DOCUMENT RETRIVAL USING LATENT SEMANTIC INDEXING FOR HINDI LANGUAGE

Jasvir [1],Vaibhav Pratap Singh[2], Amit Kumar Yadav[3]
[1] Department of Computer Science & Engineering,
[2 3] Department of Electronics and Communication
[1 2 3] Babu Sundar Singh Institute of Technology and Management, Lucknow, Uttar Pradesh. INDIA

Abstract : In this paper; topic is document retrieval using Latent Semantic Indexing for Hindi Language .In the past, information mining through web in text searching using English language or other regional language. But for Hindi language there is less SEO support is available. So, for improving the Hindi language text search and there appropriate result, here is method known as Suffix removal method. It helps to gather the information retrieval. The information extracted needs to be expressed by query, created by the user. Documents satisfying the query of the user are considered as "relevant." otherwise "non-relevant." Singular value decomposition is one of the most effective dimensional reduction scheme. It is an statistical techniques that is used in many fields, such as the principal component analysis (PCA) for image processing and face recognition we can conclude that performance of Latent Semantic Indexing has been tested for many small datasets. However, it has not been tested for a larger dataset. In our research, we focused on the performance of latent Semantic Indexing on a large dataset by changing the parameters e.g. stop word lists and term weighting schemes. A single paragraph of about 200 words maximum. For research articles, abstracts should give a pertinent overview of the work. We strongly encourage authors to use the following style of structured abstracts, but without headings: 1) Background: Place the question addressed in a broad context and highlights the purpose of the study; 2) Methods: Describe briefly the main methods or treatments applied; 3) Results: Summarize the article's main findings; and 4) Conclusion: Indicate the main conclusions or interpretations. The abstract should be an objective representation of the article, it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main conclusions.

Keywords: Information Retrieval (IR), Text Retrieval Conference (TREC), Term Weighting Scheme Tf-idf , Stop Words Latent Semantic Analysis (LSA) process, Stemming, LSI, Classical Boolean, Extended Boolean, Vector space, Probabilistic, NLP.

1. Introduction

Latent Semantic Indexing Using Hindi Language in which the input provide in Hindi language and it provides the output in Hindi or relevant form, but why it necessary let’s have look on it. At Engine when we searched or perform any “query” the algorithm remove irrelevant words from query and super-relevant outcome present for you as web or other form particular For e.g. भौतिक विज्ञान क्या है ? here for given query the algorithm remove irrelevant words by First process analyze information resources conceptually in the collection from the contained concepts (effective terms) that make up vocabulary. The output of the first process is input to the second stage, where a translation is employed and a database is created.

2. Related Research Work On Latent Semantic Indexing (LSI)

Here we focus on the existing problems Latent Semantic Indexing (LSI). There are many questions about very large datasets, stop words, and term weighting schemes in Latent semantic analysis. Here we try to answer such questions, mainly for very large datasets.

"Using latent semantic analysis to improve access to textual information" was the first research work on Latent Semantic Analysis by Dumais et al in 1988 Two different datasets were used, with the following descriptions:

- MED: The first database consisted of 1033 medical reference and titles. A 5823*1033 term-document matrix was obtained and effectiveness was evaluated against 30 queries
- CISI: The second dataset consisted of 1460 information science titles. A 5135*1460 term-by-document matrix was obtained and retrieval performance evaluated using 35 queries available with the dataset.

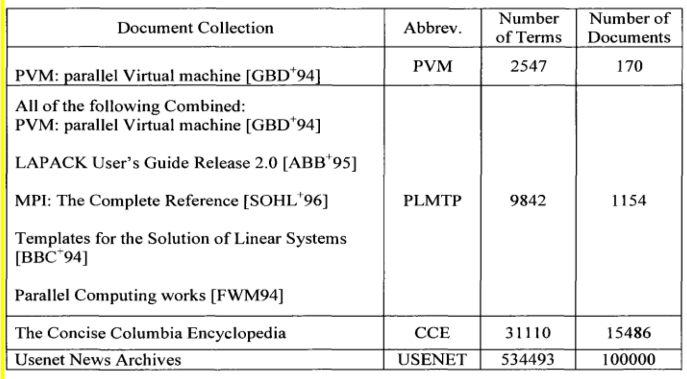
2.1 Indexing By Latent Semantic Analysis

