1$ Here $\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_{l-1}$ are all permutation matrices. ($\mathbf{A}_i$ is defined as the term-by-document matrix for the training documents in the i-th language).

Supposing that the query **q** is in the language L_1 and the document **d** is in the language L_d, when the term-by-document matrices in strong symmetry form take $\mathbf{A}_i = \mathbf{U}_i\mathbf{S}_i\mathbf{V}_i \approx \mathbf{U}_{ik}\mathbf{S}_{ik}\mathbf{V}_{ik}$ as their SVDs, the criterion of LSI for the strong symmetry form can be simplified by Theorem 2 as below.

$$
\begin{aligned}
sim(\mathbf{q},\mathbf{d}) &= \cos(\mathbf{U}_k^t \begin{bmatrix} \mathbf{q} \\ \mathbf{0} \\ \vdots \\ \vdots \\ \mathbf{0} \end{bmatrix}, \mathbf{U}_k^t \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{d} \\ \vdots \\ \mathbf{0} \end{bmatrix}) \\
&= \cos(\frac{1}{\sqrt{l}}\begin{bmatrix} \mathbf{U}_{1k}^t & \mathbf{U}_{1k}^t\mathbf{P}_1^t & \cdots & \mathbf{U}_{lk}^t\mathbf{P}_l^t \end{bmatrix} \begin{bmatrix} \mathbf{q} \\ \mathbf{0} \\ \vdots \\ \vdots \\ \mathbf{0} \end{bmatrix}, \frac{1}{\sqrt{l}}\begin{bmatrix} \mathbf{U}_{1k}^t & \mathbf{U}_{1k}^t\mathbf{P}_1^t & \cdots & \mathbf{U}_{lk}^t\mathbf{P}_l^t \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{d} \\ \vdots \\ \mathbf{0} \end{bmatrix}) \\
&= \cos(\frac{1}{\sqrt{l}}\mathbf{U}_{1k}^t\mathbf{q}, \frac{1}{\sqrt{l}}\mathbf{U}_{1k}^t\mathbf{P}_{d-1}^t\mathbf{d}) \\
&= \cos(\mathbf{U}_{1k}^t\mathbf{q}, \mathbf{U}_{dk}^t\mathbf{d})
\end{aligned}
$$

3.3 Discussion on the Two Symmetry Forms

In both symmetry forms when we make SVD analysis we can only decompose matrices of the query language and the target language instead of the several times larger matrix **W** of all the languages. Therefore the LSI approach is simplified.

Both symmetry forms have clear meanings in the context of MLIR. They embody in deferent levels the parallel structure of the training corpus. When we normalize all the term-by-document matrices $\mathbf{A}_i$ by column, each entry of $\mathbf{A}_i^t\mathbf{A}_i$ represents a similarity between two documents in the same *i*-th language. So in weak symmetry form, the similarity between documents in one language is the same with those of the matching documents in another language. That is, though corresponding documents in different language may be quantified differently, for example in vectors of different dimensions, we can simplify LSI approach as long as the consistency of similarity relationships between documents of the same language are preserved across languages. In strong symmetry form, all term-by-document matrices in different languages have the same row vectors but can be arranged in random order. It requires corresponding documents in different languages have the same quantified representations but need no alignment work. The strong symmetry form is actually a special case of the weak symmetry form.

3.4 Enhancement of Symmetry for Real Data

Naturally the term-by-document matrix from real data does not precisely conform to either of the two symmetry forms. But we find at least two ways that can help to enhance the strong form of symmetry. As we know documents are represented as vectors of term weights. Firstly we can change the conventional ways of term selection. When we use word clusters for synonyms rather than words or phrases as terms, the term-by-document matrices appear more symmetric. This is because different languages will have various amounts of synonyms to refer to the same thing or similar things, which diversify the word frequency distribution. However when we gather all these synonyms into a cluster, the frequency of the cluster appearance tends to be consistent.

Also we can choose appropriate weighting scheme, for example binary weighting, to promote the symmetry. Binary weighting ignores the details of term frequency but only cares about whether a term appears or not. Despite of being almost the simplest weighting scheme binary weighting gives a reasonable IR performance in many occasions, for example matching similar documents of different languages in MLIR.

The weak points of the above two ways are very clear. Though statistical method can help us to do word clustering job, it usually brings a lot of calculation and hence counteract with our motive of computation reduction. More precise clustering should be on a semantic basis, which practically depends on a good synonym thesaurus or even manual labor. Binary weighting leads to a loss in IR performance because it ignores details of term distribution. In fact the "strong" symmetry form has a rather strict requirement for term-by-document matrix. We hope to find easy ways for weak form symmetry enhancement. Unfortunately, we haven't so far found any effective way that can give a satisfactory result. The major problem lies in the difficulty to preserve documents similarity across different languages. Ideally if the vectors are proper semantic representation of documents and quantified properly, the similarity of a pair of documents in one language should be identical to that of a matching pair of any other language. But currently all term-frequency-based weighting schemes cannot be a good quantification of the original documents. How to represent a document on a semantic basis rather than on a pure statistical basis is one of the chief goals of our further research.

