

A “stereo” document representation for textual information retrieval

Liang Chen*, Jia Zeng

Department of Computer Science

University of Northern British Columbia

3333 University Way Prince George, B.C.

Canada V2N 4Z9

Tel:250-9605838, Fax:250-9605544

Emails: *lchen@ieee.org*, *zeng@unbc.ca*

Naoyuki Tokuda

SunFlare R & D Center, Shinjuku Hirose Bldg,

Yotsuya 4-7, Shinjuku-ku, Tokyo, Japan 160-0004

Email: *tokuda.n@sunflare.co.jp*

Abstract

Encouraged by a significant improvement over LSI (latent semantic indexing) approach in textual information retrieval of the DLSI (differential latent semantic indexing) approach which technically makes use of two term vectors for each document, we have proposed a concept of stereo, or multi-perspective, document representation, which is expected to be effective for most of textual information retrieval approaches based on vector space model. We show that the new representation based on two or more “pictures” of each document taken from different view angles contributes to the enhanced performance of textual document retrieval by enhanced capability of extracting and capturing more individualistic features of the document. A Student *t*-test on experimental results on the standard Time and ADI corpora proves that the improvements of the retrieval performances of LSI/standard term vector algorithms based on multi-perspective document representation over those based on traditional single document representation are significant.

*Corresponding Author

Keywords latent semantic index, information retrieval, differential latent semantic index, textual document

I. INTRODUCTION

It is a standard model, being called vector space model, in textual document information retrieval and classification to regard a document as a “bag of words/terms” and represent it as a term vector (Landauer, 2002). To complete a term vector representation of a document, the statistical information of terms, such as the frequencies of the terms in the document, is considered. We may also take the frequencies of terms appearing in the document collection into consideration when we construct the term vectors. A simple term vector approach extracts the similarity between two documents or most often a document and a query by measuring the angle between their term vectors.

Because of the high dimensionality of the term vectors, dimension reduction approaches, such as support vector machine and principle component analysis, etc., are very frequently incorporated. The principle component analysis in textual information retrieval and classification gives birth to the so-called latent semantic indexing (LSI) approach (Deerwester, Dumais, Landauer, Furnas & Harshman, 1990; Dumais, 1991) and the so-called differential latent semantic indexing (DLSI) approach (Chen, Tokuda & Nagai, 2003; Chen, Tokuda & Nagai, 2001).

In LSI method, a term vector of a document is projected onto a reduced dimension space. This projection is measured and used as the representation of the document, and the distance from the term vector to the space is neglected. Thus, evaluating the similarity between a pair of documents is converted to be the measurement of term vectors in terms of their projection. By doing so, however, the LSI method will lose some flexibility in grasping the special features of documents. Whereas the DLSI method not only exploits the projection of a so-called differential term vector onto a reduced dimension space, but also evaluates the distance of the vector to the space (Chen, Tokuda & Nagai, 2003; Chen, Tokuda & Nagai, 2001)¹. Therefore, by successfully capturing the special features of documents by making use of the distances of term vectors to the reduced space, which the LSI approach ignores, the DLSI approach has been shown to have distinctly improved retrieval performances over the LSI method.

In technically details, for information retrieval purpose, the DLSI approach assigns two term vectors to each document, and introduces the concepts of an intra differential term vector and an

extra differential term vector, one measuring a difference between term vectors that are associated with the same document and the other that (a difference between term vectors that are associated with different documents. As we pointed out earlier, we believe that the improvement of the DLSI method over the LSI method benefits mainly from the use of the distances of the differential term vectors to the reduced dimension space. However, we realize that, in the DLSI method, each document has two distinct term vector representations, which is different from any of the traditional vector space method. This observation leads to an intuitive question of whether by using two or more representations provides any potential benefit in document processing? In other words is it helpful if we can view a document from different perspectives? These questions drive us to the idea of stereo (multi-perspective) document representation for textual documents.

