International Journal in IT & Engineering (IJITE)
Volume 13 Issue 02, February 2026 ISSN: 2321-1776 Impact Factor: 8.256
Journal Homepage: http://ijmr.net.in, Email: irjmss@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal

DESIGN THE ARCHITECTURE FOR A TEXT-BASED WEB PAGE RANKING SYSTEM FOR A SEARCH ENGINE

Sridhar Devashish[1], Bharti Kumari[2]

[1]Research Scholar, Department of Computer science and engineering, SHEAT College of Engineering, Varanasi, UP

[2]Assistant Professor, Department of Computer science and engineering, SHEAT College of Engineering, Varanasi, UP

ABSTRACT

The most common challenge for web users is discovering fascinating and relevant information that confirms their interests on a constant basis, since the amount of information on the web grows dramatically. When a user puts a query into the web, the type of results that display as links are always the same. Furthermore, it searches the web for links based on terms found on the page, regardless of the client's needs. The research aim of this study is to propose an Architecture for a text-based web page ranking system for search engines. This paper illustrates the text based webpage ranking system for search engine. The method provides better precision and recall value than existing search engines like Ask, Google, Bing and Yahoo to rank the web pages.

Keywords: Re-ranking webpages, serch engines, retrieving webpages, tokenization, stemming etc.

INTRODUCTION

A search engine is defined as a web-based program that examines the results of a query entered by a user on the internet and returns an inventory of search results that are highly relevant to the term. It will be described as a bug that aids in locating the most relevant information sought by the user on the web. Google, Yahoo, Bing, Ask, and MSN Search are among the most well-known search engines. A search engine uses robots, bots, or spiders to explore the Web, going from page to page and website to website. They use the information acquired to create a web index for the website. A search engine comprises the following:

- A crawler that crawls through each page of a decipherable website and intends to be searched by using machine-readable text links on each page to make it ascertainable.
- An algorithm used to create index pages which are browsed.
- A matching algorithm that receives the search request, correlates it with the knowledge in the index, and provides the results.

A personalized meta-search engine recognizes users' interests, allowing the search engine to assist users in swiftly finding the information they seek. A personalized meta-search engine may organize the results based on the users' interests. Existing customized metasearch engines include savvysearch, cluty, dogpile, and IBoogle. Figure 1 illustrates the common differences between the metasearch engine and the proposed GWSP architecture.

Basically, each software employs its own algorithmic process to rank web sites, guaranteeing that only relevant results are returned for the user's inquiry. The selected question's outcome is subsequently presented on the program results page (SERP). The three major functions of an

exploration engine are crawling, classification, and matching, as indicated in the preceding article. Essentially, a web site ranking algorithm is utilized to locate user-entered queries and deliver appropriate results. Every search engine. It uses its own web site ranking system. The SERP provides the search results for the user's specified query. The main functions of a search engine are crawling, indexing, and matching.

Website rating varies amongst search engines since each has its unique ranking process. Each search engine uses its own algorithm to determine the ranking of web sites. Furthermore, because each search engine has a unique algorithm, no two search engines will provide identical results. For example, Google and Yahoo will not produce the same results for a search query.