Symmetry enhancement still does not provide a precise symmetry form. However our experiment suggest that small perturbations will not harm the precision of IR greatly. In Hongxing Zou et al., 2000 & 2002, a perturbation analysis is also given on Theorem 2.

4 Experiments on Strong Symmetry Form

4.1 Test Collection

Our experiment is based on a bilingual corpus consisting of 352 Chinese-English document pairs. All the documents are passages adopted from a bilingual column of an IT weekly newspaper *China Computer World* in a period of about six years. And all passages are introduction or comment on various new IT technologies, about 400-500 English words or 800-900 Chinese characters long. We make 2/3 of the corpus documents as training set and the other 1/3 as test set. Also we manually generate 38 queries for IR test, 19 in Chinese and another corresponding 19 in English.

4.2 Experiment Result

Chinese Word Clusters	English Word Clusters
公司,企业	company, enterprise, corporation
密码,口令	password
改变	change, alter, transform, shift
增长,增加,增添,添加	increment, add, addition, increase
路由,路由器	router
……	……

Table 1 Word Clusters

To enhance symmetry, we use word clusters as terms. For weak symmetry form, we manually generated different amount of word clusters for Chinese and English documents. But we find the symmetry is not good enough to give a reasonable performance on IR test. Further, from all the word clusters we choose 361 pairs for the two languages and now they are in strong symmetry form. *Table 1* gives some examples for word cluster pairs. In fact the term-by-document matrices of the two languages has a difference approximately 9.1% compared to their own, which shows the extent to which the strong symmetry form is violated. The difference is calculated as follows

$$D = \frac{\left\|\mathbf{A}_c - \mathbf{A}_e\right\|^2}{\left\|\mathbf{A}_c\right\| \cdot \left\|\mathbf{A}_e\right\|}$$

where $\mathbf{A}_c$ and $\mathbf{A}_e$ representing the training matrix in Chinese and in English respectively.

We use 19 Chinese queries to search for English documents and then use 19 English queries to search for Chinese documents. For comparison, we use two kinds of weighting scheme -- term frequency (TF) weighting and binary weighting. For either weighting scheme, we compare two retrieval criterions -- LSI criterion and simplified LSI criterion. The average precision-recall curves are shown in *Figure 1.*

From *Figure 1*, we find the best precision comes from our LSI criterion with TF weighting. Binary weighting give a better recall but a worse precision, because binary weighting lost much of the detail about term distribution and hence cannot discriminate as well as TF weighting. Simplified LSI (SLSI for short in *Figure 1*) criterions with both weighting schemes suffer a loss of about 5% in both precision and recall because the symmetry is not precisely kept. However the reduction in calculation and time cost is clear, as shown in *Figure 2*. We simulate the SVD process involved in the LSI approach and the simplified LSI approach for bilingual training corpus and compare the CPU time used for SVD on an $m \times m$ randomly generated matrix and for SVDs on two halves of the same matrix. The result shows that for bilingual cases the simplified LSI approach can save half of the SVD cost approximately. All the calculations are made on a PC with a P4 1.7G CPU and 256M RAM and the CPU time is given by a *Matlab* function.

The experiment results show that by manually clustering words as terms and binary weighting scheme, we can enhance the symmetry of term-by-document matrices of different languages. Although the simplified LSI approach under strong symmetry form suffers a 5% loss or so in IR performance, we approximately reduce the SVD cost to half, which is more desired in some circumstances.

