A Comparison of SVD, SVR, ADE and IRR for Latent Semantic Indexing

Wen Zhang[1], Xijin Tang[2], and Taketoshi Yoshida[1]

[1] School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Ashahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
{zhangwen,yoshida}@jaist.ac.jp
[2] Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, P.R. China
xjtang@amss.ac.cn

Abstract. Recently, singular value decomposition (SVD) and its variants, which are singular value rescaling (SVR), approximation dimension equalization (ADE) and iterative residual rescaling (IRR), were proposed to conduct the job of latent semantic indexing (LSI). Although they are all based on linear algebraic method for tem-document matrix computation, which is SVD, the basic motivations behind them concerning LSI are different from each other. In this paper, a series of experiments are conducted to examine their effectiveness of LSI for the practical application of text mining, including information retrieval, text categorization and similarity measure. The experimental results demonstrate that SVD and SVR have better performances than other proposed LSI methods in the above mentioned applications. Meanwhile, ADE and IRR, because of the too much difference between their approximation matrix and original term-document matrix in Frobenius norm, can not derive good performances for text mining applications using LSI.

Keywords: Latent Semantic Indexing, Singular Value Decomposition, Singular Value Rescaling, Approximation Dimension Equalization, Iterative Residual Rescaling.

1 Introduction

As computer networks become the backbones of science and economy, enormous quantities of machine readable documents become available. The fact that about 80 percent of business is conducted on unstructured information [1] creates a great demand for the efficient and effective text mining techniques, which aim to discover high quality knowledge from unstructured information. Unfortunately, the usual logic-based programming paradigm has great difficulties in capturing fuzzy and often ambiguous relations in text documents. For this reason, text mining, which is also known as knowledge discovery from texts, is proposed to deal with uncertainness and fuzziness of languages and disclose hidden patterns (knowledge) among documents.

Typically, information is retrieved by literally matching terms in documents with terms of a query. However, lexical matching methods can be inaccurate when they are

Y. Shi et al. (Eds.): MCDM 2009, CCIS 35, pp. 266–274, 2009.

used to match a user's query. Since there are usually many ways to express a given concept (synonymy), the literal terms in a user's query may not match those of a relevant document. In addition, most words have multiple meanings (polysemy and homonym), so terms in a user's query will literally match terms in irrelevant documents. A better approach would allow users to retrieve information on the basis of the conceptual topic or meanings of a document.

Latent Semantic Indexing (LSI) attempts to overcome the problem of lexical matching by using statistically derived conceptual indices instead of individual words for retrieval and assumes that there is some underlying or latent structure in word usage that is partially obscured by variability in word choice [2].

The rest of this paper is organized as follows. Section 2 introduces SVD and recently proposed LSI methods as SVR, ADE and IRR. Section 3 describes information retrieval, text categorization and similarity measure, which are practical applications of text mining used to examine the SVD-based LSI methods. Section 4 conducts a series experiments to show the performances of the SVD-based LSI methods on real datasets, which includes an English and Chinese corpus. Finally, concluding remarks and further research are given in Section 5.

2 SVD-Based LSI Methods

This section introduces the SVD-based LSI methods, which include SVD, SVR, ADE and IRR.

2.1 Singular Value Decomposition

The singular value decomposition is commonly used in the solution of unconstrained linear least square problems, matrix rank estimation, and canonical correlation analysis [3]. Given an $m \times n$ matrix A, where without loss of generality $m \geq n$ and $rank(A) = r$, the singular value decomposition of A, denoted by $SVD(A)$, is defined as

$$A = U\Sigma V^T \tag{1}$$

where $U^T U = V^T V = I_n$ and $\Sigma = diag(\sigma_1,...,\sigma_n)$, $\sigma_i > 0$ for $1 \leq i \leq r$, $\sigma_j > 0$ for $j \geq r+1$. The first r columns of the orthogonal matrices U and V define the orthogonal eigenvector associated with the r nonzero eigenvalues of AA^T and $A^T A$, respectively. The columns of U and V are referred to as the left and right singular vectors, respectively, and the singular values of A are defined as the diagonal elements of Σ which are the nonnegative square roots of the n eigenvalues of AA^T.

2.2 Singular Value Rescaling

The basic idea behind SVR is that the "noise" in original document representation vectors is from the minor vectors, that is, the vectors far from representative vectors.

