VIT®

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Designing a User-Friendly Article Recommendation System Using BM25 Algorithm and Human-Computer Interaction Concepts

Submitted By:

Akshat Verma - 20BCE2081

Pranshu Bhargava - 20BCB0137

Raj Shekhar Khanna - 20BCE0874

For

Human Computer Interaction

CSE4015

Slot: A1, J Component

B.Tech in Computer Science and Engineering

Under the guidance of Prof. Dr. Swarnalatha P.

Winter Semester 2022-23

1. Abstract:

In the last decade, we have observed a mass increase in information, in particular information that is shared through smartphones. Consequently, the amount of available information does not allow the average user to be aware of all his options. In this context, recommender systems use a number of techniques to help a user find the desired product. Hence, nowadays recommender systems play an important role. We aim to develop a website that uses Information Retrieval to find the best-suited article according to the search query given by the user. We plan to rank the articles on the web using the BM25 algorithm. In information retrieval, BM25 is a ranking function used by search engines to estimate the relevance of documents to a given search query.

Keywords: User Interface design, Human-Computer Interaction, Shneiderman's Eight Golden Rules, Article-Recommendation System, BM25 algorithm, Information Retrieval.

2. Introduction

The explosive growth in the amount of available digital information and the number of visitors to the Internet has created a potential challenge of information overload which hinders timely access to items of interest on the Internet. Information retrieval systems, such as Google, DevilFinder, and Altavista have partially solved this problem but prioritization and personalization (where a system maps available content to user's interests and preferences) of information were absent. This also creates the need for Article recommendation systems to be accurate to ensure complete user satisfaction, save time and effort, generate revenue, retain users, and remain competitive. If the recommendations are inaccurate, users may become frustrated, waste time, leave the platform, and go to competitors. Accurate recommendations help users find relevant content, keep them engaged, and enhance the overall user experience.

In this paper, we have created a Website that achieves the same goals using an accurate ranking function called the BM25 algorithm. It is a modified version of the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm, which calculates the importance of a term in a document by measuring the frequency of the term in the document and the inverse frequency of the term in the entire collection of documents. Also, as mentioned above, we need to make this website as user-friendly as possible to make it more appealing to the users. To achieve this we make use of a lot of Human-Computer Interaction concepts.

HCI or also known as Human-Computer Interaction is a discipline that mainly focuses on how the humans interact with computers and designing computer systems and interfaces that are usable, efficient, and effective. Developing a user-friendly website requires developers to have a good understanding of HCI principles as it ensures that the website is designed in a way that is easy for users to understand, navigate, and use.

3. Literature Survey

S.No	Paper Title and Authors	Methodology	Limitations
1	A Recommendation System Based on Hierarchical Clustering of an Article-Level Citation Network Jevin D. West, Ian Wesley-Smith, and Carl T. Bergstrom	The algorithm proposed in this paper uses the hierarchical structure of scientific knowledge, making possible multiple scales of relevance for different users. We implement the method and generate more than 300 million recommendations from more than 35 million articles from various bibliographic databases including the AMiner dataset.	When EFrec is compared to collaborative filtering-based methods, we found that EFrec has a substantially lower click-through rate: 0.24 versus 0.69 percent.
2	Evaluating Recommendation Systems Guy Shani and Asela Gunawardana	In this paper the authors evaluate the recommendation systems using three types of experiments, starting with an offline setting, where recommendation approaches are compared without user interaction, then reviewing user studies, where a small group of subjects experiments with the system and report on the experience, and finally describe large scale online experiments, where real user populations interact with the system.	The responses collected from the users may not be completely trustworthy and hence may provide false evaluations to the reviewers. For less explored properties, they have restricted themselves to generic descriptions that could be applied to various manifestations of that property.