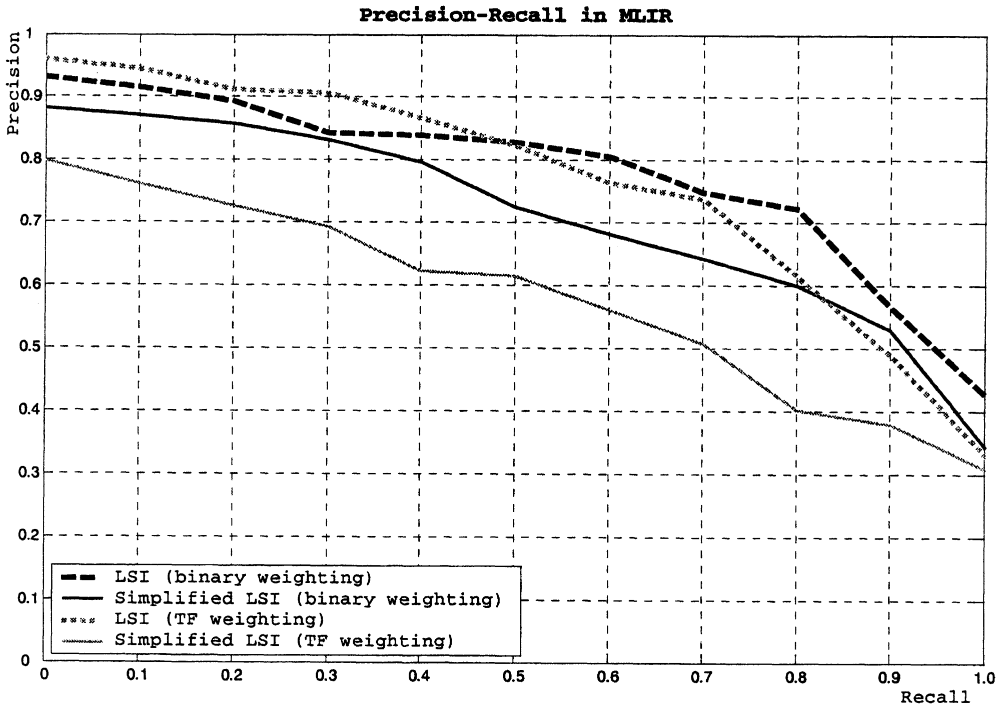

Figure 1 Precision-Recall for Multi-linguistic IR

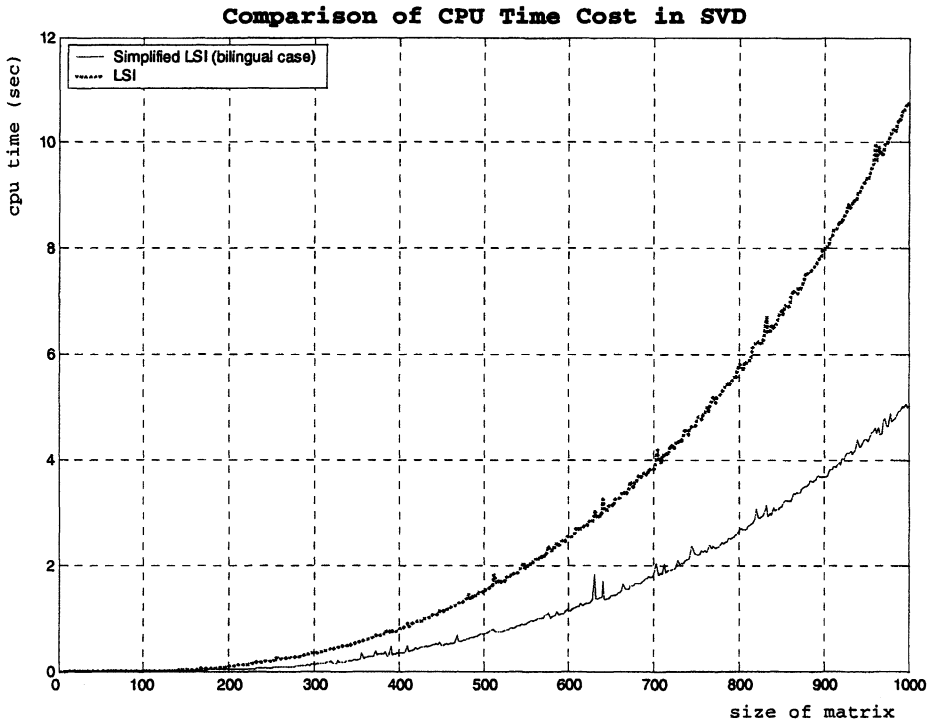

Figure 2 Reduction of CPU Time Using Simplified LSI

5 Conclusions and Future Work

How to avoid or reduce the SVD cost is the major problem of LSI approach, especially for multi-linguistic IR. Theoretically when the term-by-document matrices of training corpus are in either of two symmetry forms

(1) weak form: the similarity between documents in one language is the same with those of the matching documents in other languages

(2) strong form: the vectors representing matching documents in different languages are the same but they can be arranged in arbitrary order to compose the term-by-document matrix

LSI approach can be simplified by reducing the sizes of matrices for SVD analysis. To enhance the symmetry for real data, we propose two ways – word clustering and binary weighting. Our experiment suggests they are two feasible ways to enhance the strong form symmetry to some extent for a reasonable IR performance. But how to reach the weak symmetry form, which is more generalized and hence seems more promising in practical use, needs further research.

References

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

Douglas William Oard. 1996. Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications. *PhD thesis, University of Maryland, College Park, August 1996.*

Hongxing Zou, Dianjun Wang, Qionghai Dai, and Yanda Li. 2000. SVD for row or column symmetric matrix. *Chinese Science Bulletin*, Vol.45, No.22, pp.2042-2044.

Hongxing Zou, Dianjun Wang, Qionghai Dai, and Yanda Li. 2002. Singular value decomposition for unitary symmetric matrix. *Chinese Journal of Electronics.* (Tracking number: 02-730; Accepted for publication).

Y. Yang, J. Carbonell, R. Brown, and R. Frederking. 1997. Translingual information retrieval: Learning from bilingual corpora. *Artificial Intelligence Journal special issue: Best of IJCAI-97.*