Thus, we need to augment the influence of representative vectors and reduce the influence of minor vectors in the approximation matrix [4]. Following this idea, SVR adjusts the differences among major dimensions and minor dimensions in the approximation matrix by rescaling the singular values in Σ. The rationale of SVR can be explained as equation 2.

$$A = U\Sigma^{\alpha}V^{T} \tag{2}$$

We can see that the difference of SVR in equation 2 with SVD in equation 1 is that the singular values in Σ are added with an exponential as α. That is, we can regard $\alpha = 1$ is the case in SVR for SVD. If we want to enlarge the differences among major dimensions and minor dimensions, then Σ can be properly adjusted with α more than 1. Whereas, Σ can be adjusted with α less than 1. With this method, the vectors with major semantics in documents can be augmented to distinguish themselves from noisy vectors in documents significantly.

2.3 Iterative Residual Rescaling

Most contents in this Section can be regarded as a simplified introduction of reference [5]. Briefly, IRR conjectures that SVD removes two kinds of "noise" from the original term-document matrix: outlier documents and minor terms. However, if the concentration is on characterizing the relationships of documents in a text collection other than looking for the representative documents in the text collection, that is, we do not want to eliminate the outlier documents from text collection, then, IRR can exert great use of retaining the outlier documents in the approximation matrix while eliminating the minor dimensions (terms).

In details, two aspects in IRR make it different with SVD. The first one is that the document vectors will be rescaled by multiplying a constant which is the exponential to the Euclidian length of the vectors, respectively, with a common rescaling factor. By this method, the residual outlier documents after subtraction from major eigenvectors will be amplified longer and longer. The second difference of IRR from SVD is that only the left eigenvector with the largest eigenvalue will be retained as a basis vector in each of the iterations, and subtracted from the original matrix to produce the residual matrix. With these two differences, the outlier document vectors will become major vectors in the residual matrix and extracted as basis vectors to reconstruct the approximation matrix.

2.4 Approximation Dimension Equalization

Based on the observation that singular values have the characteristic of low-rank-plus-shift structure [6], ADE flattens out the first *k* largest singular values with a fixed value, and uses other small singular values to relatively equalize the dimension weights after SVD decomposition.

ADE extends the ability of SVD to compute the singular vectors and values of a large training matrix by implicitly adding additional ones with relatively equal weights to realize "extrapolating" the singular values [7]. With this method, ADE intends to improve the performance of information retrieval because document vectors will be flattened to become more similar to each other than before. In essence, we can regard

ADE as a method of reducing the discriminative power of some dimensions while enlarging the differences of other dimensions with minor singular values, so that document vectors in a certain range will seem more similar after the ADE process, while maintaining the differences between documents in this range and other documents outside this range.

More specifically, ADE equalizes the singular values in Σ of approximated SVD matrix for term-document matrix. For a matrix A with singular values Σ as shown in Equation 3, and a number $k < r$, we define

$$\tilde{I}_k = I_k + \frac{1}{\sigma_k}\Sigma - \frac{1}{\sigma_k}\Sigma_k \tag{3}$$

This diagonal matrix is illustrated graphically in Figure 1. After obtaining $\tilde{I}_k$, we use it to replace Σ_k to approximate the term-document matrix by Equation 4.

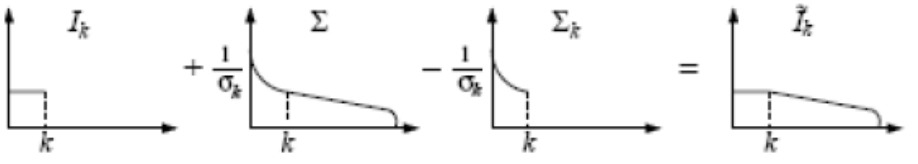

Fig. 1. Combining dimension weights to form $\tilde{I}_k$

$$A_k = U_k \tilde{I}_k V_k^T \tag{4}$$

3 Experiment Design

In this section, parameter settings for above SVD-based LSI methods are specified and we describe information retrieval, text categorization and similarity measure for evaluation of indexing quality.