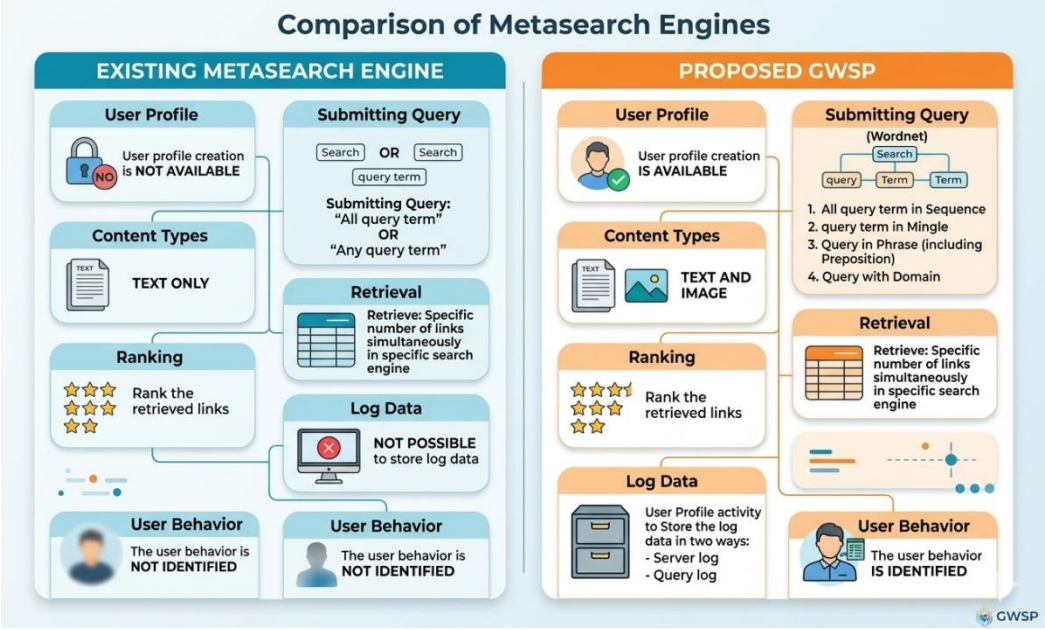

FIG. 1: FEATURE DIFFERENCES BETWEEN EXISTING METASEARCH ENGINE AND PROPOSED GWSP

There are two variables influencing the search engine ranking process: They are

1. On Page Factors,
2. Off Page Factors

ON-PAGE FACTORS

Essentially, on-page elements will operate on the web pages that are there in the website with the purpose of making the website search engine friendly and also to aid enhance the search engine's ranking process for the keywords searched. On-page criteria are critical for each website in order to rank the web page; publishers manage these elements. On-page criteria include title tags, meta tags, and website page content. The following are some on-page characteristics that influence the ranking process of web pages.

Proper Content of the page- The website's content should be unique and relevant to the theme. The website displays items related to the user's search and assists the searcher in providing a

positive experience when browsing the Web. In truth, search engines reject websites that are not regularly updated with up-to-date material. The webmaster will develop a website that has logical relationships between the material and keywords relevant to the information that is available on the web page, which aids in the search process on the web.

Title Tags- It provides search listing on the top of web browser and it is an HTML tag which is used to describe the text that present within the <title> tag. This title should give an accurate details about the content of the web page. Basically, each and every website and web pages may contain title tags and it is unique.

URL Structure- Figure 3 depicts an important part of search engine optimization: the URL structure contains main keywords such as protocol, host name or domain name, a directory, file name, and query arguments.

Meta Description Tag: This tag aims to offer a quick summary of the material found on the website or web page. Search engines will use these tags to assist users find reputable information. Figure 4 shows the Meta description tag.

XML Sitemap: The XML site map allows web crawlers to index the list of hyperlinks and URLs that are present on a website. Thus, having an XML sitemap for a website is a smart idea.

FIG. 2: TITLE TAG USED IN A HTML PAGE STRUCTURE

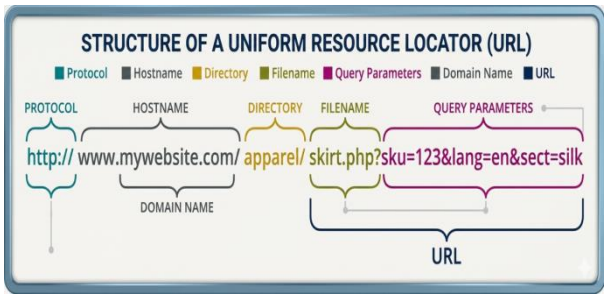

FIG 3: SEARCH BASED ON THE URL

FIG. 4: META DESCRIPTION TAG USED IN A HTML PAGE

OFF PAGE FACTORS

Aside from the constraints of the web page structure, off-page elements contribute to the improvement of the ranking process. Website publishers do not have direct influence over off-page ranking variables. The following are some off-page factors that influence the ranking of web pages.