Human experience has shown that multi-perspective descriptions have always been beneficial in visual recognition and acoustic recognition. We know that two eyes are required for stereographic recognition of objects, which gives our sense of depth perception. As well, two ears are necessary for distinguishing minor differences between music styles. This leads to our belief that perceiving objects from two or more different angles is always an advantage. In addition to the daily life experience, there are some pieces of evidences that multi-perspective descriptions are also useful in information research, besides the DLSI applications that uses two term vectors for a same document. Researchers in the area of face recognition has realized that the system usually gets better performance if, in the database, each person is associated with more than one picture (Wu & Zhou, 2002). As well, data fusion, a method which is based on the idea of integrating many answers to a question into a single best answer, has also indicated success of its application in meta-search (Montague, 2002). Meta-search combines the results from many different search engines to produce a single list. Each of these search engines may exploit different search strategies/algorithms, thus we can interpret it in a way that each document is viewed from different ways by many different search strategies / algorithms². Experiments have shown that meta-search can significantly improve the raw performance of the input search engines (Montague, 2002). Researchers have also achieved success in information retrieval for structured documents such as HTML documents, by using language models estimated from multiple structural information, such as in-links, titles, URLs, out-degrees, presented in the structured documents (Ogilvie & Callan, 2003; Ogilvie & Callan, 2004; Zhang, Song, Lin, Ma,

Jiang, Jin, & et. al., 2003).

Having observed a multiple of evidences supporting the idea of using multiple algorithms and using multiple structural information, we propose here this idea of stereo (multi-perspective) representation for textual documents, which are considered as “bags of words/terms and are not required to have any “structural information, for information retrieval purpose. We are urged to find out if it is always helpful to observe one document in two or more angles by assigning a textual document with two or more term vectors based on the partitioning the “bag of words into two or more bags so that the measuring the similarity of a query and a document becomes the integrating of the values of similarities between the query and the multiple “bags of words of the document.

The concept of stereo document representation is wider than the idea of DLSI approach. While the DLSI requires us to locate an independent portion for each document in the document collection, such as a summary or an abstract of the document, the stereo representation does not put any such constraints. More importantly, DLSI approach is an individual approach that technically requiring two term vectors for each documents; the stereo document representation here is a general principle that we expect to be able to be applied/adopted by many textual document retrieval approaches based on vector space model.

The purpose of our paper is to show that the method of stereo (multi-perspective) document representation which we proposed, can actually improve information retrieval performances. The rest of this paper is organized as follows: in Section II, we will provide the detailed description of the stereo document representation and the approach for using the stereo document representation in information retrieval. The experiments will be given in Section III, and in the conclusion we will present a summary of the paper as well as projecting our future research directions.

II. METHODS

As it was explained in Section I, the main idea of a multi-perspective document representation is to use two or more term vectors for each document. In this section we will first introduce the traditional term vector model for information retrieval, followed by the description of the multi-perspective document representation. In order to facilitate information retrieval, a document representation should be accompanied by a similarity measurement that is vital in measuring the

relevance of documents. Therefore we will also introduce the similarity metrics for a query and a document using the multi-perspective representation.

A. *Traditional Vector-Space Model*

Each document in a data collection can be represented as a term vector by exploiting the frequencies of the terms appearing in the document. Consider a data collection of d documents and t terms. For each document j , we assign a vector $(a_{1j}, a_{2j}, \dots, a_{tj})^T$, where a_{ij} reflects the importance of term i in representing the semantics or meaning of document j .

A query can be regarded as a short document, thus is also represented as a term vector.

In the traditional vector-space model, the similarity between a document and a query is measured by the cosine of the angle between the term vectors of the document and the query, or the angle between the projections of these term vectors in a dimension-reduced space.

B. *“Stereo” document representation*

We believe that, from the information retrieval point of view, it is not necessary for us to “read” the entire context of a document before we make a decision on its similarity to a posed query. *Reading one part of a document may be enough for judging if the document is relevant to a query; reading two or more parts can enhance the confidence of a decision.* Our stereo document representation requires the use of different ways to extract information out of one document. This can be interpreted as applying different perspectives on the document. In order to do so, we split the document into different sub-files, where each sub-file conveys partial information of the original document.

However, we do not have a fixed definition on how to create sub-files. We only require that a sub-file created should contain the information in the form of a “word” bag that is rich enough to represent the original document such that it can be used for information retrieval purposes. A sub-file of a document can be the abstract of the document (if an abstract is available), or for example odd numbered sentences. The sub-files of the same document may contain overlapping sentences, just as pictures of a person taken from different angles may have overlapping parts.

