A Comparative Study of Information Retrieval Models: Latent Semantic Analysis (LSA) vs. Vector Space Model (VSM)

Bhumika Gupta[1†] and Pranjul Dubey[2†]

[1*]Electrical Engineering Department, IIT Madras, Chennai, India.
[2]Engineering Design Department, IIT Madras, Chennai, India.

Contributing authors: ee21s060@smail.iitm.ac.in;
ed21s011@smail.iitm.ac.in;
[†]These authors contributed equally to this work.

Abstract

Information retrieval (IR) is the process of retrieving relevant information from a large corpus of documents. Vector Space Model (VSM),Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) are some of the popular classical IR models. VSM is a simple and effective model for Information Retrieval, but it has limitations in handling synonyms and homonyms which to some extent can be tackled by LSA. LDA is also known to represent the documents as random mixtures over latent topics, where each topic is characterized by a distribution over words[1]. In this work, we overcome some of the limitations of VSM by using LSA and by using query expansion. We finally compare the performance of these IR models quantitatively.

Keywords: Information Retrieval System, IR, TF-IDF, VSM, LSA, Query Expansion

1 Introduction

In recent years, the exponential growth of digital information has posed significant challenges for effective information retrieval. Natural Language Processing (NLP) techniques have emerged as powerful tools for organizing, searching, and retrieving information from vast textual data. Among the numerous NLP models, Latent

Semantic Analysis (LSA) and the Vector Space Model (VSM) have gained considerable attention due to their ability to capture semantic relationships and improve the accuracy of information retrieval systems.

The primary goal of information retrieval is to retrieve relevant documents given a user's query. The traditional Vector Space Model represents documents and queries as high-dimensional vectors in a vector space. However, the Vector Space Model relies on the exact matching of query terms with document terms, which often leads to limitations in capturing semantic similarity and handling synonymy and polysemy. To address these limitations, LSA was introduced as a statistical model that leverages the latent semantic structure of the document collection.

LSA, also known as Singular Value Decomposition (SVD), employs a matrix factorization technique to discover latent semantic relationships among words and documents. By reducing the dimensionality of the term-document matrix, LSA captures underlying patterns and identifies hidden semantic similarities. This enables LSA to handle synonymy and polysemy by mapping words into a latent semantic space. Consequently, LSA has shown promising results in various information retrieval tasks.

On the other hand, the Vector Space Model remains a widely used baseline model in information retrieval. It represents documents and queries as vectors and measures the similarity between them using techniques such as cosine similarity. The Vector Space Model's simplicity and efficiency make it an attractive choice for many applications. However, its reliance on term-frequency information without considering semantic relationships can limit its retrieval effectiveness, especially in the cases of ambiguous queries or documents.

This work aims to conduct a comprehensive comparative study of LSA and the Vector Space Model in the context of information retrieval. We seek to investigate their strengths, weaknesses, and performance across various retrieval tasks. By examining the effectiveness of these models in terms of retrieval accuracy, precision, recall, and other evaluation metrics, we aim to provide insights into their suitability for different application domains.

Furthermore, we will explore the impact of different factors on the performance of LSA such as the choice of dimensionality reduction techniques and query expansion methods.

2 Problem Statement

We have the Cranfield dataset at our disposal. It belongs to the category of typical classical datasets used for information retrieval. There are 225 queries and 1400 documents in the dataset. A collection of pertinent papers and their accompanying relevance scores are provided for each query. The relevancy score indicates if a text fully answers the question, merely serves as an example, or only alludes to a few

pertinent topics.

This work aims to develop a simple information retrieval system that overcomes the shortcomings of the existing Vector Space Model (VSM)-based search engine In the existing VSM-based IR, the document vector is represented by TF IDF scores, and the retrieved documents are sorted in descending order of cosine similarity to the search query terms. We show the improvement in evaluation metrics and IR performance by using LSA-based IR model. To assess the success of retrieval, we compare and contrast the performance of our improvised search engine with that of the existing search engine using the evaluation metrics nDCG, Mean Precision, Mean Recall, and mean F-score.

3 Limitations of VSM-based IR

While analysing the performance of the basic VSM-based IR system (in assignment 2), we found out the following limitations.

- Assumes the terms to be independent. Hence, if the query phrase contains the term "love" it will only match the documents containing "love" and leave the documents containing "affection".
- Calculation intensive.
- Update is not easy. Requires idf calculation all over again.
- Lengthy documents get poor similarity scores even though they maybe highly relevant.

4 Related Work

Information retrieval has been approached in a variety of ways. Cleaning is the first and most important stage in information retrieval. An IR system's performance can be significantly enhanced by using effective pre-processing techniques [1] [2]. We applied the identical preprocessing methods as in assignment 1a. Additionally, the preprocessed data is transformed into a numerical format. Probabilities or vectors could be used. The same standards are used for queries. The papers could be ranked according to the queries using a similarity score.