3.1 Parameter Setting

For SVD, SVDC and ADE, the only required parameter for them to compute latent subspace is preservation rate, which is equal to $k / rank(A)$, where k is the rank of the approximation matrix. In most cases of a term-document matrix A, the number of index terms in A is much larger than the number of documents in A, so we can use the number of documents in A to approximate $rank(A)$ for computation simplicity. Moreover, the preservation rate of ADE is the proportion of singular values in Σ to be equalized. For example, if the preservation rate is set as 0.1 for ADE, then 10 percent of

singular values in Σ with the largest values will be equalized by replacement by an identity matrix. For IRR and SVR, besides the preservation rate, they further need another parameter, a rescaling factor, to compute the latent subspace. To compare document indexing methods at different parameter settings, preservation rate is varied from 0.1 to 1.0 in increments of 0.1 for SVD, SVR and ADE. For SVR, its rescaling factor is set to 1.35, as suggested in [4] for optimal average results in information retrieval. For IRR, its preservation rate is set as 0.1 and its rescaling factor is varied from 1 to 10, the same as in [5]. The preservation rate of IRR is set as 0.1 because R_s will converge to a zero matrix when i increases. That is, the residual matrix approaches a zero matrix when more and more basic vectors are subtracted from the original term-document matrix. Consequently, all the singular vectors extracted at later iterations will be zero vectors if a large preservation rate is set for IRR.

3.2 Information Retrieval

In this research, for English information retrieval, 25 queries, which are uniformly distributed across the 4 categories, are developed to conduct the task of evaluating the semantic qualities of the SVD-based LSI methods. For Chinese information retrieval, 50 queries, which are uniformly distributed across the selected 4 categories, are designed for evaluation.

3.3 Text Categorization

In the experiments, support vector machine with linear kernel is used to categorize the English (Chinese) documents in the corpora. One-against–the-rest approach is used for multi-class categorization and three-fold cross validation is used to average the performance of categorization.

3.4 Similarity Measure

The basic assumption behind similarity measure is that similarity should be higher for any document pair relevant to the same topic (intra-topic pair) than for any pair relevant to different topics (cross-topic pair).

In this research, documents belonging to same category are regarded as having same topics and documents belonging to different category are regarded as cross-topic pairs. Firstly, all the document vectors in a category are taken out and document pairs are established by assembling each document vector in the category and another document vector in the whole corpus. Secondly, cosine similarity is calculated out for each document pair and then all the document pairs are sorted descending by their similarity values. Finally, formula 5 and 6 are used to compute the average precision of similarity measure.

$$precision(p_k) = \frac{\#\text{ of intra - topic pairs } \mathrm{p_j} \text{ where } \mathrm{j \le k}}{\mathrm{k}} \tag{5}$$

$$average\_precision = \frac{\sum_{i=1}^{m} p_i}{\mathrm{m}} \tag{6}$$

Here, p_j denotes the document pair that has the ith largest similarity value of all document pairs. k is varied from 1 to m and m is the number of total document pairs. The larger is the average precision, the more document pairs, in which documents are belonging to the same category, will have larger similarity values than documents pairs in which documents are in different categories. Because documents can have similarities for their similar contents or their statistical properties of identifying its categories, similarity measure is employed to measure the semantic quality and statistical quality of indexing terms synthetically.

4 Results of Experiments

This section describes the experimental results of SVD, SVR, ADE and IRR on three kinds of text mining tasks: information retrieval, text categorization and similarity measure.

4.1 The Corpora

The English corpus, Reuters-21578 distribution 1.0 is used for performance evaluation of our proposed method, which is available online (http://www.research.att.com/~lewis) and can be downloaded freely. It collects 21,578 news from Reuters newswire in 1987. Since 1991, it appeared as Reuters-22173 and was assembled and indexed with 135 categories by the personnel from Reuters Ltd in 1996. In this research, the documents from 4 categories as “crude” (520 documents), “agriculture” (574 documents), “trade” (514 documents) and “interest” (424 documents) are assigned as the target English document collection. That is, 2,042 documents from this corpus are selected for evaluation. After stop-word elimination and stemming processing, 50,837 sentences and 281,111 individual words are contained in these documents.

As for the Chinese corpus, TanCorpV1.0 is used as our benchmark dataset, which is available in the internet (http://www.searchforum.org.cn/tansongbo/corpus.htm). On the whole, this corpus has 14,150 documents with 20 categories from Chinese academic journals concerning computer, agriculture, politics, etc. In this dissertation, documents from 4 categories as “agriculture”, “history”, “politics” and “economy” are fetched out as target Chinese document collection. For each category, 300 documents were selected randomly from original corpus so that totally 1,200 documents were used which have 219,115 sentences and 5,468,301 individual words in sum after morphological analysis.