Assume we have a data collection of n documents D_1, D_2, \dots, D_n . We extract the information of each document from p different perspectives so that there are p term vectors, namely,

$D_{j_1}, D_{j_2}, \dots, D_{j_p}$ for each document D_j . As a total, we get np term vectors for the document collection.

C. Similarity measurement

Although the similarity of two term vectors is usually measured as the angle in (reduced or not reduced) term vector space, we shall define a similarity metric for measuring the similarity between a query and an entire document D_j . There are indeed many different ways to define the similarity between a document and a query. Comparison of the effectiveness of different definitions of similarity metrics is beyond the purpose of this paper; among many possibilities, we merely adopt the following two definitions in the paper, leaving the examining of the other definitions to future research.

The first approach is the most direct strategy. It takes an average of the similarities evaluated between different term vectors and a query. Suppose the similarities between term vectors D_{j_i} , $i = 1, 2, \dots, p$, of document D_j and query q are $\text{sim}(D_{j_i}, q)$, we define the similarity between D_j and q as:

$$\text{sim}(D_j, q) = \frac{\sum_{i=1}^p \text{sim}(D_{j_i}, q)}{n_j}. \quad (1)$$

Another strategy is based on the idea that we interpret document j 's sub-files to be the member documents in cluster j . Therefore the retrieving process can be converted into a classification problem. We can then apply a voting approach proposed by Cohen and Hirsh (1998), which uses the noisy-or operation to combine the similarity values of all the documents in one cluster to arrive at one single value per class. Therefore, the similarity between D_j and q is defined as:

$$\text{sim}(D_j, q) = 1 - \sum_{i=1}^p (1 - \text{sim}(D_{j_i}, q)). \quad (2)$$

III. EXPERIMENTS

We have conducted our experiments on two standard document collections, TIME and ADI, where queries and relevance judgment are readily available. In our experiments, we will compare the retrieval performances of the standard term vector approach and the term vector approach which uses stereo term vector representation. As well, we have compared the performances of standard LSI approach and the LSI approach using the stereo term vector representation.

A. Data collections

The TIME collection consists of articles from Time magazine's world news section in 1963. ADI is a test collection of document abstracts from library science and related areas. Table I shows some characteristics of these data collections.

(Put Table I here)

As in most IR systems, we applied pre-processing on the textual collections. We used a negative dictionary, SMART³ stop list to remove common words ("and", "the", etc).

In order to make a fair comparison with the standard LSI method, which has been reported to be used on the same test collections by Dumais (1991), the words that are not in the SMART stop list and occur more than once in the collection are included as effective terms for constructing the term vectors of documents. We split each document into two sub-files and filter the text with the list of effective terms. Following Deerwester, Dumais, Landauer, Furnas & Harshman (1990), we only use the frequencies of terms in a file or a sub-file to construct the term vector for a document. In other words, we regard the global weight of each term to be 1 in constructing term vectors.

B. Evaluation Criteria

The performance of an IR system is often summarized in terms of precision and recall. *Precision* is the number of relevant documents retrieved over the number of all the retrieved documents. *Recall* is the number of relevant documents retrieved over the number of all the relevant documents. We have evaluated our system by the combination of both parameters. It is called *average precision*, which is the average of the precision values over certain recall levels. Following Dumais (1991), in our experiment, the average precision over the three recall levels of 0.25, 0.50, 0.75, has been used in our experiments.

C. Experimental Results

We begin with the setting up of term vectors, where each element in a term vector is the frequency with which a term occurs in a document.

Due to the fact that the documents in the collections are actually quite short and no abstracts are available, we construct sub-files of a document in the following way. Let us denote o to be

the number of overlapping sentences, and denote p to be the number of perspectives. For every $o + p$ sentences in a document, we assign o sentences to all the sub-files, and then assign the remaining p sentences equally into each sub-file. We define r to be the overlapping rate $\frac{o}{o+p}$, which represents the “size” of the overlapped content in the sub-files.