In 1990 Deerwester published a paper titled "Indexing by latent semantic analysis" which shows indexing and retrieval using latent semantic analysis. Here we try to show the use of singular value decomposition. Datasets with queries from the Medlarscollection was used. Stop words were removed but stemming was not done on dataset. After reviewing following flaws were found in their work: LSI was evaluated using small dataset. It was a poor choice because the dataset contains similar documents. Some queries were poorly stated. In the first experiment, 13% average improvement in result claimed. In another experiment, 50%improvement in retrieval performance was claimed.

The above experiment was tested with LSI and result was compared against a term matching method, and the vector method. The SMART and Voorhees systems were used for these purpose information retrieval systems. They have different indexing, term weighting, and query processing procedure. A total of 439 stop words were applied to reduce terms. However, stemming was not used in this experiment. They found that latent semantic indexing method is better than simple term matching in one case and equal in another case when considering two different datasets.

2.2 Large-Scale Information Retrieval Using Latent Semantic Indexing

Todd A. Letsche [Letsche 1996] used the datasets shown in Table 3.1 in his master's thesis at the University of Tennessee, Knoxville, USA.



Document Collection	Abbrev.	Number of Terms	Number of Documents
PVM: parallel Virtual machine [GBD$^+$94]	PVM	2547	170
All of the following Combined: PVM: parallel Virtual machine [GBD$^+$94] LAPACK User's Guide Release 2.0 [ABB$^+$95] MPI: The Complete Reference [SOHL$^+$96] Templates for the Solution of Linear Systems [BBC$^+$94] Parallel Computing works [FWM94]	PLMTP	9842	1154
The Concise Columbia Encyclopedia	CCE	31110	15486
Usenet News Archives	USENET	534493	100000

dataset, no improvement was claimed. Similarity of documents and poor queries caused very poor performance Interventionary studies involving animals or humans, and other studies require ethical approval must list the authority that provided approval and the corresponding ethical approval code.

By reviewing following issues were found:

They removed stop words and for stemming they used Porter's stemmer and reduce the terms. However, they do not mention length of the stop word list. They used Latent Semantic Analysis to search Whole WWW. A dataset of 100,000 documents from USENET was used. However, Letsche claimed that the system was unable to search the dataset. Latent semantic Indexing achieved nearly 30% better performance than Lexical based analysis. This technique does not use singular value decomposition or cosine similarity.

2.3 Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing

Dumais used 2,482 documents for his experiment. They used English and French documents. Here queries written in one language can retrieve other language documents along with the document in original language. Following points were found in reviewing his paper: First of all they have neither used standard dataset nor they have removed stop words or stemmed the words. He compared Latent Semantic Indexing result with human retrieval. In comparison to human text, the performance of this Latent Semantic Indexing system was about 10% poor but 15% better for top 10 retrievals.

2.4 Latent Semantic Indexing-Based Intelligent Information Retrieval System For Digital Libraries

In 2006, in his paper uses 116 journals available in the VIT Vellore India as their own dataset. Following points were noted while reviewing his work: He used small dataset, data set and queries were not standard. There is no indication of the improvements in the performance. They used Porter's algorithm for stemming purpose and did not remove stop words. In this paper they proposed that Latent Semantic Indexing was superior than standard vector space model.

2.5 Singular Value Decomposition And Rank K Approximation

In 2005 Bast & Majumder paper titled "Why Spectral Retrieval Works" studied SVD method by changing the value of K on three different datasets. They wanted to find acceptable value of K. Using this method they choose some values of K and measured the performance of Latent Semantic Indexing. following dataset and corresponding queries were used:

- Time Collection contains 425 documents and 3882 unique terms.
- Reuters Collection contains 21,578 documents and 5701 unique terms.
- Assumed Collection contains 233,445 documents and 99,117 unique terms

Findings of this paper are as follows:

The dataset was large. By varying K’s value it was found that K does not play an important role in the retrieval performance. Larger value of K reduces the relatedness of word pairs.