Building Quality Link Building- Building quality links will improve the website's ranking with search engines. If reputable websites are linked to the website, it signifies that the website contains good and valuable material.

Forum Posting- Forum postings are used as online discussion websites where a group of individuals may exchange their perspectives on a certain issue. It might be available to the public or limited to a certain group. Participants in the debate may put topic-related links in their signatures.

Social Networking (SN)- Social networking is used to distribute and promote content from websites on social media platforms. It may raise awareness among users using social media sites such as Facebook, LinkedIn, Twitter, and Tumblr. It is really beneficial in terms of increasing website visitors, which helps to expand the network and mix with others. Figure 5 provides an example of social networking.

FIG. 5: EXAMPLES FOR SOCIAL NETWORKING

Videos: Video is one of the intriguing factors that can increase a website's search engine rankings. Essentially, it may give the information required by the user in an appropriate manner.

Blogging: Blogs are informational websites that display content in a certain sequence, with the user's most recent entries appearing first. It helps to generate fresh links to the website. Blogging is an excellent technique to focus online visitors at no cost.

RE-RANKING WEBPAGES

There are several web crawlers available now to recover data, including Google, Yahoo, Amazon, Bing, Ask, and others. The customer enters the inquiry after recovering the connection from the web crawler. The connections use re-positioning computation to recover the optimal connection provided to the client. Website page repositioning plays an important role in web frameworks. Palatable and relevant information disclosures from online usage information representation for engaging website suggestions are both challenging and necessary.

A ranking calculation identifies the watchwords in a web record or page. The web pages' positions are determined using precise criteria. For example, if the inquiry watchwords appear

near the top of a site, such as in the feature or the first few parts of content, it is assumed that the page is gradually relevant to the subject. Recurrence is a key aspect in assessing significance.

A web crawler studies the frequency with which watchwords appear in conjunction with other terms on a website page. A page with a greater frequency of catchphrases is considered more essential than other site pages. No two web indexes handle placement in the same way since each may use a different criterion for page positioning. For example, some web crawlers may collect more site pages and then list them, whilst others may file website pages on a regular basis, with less pages every time.

As a result, no web crawler has the identical set of website pages to search through. These typically cause contrasts when viewed in terms of results. Unfortunately, web index producers must be cautious with website administrators that attempt to misuse its positioning criteria in order to improve the positioning of certain sites for personal gain. As a result, their placement plans must be constantly updated.

The customer submits an inquiry; in the online stage, the client's query is first fed into an inquiry extension, which detects and eliminates ambiguity in the question before passing it to the internet searcher. Web crawlers get results from prolonged inquiries and proceed to the suggested re-positioning framework, which re-ranks the recovered results by breaking down the connection of a single web crawler.

Clients are becoming increasingly reliant on the online search tool placement strategy to locate information relevant to their needs. Frequently, the customer want to make data decisions based on the top-ranked outcome rather than report details. The thoroughly positioned report is more deceptive to the customer and more well-known. As a result, the placement framework identifies the most relevant link structure among the various web crawlers.

Web indexes now employ catchphrase-based inquiry types such as single various and expression finding approaches. These systems have a few demands.

1. At the season of pursuit process client can't express their aim
2. Watchword coordinating procedure is hard to distinguish the importance of the catchphrases.
3. Most coordinating watchword is eluded as best positioned outcome
4. Less quality positioning of website pages dependent on their fame of the web record.
5. New report isn't having high position esteems.

The desired findings are specifically relevant to the client's question. This framework assigns a high priority to the new page for rank progression. First, determine the best N results from the various web crawlers based on the customer inquiry. The following step is to compare the connections of various web crawler results and assign re-positioning characteristics. The underlying positioning quality is determined by the imperative score for the connections. Finally, by combining the imperative and likeness scores, additional positioning traits emerge.