In order to maintain the sub-files to be reasonable sized, the overlapping rates we set for these two collections are slightly different, due to the difference of the average size of documents in the collections (see Table I). For the TIME collection, we choose $r = 1/2$. For the ADI collection, we choose $r = 5/7$.⁴

1) *Standard term vector approach versus the term vector approach using the multi-perspective representation:* We have conducted experiments with the standard term vector approach, which uses only one term vector for each document. As well, we have also experimented with the modified term vector approaches which use the stereo term vector representation. The results are shown in Table II. Based on the similarity metrics computed by Eqs. 1 and 2, the two schemes of the modified term vector approaches are denoted as S-term-vector-A and S-term-vector-B respectively. Table II demonstrates that the performance of the standard term vector approach can be significantly improved by using the multi-perspective document representation.

(Put Table II here)

2) *The standard LSI approach versus the LSI approach based on the multi-perspective representation:* We have applied the stereo term vector models on the LSI approach using both TIME and ADI collections. We denote the modified LSI approaches as the SLSI (Stereo-LSI) approach in general. Based on the similarity metrics computed by Eqs. 1 and 2, the two variants of SLSI approaches are denoted as SLSI-A and SLSI-B respectively.

The LSI approach uses a singular value decomposition (SVD) algorithm to decompose the term-by-document matrix into a dimension-reduced space. Therefore the similarities between documents and queries can be calculated as the cosine of the angles between their projections on the reduced space. The dimension-reducing rate, which is defined as the dimension of the dimension-reduced space over the entire rank of the term-by-document matrix, is a very important parameter for the LSI approach. Therefore, we have conducted the experiments with different dimension reducing rates. The results on TIME and ADI collections are shown in Figures 1 and 2. The figures clearly indicate that the performance of the LSI approach can be significantly improved by using the multi-perspective document representation.

(Put Fig.1 here)

(Put Fig.2 here)

3) *Test for the Significance Level of Experiments:* We have conducted t-tests to examine if the improved performance of the stereo model remains to be significant in order to ensure that the robustness of the result has not just come about by chance. We exploit the tripled precision values for all queries at all recall levels on each dimension reducing rate using standard LSI/standard term vector approach, and the modified LSI/term vector approaches using the similarity metrics of Eqs. 1 and 2. The paired-sample t-test has been applied to test the null hypothesis that the LSI/simple term vector approach and each modified LSI/stereo term vector approach do not have different performances on each dimension-reducing rate. It should be note that a simple term vector approach is actually equivalent to an LSI approach at the dimension reducing rate of 100%. The levels of significance, which are used to indicate whether the improved results of our experiments have come about through mere chance, are shown in Figs. 3 and 4. We can see that, the improved performances of the modified approaches using the stereo document representation over the original approaches are significant in most of the cases.⁵

(Put Fig.3 here)

(Put Fig.4 here)

We have conducted further t-tests on the tripled samples for all the dimension reducing rates (see Table III). The results show that the improved performances of modified LSI/(term vector) approaches using either Eqs. 1 or 2 over the original LSI/term vector approach are significant, way beyond the 10^{-7} level and the 10^{-23} level for both of the ADI and TIME data collections respectively. These results indicate that, although the significance of the improvement of the stereo document representation in *some* dimension reducing rates are not significant enough, which implies that there are big chances that the improvements have come about through mere coincidences at these dimension reducing rates; by considering the overall performance, in either the TIME collection or the ADI collection, the probability that the improvements come about by chance is extremely small.⁶ Consequently, we can claim that, these t-tests show that using stereo document representation can significantly improve the performance of IR systems.

(Put Table III here)

IV. CONCLUSION

We introduced and developed the multi-perspective model into textual document representation which is expected to be adopted by many information retrieval approaches based on the vector space model. Our experimental results on the standard data collections (TIME and ADI), and the results of the statistical analysis (t-tests) conducted on them, have indicated the effectiveness for the term vector approach and LSI approach using the multi-perspective representations. At our present stage of development, however we still do not know if the observed improvement in the present level of corpus size may decrease or disappear as the corpus size increases. Experiments on large corpus, such as TERC data, will be necessary for further testing.

Among the different sub-files used to generate the multiple stereo term vectors of a document, we have proposed a certain overlap. We believe that an overlapping is especially important for small sized documents. Otherwise, the sub-files may be too small to represent the meaning of the entire document. However, we do not know exactly how the overlapping rate affects the information retrieval performance. Experiments on much larger data collections may shed more light on elucidating the various aspects of document information retrieval performances including an optimal partitioning and selection of sub-files from documents.

