Web Information Retrieval using Vector Space Model and Docu-Tally Metric

1st Nitish Chaturvedi
Department of Networking and Wireless Communication
SRM Institute of Science and Technology, Chennai
Chennai, India
nc7219@srmist.edu.in

2nd Eshanika Ray
Department of Networking and Wireless Communication
SRM Institute of Science and Technology, Chennai
Chennai, India
er2744@srmist.edu.in

3rd Meenakshi K*
Department of Networking and Wireless Communication
SRM Institute of Science and Technology, Chennai
Chennai, India
meenaksk@srmist.edu.in

2023 International Conference on Computer Communication and Informatics (ICCCI) | 979-8-3503-4821-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCI56745.2023.10128258

***Abstract*—Regardless of the type of data set, it is frequently challenging to sift through the vast amount of data that is available on the Internet as a result of technological improvements. To deal with challenges mentioned above, we have come up with a ranking approach which is computed using NLP and vector space model. The approaches used for information retrieval start with a basic machine learning model and progress to multi-stage architectures and frameworks like language modelling and term matching. The main goal of this work is to use a standard retrieval process to glean insights from large amounts of data, which is the problem we are aiming to solve. The method utilised in the study is latent semantic analysis, which takes advantage of the semantic aspects at play and can be used to glean insights from lengthy texts.**

***Index Terms*—information retrieval, language modelling, latent semantic analysis, vector space model**

I. INTRODUCTION

Rapid breakthroughs and the introduction of new technologies have led to a gigantic flow of data which contains significant information that can be utilised to raise the calibre of services. The information retrieval process is basically a procedure for evaluating information from massive amounts of data.

Any individual conducting the research has a myriad of options for dealing with the issues of information retrieval. Practically, all search engines focus on words rather than novel techniques, the vast amount of information flow and the increase in data and web-pages exacerbates the challenges that arise during information retrieval from recovering relevant and trustworthy information. When we are using a extraction system, to search for specific information, an individual is only required to provide a finite key phrases to get down to the search.

When narrowing down the search results might be ineffective at times, and their quantity varies in thousands. Despite the fact that retrieval system has evolved as a field of research, the challenges in developing an efficient retrieval system hasn't been addressed and looked upon. As an effort to make the information retrieval system more accessible, an interface was created that communicates directly with the user that allows them to discover pertinent knowledge without any human intervention.

The algorithm computes the rank of a particular information by processing the estimating the document's score using the docu-tally metric. Setting the endpoints for the docu-tally extracts the insights from a particular piece of information. Instead of solely focusing on the keywords, the approach searches using the tokenization which breaks the sentences into tokens and remove all extraneous data only to get the necessary details.

II. Related Works

Singh et al [1] in his research study discussed various IR models and had kept various search engines as the benchmarks for testing.To assess the performance of an information retrieval model, we need a document repository, and comprehensive description of user queries, and an accurate assessment for every query-document combination. Yusrandi et al [2] implemented the vector space model in which retrieval is done by calculating the distance between keywords and documents by presenting them in a vector format. Zheng [3] devised the application of natural language processing from low level to high level and proved that NLP is suitable for obtaining better accuracy. Fitzgerald et al [4] worked to find out that IRS capabilities are indexing design dependent, thus introducing hybrid indexing methods. Adding on that, they found a hybrid indexing method, produced efficient outputs and can be used as an alternative for the indexing method in future scope. Uzun et al [5] in their study proposes UzunExt, a breakthrough technique that extracts content quickly using string methods and additional information without developing a DOM Tree.

Husain et al [6] came up with a novel solution to retrieve relevant code using language modeling. It contains corpus, which contains 6 million functions from open-source code spanning six different programming languages. Kherwa et al [7] aims to find hidden relationships in documents for better understanding of relation between terms in a dataset. Arora et al [8] the recall and precision technique are used to evaluate the efficacy of information retrieval systems. Hoffman [9] came up with an improved iteration of latent semantic analysis, which states that proposed study can be derived from mixture decomposition, and to avoid overfitting, generalization of maximum likelihood by tempered EM was recommended.

A. Information Retrieval on Web

On the lines of massive development of information which is available on the Internet, the World Wide Web has become a remarkably effective platform for data storage and retrieval. In recent years, the World Wide Web has grown at an exponential rate. It is estimated that there are around 30-50 billion pages on the Internet, and this number has recently surpassed 2 trillion. Out of the mentioned number, only half of them provide the authentic information, rest are either irrelevant or duplicated from some source. This creates an opportunity for a system which can be used to extract information in a particular manner.

