Very Low-Dimensional Latent Semantic Indexing for Local Query Regions

Yinghui Xu Kyoji Umemura
Toyohashi Unversity of Technology Dept. of Information and Computer Sciences
1-1, Hibarigaoka, Toyohashi, Aichi,Japan
xyh@ss.ics.tut.ac.jp umemura@tutics.tut.ac.jp

Abstract

In this paper, we focus on performing LSI on very low SVD dimensions. The results show that there is a nearly linear surface in the local query region. Using low-dimensional LSI on local query region we can capture such a linear surface, obtain much better performance than VSM and come comparably to global LSI. The surprisingly small requirements of the SVD dimension resolve the computation restrictions. Moreover, on the condition that several relevant sample documents are available, application of low-dimensional LSI to these documents yielded comparable IR performance to local RF but in a different manner.

1 Introduction

The increasing size of searchable text collection poses a great challenge to performing the information retrieval (IR) task. Latent Semantic Indexing (LSI) is an enhancement of the familiar Vector Model of IR. It satisfies the IR task through discovering corpus-wide word relationship based on co-occurrence analysis of a whole collection. LSI has been successfully applied to various document collections and has achieved favorable results, sometimes outperforming VSM (Dumais, 1996). However, the principal challenges to applying LSI to large data collections are the cost of computing and storing SVD.

Local analysis of the information in a set of top-ranked documents for the query is one promising way to solve the computationally demanding IR task for a large collection. To solve the computational complexity of LSI, David Hull introduced one interesting method, local LSI, for routing problems(Hull, 1994). The basic idea is: apply the SVD to a set of documents known to be relevant to the query; then all the documents in the collection can be folded into the reduced space of those relevant documents. By concentrating on the local space around the query results, we may be able to compute using flexible and efficient LSI algorithms.

In this paper we put much emphasis on local dimensionality analysis of the local query regions filled with relevant documents. In ideal experimental cases, local LSI involves only the documents known to be relevant to the query. To our surprise, in most of our experiments, local LSI obtains its best IR performance using just one or two SVD dimensions. These interesting results moved us to try performing local LSI with one or two SVD dimensions on the top return sets of VSM in ad-hoc IR experiments. We found that this worked surprisingly well. In a practical setting, local LSI may be regarded as a variation of pseudo relevance feedback (RF). Therefore, the comparative results with local RF are provided in this paper as well. The experiments show that local LSI with one or two SVD dimensions can contribute to expanding the query information in a manner different from traditional local RF.

This paper is organized as follows: Section 2 reviews existing related techniques. Section 3 describes the implementation architecture of the ex-

periments and gives the experiment results. Section 4 explains the result and points out characteristic of the local LSI. Section 5 draws the conclusions.

2 Related works

2.1 Latent Semantic Indexing

Latent semantic indexing (Berry et al., 1999) is one kind of vector-based query-expansion methods that use neither terms nor documents as the orthogonal basis of a semantic space. Instead, it computes the most significant orthogonal dimensions in the term-document matrix of the corpus, via SVD, and projects documents into the low rank subspace thus found. LSI then computes semantic similarity based on the proximity among projected vectors.