It is generally understood that in information retrieval, the performance normally increases with the size of the documents. This does not seem to agree with our improved precisions of the stereo approach, since each perspective of a document introduced in our scheme is necessarily smaller than that of the document. However, we would like to point out that the total size of all perspectives in the stereo approach is actually *not* smaller than that of the original document. We believe that it is the final “fusion” process on each perspective that improves the performance. As it was pointed out by Landauer (2002) , “Matching queries to documents in such a way as to satisfy human searchers that the document has the semantic content they want involves an emulation of human comprehension”. Thanks to genetically endowed intuition, human beings do not have to read an entire document, especially when the documents have many pages, before we can make a decision on whether the document is what we are searching. We believe that the improvement of the stereo approach over the non-stereo approach should be more significant for longer documents rather than for shorter ones. The curves of the significance level illustrated in Fig.3 for the experiments on the relatively longer document collection TIME and in Fig.4

on the relatively shorter document collection ADI, seem to support this idea. However, further experiments on different data collections, especially on larger data collection are necessary to confirm or deny it.

When using the stereo document representation, a strategy should be developed for estimating the similarity between a query and a document represented by multiple term vectors. This paper has introduced two similarity metrics for this purpose, although many other metrics may also be defined. Further research on other strategies is an interesting topic.

The vector space model has not only been used in information retrieval, but also been adapted and explored to a variety of other applications (see e.g., (Landauer & Dumais, 1997)). One such application relates to the quantitative assessment of semantic content within written text (Shapiro & McNamara, 2000; Hu, Cai, Graesser, Louwerse, Penumatsa, Olney, & *et al*, 2003). Our stereo document representation seems to be appropriate for lengthy documents and because of this, we should explore the possibility of its application to these problem areas involving reasonable sized texts.

ACKNOWLEDGMENT

We are most grateful to Dr. Susan Dumais for the detailed explanation on her paper (Dumais, 1991). We wish to thank the anonymous referees for making a number of helpful and constructive comments and suggestions. We also wish to thank the editor-in-chief Dr. Don Kraft for his encouragement. The research of the first two authors is supported by an NSERC discovery grant of Canada.

Footnotes

1. In paper (Chen, Tokuda & Nagai, 2003), the “differential term vector” was erroneously referred as “differential document vector”.
2. Although it is the unlikely that the data collections of the different searching engines are exactly equivalent, that is, a document might not appear in the data collections of some searching engines.
3. SMART stop list is a negative dictionary used in the vector-model based IR system SMART (Salton, 1971). It can be downloaded from: <ftp://ftp.cs.cornell.edu/pub/smart/english.stop> .
4. To simplify the preprocessing, we regard a raw of a document as a sentence. In the TIME data collection, for every 4 sentences in a document, we assign 2 sentences to all the sub-files and then assign the remaining 2 sentences evenly into each sub-file; In the ADI data collection, for every 7 sentences in a document, we assign 5 sentences to all the sub-files and then assign the remaining 2 sentences evenly into each sub-file.
5. The significance levels for experimental results of ADI data collection, are always higher than those of TIME data collection. One main reason is that the number of standard queries for ADI is smaller than that for TIME data collection. The other reason is that the size of each document for TIME is always longer than those of ADI as will be discussed in the conclusion section.
6. The result is actually within our expectation. We can justify this by analyzing two common examples of the applications of the “stereo” concepts: “stereo” audios and “stereo” videos, which are the origins of our “stereo” method. As we may have noticed, the improvement of the perception by using two eyes / ears over single eye / ear may not be significant (i.e., it is negligible so that it could not be ruled out as coincidence) in each of many snapshots / short time periods. However, for a long run, the difference of two eyes/ears and single eye/ear is certainly NOT negligible, thus cannot be taken as a coincidence, i.e., in a statistical sense, significantly different. Using the terminology of statistics, the concept of “significance” is interpreted to be the chance by which the improvements merely come from coincidence. Although the improvements of a stereo model using “stereo” audio/video/document applied on some samples or some parts of a sample set might not be significant, the chance by which the stereo model maintains the improvements (even if “not significant” in some cases) in all these samples / all parts of the sample set, merely come from coincidence should be very small. This actually indicates that the advantage of the stereo method is significant.