4.2 Results on Information Retrieval

We can see from Figure 2 that obviously, on Chinese information retrieval, SVD has the best performance among all the SVD-based LSI methods. Meanwhile, on English information retrieval, SVR outperforms all other SVD-based LSI methods. It seems that language type or document genre of the corpus has a decisive effect on performance of SVD and SVR in information retrieval. The semantic quality of SVD is improved by SVR on Chinese documents, while it is worsened by SVR on English documents. That is to say, the effectiveness of augmenting singular values in Σ to

improve semantic quality of document indexing completely depends on the specific documents to be retrieved. The performance of ADE is very stable on Chinese information retrieval at a lower level while on English information retrieval, its local maxima occur at the limits of preservation rates. Its stable performance illustrates that the singular values of ADE are indistinguishable in value from each other even at the preservation rate 0.1. However, its erratic performances in English information retrieval indicate that the semantic quality of ADE is greatly influenced by the number of singular values to be equalized. IRR, on both Chinese and English retrieval, has the poorest performance among all the SVD-based LSI methods. This outcome illustrates that document vectors indexed by IRR do not have the competitive capacity to capture semantics from documents.

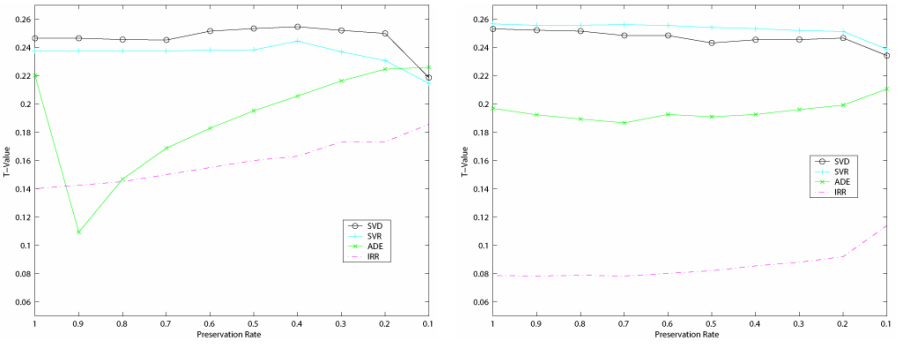

Fig. 2. Performances of SVD-based LSI methods on English (left) and Chinese (right) information retrieval

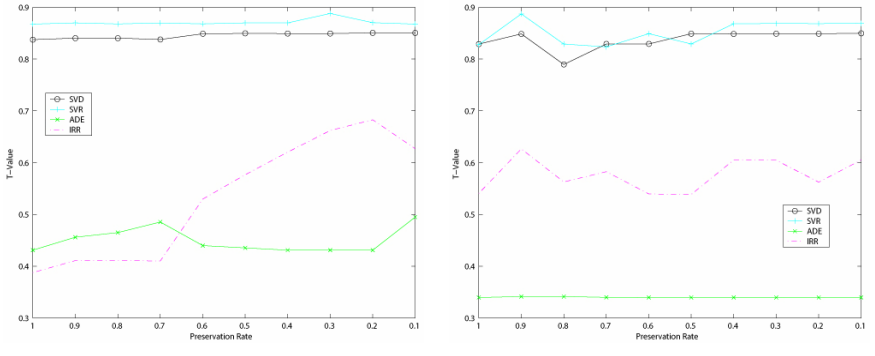

Fig. 3. Performances of SVD-based LSI methods on English (left) and Chinese (right) text categorization

4.3 Results on Text Categorization

We can see from Figure3 that also SVD and SVR outperform other SVD-based LSI methods on both Chinese and English text categorization. On English corpus, SVR is better than SVD while on Chinese corpus, they have comparable performances. The

better performance of SVR over other SVD-based indexing is in that it augments the differences between singular values in Σ. These differences are made by adding an exponential more than 1.0 to the singular values in Σ. Further, it can be deduced that statistical quality of an indexing method can be improved by increasing differences between its singular values in SVD when matrix decomposition is completed. Although ADE and IRR are obviously worse than the other three SVD-based methods on Text Categorization, there are some interesting behaviors in their performances. Regarding the Chinese corpus, IRR outperforms ADE overwhelmingly, but the outcome is the opposite regarding English corpus, where IRR peaks in performance when its rescaling factor is set as 2.0.

4.4 Results on Similarity Measure

We can see Figure 4 that SVD has the best performance on both Chinese and English corpus. SVR ranks the second among all SVD-based LSI methods. That means SVR can appropriately capture relationships between documents and their corresponding categories, but it cannot characterize relationships among documents in a collection excellently. As for ADE on both Chinese and English Similarity Measure, local maxima occur in performance at preservation rates 0.1 and 1.0. At preservation rate 0.1, ADE changes very few singular values in Σ, and at preservation rate 1.0, all the singular values more than 0 in Σ are equalized as 1.0. The results of ADE on Similarity Measure indicates that the best performance of ADE can only occur at two possible preservation rates: the rates 1.0 or 0.0. For IRR, its performance on Similarity Measure is kept stable across all rescaling factors from 1.0 to 10 on both Chinese and English corpus. Thus, we can conclude that for IRR, its rescaling factor is not the dominant factor influencing its capacity on Similarity Measure.