LSI uses SVD to factor the term-document training matrix *A* into three factors: $A = U\Sigma V^T = Udiag(\sigma_1, \sigma_2, \cdots, \sigma_n)V^T$ Where $U = (u_1, u_2, \cdots, u_m) \in \Re^{m\times m}$ and $V = (v_1, v_2, \cdots, v_n) \in \Re^{n\times n}$ are unitary matrices (i.e. $U^TU = I, V^TV = I$), whose columns are the left and the right singular vectors of *A* respectively, $\Sigma \in \Re^{m\times n}$ is a diagonal matrix whose diagonal elements are non-negative and arranged in descending order ($\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$), and $p = min(m, n)$. The values $\sigma_1, \sigma_2, \cdots, \sigma_p$ are known as the singular values of *A*, and are the square roots of the eigenvalues of AA^T and A^TA.Suppose the rank of *A* is *r*, then $r \leq p$ and only $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are positive, while the remaining *(p-r)*, if *r<p*, singular values are zero. In LSI retrieval, researchers are only concerned with the first *r* singular values of *A*. LSI uses the structure from SVD to obtain the reduced-dimension form of the training matrix A as its "latent semantic space." Notation for $k \leq r$, defines the reduced-dimension form of *A* to be $A = U\Sigma V^T = Udiag(\sigma_1, \sigma_2, \cdots, \sigma_k, 0, \cdots, 0)V^T$. That is, A_k is obtained by discarding the *r-k* least significant singular values and the corresponding left and right singular vectors of *A* (since they are now multiplied by zeros). Then, the first *k* columns of U that correspond to the *k* largest singular values of *A* together constitute the projection matrix for LSI: $Sim(\vec{d}, \vec{q}) = (A_k^T\vec{d}) \bullet (A_k^T\vec{q})$. Analogous to VSM, the vector representation of a document is the weighted sum of the vector representation of its constituent terms. For document vector d_i and query vector q_i, $A_k^T\vec{d}$ and $A_K^T\vec{q}$ are now the LSI vector representations of that document and query, respectively, in the reduced-dimension vector space. This process is known as "folding in" documents (or queries) into the training space. Actually, LSI assumes that the semantic associations among terms can be found through this one-step analysis of their statistical usage in the collection, and they are implicitly stored in the singular vectors computed by SVD.

2.2 Relevance Feedback

A feedback query creation algorithm developed by Rocchio (Rocchio, 1971) in the mid-1960s has, over the years, proven to be one of the most successful profile learning algorithms. The algorithm is based upon the fact that if the relevance for a query is known, an optimal query vector will maximize the average query-document similarity for the relevant documents, and will simultaneously minimize the average query-document similarity for non-relevant documents. Rocchio shows that an optimal query vector is the difference vector among the centroid vectors for the relevant and non-relevant documents. $\vec{Q}_o = \frac{1}{R}\sum_{D\in Rel.}\vec{D} - \frac{1}{N-R}\sum_{D\notin Rel.}\vec{D}$ where *R* is the number of relevant documents, and *N* is the total number of documents in the collection. Also, all negative components of the resulting optimal query are assigned a zero weight. To maintain focus of the query, researchers have found that it is useful to include the original user-query in the feedback query creation process. Also, coefficients have been introduced in Rocchio's formulation, which control the contribution of the original query, the relevant documents, and the non-relevant documents to the feedback query. These modifications yield the following query reformulation function: $\vec{Q}_n = \alpha \times \vec{Q}_o + \beta \times \frac{1}{R}\sum_{D\in rel}\vec{D} - \gamma \times \frac{1}{N-R}\sum_{D\notin rel}\vec{D}$ In this paper, the experiment results based on the local RF were performed for comparing with the results of Local LSI. The terms in the query are reweighted using the Rocchio formula with $\alpha : \beta : \gamma = 1 : 1 : 0$. As for the local information relevant to the query, they were obtained by extracting several top-ranked documents through the VSM retrieving process in the experiments. Jiang has ever used the similar experiments (VSM+LSI) for Local LSI in his pa-

per "Approximate Dimension Reduction at NTCIR" (Fan and Littmen, 2000).

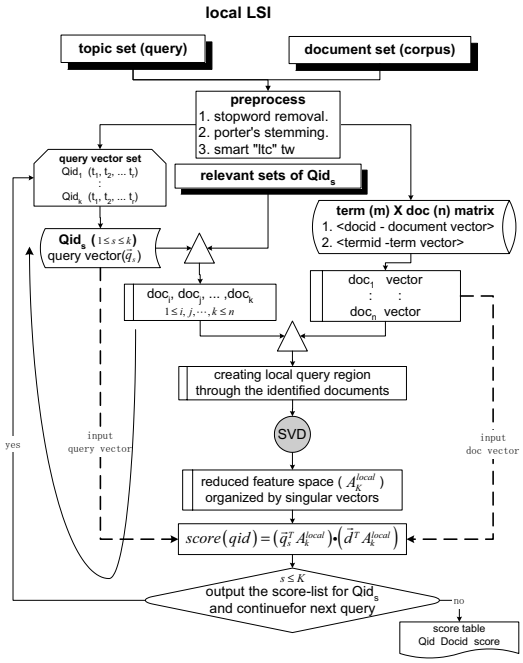

Figure 1: Implementation architecture

3 Experiment Set-up and Results

3.1 Implementation Architecture

Figure 1 shows overall the architecture of the experiment. The procedure is described as follows:

1. Indexing the document collection and query sets

2. Given a query, retrieving some document by the relevant sets. In some cases the relevant sets are derived from the known relevant documents and in other cases we regarded the top returned documents as the relevant sets.