Even though query expansion is usually effective in improving recall, [14] also showed that using similarity-based query expansion can improve precision. Unfortunately, they do not provide results of applying this technique standalone; their precision figures include other methods as well.

Other known attempt at Cranfield collection is [9]. This work employed four vector-based IR methods: vector space model (VSM), latent semantic indexing (LSI), generalized vector space model (GVSM), and approximate dimension equalization (ADE) (essentially a combination of the prior three). Their achieved precisions varied from 21.7% (GVSM) to 41.2% (LSI).

4.1 Vector Space Model (VSM)

The Vector Space Model (VSM) is a widely used IR model that represents documents and queries as vectors in a high-dimensional space. In this model, each document and query is represented by a vector where each dimension corresponds to a unique term in the vocabulary. The magnitude and direction of the vectors are used to calculate the similarity between documents and queries. The most common similarity measure used is the cosine similarity, which calculates the cosine of the angle between the vectors. Higher cosine similarity indicates a higher degree of similarity between the document and query.

To apply the VSM, several steps are involved. First, the text data is preprocessed by removing stop words, stemming, and other normalization techniques. Then, the term frequency-inverse document frequency (TF-IDF) scheme is used to calculate the weight of each term in the document or query, which emphasizes rare and important terms. Finally, the document and query vectors are constructed by assigning the TF-IDF weights to the corresponding terms. The cosine similarity between the document and query vectors is computed, and the documents are ranked based on their similarity scores.

The VSM has been extensively studied and applied in various NLP tasks, including information retrieval, text classification, and document clustering. It has demonstrated good performance in many applications due to its simplicity and effectiveness in capturing term relationships and document similarities. Many studies have focused on improving different aspects of the VSM, such as term weighting schemes, query expansion techniques, and relevance feedback methods, to enhance its performance and address its limitations. Some notable works in this area include Salton and McGill's pioneering work on VSM for IR [1], and subsequent research on extensions and variations of the VSM, such as the Okapi BM25 ranking function [2] and the SMART Information Retrieval System [3].

4.2 Latent Semantic Analysis (LSA)

The term-document matrix's low-rank approximation is discovered by LSA [1]. The corpus's terms (individual words) are visualized as concepts. In terms of the concept space, the document is represented by the column vectors of VT. The diagonal matrix's entries, which are ordered in non-increasing order from top to bottom and left to right, are used to assign weights to each notion. As a result, the notions with higher weights are indicated by them. We are essentially creating a k-rank approximation of the original term-document matrix since we only take the top k singular values. By combining the dimensions of phrases with comparable meanings, the low-rank approximation solves the issue of synonymous terms.

LSA makes the assumption that words used in related contexts have meanings that are comparable and, therefore, are synonymous. Due to the fact that dimensions of

words with comparable meanings are added, polysemous words whose dimensions are in the same direction as the pertinent context help to somewhat ameliorate the problem of polysemy.

4.3 LSA with Query Expansion using NLTK WordNet

The Latent Semantic Analysis (LSA) model is a popular approach for information retrieval (IR) in Natural Language Processing (NLP). It leverages the power of semantic analysis to capture latent relationships between terms and documents. In the context of IR, LSA works by representing documents and queries as vectors in a high-dimensional space and measuring their similarity based on their cosine similarity. However, to address the issue of vocabulary mismatch and to enhance the retrieval performance, query expansion techniques can be employed. One such technique is using WordNet, a lexical database in NLTK (Natural Language Toolkit), for expanding the query.

Query expansion using NLTK WordNet involves identifying synonyms, hypernyms, and hyponyms of the original query terms. By accessing the WordNet lexical database, the LSA model can expand the query terms to include related words with similar or broader meanings. This expanded query can then be used to retrieve more relevant documents that may not have been captured by the original query. The expanded query is represented as a vector in the LSA space, and the retrieval process is performed using the cosine similarity measure as before. The incorporation of WordNet-based query expansion can improve the recall and precision of the LSA model, enabling it to retrieve more relevant documents and better serve the information needs of the user.

5 Methods

The preprocessing steps used in the models are described in the respective model sections below. The evaluation metrics used for evaluating the model performances are the snippets taken from assignment 2 which include nDCG, mean Precision, mean Recall, and mean F1-Score.

5.1 Baseline Model- Vector Space Model (VSM)

For the implementation of VSM, we processed the documents and the queries using basic preprocessing steps such as word tokenization, stopword removal, and stemming. The Cranfield dataset is then converted into a document-term matrix, where each row represents a document and each column represents a term. The term frequency-inverse document frequency (tf-idf) values for each term in the document-term matrix is calculated. Each query as a vector of tf-idf weights for each term in the vocabulary is represented and the cosine similarity between the query vector and each document vector in the dataset is calculated. The documents are then ranked in decreasing order of similarity score and the most relevant documents are retrieved.