OBJECTIVES OF THE STUDY

- To gather information regarding various web search engines and web personalization techniques using machine learning methods.
- To study text-based webpage ranking systems for search engines.

➢ To investigate various search engine techniques for retrieving phrase-based information using machine learning algorithms.

REVIEW OF LITERATURE

Haveliwala [1] Page Rank was determined to be computed for extremely large subgraphs of the web on devices with limited main memory. He employed personalized Page Rank scores to modify topic-sensitive online search. He concluded that the use of customized Page Rank scores would improve online search, but the number of hub vectors (e.g., the range of appealing web material used in a bookmark) was limited to sixteen due to computational constraints.

Jeh et al. [2] developed customized online searches that allowed users to adjust the global Page Rank algorithm using their bookmarks or homepages. Their work is primarily focused on worldwide relevance, taking use of the web's linking structure.

Phyo Thu Thu Khine et al. [3] highlight the sharpness of keywords in relational databases to boost the speed of finding needed keywords. A user does not want the data from the database structure or SQL. A user submits a list of keywords, and the system finds and ranks the relevant documents based on their prevalence. Three systems, DISCOVER, BANKS, and DBX, employ keywords to search relational databases. A query specifies a set of keywords and searches the relational database for them. Compartmentalization relative Database: each price in this column is considered a low-text document that may be utilized for keyword-based search. Query Cleaning: The system accepts a question as input and returns a 'clean' query result. This may be accomplished by removing the stop words from the question, which are useless. As a result, the outcomes may not meet the user's expectations.

Sumit Bhatia et al. [4] created a method to automatically classify online search queries into five categories: low frequency high entropy, low bandwidth low entropy, high frequency, low entropy, and high frequency and high entropy. This research employs commercial search engine user queries and click entropy. The query is then reformulated to get better results, both query-based and click-based.

Zhou Hui et al. [5] emphasize the Search Engine Work Principle. This analytical study effort focuses on the following topics: search engine optimization, search engine principals, search ranking factors, and web site search engine optimization techniques. Search Engine optimization refers to web site authoring, web page content, and web structure, which outline how Search Engines operate on web content ranking methods. Search Engine encompasses the following three categories: Full-text Search Engine, Meta Search Engine, and Directory Search Engine. Issue affecting search Ranking is determined by three important factors: search matching, density, distribution, and webpage tag labels.

Robin Sharma et al. [6] suggests the design of linguistic data retrieval to improve search results. The formula is intended for the linguistic indexing of web pages. Semantic web search is the search for a specific meaning for a user's query. The meaning of the question is concealed inside the question itself. Words like "What" Why, once influences the question, which signifies terrible. Traditional searches omit or ignore interrogational terms from questions, and page rankings are based on major keywords. The search goes to an ancient search, which does not appear to satisfy users.

John B. Killoran, [7] focuses on two important points: 'key concepts' and 'essential lessons' for increasing website exposure through the use of Search Engine Optimization techniques. Skilled professionals create material on the internet for themselves and/or customers on websites they manage. They utilize the website title, key phrase, website name, or the name of the entity that controls it. Search rankings force website developers to constantly evaluate the precise measure of their competitive fitness and gaps for each particular query.

S.G. Choudhary et al. [8] provide certain semantic methodologies and present entirely alternative algorithms for page ranking and metaphysics. The algorithms supported linguistics. Search can be done in the following ways: PSSE: custom linguistics. The search engine uses the user's profile, ranking score, and metaphysics to calculate tailored issues, which helps to generate a large number of customized results. Google Search Engine uses user profiles to tailor search results based on their location and prior search interest results. PSSE is designed in two parts: offline and online. The offline portion is made up of crawl and preprocessing procedures. The online element contains the question and answer procedure, as well as the rating of results.