3. performing the singular value decomposition on documents identified in 2.

4. Only a few dimensions for the LSI are retained.

5. Projecting the document vectors and the query vector into the user-cared feature space, and then using the standard Cosine measure to get the final score for this query.

6. Back to the step 2 and continue the analysis on the next query in the same way.

Step 1 is pre-processing procedure for IR system. Only the tag removal, upper case characters transverse, stoplist removal and Porter's stemming were adopted in this phase (Frakes and Baeza-Yates, 1992). Next, the smart "ltc" term weighting scheme (Salton and McGill, 1983) was used to compute the entries of the term document matrix for the collection and entries of the query vector. The second step can be regarded as filter container. In this paper, the three kinds of routine schemes were performed. In the first case, the local space for each query was represented simply by all document vectors, which have already been judged to be relevant (appearing in the relevant judgment file). We note that although it is an ideal case, it may form a useful upper bound on performance. In the second case, we assume the condition that the user provides a reasonable number of relevant documents. In the third case, the local space for each query was built on the top return sets of VSM. The use of the top returned items from VSM is similar to blind feedback or pseudo RF.

3.2 Characteristic of test collection

There are three test collections in our experiments. Two of them, Cranfield and Medlars, are small. The third one is a large-scale test collection, NACSIS. The Cranfield corpus consists of 1,400 documents on aerodynamics and 225 queries, while Medlars consists of 1,033 medical abstracts and 30 queries. Although these two collections are very small, they were used extensively in the past by IR researchers. As for the NACSIS test collection for the IR 1 & 2 (NTCIR 1&NTCIR 2) (Kando, 2001), these documents are abstracts of academic papers presented at meetings hosted by 65 Japanese scientists and linguists. In our experiments, the English Monolingual IR was performed. This collection consists of approximately 320,000 English documents in NTCIR-1 and NTCIR-2.

3.3 Local Routine Experiments (Ideal Case)

We first present the experimental results on the ideal condition. The document vectors already judged to be relevant to the query were used. SVD calculation are performed on the local region organized by

Table 1: Results on the Cran., Med. and NTCIR are shown in terms of ave. precision, precision at document cutoff of 10. Results of the local LSI experiment based on three different SVD dimensions were provided.

	Cranfield			Medlars			NTCIR (E-E) (D run)		
	K	Avr. P-R	R-p	K	Avr. P-R	R-p	K	Avr. P-R	R-p
VSM	-	0.4148	0.3885	-	0.5306	0.5359	-	0.212	0.2277
		+0%	+0%		+0%	+0%		+0%	+0%
G. LSI	200	0.4543	0.4180	80	0.6680	0.6648	-	-	-
		+9% +0%	+8% +0%		+26% +0%	+25% +0%	-	-	-
L. LSI	1	0.8833	0.8243	1	0.8946	0.8139	1	0.6997	0.6508
		+113% +95%	+112% +97%		+69% +34%	+52% +22%		+230%	+186%
	2	0.8607	0.8185	2	0.8769	0.8035	2	0.7062	0.6314
		+108% +90%	+108% +96%		+67% +32%	+52% +20%		+233%	+177%
	3	0.8585	0.8102	3	0.8726	0.8019	3	0.6934	0.6293
		+107% +90%	+108% +96%		+68% +30%	+51% +20%		+228%	+176%