REFERENCES

- [1] Chen, L., Tokuda, N. & Nagai, A. (2003). A new differential lsi space-based probabilistic document classifier. *Inform. Process. Lett.*, 88, 203–212.
- [2] Chen, L., Tokuda, N. & Nagai, A. (2001). Probabilistic information retrieval method based on differential latent semantic index space. *IEICE Trans. on Information and Systems*, E84-D, 910–914.
- [3] Cohen, W. & Hirsh, H. (1988). Text categorization using WHIRL. In R. Agrawal & P. Stolorz (Eds.) *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*(pp. 169–173). Menlo Park, California: AAAI Press.
- [4] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41, 391–407.
- [5] Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23, 229–236,
- [6] Hu, X., Cai, Z., Graesser, A. C., Louwerse, M. M., Penumatsa, P., Olney, A. & the Tutoring Research Group (2003). An improved LSA algorithm to evaluate student contributions in tutoring dialogue. In G. Gottlob & T. Walsh (Eds.), *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*(pp. 1489–1491). San Francisco: Morgan Kaufmann.
- [7] Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. In N. Ross (Ed.) *The psychology of learning and motivation*, vol. 41 (pp. 43–84). New York: Academic Press.
- [8] Landauer, T. K. & Dumais, S. T. (1997). A solution to plato’s problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240,
- [9] Montague, M. (2002). Metasearch: Data fusion for document retrieval. Ph.D. dissertation, Dartmouth College, Hanover, New Hampshire.
- [10] Ogilvie, P. & Callan, J. (2003). Combining document representations for known item search. In C. Clarke, G. Cormack, J. Callan, D. Hawking & A. Smeaton (Eds.) *Proceedings of the Twenty Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 143–150). Toronto, Canada: ACM Press.
- [11] Ogilvie, P. & Callan, J. (2004). Combining structural information and the use of priors in mixed named-page and homepage finding. In E. M. Voorhees & L. P. Buckland (Eds.) *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)* (pp.177-184). Gaithersburg, MD: National Institute of Standards and Technology.
- [12] Salton, G. (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Englewood Cliffs, New Jersey: Prentice Hall.
- [13] Shapiro, A. & McNamara, D. (2000). The use of latent semantic analysis as a tool for the quantitative assessment of understanding and knowledge,” *Journal of Educational Computing Research*, 22, 1–36.
- [14] Wu, J. & Zhou, Z.-H. (2002). Face recognition with one training image per person. *Pattern Recognition Letters*, 23, 1711–1719.
- [15] Zhang, M., Song, R, Lin, C., Ma, S., Jiang, Z., Jin, Y., Liu, Y. & Zhao, L. (2003). THU TREC 2002: Web Track Experiments. In E. M. Voorhees & L. P. Buckland (Eds.) *Proceedings of the Eleventh Text REtrieval Conference* (pp.586-590). Gaithersburg, MD: National Institute of Standards and Technology.

TABLE I
CHARACTERISTICS OF DATA COLLECTION

	TIME	ADI
number of documents	425	82
number of queries	83	35
average number of documents relevant to a query	4	5
vocabulary size	20959	1513
average document size	588.12	66.72

TABLE II
PERFORMANCES OF STANDARD TERM VECTOR APPROACH AND THE TERM VECTOR APPROACH ADOPTING
MULTI-PERSPECTIVE REPRESENTATION

Document Collection	Approach	Average Precision
Time	term vector approach	57.92%
	S-term-vector-A approach	58.95%
	S-term-vector-B approach	58.87%
ADI	term vector approach	28.28%
	S-term-vector-A approach	30.95%
	S-term-vector-B approach	30.66%

TABLE III
SIGNIFICANCE LEVEL FOR T-TEST ON ALL THE SAMPLES ON TIME AND ADI COLLECTIONS

Significance Level	TIME	ADI
Improved performances of modified LSI/term vector approach using Eq. 1 over the original LSI/term vector approach	7.95×10^{-26}	4.29×10^{-9}
Improved performances of modified LSI/term vector approach using Eq. 2 over the original LSI/term vector approach	2.13×10^{-24}	1.13×10^{-8}