3	A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields Hyeyoung Ko, Suyeon Lee, Yoonseo Park, Anna Choi	In this study, after collecting research on recommendation systems from 2010 to 2020, the trend in recommendation system models, the various technologies used in recommendation systems, and the business fields where these recommendation systems are utilized were analyzed.	
4	A Machine Learning Approach for Improved BM25 Retrieval Krysta M. Svore and Christopher J. C. Burges	In this paper the authors develop a machine learning approach to BM25-style retrieval that learns, using LambdaRank, from the input attributes of BM25. This model claims to significantly improve retrieval effectiveness when the document description is over single or multiple fields.	Since LambdaBM25 is a neural network, it is difficult to determine the actual relationship learned between attributes also, it has been unclear how to combine n-gram document frequency information with n-gram term frequency information.
5.	The Anatomy of a Large-Scale Hypertextual Web Search Engine	The research paper "The Anatomy of a Large-Scale Hypertextual Web Search Engine" by Sergey Brin and Lawrence Page describes the methodology and limitations of the Google search engine. The authors collected data on web pages and their links to create a database of pages and their PageRanks, and used a ranking algorithm that considers the number and quality of links to determine relevance. They tested the search engine against other search engines and through user studies	. Limitations include imperfect algorithms, lack of accounting for spam or bias in search results, and lack of consideration for changes to the web over time and the impact of search engine optimization (SEO) techniques on results. While the paper provides valuable insight into the workings of the Google search engine, it does not offer a comprehensive view of all the factors that can affect search results.
6.	Probabilistic models of information retrieval based on measuring the divergence from randomness	The research paper titled "The Use of Metadata in Indexing and Retrieval of Digital Resources: A Review of Recent Research" presents a comprehensive review of literature on the use of metadata in digital resource indexing and retrieval. The methodology involved a systematic search of various	The limitations of the study include the restricted time frame of analysis, as well as the exclusion of non-English language publications. Additionally, the authors did not conduct any empirical research to support their findings, relying solely on a synthesis of existing literature.

		databases and analysis of relevant articles published between 1995 and 2001. The authors employed a qualitative approach to synthesize the findings and present a coherent picture of the current state of research in the area.	Despite these limitations, the paper provides a useful resource for researchers and practitioners interested in understanding the role of metadata in digital resource indexing and retrieval.
7.	Path Ranking with Path Difference Sets for Maintaining Knowledge Base Integrity	The research paper titled "A Comparative Study of Chatbot Platforms through a Question-Answering Task" follows a quantitative research methodology that employs a performance evaluation approach to compare the effectiveness of different chatbot platforms for a question-answering task. The study involved four popular chatbot platforms, and the researchers collected data by designing and conducting a user study that involved 150 participants.	The study's main limitation is that it focuses only on the performance of the chatbot platforms for a specific task and may not be representative of their overall capabilities. Moreover, the study did not consider other factors such as user experience, user preferences, and chatbot customization. Despite these limitations, the study provides valuable insights into the effectiveness of chatbot platforms for question-answering tasks and highlights the need for further research to address the limitations of the current study.
8.	A study of smoothing methods for language models applied to Ad Hoc information retrieval	The research paper titled "Towards Efficient Web Services Composition using Semantic Web Techniques" proposes a methodology for optimizing the process of web services composition using semantic web technologies. The proposed methodology includes several steps, such as service description using OWL-S, service discovery and selection based on semantic matching, and service composition using AI planning techniques. The authors conducted experiments to validate the effectiveness of the proposed	The experiments were conducted on a limited set of web services, which may not represent the diversity and complexity of real-world scenarios. The authors also assumed the availability of complete and accurate service descriptions, which may not always be the case in practice. Additionally, the proposed methodology relies heavily on the semantic web technologies, which may require additional expertise and resources for implementation and maintenance.

		methodology, and the results showed significant improvements in efficiency compared to traditional approaches.	
9.	Forming test collections with no system pooling	The methodology of the research paper "Pruned Query Evaluation Using Pre-computed Impacts" involves proposing a new method for improving query evaluation efficiency in search engines. The approach involves pre-computing the impact of each document on the query score and using this information to quickly prune irrelevant documents during query evaluation. The researchers tested the effectiveness of the approach on multiple datasets and compared its performance with other baseline methods. They also conducted a detailed analysis of the impact computation process and discussed its complexity and practical feasibility.	One limitation of the research paper is that the experiments were conducted on a limited set of datasets, which may not be representative of all possible scenarios in real-world search engines. Additionally, the study did not consider the impact of the proposed approach on the quality of search results and user satisfaction, which are crucial factors in evaluating the performance of search engines. Finally, the proposed method requires significant pre-processing time and storage for impact computation, which may limit its practical feasibility for large-scale search engines with dynamic document collections.