these relevant documents with respect to its query. The IR performance of VSM and global LSI were regarded as the baseline for comparison. As for the NTCIR collection, English-English Monolingual IR was performed and we only extracted the "D" (Description) field of the topic as the query. Due to its large size, only the result of VSM is the baseline. Additionally, to observe the influences of SVD factors on the IR performance for local LSI experiments, results based on LSI dimension from 1 to 3 were also provided for comparison. As we expected, the majority of experimental studies are directed towards obtaining better solutions for the local routine LSI method. In table 1, K represents the SVD dimension for LSI analysis. As for the k value of global LSI, it is the parameter by which LSI yields the best IR performance. The improvement in the average precision of local routine LSI is 113, 69 and 233 percent better than that of VSM on Cranfield, Medlars and NTCIR test collections respectively. The improvement in average precision of local routine LSI is 95 and 34 percent better than that of global LSI on Cranfield with 200 SVD dimensions and Medlars with 80 SVD dimensions respectively. Moreover, in the case of SVD factors equal to 1, we obtain the best IR performance among all cases on the Cranfield and Medlars. While the NTCIR collection obtained its best IR performance with 2 SVD dimensions, there is only a slight difference between the case with 1 singular vector and the case with 2. Such small numbers caught our attention, since they indicate that there is a nearly linear surface in the local region and that the dominant SVD dimensions can capture such surface and yield a good IR performance for local LSI analysis.

To clarify how the local LSI space influences IR performance, we projected the document vectors onto the extracted local routine LSI space and figured out the distribution in figure 2. The data of plots are based on one query from the Medlars collection. Only the largest singular vector was used for the left plot, and the two largest were used for the right. Based on the plots, we find that these dimensions do not vary significantly for the non-relevant documents, Thus, they tend to cluster around the origin. On the other hand, the relevant document space illustrates that local SVD factors are designed to capture their structure.

Since the pre-judged set of documents is generally not available for the ad-hoc query, In this paper, to investigate the efficiency of local LSI using very low dimensions, we continue to do some experiments using different numbers of relevant documents, which were selected from the relevant judgment file. The comparative results based on four cases in which the SVD factors equal 1, 2 and 3, respectively, were shown in Table 2. The second column is the condition, which means that the number of relevant documents belonging to the analyzing object (query) should exceed the value in table. Column 3 "#qry"

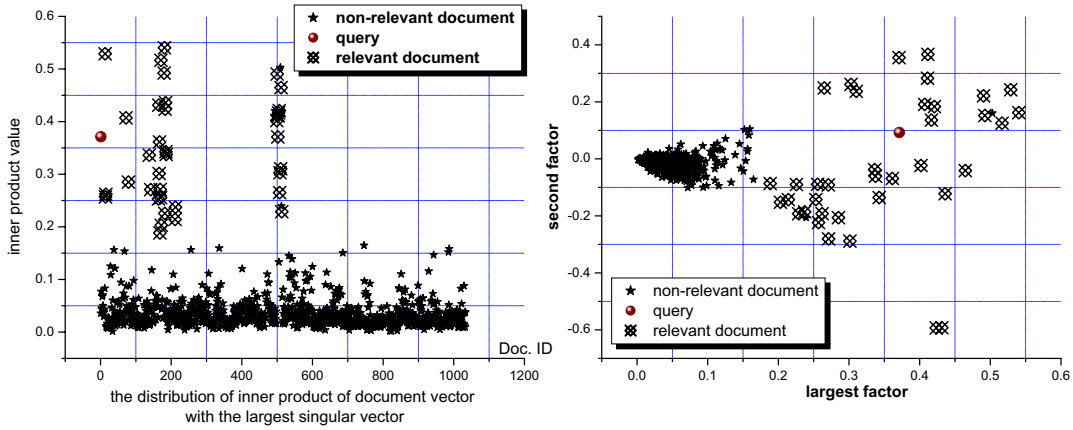

Figure 2: Medlars: document collection distribution after represented by the Local region singular vectors. For the left figure the X-axis is doc.ID and Y are the inner products of the doc vector with the largest singular vectors. X and Y coordinates on the right are the inner product of the document vectors with the first and second largest singular vector, respectively.

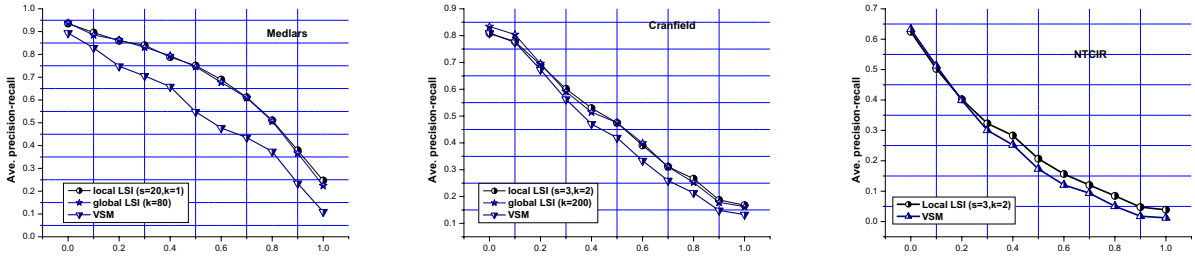

Figure 3: the Ave. precision-recall comparison plots between the best run of local LSI with the baseline VSM and global LSI.