The retrieval models assist users in completing search tasks by obtaining a small number of pertinent pages from large corpus of texts. The relevance of a particular retrieval model can be evaluated by the response it gives for a given query.

B. Latent Semantic Analysis

The process of determining and capturing the contextual-usage meaning of words using statistical computations on a large corpus of literature is known as latent semantic analysis. The idea behind this is creating a relationship between a piece of information and the terms contained within. It works on traditional mathematical techniques i.e. singular decomposition technique, to extract insights from an unstructured data between document and the terms within.

LSI has a number of disadvantages, the most significant of which being its failure to capture polysemy in which when represented in a vector form yields an average of all the word's meanings. Apart from this, it is widely believed that semantic analysis can help with search engine optimization, which is not very effective and the whether or not, it can be used is still a problem to work upon.

III. Algorithm

For retrieving useful information from the document, the pathway which has been the prime focus of your study is analyzing the query. The query is entered as an input in the machine translated form for the query analysis. Query Analysis is a process which is primarily used in databases having a huge number of queries which are optimized further for an enhanced performance.

Query analysis starts with the queries being sequenced in a vector format. Tokenization, being one of the initial steps in the NLP pipeline, is a process in which particular text is broken into words, and each of them is referred to as a token.

In tokenization, each token is checked to make sure it is not a stop word. If the token turns out to be a stop word, the attestation is then moved to the next token. Post the removal of stop words from tokenization and after the tokens being assigned, next up the process is stemming. Stemming is a process in which we reduce words to the most basic form. A stemmer is assigned to each token to remove the extra letters. After this tagging is initiated, in which each token of speech is recognised. After tagging, chunking is done as a part of collecting information in parts and joining it to make a sentence. Hence, after this, by doing the query analysis by NLP, the content and the intent of the document is extracted.

Latent Semantic Analysis is based on the rules of tagging, which is used to construe a piece of

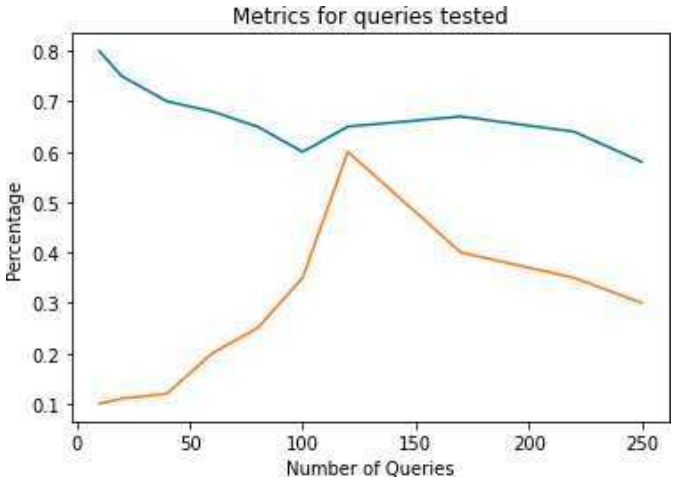

Fig. 1. Graphical representation of average f-score and the discounted cumulative gain

document. To make sure that the comprehension is enhanced, it is then converted into a document matrix and WordNet is used to find synonyms and thesaurus. Post the conversion of the document matrix, it is then transformed into a lower-dimensional matrix which eases out the process. Each sentence is then semantically parsed, and the notion is extracted.

The document is then indexed based on the query using the measure of similarity and intervals between two cosines, and the intensity of the concept is ascertained by the number of times a certain word occurs in a given piece of text. The semantic weight, which is quite often a 0 or 1, is then processed.

The docu-tally score, which is a metric based on the cosine distance and the semantic weight, is used to establish the overall rating of each page. Only after documents are classified, the index is calculated based on the extent of relevancy, and the score of every document is assessed using the docu-tally score. The document is retrieved only after the docu-tally margin limit is crossed. The documents are sorted in descending order by docu-tally score, culminating with the necessary documents showcased at the upper end.