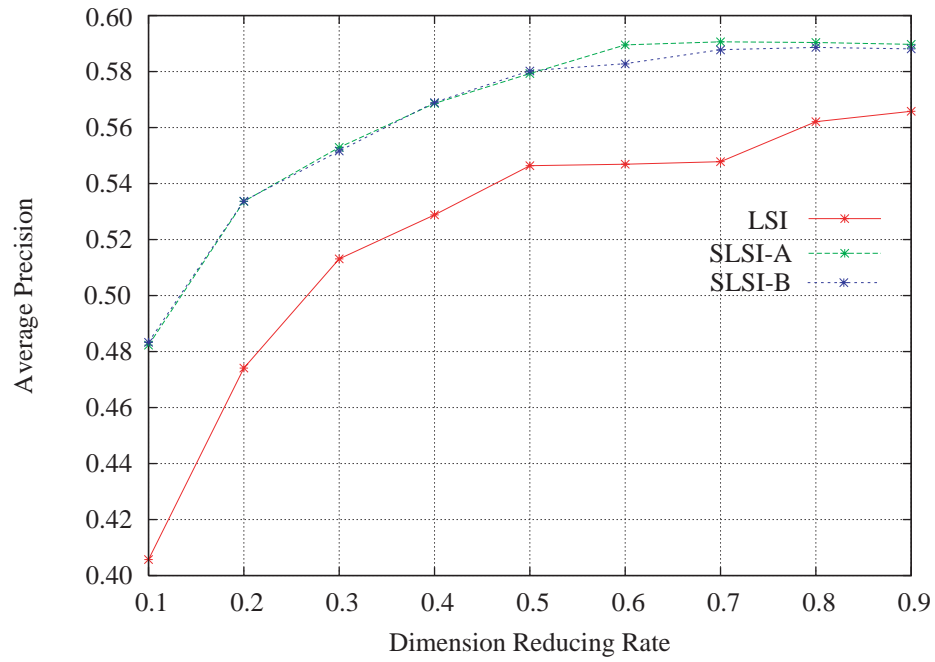


Fig. 1. Performances on TIME using LSI, SLSI-A and SLSI-B

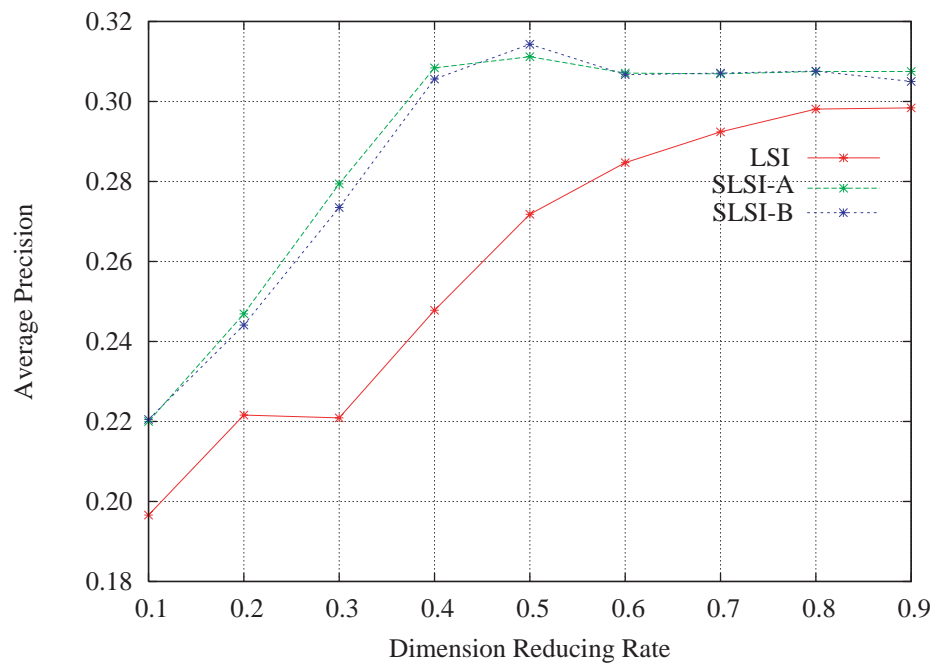


Fig. 2. Performances on ADI using LSI, SLSI-A and SLSI-B

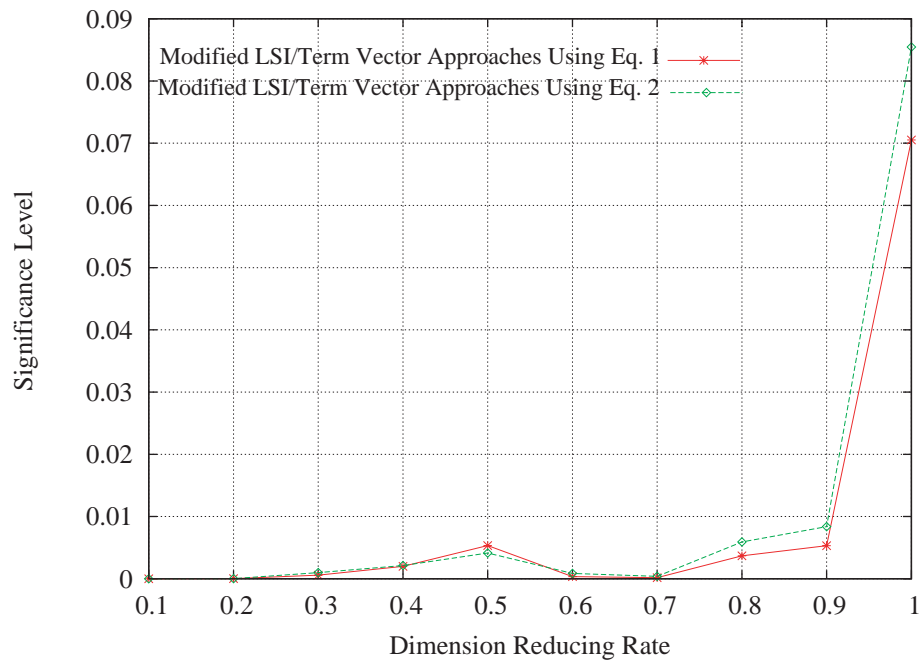