indicates the numbers of queries in the test collection which satisfy the condition appearing in the second column. The fourth column gives the parameter indicating the number of relevant documents to be used for creating the local space of the correspondent query. As we expected, local LSI using one or two SVD dimensions built from the first two singular vectors resulted in the best IR performance in the partially ideal experiments. The comparison of the results was shown in the Table 2.

We know that the most important step in LSI is the phase of SVD. It requires $O(k \times nz^2)$ to find the k leading eigenvectors. The parameter nz is the non-zero entries of the term-by-document matrix. These requirements are unacceptably high for document data sets as the non-zero entries number tens of thousands. According to the LSI analyzing procedure, it includes the SVD phase and the subsequent projecting treatment. For global LSI, the computation complexity can be evaluated by: $O(nz^2k + \#qry \times k^2 \times nz^2 \times qnz^2)$ $k = (100 \sim 300)$ While our approach can be estimated by: $O(\#qry \times [(nz_{loc}^2 k_{loc}) + (k_{loc}^2 \times nz^2 \times qnz^2)])$ $k = 1$ *or* 2 In the above equation, "nz_{loc}" represents the non-zero entries of the local query region. "qnz" are non-zero entities in the query vector. The value of "nz_{loc}" varies with the number of known relevant documents. Note that the difference between these two equations shows clearly that local LSI on small SVD dimensions is much easier to compute than global LSI. According to our observation, it is particularly fast when computing only the largest singular value.

Based on the above experiments, the interesting results and the power of the two largest singular vectors prompted us to try putting the local LSI with one

or two singular dimensions into the practical experiments. In this paper, we used the simplest and most efficient VSM method as the initial retrieval step for extracting the relevant information around the query. We assume that the top-ranked documents obtained by VSM are relevant documents. The details are introduced in section 3.4.

3.4 Ad-hoc local LSI experiment

In this experiment, we note that using the top returned items from VSM is sometimes called blind feedback or pseudo RF. Hence, we borrow the idea of local RF. The expanded query representation was obtained by combining the original query vector with its projecting result on the local SVD dimensions. The equation for expanding the scheme is as follows: $\vec{q}_{new} = \vec{q}_{ori} + A_k^{loc}(A_k^{loc})^T\vec{q}_{ori}$ In the equation:

$$A_k^{loc} = U_k^{loc}\Sigma_k^{loc}(V_k^{loc})^T$$

$$Sim(\vec{d}, \vec{q}_{new}) = \vec{d}\bullet(\vec{q}_{ori} + U_k^{loc}(\Sigma_k^{loc})^2(U_k^{loc})^T\vec{q}_{ori})$$

As for the parameter *k*, representing the SVD dimensionality of the local region, we set its value equal to 1 or 2 in this experiment. At first, to show that local LSI on small dimensions works well is a practical case clearly, we gave the comparable plots between local LSI with the baseline VSM and global LSI. The 11ppt. average precision recall plots of local LSI were figured out for the three test collections in figure 3. The symbol s in the figure represents the sample size and *k* represents the SVD dimension. To our satisfactions, local LSI based query expansion method does much better than VSM and more closely approaches the global LSI.

Next, to investigate the effectiveness of low dimensional LSI on local query region in restructuring the user cared information space, local RF with Rocchio's weights $\alpha : \beta : \gamma = 1 : 1 : 0$, as in Xu and Croft (Xu and Croft, 2000), was used for comparison. Both of them were used on the same sample documents. The difference between them is a twofold one. First, the standard RF formula shown in section 2.2 make use of weighting parameters for query expansion, while our approach does not. Secondly, different combination object was used. The local RF experiment performed in this paper makes use of the centroid of the top s returned document vectors. In our approach, we combine the original query vector with its projecting results on the low local SVD space.