5.2 Latent Semantic Analysis (LSA)

LSA is a statistical technique that identifies the underlying relationships between terms and documents. It uses a matrix factorization technique to represent the dataset as a lower-dimensional space. For implementing LSA for information retrieval on the Cranfield dataset we converted the Cranfield dataset into a document-term matrix and then computed the singular value decomposition (SVD) of the document-term matrix to identify the underlying relationships between terms and documents. Each query is then represented as a vector in the lower-dimensional space. The cosine similarity between the query vector and each document vector in the lower-dimensional space is then obtained to rank the relevant documents for successful retrieval. In this work, we explored different numbers of the reduced dimensions for truncated SVD so as to study their effect on the retrieval performance.

5.3 Latent Semantic Analysis (LSA) with Query Expansion

For this implementation, We keep the components of the LSA fixed at 350, and expand the query terms by using synonyms of the existing query terms using wordnet from nltk library. We experiment with addition of different number of query terms to be added such as 3 and 8 for the terms selected at random from the query terms. We then follow the same steps as in LSA for document retrieval.

6 Results

The obeservations on the evaluation metrics are:

- Precision: decreases with k
- Recall: non-decreasing function with respect to k
- F1-Score : increases then stabilises as it depends on both precision and recall. F1 measures gives equal weightage to precision and recall.
- nDCG : takes into account the graded relevance values as we have in the given dataset .

The results of the comparative study are as follows:

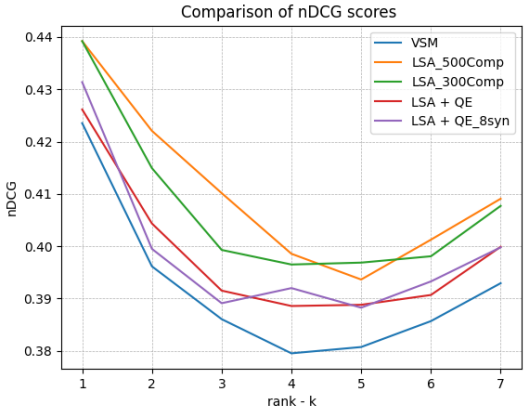

Fig. 1 The figure shows the comparison of nDCG Scores for Baseline VSM, LSA with 500 components, LSA with 300 components, LSA with 350 components with 3 synonym terms added to expand query and LSA with 350 components with 8 synonym terms added to expand query.

The nDCG Score is the best for LSA with 500 components which is expected. As more latent dimensions are better capable of representing the actual concepts in latent space. The LSA with query expansion model's performance is somewhere in between the VSM and LSA.

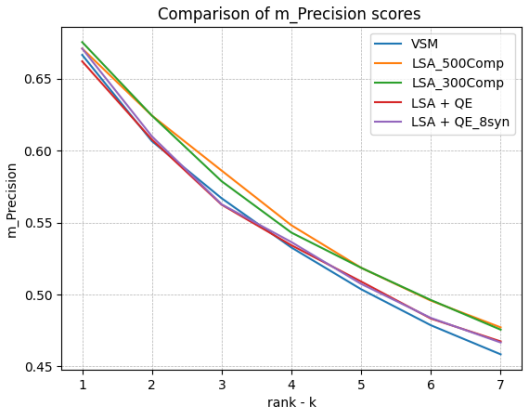

Fig. 2 The figure shows the comparison of mean Precision Scores for Baseline VSM, LSA with 500 components, LSA with 300 components, LSA with 350 components with 3 synonym terms added to expand query and LSA with 350 components with 8 synonym terms added to expand query.

The mean Precision Score is the best for LSA with 500 components which is expected. As more latent dimensions are better capable of representing the actual concepts in latent space. Also,it is known that more the number of components better the precision.

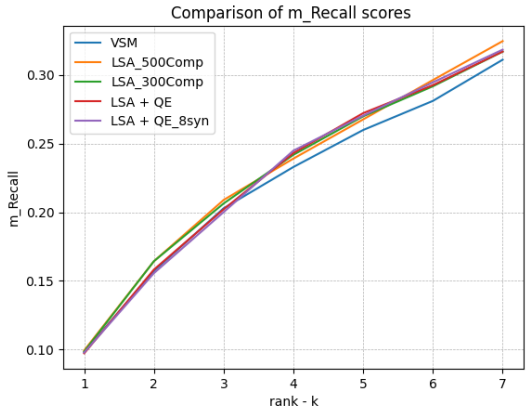

Fig. 3 The figure shows the comparison of mean Recall Scores for Baseline VSM, LSA with 500 components, LSA with 300 components, LSA with 350 components with 3 synonym terms added to expand query and LSA with 350 components with 8 synonym terms added to expand query.