2.6 Singular Value Decomposition For Text Retrieval

In 2001 Husbands' paper entitled "On the Use of the Singular Value Decomposition for Text Retrieval" explored Latent Semantic indexing performance for large datasets. Following datasets were used:

- MEDLINE/ MED collection: This dataset is a small dataset.
- TREC-6 Dataset: A collection with 115,000 terms and 528,155 documents. This is a large dataset.
- NPL Dataset: NPL dataset contains 11,429 documents and it is a small dataset.

They compare the performance of Latent Semantic Indexing with term matching. For MED data set, Latent Semantic Indexing showed better performance than term matching technique. However, the large TREC-6 dataset Latent Semantic Indexing (K=200) showed poor retrieval.

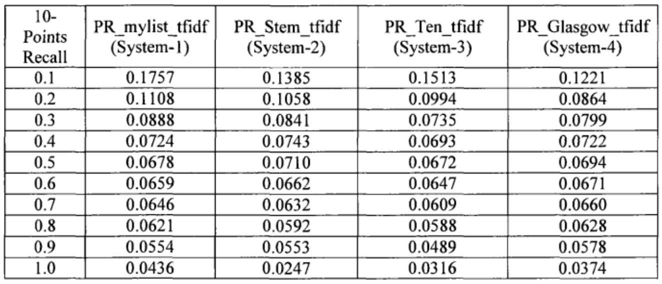
3. Results

The following sections the result of the evaluation

3.1. Result: Experiment-1

The 10-point interpolated precision of four different systems.

The recall precision graph based on the data is shown in



10-Points Recall	PR_mylist_tfidf (System-1)	PR_Stem_tfidf (System-2)	PR_Ten_tfidf (System-3)	PR_Glasgow_tfidf (System-4)
0.1	0.1757	0.1385	0.1513	0.1221
0.2	0.1108	0.1058	0.0994	0.0864
0.3	0.0888	0.0841	0.0735	0.0799
0.4	0.0724	0.0743	0.0693	0.0722
0.5	0.0678	0.0710	0.0672	0.0694
0.6	0.0659	0.0662	0.0647	0.0671
0.7	0.0646	0.0632	0.0609	0.0660
0.8	0.0621	0.0592	0.0588	0.0628
0.9	0.0554	0.0553	0.0489	0.0578
1.0	0.0436	0.0247	0.0316	0.0374

Table 3.1-10-point Interpolated Precision of Four Different Systems

In the Table value of 0.1 represents the top 10% documents that are relevant. As an example precision with top 10% of the documents is 17.57%. This value is calculated using interpolating the precision values of all queries used for this thesis at the standard recall value 0.1. We can compare retrieval systems with different parameters in terms of precision in different standard recall points e.g., 0.1, 0.2 etc. For example, at recall point 0.3 means, the precision values for system-1 is 8.88% and for system-3 is 7.35%. So, if we subtract (8.88-7.35) %=1.53%, means that system-1 shows 1.53% better retrieval performance than sytem-2 for the op30 % retrieved documents.

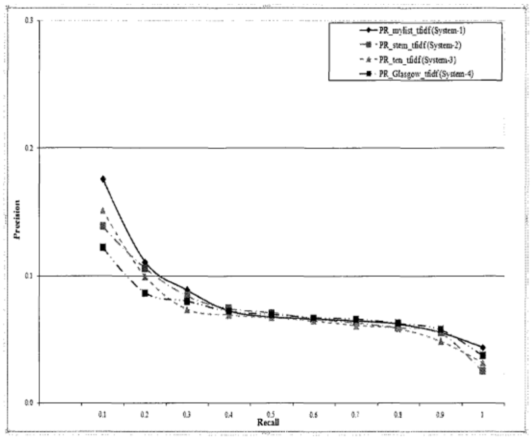
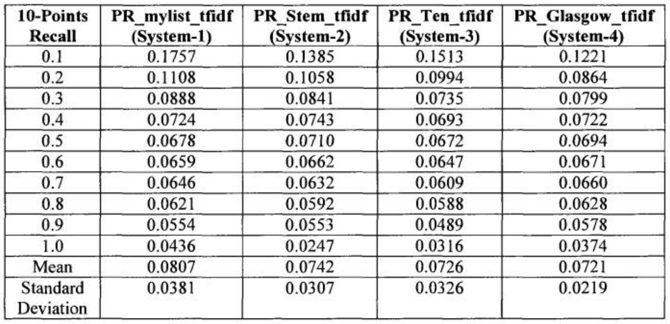