Table 3 shows these results in terms of varying feedback size with one or two SVD dimensions. The first column "sample size" in the table is the value of s according to which we would select the top rank documents . We see that local LSI outperforms local RF for most combinations of sample size and one or two SVD dimensions in the experiment on Medlars. The best run on Medlars using local LSI is 8.4% better than the best in local RF. As for the best run on Cranfield and on NTCIR, local LSI got comparable results with the local RF. In the experiments, we note that with the increasing of sample size, the precision of local LSI decreased more than that of local RF. Based on our analysis, there are two reasons for this. First, In the VSM based local LSI experiments, we assume that the top s documents from the initial retrieval by VSM are relevant, although that assumption does not always hold. In the case where the dominant components of the top s return sets are non-relevant, the maintained SVD dimensions would deviate from the orientation that we preferred. This will influence the following projection procedure greatly. The average precision-recall results of VSM on Cranfield and NTCIR is 0.38 and 0.21, respectively. Neither one is ideal. The second factor is the characteristic of the test collection. The number of relevant documents for query sets ranges from 2 to 40 and from 3 to 170 for the Cranfield and NTCIR, respectively. With such wide range of query sets, some queries don't have enough relevant documents for this strategy to be feasible. Therefore, from the experiment results, it is still reasonable for us to believe that if several relevant sample documents of a query are available, low-dimensional local LSI will be able to achieve comparable performance to local RF.

4 Analysis and discussion

4.1 Local dimensions

One important variable for LSI retrieval is the number of dimensions in the reduced space. In this paper, we found that one or two SVD dimensions are able to represent the structure of the local region that corresponds to the user's interests. The first two largest singular vectors will represent the two major

Table 2: Ave. precision-recall comparing results based on different SVD factors.

Coll.	Cond.	#qry	#sel. Rel.	SVD fact.	Ave. P-R
Cran.	>15 (#rel)	27	10	1	0.6857
				2	0.6667
				3	0.6654
			5	1	0.5749
				2	0.5692
				3	0.5641
Med.	>15 (#rel)	25	10	1	0.7945
				2	0.8007
				3	0.7952
			5	1	0.7160
				2	0.7142
				3	0.7137
NTCIR	>15 (#rel)	23	10	1	0.3899
				2	0.3987
				3	0.3967
			5	1	0.2917
				2	0.2913
				3	0.2883

interests. The local SVD dimensions built on them have the ability to absorb the interests of a query and have no interest in the non-relevant information. It indicates that there is near linear surface in the local query region. That is why local LSI works well on small dimensions, especially on the condition that there is only one dominant interests in the query. Of course, in cases where there is much noisy information in the local region, the SVD dimension may fail to satisfy the true needs of the user. Finally, based on the experiments in this paper, we would like to point out that for performing SVD on a particular local query region, the requirement of the SVD dimension should not be demanding. In our opinion, 2 or less is sufficient to obtain ideal IR performance.

4.2 Size of local region

The size of a local region is also one important parameter for local LSI. We did not do much analysis on how to determine the best size of the local region for local LSI. In the absence of any clear guidelines now, we merely offer some suggestions and an analysis. The local region should be large enough so that it will contain more relevant information. However, there are also several reasons why the local region should not be too large. Adding a large number of non-relevant documents of marginal value will only increase the number of LSI factors needed to describe the local region without improving their quality, and this will only degrade the IR performance. Therefore, as for the size of the local region, it is a tradeoff. According to the experimental results and the analysis in section 3.4, since the local LSI does well on one or two SVD dimensions, so as to avoid influences of non-relevant information brought by more involved documents, it is better to restrict the size of a local region below 30. In the experiments on Medlars, local LSI produced its best run at 20 top return documents. Of course, the threshold for the size of local region should be collection-dependent and experiment-determined. It may also be possible to set the threshold by the performance of the initial retrieval method, but we have not yet analyzed this.

4.3 Advantages

Finally, we would like to point out the advantages of low-dimensional LSI analysis for local query region. Our results compared with VSM and global LSI show clearly that local LSI with low dimensions performs much better than VSM under some sample sets and achieves the comparable IR performance to global LSI. Additionally, because the largest singular vectors are essential for retrieval performance on the local query regions, local LSI approaches the computational complexity of global LSI by using such small SVD dimensions. Despite the fact that local LSI has increased the cost of separate SVD computation for each query, the relative modest requirements of SVD dimension make it feasible for large scale IR task.