The mean Recall Score is almost the same for LSA with 500 components, LSA with 300 components and LSA with Query Expansion which is expected. As the Query expansion is aimed at increasing the recall by retrieving more relevant documents.

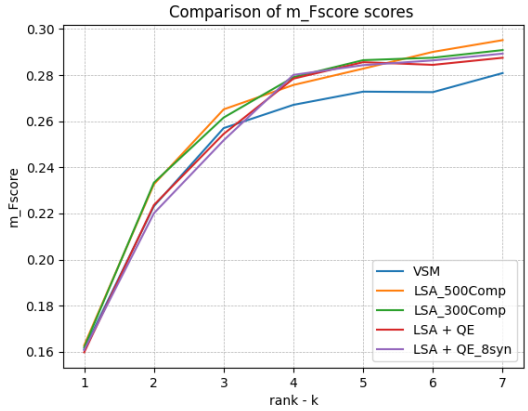

Fig. 4 The figure shows the comparison of mean F1 Scores for Baseline VSM, LSA with 500 components, LSA with 300 components, LSA with 350 components with 3 synonym terms added to expand query and LSA with 350 components with 8 synonym terms added to expand query.

The F1 Scores are almost same for LSA, LSA with Query Expansions. The F1 Score for VSM is not at par with these models.

7 Discussion

The results demonstrate the superiority of Latent Semantic Analysis (LSA) Information retrieval over the Vector Space Model (VSM). The LSA when supplemented with query expansion performed better than VSM but not better than LSA. This can be attributed to the fact that the Cranfield dataset contains documents majorly from one domain (scientific excerpts from aerospace) so query expansion may have become redundant as LSA inherently captures the latent relations between the words.

8 Conclusion

This work presented a comparative analysis of the Vector Space Model (VSM), Latent Semantic Analysis (LSA), and LSA with query expansion for information retrieval on the Cranfield dataset. The evaluation metrics, including nDCG, mean precision, mean recall, and mean F-scores demonstrated that LSA outperformed VSM. The use of LSA, which incorporates semantic analysis, allowed for a better representation of the underlying semantic relationships between terms and documents. This improved representation led to more accurate retrieval results, resulting in higher nDCG scores, indicating better ranking quality of the retrieved documents. Additionally, LSA achieved higher mean precision, mean recall, and mean F-scores, indicating a better balance between precision and recall and a higher overall retrieval performance compared to VSM. Furthermore, the study explored the effectiveness of query expansion using LSA with NLTK WordNet.

Overall, the findings of this study highlight the superiority of LSA over VSM and the added benefit of query expansion using NLTK WordNet in improving the retrieval effectiveness. These results contribute to the understanding of the strengths and limitations of different IR models and provide insights into the potential enhancements that can be achieved through the utilization of semantic analysis and query expansion techniques.

Acknowledgments. We appreciate Prof. Sutanu Chakraborti for giving us the chance to work on this project and for giving us the resources we needed to create an all-encompassing information retrieval system. We also appreciate the TAs of the NLP course for their assistance in resolving our problems. Throughout the course, we had a lot of fun and learned a lot. We also want to thank our classmates for bringing up some really insightful talks in class that helped us with our assignment.

9 References

[1] Salton, G., McGill, M. J. (1986). Introduction to modern information retrieval. McGraw-Hill.

[2] Robertson, S. E., Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 232-241.

[3] Singhal, A., Buckley, C., Mitra, M. (1996). Pivoted document length normalization. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 21-29.

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

[5] Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

[6] Manning, C. D., Raghavan, P., Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

[7] Bird, S., Klein, E., Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.

[8] Natural Language Information Retrieval TREC-3 Report. Tomek Strzalkowski, Jose Perez Carballo, Mihnea Marinescu.

[9] Approximate Dimension Equalization in Vector-based Information Retrieval. Fan Jiang, Michael L. Littman.

[10] Harnani Mat Zin, Norwati Mustapha, Masrah Azrifah Azmi Murad, and Nurfadhlina Mohd Sharef, "The effects of preprocessing strategies in sentiment analysis of online movie reviews", AIP Conference Proceedings 1891, 020089 (2017) https://doi.org/10.1063/1.5005422

[11] Haddi, Emma, et al. 'The Role of Text Preprocessing in Sentiment Analysis.' Procedia Computer Science, vol. 17, 2013, pp. 26–32. DOI.org (Crossref), doi:10.1016/j.procs.2013.05.005.

[12] WordNet – a lexical database for the English language http://wordnet.princeton.edu