Shikha Goel et al. [9] present the technique to Search Engine Analysis that is based on page level keywords. Page level keywords are the keywords present on each pages of a website. Page level keywords are an ineffective way to maintain the connectivity of Search Engine results. A user creates a query, and a search engine designer creates a database for it. The queries are then executed by users to compute page level keywords, as well as unit calculations. Keywords are word groupings that users use to look for things on search engines.

Arooj Fatimah et al. [10] highlight the limitations of traditional data retrieval systems such as keyword-based search and emphasize the power of semantic search. Traditional data retrieval systems require particular types of questions from users, but users do not need to worry about negative pre-defined structures. This research developed a strategy for excellent data retrieval and user expertise in linguistics search.

Sachin Pardeshi [11] discusses the current methods for personalizing web search, with a focus on the study, comparison, and implementation of several customized web search methodologies that are extensively utilized today. The increasing abundance of information causes overload on the internet. The search engine helps to retrieve information from the internet.

Lei Duan et al. [12] investigated the uncommon characteristics known as outlying aspects of numerical data. Based on outlier analysis research, they attempted to extract anomalous behaviour of objects in comparison to existing data. They contrasted the concept of outlyingness in subspaces. They ranked it using a probability-based technique. Their investigation looked at both actual and synthetic data sets. The strategy is proven to be more efficient in identifying outlying aspects in dimensional numeric data sets.

Vajenti Mala and colleagues [13] have presented a system that combines the two strategies, semantic and keyword, based on online searching and retrieval information, similar to a new search engine. Semantic search engines and keyword search engines are chosen depending on the precision ratio and natural language queries. The semantic search engines picked are Wikipedia, Google, and Yahoo, while the keyword search engines are Hakia, Bing, and Duck Duck Go. Finally, the performance evaluation of relevant and non-related documents outperformed semantic and keyword search engine features.

Luigi Laura et al. [14] developed a search engine for semantic illegal content hunter (SICH), which is used to automatically detect unlawful information on the internet. The SICH system is divided into three sections: crawler, indexer, and query processor. It produces superior results while searching, downloading, filtering, grouping, and organizing illicit information found in unstructured text documents.

Hikmat A et al. [15] suggested a Profile-Based Semantic Method with Heuristics for Web Search Personalization. The search results do not match user preferences since the search is keyword-based rather than semantic-based. Exploiting user profiles through the use of semantic web technologies for customization may result in a step forward in future retrieval systems.

Amelec Viloriaa et al. [16] proposed the ACVPR algorithm and metasearch system design. This search tool uses sophisticated and advanced technologies such as the semantic web and Hadoop2 big data analysis to forecast meta keywords to broaden the user's search query and handle a vast number of web links provided by background search engines. The suggested knowledge representation system may recreate domain models found in a variety of hypermedia systems.

Jesus Silvaa et al. [17] offer the ACVPR algorithm and architecture of a metasearch system, IMSS-P, for the construction of a surveillance technology and competitive intelligence system for the SMEs. The system consists of two primary elements. The first one assists the user in collecting prerequisites for the search process, resulting in more complete search keys than those utilized by search engines. The second module is in charge of web mining, which involves investigating links from URLs returned by the public interfaces of the most popular search engines.

Mhawi et al. [18] developed a revolutionary advanced document indexing technique (ADIM) that incorporates an evolutionary algorithm. The suggested IRS consists of three major phases. The first stage (i.e., the advanced documents indexing method) involves preprocessing, which consists of two steps: dataset document reading and the advanced documents indexing method (ADIM), which results in two tables. The second stage is the query search algorithm, which generates a list of words or keywords and retrieves relevant articles.

ARCHITECTURE FOR TEXT MINING BASED WEBPAGE RANKING SYSTEM

The suggested algorithm ranks webpages based on commonalities across search engines such as Google, Yahoo, Bing, and Ask. This search engine searches for results based on user queries using a web crawler. It generates multiple online links that may or may not be related to the search query. Text mining algorithms are used to search for commonalities in online pages and generate clusters in various search engines. After clustering, use the ranking function to discover the top 'n' webpages, and then categorize the pages based on their rank. Figure 6 illustrates the general flow of the suggested technique.