Compared with the local RF method, both the local LSI and local RF achieve better results by providing high-centralized relevant information in the local region. Provided that relevant sample documents are used with the same number, local RF is able to make use of the combination of document vectors and a heuristic procedure to improve IR performance, while local LSI makes use of SVD to extract the useful information from the information space. In some sense, this SVD method is more comprehensive than local RF.

Table 3: comparative results of Local LSI and Local relevance feedback on the local region organized by the return sets of VSM on Med., Cran. and NTCIR, respectively. The SVD dimension value for the local LSI is the one from which the best IR performance was obtained at the specific sample size.

#ss.	svd	11 ppt. Ave. P-R		R-p	
(s)	fac.	LLSI	LRF	LLSI	LRF
3	2	0.5858	0.5977	0.5760	0.5816
5	1	0.6417	0.6243	0.6300	0.6198
10	1	0.6577	0.6152	0.6431	0.6093
20	1	0.6764	0.6044	0.6393	0.5845
30	1	0.6598	0.5854	0.6246	0.5699
40	2	0.6514	0.5776	0.6157	0.5722
#ss.	svd	11 ppt. Ave. P-R		R-p	
(s)	fac.	LLSI	LRF	LLSI	LRF
3	2	0.4524	0.4528	0.4206	0.4186
5	2	0.4443	0.4403	0.4203	0.4145
10	2	0.4357	0.4327	0.3981	0.3988
20	2	0.3993	0.4269	0.3571	0.3870
30	2	0.3782	0.4252	0.3345	0.3915
40	2	0.3464	0.4232	0.3106	0.3873
#ss.	svd	11 ppt. Ave. P-R		R-p	
(s)	fac.	LLSI	LRF	LLSI	LRF
3	2	0.2367	0.2346	0.2380	0.2297
5	2	0.2292	0.2302	0.2341	0.2347
10	2	0.2119	0.2249	0.2205	0.2404
20	2	0.1728	0.2110	0.1800	0.2203
30	2	0.1458	0.2026	0.1575	0.2208
40	2	0.1404	0.1978	0.1470	0.2171

5 Conclusion and future work

In this paper, the results show that very low-dimensional LSI on the local query region performs IR task well. Such small dimensional requirements of local LSI make it more attractive, enabling us to better address the computation complexity. We can perform the low-dimensional LSI on several known relevant document spaces to obtain significant improvements in retrieval performance. Moreover, provided that several relevant sample documents are available, local LSI using small dimensions obtains results comparable to the local RF although in a different manner. Our future work will:

1. Continue to study the optimal size of local region for local LSI so as to automatically determine it.
2. Find a more efficient initial retrieval method for obtaining high quality sample sets of each query.

Acknowledgement

This work was supported in The 21st Century COE Program "Intelligent Human Sensing", from the ministry of Education, Culture,Sports,Science and Technology.

References

M. W. Berry, Zlatko Drmax, and Elizabeth R. Jessup. 1999. Matrix, vector space, and information retrieval (technical report). *SIAM Review*, 41:335–362.

S. T. Dumais. 1996. Using for information filtering: Trec-3 experiments. In *In Donna K. Harman, editor, The 3rd Text Retrieval Conference (TREC-3)*, pages 282–291. Department of Commerce, National Institute of Standards and Technology.

J. Fan and M. L. Littmen. 2000. Approximate dimension equalization in vector based information retrieval. In *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan-Kauffman.

W. B. Frakes and R. Baeza-Yates. 1992. *Information retrieval - Data Structure Algorithms*. Prentice Hall, Englewood Cliffs, New Jersey 07632.

D. Hull. 1994. Improving text retrieval for the routing problem using latent semantic indexing. In *In proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 282–291. Association for computing Machnery.

N. Kando. 2001. Clir syetem evaluation at ntcir workshop. National Information (NII) Japan.

J. Rocchio. 1971. Relevance feedback in information retrieval. In *The Smart Retrieval System-Experiments in Automatic Document Processing*, pages 313–323. Englewood Cliffs, NJ, 1971, Prentice-Hall, Inc.

G. Salton and M. J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY.

J. Xu and W. B. Croft. 2000. Improving the effectiveness of informational retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS)*, 18(1).