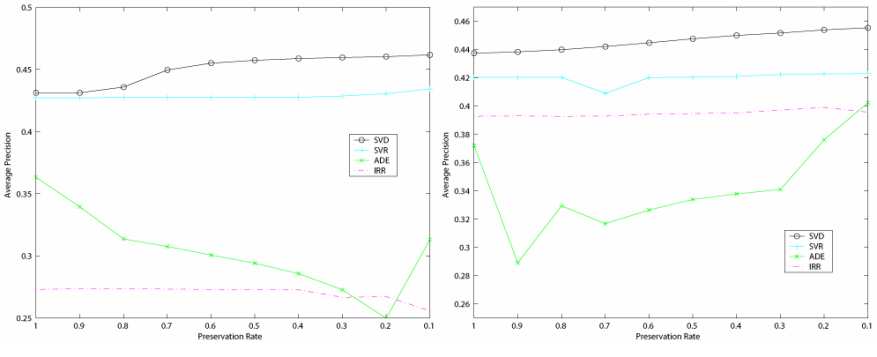

Fig. 4. Performances of SVD-based LSI methods on English (left) and Chinese (right) similarity measure

5 Concluding Remarks

In this paper some experiments are carried out to examine the effectiveness of SVD-based LSI methods comparatively on text mining with two corpora as a Chinese and an English corpus. The experimental results demonstrate that SVD and SVR are

also still better choices than other methods for latent semantic indexing. ADE and IRR can not derive satisfying performances in practical applications of text mining, because of great differences between approximation matrix and original term-document matrix in Frobenius norm.

Although the experimental results have provided us with some clues on latent semantic indexing, a generalized conclusion is not obtained from this examination. Our work is on the initial step and more examination and investigation should be undertaken for more convincing work.

One of research directions supporting text mining is document representation [8]. In order to represent documents appropriately, we should improve not only the statistical quality but also the semantic quality of document indexing. Thus, more attention will be concentrated on the areas of semantic Web and ontology-based knowledge management [9], especially on the work that employs ontology to describe the existing concepts in a collection of texts in order to represent documents more precisely and explore the relationships of concepts from textual resources automatically.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grant No.70571078 and 70221001 and by Ministry of Education, Culture, Sports, Science and Technology of Japan under the "Kanazawa Region, Ishikawa High-Tech Sensing Cluster of Knowledge-Based Cluster Creation Project".

References

1. White, C.: Consolidating, accessing and analyzing unstructured data, http://www.b-eye-network.com/view/2098
2. Berry, M.W., Dumais, S.T., O'Brien, G.W.: Using linear algebra for intelligent information retrieval. SIAM Review 37(4), 573–595 (1995)
3. Golub, G.H., von Loan, C.F.: Matrix Computations, 3rd edn., pp. 72–73. The John Hopkins University Press (1996)
4. Yan, H., Grosky, W.I., Fotouhi, F.: Augmenting the power of LSI in text retrieval: Singular value rescaling. Data & Knowledge Engineering 65(1), 108–125 (2008)
5. Ando, R.K.: Latent Semantic Space: Iterative Scaling Imrpoves Precision of Inter-document Similarity Measurement. In: Proceedings of SIGIR 2000, pp. 216–223 (2000)
6. Zha, H., Marques, O., Simon, H.D.: Large scale SVD and subspace-based methods for information retrieval. In: Ferreira, A., Rolim, J.D.P., Teng, S.-H. (eds.) IRREGULAR 1998. LNCS, vol. 1457, pp. 29–42. Springer, Heidelberg (1998)
7. Jiang, F., Littman, M.L.: Approximate Dimension Equalization in Vector-based Information Retrieval. In: Proceedings of the Seventh International Conference on Machine Learning (ICML 2000), pp. 423–430 (2000)
8. Zhang, W., Yoshida, T., Tang, X.J.: Text classification based on multi-word with support vector machine. Knowledge-based Systems 21(8), 879–886 (2008)
9. Zhang, W., Yoshida, T., Tang, X.J.: Using Ontology to Improve Precision of Terminology Extraction from Documents. Expert Systems with Applications (2009) (in press)