RETRIEVING WEB PAGES

The principal portions of the SERP feature a list of web sites received from the web in response to the user query, but the resultant page may also include web pages related with the adverts. In general, SERP results are classified into two types: organic search, which retrieves sites using the search engine's algorithms, and sponsored search, which retrieves pages that are adverts. The resulting web sites are generally ordered based on their relevancy to the user's query.

Each result presented on the SERP often comprises the title of the web site, a link to the real web page on the web, and a brief explanation indicating that the retrieved page may match the text entered in the term when searching. In the event of sponsored results, the Advertiser will pick which material to display.

FIG. 6: ARCHITECTURE FOR TEXT MINING BASED SEARCH ENGINE

PREPROCESSING

Many text mining applications rely heavily on pre-processing techniques to remove noisy input. This is a first phase in the text mining process. In this study, three preprocessing approaches were utilized to get SERP from multiple search engines: tokenization, stop word removal, and stemming.

TOKENIZATION

Tokenization is the process of separating a line of text into words, phrases, symbols, or other significant structures known as tokens. The token collection serves as input for extra processing, such as parsing or text mining. Tokenization is a key part in creating lexical analysis, which is valuable to both linguistics and computer science. Figures 7 show the tokenization of entire web page text, the operator separator, and the transition to the next phase, Stop word elimination.

STOP WORD

Natural language includes terms that are widely used in stop words. The reason that specific stop-words must be removed from an input text is because they make the input text look weighted and less significant to analysts. Eliminating stop words instantly reduces the dimensionality of term space. The articles, prepositions, and pronouns in the text document containing generic terms do not produce a document. The basic step is to eliminate stop words from documents that do not appear to be computed as keywords in text mining applications (TMA). To, a, the, of, and from are examples of terms that appear often in a group but do not discriminate for process. Stop words reduce the index size. The goal of information retrieval has been to limit or eliminate the usage of stop words. Using improved index compression and weighting stop terms depends on query processing. Figure 8 demonstrates how to delete stop words such as of, the, then, that, we, are, as, and at.