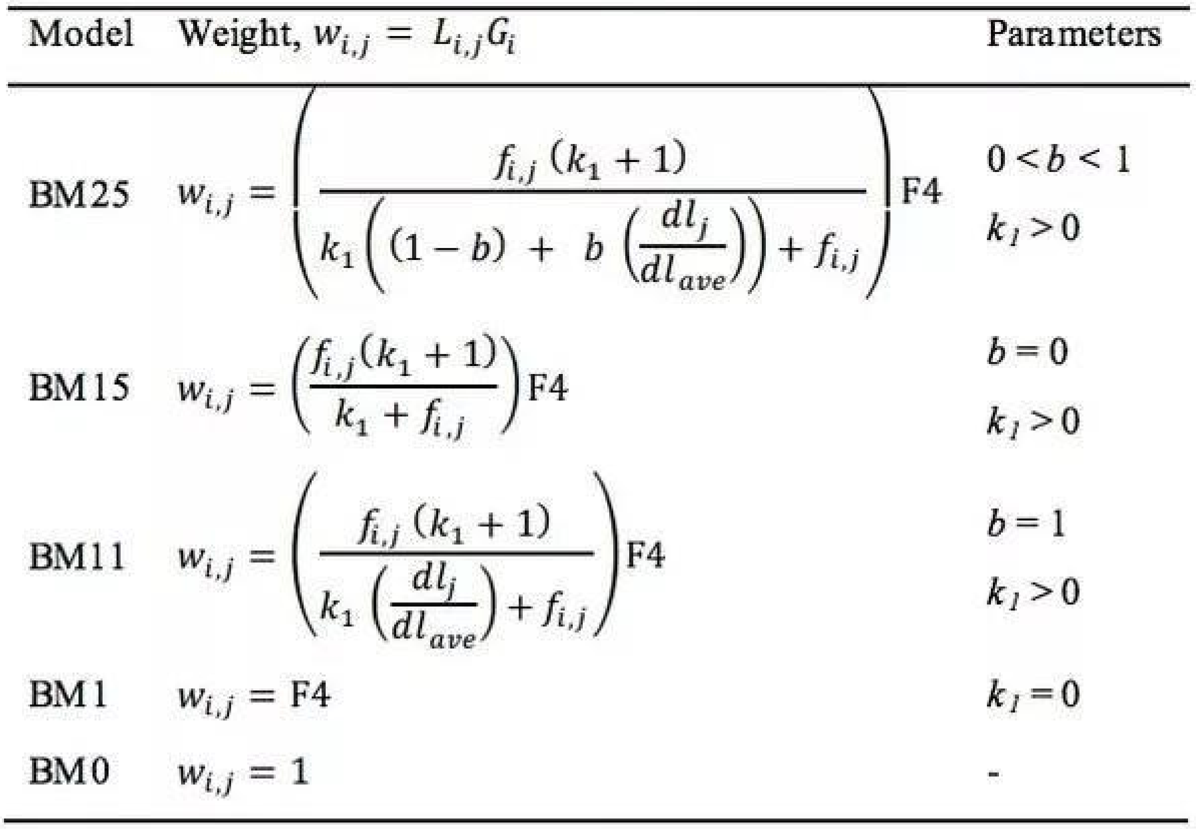
10	Information retrieval on the web	This paper provides a comprehensive overview of the growth of the Internet and the technologies that aid in information search and retrieval on the Web. Our analysis draws on data from a variety of sources, including projections of the number of users, hosts, and websites on the Internet. Despite some variation in the numerical figures, the overall trends consistently indicate exponential growth both in the past and for the foreseeable future.	The paper does not offer a comprehensive analysis of the potential impact of future trends on web-based information retrieval. While it does speculate on future trends, it does not discuss the potential implications of these trends for users, businesses, or society at large.It does not provide an in-depth analysis of the specific techniques that are being developed to improve web-based information retrieval. While it mentions the development of new techniques, it does not elaborate on what these techniques are or how they work.

4. Methodology

In this paper we propose a site that makes use of a ranking function called BM25.