Fig. 3. Significance Level of modified LSI/Term Vector Approaches on TIME Data Collection

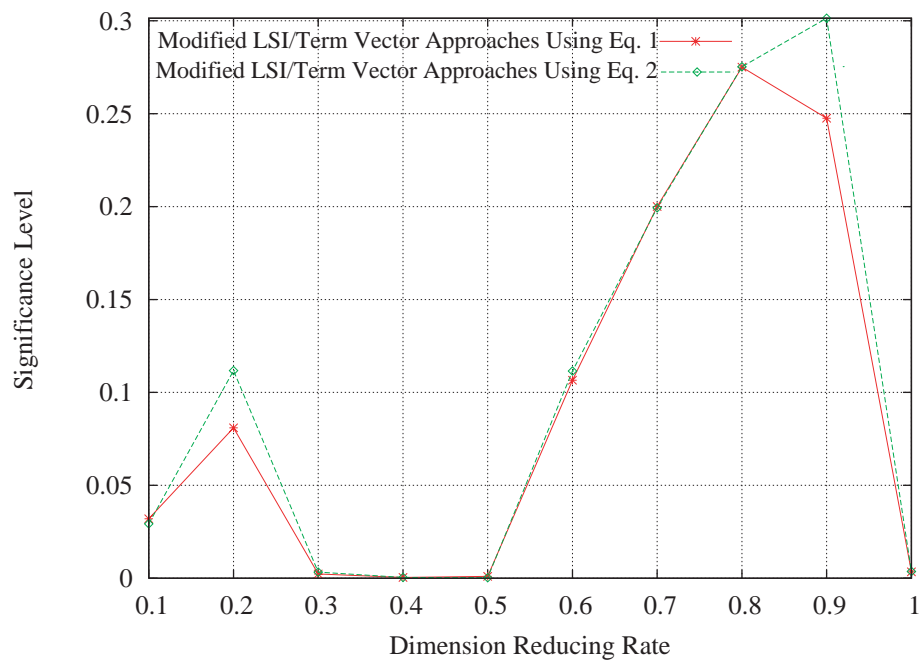


Fig. 4. Significance Level of modified LSI/Term Vector Approaches on ADI Data Collection