IV. Results

The underlying mechanisms is substantiated using standard data sets, each with 50 samples collected. So 250 queries are accumulated, trials are undertaken, and the offline metric is computed. The figure[1] below illustrates the average f-score and the discounted cumulative gain for the number of queries tested:

The precision, recall, F-measure, and fallout observed for the suggested method as well as a few

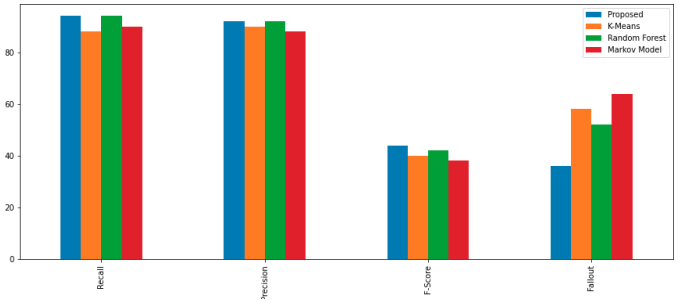

Fig. 2. Performance of proposed model w.r.t other ML models in terms of standard metrics

previous mechanisms such as retrieval based on SVM, Random Forest, and the Markov Model are shown in the figure[2] above. The obtained results suggest that the proposed mechanism outperforms the other two in terms of precision, recall, and fallout, as well as a suitable F-score.

V. Conclusion

NLP's impact on information retrieval tasks has primarily been mostly an expectation instead of reality. There hasn't been much development around it, but slowly researchers are finding their ways in the intersection of NLP and information retrieval . In this study, we have outlined our recent research in information retrieval using NLP, which have had varying results. We explored the significance of NLP and IR prior to actually outlining our initial efforts on leveraging semantic analysis to derive dependencies. Post exploring, we reassessed the impact that NLP could play in IR activities and decided to focus our efforts on deploying NLP.

The exploration of various other methods led us to use the latent semantic analysis carried out in the research in order to extract relevant information from the standard dataset like Cranfield and CISI is available on the internet. Rather than accessing documents based upon keywords, the method utilizes semantic analysis to decipher important information. In the proposed approach, we have used the metric docu-tally score, which ranks the document by assigning scores that represent how relevant the information is. The method proposed in the study is quite efficient in proving the fact that it is at par with other machine learning algorithms in extracting the information.

References

[1] Singh JN, Johri P, Kumar A, Singh M (2022) Web Information Retrieval Models Techniques and Issues: Survey. 2022 2nd International Confer- ence on Advance Computing and Innovative Technologies in Engineer- ing (ICACITE). https://doi.org/10.1109/icacite53722.2022.9823707

[2] Yusrandi, Muladi, Rosyid HA, Mahamad AK (2021) Document Search in Information Retrieval System Using Vector Space Model. 2021 7th International Conference on Electrical, Electronics and Information En- gineering (ICEEIE). https://doi.org/10.1109/iceeie52663.2021.9616735

[3] Zheng Z (2021) Natural language processing and information retrieval system based on BP neural network. 2021 IEEE 4th International Conference on Information Systems and Computer Aided Education (ICISCAE). https://doi.org/10.1109/iciscae52414.2021.9590737

[4] Andrew K (2021) The hybridised indexing method for research-based information retrieval - Kyle Andrew Fitzgerald, Andre Charles de la Harpe, Corrie Susanna Uys, 2021. In: Journal of Information Science. https://journals.sagepub.com/doi/abs/10.1177/0165551521999800.

[5] Uzun E (2020) A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages. IEEE Access 8:61726–61740. https://doi.org/10.1109/access.2020.2984503

[6] Husain H, Wu H-H, Gazit T, Allamanis M, Brockschmidt M (2019) CodeSearchNet Challenge: Evaluating the State of Semantic Code Search. arXivorg. https://doi.org/10.48550/arXiv.1909.09436

[7] Kherwa P, Bansal P (2017) Latent Semantic Analysis: An Approach to Understand Semantic of Text. 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC). https://doi.org/10.1109/ctceec.2017.8455018

[8] Arora M, Kanjilal U, Varshney D (2016). Evaluation of information retrieval: precision and recall. International Journal of Indian Culture and Business Management. 12. 224. 10.1504/IJICBM.2016.074482.

[9] Hofmann T (2013) Probabilistic Latent Semantic Analysis. arXivorg. https://doi.org/10.48550/arXiv.1301.6705

[10] A, G., & A, K. K. (2022). Predicting Malicious Node Behavior in Wireless Network Using DSR Protocol and Network Metrics. International Journal of Computer Communication and Informatics, 4(1), 1-10. https://doi.org/10.34256/ijcci2211