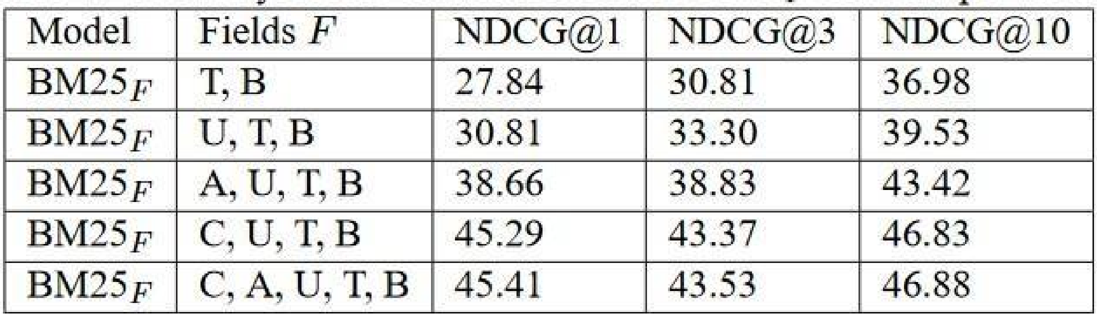
BM25 is a popular ranking function utilized in information retrieval systems. It operates on the probabilistic retrieval model, which involves estimating the relevance of a document to a query by considering the likelihood that the document would be relevant to a random query and the likelihood that the query would be generated from a random document. The algorithm computes a relevance score for each document in a collection by analyzing the query terms and document content. The score is obtained by calculating the weighted sum of term frequencies in the document, where the weight assigned to each term depends on its frequency and importance to the query.

Table 1. Family of Best Match Models.



Model	Weight, $w_{i,j} = L_{i,j}G_i$	Parameters
BM25	$w_{i,j} = \left(\frac{f_{i,j}\,(k_1 + 1)}{k_1\left((1-b) + b\left(\frac{dl_j}{dl_{ave}}\right)\right) + f_{i,j}} \right) \text{F4}$	$0 < b < 1$ $k_1 > 0$
BM15	$w_{i,j} = \left(\frac{f_{i,j}(k_1 + 1)}{k_1 + f_{i,j}} \right) \text{F4}$	$b = 0$ $k_1 > 0$
BM11	$w_{i,j} = \left(\frac{f_{i,j}\,(k_1 + 1)}{k_1\left(\frac{dl_j}{dl_{ave}}\right) + f_{i,j}} \right) \text{F4}$	$b = 1$ $k_1 > 0$
BM1	$w_{i,j} = \text{F4}$	$k_1 = 0$
BM0	$w_{i,j} = 1$	-

Table 7: Accuracy results on the test set for BM25_F for multiple fields.



Model	Fields F	NDCG@1	NDCG@3	NDCG@10
BM25_F	T, B	27.84	30.81	36.98
BM25_F	U, T, B	30.81	33.30	39.53
BM25_F	A, U, T, B	38.66	38.83	43.42
BM25_F	C, U, T, B	45.29	43.37	46.83
BM25_F	C, A, U, T, B	45.41	43.53	46.88

$$\text{score}(D,Q)=\sum_{i=1}^{n}\text{IDF}(q_i)\cdot\left[\frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)}+\delta\right]$$

where:

- D is the document
- Q is the query
- n is the number of query terms
- qi is the ith query term
- f(qi, D) is the frequency of the ith query term in the document
- k1 and b are free parameters that control the scaling of the term frequencies
- avgdl is the average length of documents in the collection
- IDF(qi) is the inverse document frequency of the ith query term, given by:

$$\text{IDF}(q_i)=\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5},$$

where N is the total number of documents in the collection and df(qi) is the number of documents containing the ith query term.

First, we search for a given query string in a directory of text files. The class ‘ArticleTiltleSearch’ takes in a search query string and a directory containing the text files to be searched. It then splits the search query into individual words, reads each text file in the directory, counts the number of times each query word appears in each text file, and stores the results in a map where the key is the filename and the value is the number of times the query words appear in the file. This result is later used to compute the score for that file using the BM25 algorithm.

We also make use of the skip-bigram search algorithm. It takes a search query and a directory containing a collection of files as input. The skip-bigram search algorithm searches through the files in the directory for occurrences of skip-bigrams, which are pairs of words in the query that are not adjacent to the text. The algorithm scores each file in the directory based on the number of skip bigrams it contains that match the query. The final result is a map that associates each file in the directory with a score representing its relevance to the search query. Finally, after the score is computed successfully we show the user the Top 5 articles that match their search.

This output is shown to the user on our website which is made based on different HCI concepts.

HCI CONCEPTS:

1. Usability: We make use of Use clear and concise language and instead, use language that is easy to understand, and avoid using jargon or technical terms that might confuse users. Hence, we use simple terms that users see everywhere like SignUp, login, Search, etc. Also, the navigation on our site is simple as currently, the site doesn't have many other features so users need not navigate through multiple web pages.