Figure : Recall-Precision Graph for Experiment-1

The system-1 with stop word list provides best result when compared to the other systems. For top 10% retrieval, system-1 shows 5.37% better performance than system-4, 3.68% better performance than system-2, and 2.44% better performance than system-3. However, after top 40% retrieval all the systems showed almost the same retrieval performance. In system-2, we just used Porter's stemmer without removing stop words. It can be seen that the performance of system-2 compared to system-4 is 1.64% better. From the above, it is concluded that in case of latent Semantic Indexing based information retrieval, the stop words is depends upon input data. From the above result we can see thatstop words reduce performance in case of Latent Semantic Indexing-based information retrieval with large dataset.

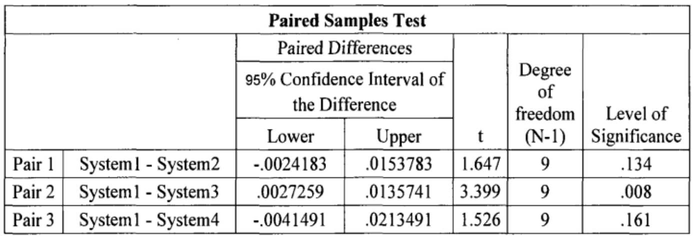
3.2 Paired Sample T-Test

We calculated the t-test using SPSS represents recall-precision of all systems and means and standard deviations of precision. Table 5.5 shows output of the t-test. We compared system-1 with other systems in terms. We found three different level of significance for every system. We found that the t-test produced a result below the threshold value. If we apply t-test to compare different systems, paired sample t-test results indicate minor differences for all cases.



10-Points Recall	PR_mylist_tfidf (System-1)	PR_Stem_tfidf (System-2)	PR_Ten_tfidf (System-3)	PR_Glasgow_tfidf (System-4)
0.1	0.1757	0.1385	0.1513	0.1221
0.2	0.1108	0.1058	0.0994	0.0864
0.3	0.0888	0.0841	0.0735	0.0799
0.4	0.0724	0.0743	0.0693	0.0722
0.5	0.0678	0.0710	0.0672	0.0694
0.6	0.0659	0.0662	0.0647	0.0671
0.7	0.0646	0.0632	0.0609	0.0660
0.8	0.0621	0.0592	0.0588	0.0628
0.9	0.0554	0.0553	0.0489	0.0578
1.0	0.0436	0.0247	0.0316	0.0374
Mean	0.0807	0.0742	0.0726	0.0721
Standard Deviation	0.0381	0.0307	0.0326	0.0219

Table 3.2-10-point Interpolated Precision, Mean and Standard Deviation of Four Systems



Paired Samples Test						
		Paired Differences		t	Degree of freedom (N-1)	Level of Significance
		95% Confidence Interval of the Difference				
		Lower	Upper			
Pair 1	System1 - System2	-.0024183	.0153783	1.647	9	.134
Pair 2	System1 - System3	.0027259	.0135741	3.399	9	.008
Pair 3	System1 - System4	-.0041491	.0213491	1.526	9	.161

Table 3.2 : Comparison among Systems in Terms of T-testg

4. Discussion

Relevance judgment is an important part of Information Retrieval dataset. This help in measuring precision and indicating which documents are relevant to which query. In these sections we discuss the background of TREC-8 Relevance Judgment.