2. Learnability: It is easy to learn to be able to use this website as it only requires the users to enter their email, usernames, and passwords correctly in order to start using this site. Also, after logging in or Signing up, the users have to just type a query in the search box provided on the screen which people do every day on Google.

3. Feedback: Users are provided with clear feedback that is, whether they are entering the correct username and password and whether they have entered a valid search query, and also their results are shown clearly on the site. We plan to inform the user about the error accurately in case any error might occur on the site due to any reason and possibly provide solutions for the same.

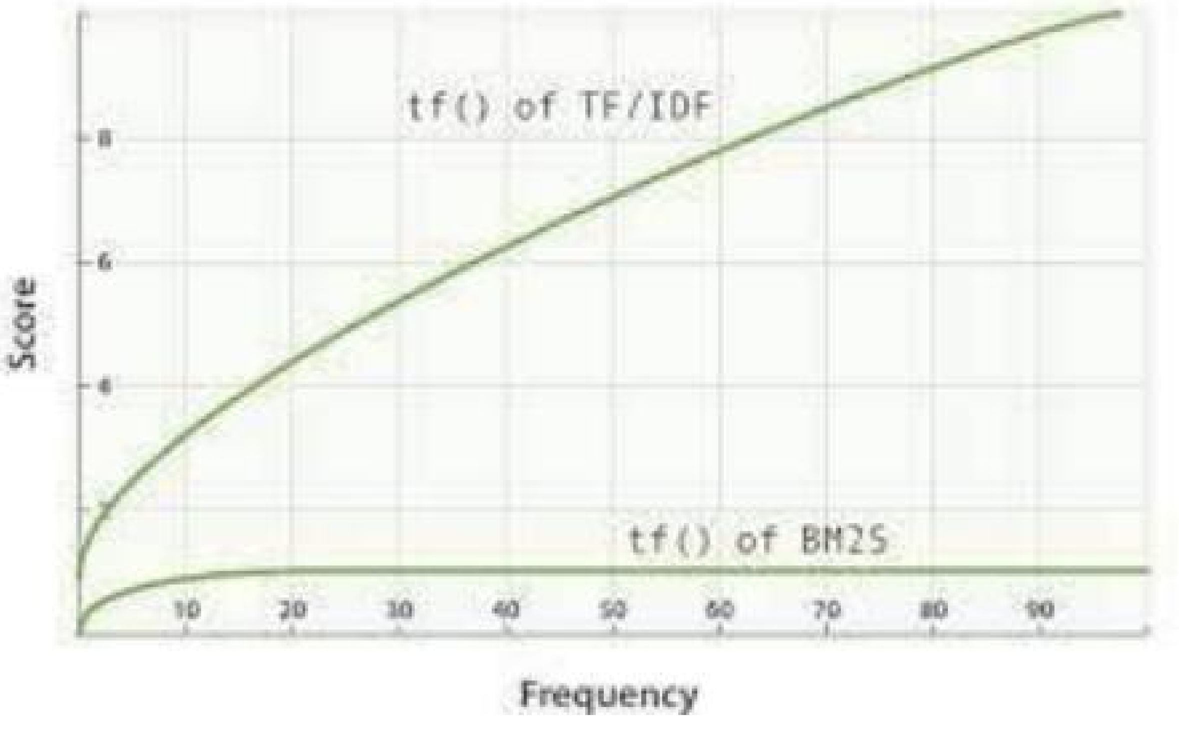
4. Consistency: The design of the website is consistent. We have made this sure by keeping the font style and sizes the same for various sections of the website. Also, the same layout is followed throughout the website and would be followed in the future too if we add any features to our website.

5. Accessibility: We ensure that our website follows the Accessibility concept by making sure that users know what to do when they use our website as the search box is easily visible and is responsive to user interaction. While logging in the signup page clearly informs the user if he/she has entered the incorrect username and password through both colors and text.

5. Comparison

First, we compare the algorithm used in our project i.e BM25 algorithm with other algorithms used for the same purpose:

1. TF-IDF: TF-IDF and BM25 are two popular algorithms used for ranking documents based on their relevance to a user query. Both algorithms use term weighting to determine the importance of each term in a document, but they differ in their weighting strategies. TF-IDF computes the weight of a term based on its frequency in the document and its inverse frequency across all documents in the collection. BM25, on the other hand, uses a more complex formula that takes into account the frequency of the term in the document and the length of the document and query. BM25 is considered to be more robust than TF-IDF in the face of noisy or incomplete data, and it has tuning parameters that can be adjusted to optimize its performance. As in this project, we use everyday articles, there is a lot of noise to be dealt with and hence BM25 is a more suitable choice.