Relevance Judgments are important for test collection. It is necessary for every topic to create a list of relevant documents. TRECs uses pooling method to assemble the relevant document. Here a pool of relevant documents is created by using samples of documents selected by research groups. The pool is judge by the human, who makes simple yes no judgment for each document in the pool. Un-judged documents are considered as un-relevant. This sampling technique is valid since other systems uses ranked retrieval methods, where documents most likely to be relevant returned first.

This method can be concluded as follows:

1. To have researchers and companies to participate.
2. To run all 50 question or query from first to last their system (TREC-8, Topic 401-450).
3. To submit raw retrieval results (Take the top 100 highest ranked documents from each topic e.g., TREC-8: 7100 documents).
4. To combine them keen on the applicant set e.g., TREC-8: 1736 documents.
5. To contain person appraiser judge application of each document.
6. To evaluate results and compile of performance statistics (e.g., TREC-8: 94
Documents) .

5. Conclusions

The Information Retrieval was born in the early 1950s and over the last decade, the field has grown considerably. Several Information Retrieval systems are used by a different no of users. Retrieving information from textual digital data is a challenge to the users where the user is working on a different language, and an Information Retrieval system is a solution to this challenge.

we introduced Latent Semantic Indexing research and presented the problem statement.

We discussed the Latent Semantic based Information Retrieval system. Here we also covers different term weighting schemes, ranked based similarity measurement technique, singular value decomposition, and the idea of recall-precision and t-test.

We survey was done on Latent Semantic Indexing which helped us to identify the research questions on latent Semantic based Information Retrieval System.

We described the characteristics of dataset, and the pre-processing techniques. To address the research questions, we used our own stop words lists for Hindi language and applied Porter's stemmer, with SVD over LSI.

We showed the experimental design, along with the results, its findings. All the results are on the top 10% retrieval. We performed two experiments to measure the performance of Latent Semantic Indexing in case of Hindi dataset. The experiments are given below briefly:

• Different stop word list affects the experiment.

• Different term-weighting schemes also affect the experiment.

In the first experiment, retrieval performance was tested of Latent Semantic Indexing using different stop word lists and without stop words removal. The stop word list we created 5.37% better performance and 4% better than without removing stop word.

In the second experiment, we used 3 different schemes of term-weighing. Term frequency- inverse document frequency schemes showed 9% better performance, and log-entropy showed 5.54% better performance than raw term frequency. Also, term frequency- inverse document frequency showed 3% better performance than log-entropy. In summary, term frequency- inverse document frequency showed better performance. In general, Latent Semantic Indexing has advantages as follows:

1. Both terms and documents are explicitly represented in the same space.

2. Queries and new documents can easily be added.

3. Latent Semantic Indexing uses Singular Value Decomposition to reduce dimension and to remove noise.

4. Latent Semantic Indexing is able to handle polysemy and synonymy.

Possible Extensions of the Experiments

LSI is being used in many of IR and text processing applications. Here we studied the accuracy of Latent Semantic Indexing in document retrieval of Hindi Language. Following are some research works related to this study that might be used in future:

1. Different Weighing scheme can be used

2. New Stop word list can be considered

3. We have worked upon Hindi language, other language can also be considered..

References

[1] [Al-Maskari et al 2008] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. Relevance Judgments between TREC and Non-TREC. Assessors. SIGIR' 08, Singapore, July 20-24, 2008.

[2] [Anderson et al 2009] Anderson, David Ray, Sweeney, Dennis J., and Williams, Thomas Arthur (2009). Statistics for business and economics. 11th ed. Cengage Learning Inc. lorence, KY

[3] [Baeza-Yates &Neto 1999] Ribeiro- Ricardo Baeza-Yates and BerthierRibeiro-Neto. Modern Information Retrieval. Addison Wesley Longman Publishing Co. Inc., first edition, 1999.

[4] [Berry &Dumais 2008] Berry Michael W. &Dumais Susan. Latent Semantic Indexing Web Site, http://www.cs.utk.edu/~lsi/, accesstime: Jan 2008.

[5] [Berry et al 1993]M. Berry, T. Do, G. O'Brien, V. Krishna, and S. Varadhan.Svdpackc (version 1.0) user's guide. Technical Report No.CS-93-194, University of Tennessee, May 1993.