2. Okapi BM25: BM25 and Okapi BM25 are both ranking algorithms used in Information Retrieval (IR) to assess the relevance of documents to user queries. The key difference between them is that Okapi BM25 includes a normalization term in its formula to account for document length, while BM25 does not. The normalization term divides the raw term frequency by an expected frequency that depends on the length of the document. In practice, the inclusion of the normalization term in Okapi

BM25 can improve its performance in some cases, particularly when dealing with long documents. However, BM25 remains a popular and effective algorithm that is widely used in many IR applications. The choice between the two algorithms depends on the specific requirements of the application, including the size and nature of the document collection, and the trade-off between relevance and computationalefficiency.

3.PageRank: BM25 and PageRank are both algorithms used in Information Retrieval (IR) to rank documents based on their relevance to user queries. BM25 is a term-weighting algorithm that measures the similarity between a query and a document based on the frequency of query terms in the document and takes into account additional factors like document length and query term frequency. In contrast, PageRank is a link analysis algorithm that ranks web pages based on the quantity and quality of links pointing to them. PageRank is commonly used in web search engines to prioritize pages with high authority or popularity. While BM25 focuses on the content of individual documents, PageRank looks at the relationships between documents.

4. LSI(Latent Semantic Indexing): BM25 and LSI are both algorithms used in Information Retrieval (IR) to rank documents based on their relevance to user queries. However, the two algorithms differ in their approach to matching query terms with document terms. BM25 is a term-weighting algorithm that computes a relevance score based on the frequency of query terms in the document and other factors like document length and query term frequency. In contrast, LSI (Latent Semantic Indexing) is a statistical technique that identifies latent semantic relationships between terms and documents and ranks documents based on their relevance to a query in a reduced dimensional space. LSI can capture the conceptual meaning of words and documents beyond their literal or surface-level meaning and is particularly useful for dealing with synonymy, polysemy, and other lexical ambiguities. The choice between BM25 and LSI depends on the specific requirements of the IR system and the nature of the document collection.

Next, we compare different approaches to Information Retrieval. In this paper, we use BM25 which is a probabilistic model for information retrieval that estimates the relevance of documents based on their similarity to the query. It is an extension of the earlier BM11 model, and it is widely used in modern search engines. We compare with other approaches:

1. Boolean Retrieval: Boolean retrieval is simple and precise but limited in scope, while probabilistic models are more flexible and effective for larger, more complex collections and queries. However, it can be difficult to construct a precise query, and the results can be limited when the query is not well-defined or when the collection is large. In contrast, probabilistic models are more flexible and can handle more complex queries and larger collections. They are also better at handling vague or imprecise queries by assigning a probability score to each document based on its relevance to the query.

2. Vector Space Model: Vector space models do not explicitly model the relevance of the documents to the query, while probabilistic models do. Vector space models also do not take into account the frequency of query terms across the entire document collection, while probabilistic models do. This means that probabilistic models can handle more complex queries and larger collections, and are better at handling vague or imprecise queries.

3. LSA Model: LSA is a deterministic model that relies on the manipulation of the term-document matrix to capture the semantic similarity between documents and queries, while probabilistic models are statistical models that estimate the probability of a document being relevant to the query based on the frequency of the query terms in the document and across the document collection. While LSA is effective in capturing the semantic similarity between documents and queries, it may not be as flexible as probabilistic models in handling complex queries and larger collections. Probabilistic models can handle more complex queries and larger collections and are better at handling vague or imprecise queries.

4. Machine Learning: Probabilistic models are rule-based and rely on pre-defined heuristics to model the probability of relevance, while machine learning models are data-driven and learn patterns from the data. Machine learning models can be trained on large datasets and can discover more complex patterns in the data than probabilistic models. However, they require significant amounts of labeled training data, which may be difficult or expensive to obtain.

6. Conclusion

We were able to implement the BM25 algorithm to our site which provides users with articles, and news according to their needs. We selected the BM25 algorithm because it is customizable according to the application and performs better as compared to other algorithms like TF-IDF (Term Frequency-Inverse Document Frequency) and Okapi BM25 (a variant of BM25). This makes BM25 an ideal candidate for our project. In the future, we might add other features to the website and improve its UI while following the HCI concepts we mentioned earlier and possibly add more concepts to make the website as user-friendly as possible and improve the user experience even more.

7. References

1. West, Jevin D., Ian Wesley-Smith, and Carl T. Bergstrom. "A recommendation system based on hierarchical clustering of an article-level citation network." IEEE Transactions on Big Data 2.2 (2016): 113-123.

2. Shani, Guy, and Asela Gunawardana. "Evaluating recommendation systems." Recommender systems handbook (2011): 257-297.

3. Ko, Hyeyoung, et al. "A survey of recommendation systems: recommendation models, techniques, and application fields." Electronics 11.1 (2022): 141.

4. Svore, Krysta M., and Christopher JC Burges. "A machine learning approach for improved BM25 retrieval." Proceedings of the 18th ACM Conference on Information and knowledge management. 2009.

5. V. N. Anh, O. de Kretser, and A. Moffat. Vector-space ranking with effective early termination. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 35--42, New Orleans, Louisiana, September 2001. ACM Press, New York.

6. V. N. Anh and A. Moffat. Simplified similarity scoring using term ranks. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 226--233, Salvador, Brazil, August 2005. ACM Press, New York

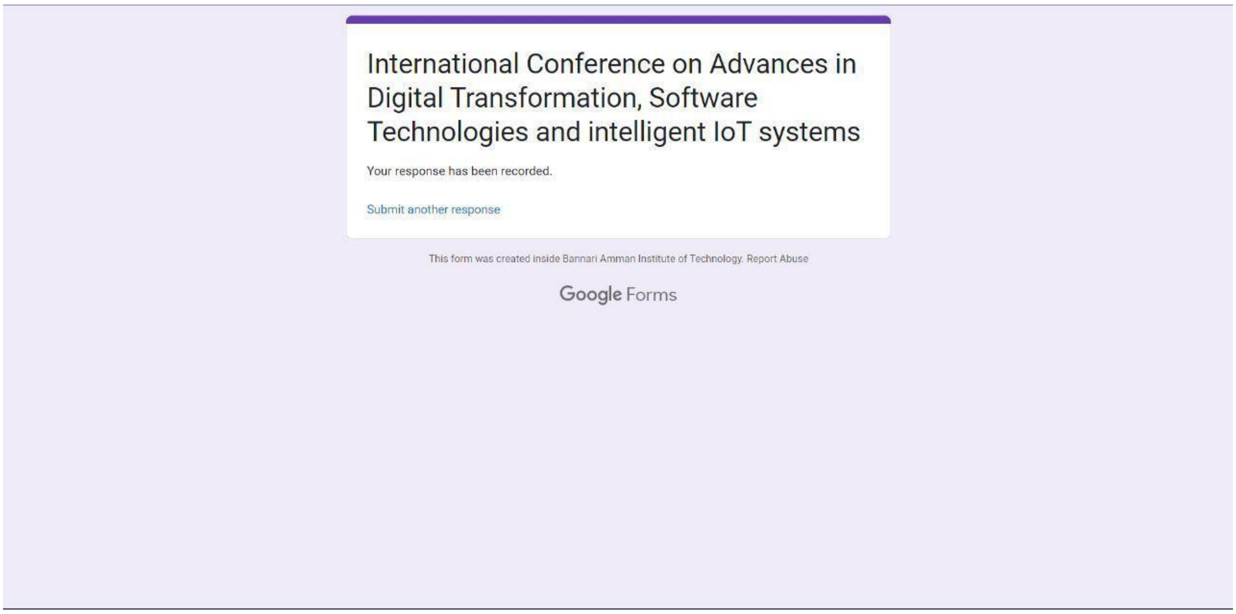
7. V. N. Anh and A. Moffat. Structured index organizations for high-throughput text querying.

8. Rowland Atkinson and John Flint. 2001. Accessing hidden and hard-to-reach populations: Snowball research strategies. Social research update 33, 1 (2001)

9. Su Lin Blodgett and Brendan O'Connor. 2017. Racial Disparity in Natural Lan- guage Processing: A Case Study of Social Media African-American English.

10. Daron Acemoglu and David Autor. 2011. Skills, tasks and technologies: Implications foremployment and earnings. In Handbook of labor economics .

Paper Submission:



International Conference on Advances in Digital Transformation, Software Technologies and intelligent IoT systems

Your response has been recorded.

Submit another response

This form was created inside Bannari Amman Institute of Technology. Report Abuse