[6] [Chakravarthy&Netserf1995]A. Chakravarthy and K. Haase. Netserf. Using semantic knowledge to find internet archives. In Proceedings of the18th Annual International ACM SIGIR Conference onResearch and Development in Information Retrieval, Seattle, Washington, USA, 1995. Pp. 4-11.

[7] [Chen et al 2001]L. Chen, N. Tokuda, and A. Nagai. Probabilistic information retrieval method based on differential latent semantic index space. IEICE Trans, on Information and Systems, E84-D (7):910-914, 2001.

[8] [Chen et al 2003]L. Chen, N. Tokuda, and A. Nagai. A new differential Isispace-based probabilistic document classifier. Information Processing Letters, 88:203-212, 2003.

[9] [Chen et al 2004]L. Chen, J. Zeng, and J. Pei. Classifying noisy and in complete medical data by a differential latent semantic indexing approach. In Data Mining in Biomedicine. Springer Press, 2004.

[10] [Chen et al 2005]L. Chen, J. Zeng, and N. Tokuda. A "stereo" document representation for textual information retrieval. Journal ofthe American Society for Information Science and Technology, 2005.

[11] [Chen et al 2006]Liang Chen, JiaZeng and NaoyukiTokuda. A "Stereo "Document Representation for Textual Information Retrieval. Journal of the American Society for InformationScience and Technology (JASIST). 57:6, pp. 768—774.2006.

[12] [Cohen & Hirsh 1998] W. Cohen and H. Hirsh. Text categorization using whirl. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pages 169-173,New York, 1988

[13] [Croft et al 2009] W. Bruce Croft, Donald Metzler, and Trevor Strohman. Information Retrieval in Practice. Addison Wesley, 2009.

[14] [Deerwester et al 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T.K., and Harshman, R. A. Indexing by latent semanticanalysis. Journal of the American Society for Information Science, 1990, 41(6), pp. 391-407.

[15] [Dumais et al 1988] Dumais S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. Using latent semantic analysis to improve access to textual information. In Proceedings of CHI'88:Conference on Human Factors in Computing, New York: ACM, 1988. pp. 281-285.

[16] [Dumais et al 1996] Susan T Dumais, Thomas K Landauer, Michael L Littman. Automatic Cross Linguistic Information Retrieval using Latent Semantic Indexing. SIGIR 1996.

[17] [Dumais et al 1997] Susan T. Dumais, Todd A. Letsche, Michael L. Littman,Thomas K. Landauer. Automatic Cross-Language Retrieval Using Latent Semantic Indexing. AAAI-97Spring Symposium Series: Cross- Language Text andSpeech Retrieval, 1997. Stanford University, pp. 18-24.

[18] [Dumais 2004] Dumais, S. Latent Semantic Analysis. ARIST Review of Information Science and Technology, vol. 38, 2004,Chapter 4.

[19] [Fisher & Yates 1995] R.A. Fisher and F. Yates. 6th ed. Statistical tables forbiological, agricultural and medical research. London:Longman Group, 1995.

[20] [Frakes& Yates 1992] W. Frakes and R. Baeza-Yates. Information Retrieval:Data Structures and Algorithms. Prentice Hall, 1992.

[21] [Green grass 2001] E. Greengrass. Information retrieval: A survey. Technical Report DOD Technical Report TR-R52-008-001, 2001.

[22] [Giles et al 2001] Justin T. Giles, Ling Wo, Michael W. Berry. GTP(General Text Parser) Software for Text Mining. CRCPress 2001.

[23] [Harman 1995] D. Harman. Overview of the third text retrieval conference.Third Text REtrieval Conference (TREC-3), pp. 1-19.National Institute of Standards and Technology SpecialPublication 500-207, 1995.

[24] [Harter 1986] S. Harter. Online Information Retrieval: Concepts,Principles, and Techniques. Academic Press, Inc., 1986.

[25] [Hull 1994] D. Hull. Improving text retrieval for the routing problem using latent semantic indexing. 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 282-291, 1994