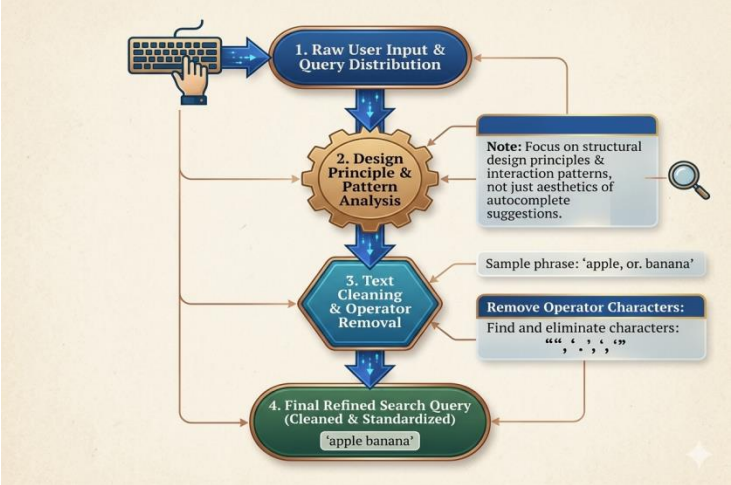

FIG. 7: TOKENIZATION PROCESS

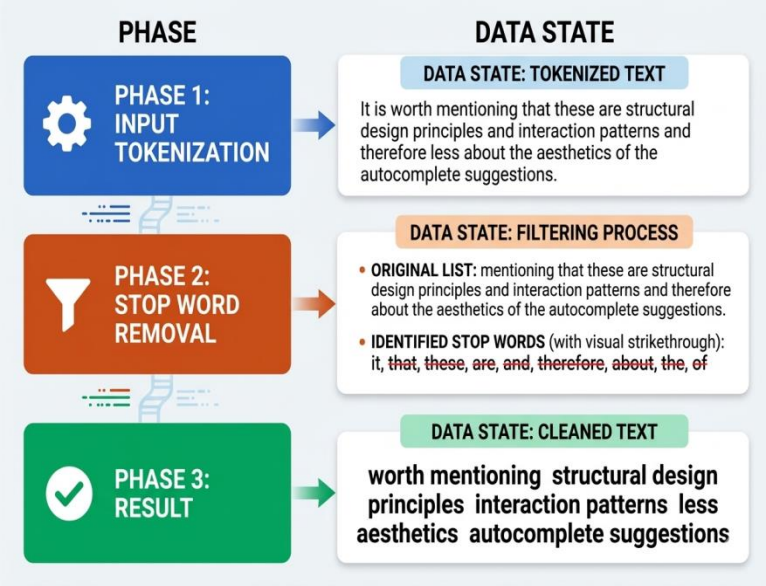

FIG 8: REMOVAL OF STOP WORDS

STEMMING

Conflation is an alternate word for stemming. This is done to lessen the discrepancies between words caused by modulation or derivation from a similar stem. This approach is used to determine the root (stem) of a word. Select, chosen, selecting, and selections are examples of terms that can be derived from the word'select'. The goal of this technique is to eliminate multiple suffixes, reduce the number of words, create similar stems, and save time and memory space. The stemming process enhances efficacy by identifying similarities between the query and the relevant document. Users looking for the term "swimming" may be interested in documents containing the phrase "swim". It reduces the term index by 17% and uses lossy compression. Porter's stemming technique was employed in this study to eliminate suffixes such as "ing", "ion", "ious", and so on. Figures 9 depict the stemming procedure for determining the specific keywords.

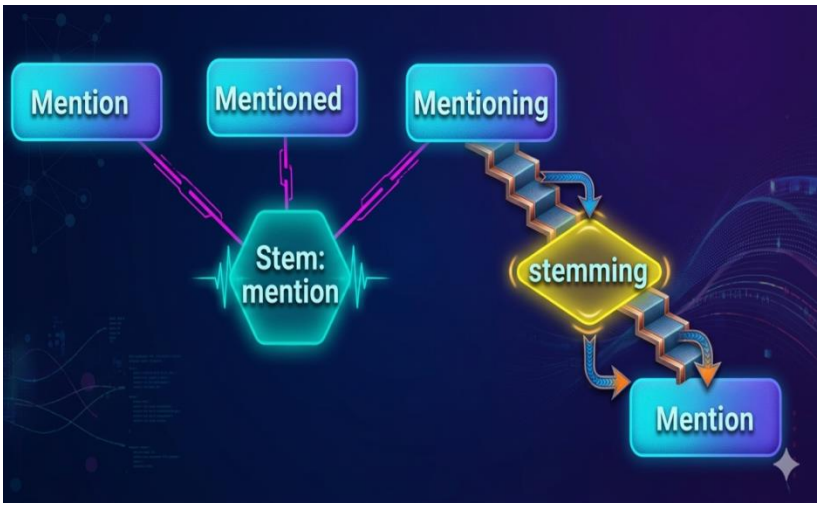

FIG. 9: EXAMPLES FOR STEMMING PROCESS

CLUSTERING

Normal online search engines have an issue in that they list a large list of documents from the search results, making it difficult for the user to identify the correct document. To address this issue, a clustering approach is utilized to assist the user in locating important search results as clusters based on the similarities observed in the resulting web pages. Such similarities can be detected using the extracted key characteristics. Thus, the similarity of the main traits was examined, and clusters were constructed based on the features. In this study, the LBG technique is used to frame clusters of comparable web pages.

LBG ALGORITHM

LBG algorithm named from Linde, Buzo, Gray is used to create cluster of similar web pages. This algorithm works similar to the K-means clustering method with given set of input values s= $\{x_i \in R^d | = 1,2,\ldots . ,n \}$ as input and creates a subset of center vectors C = $\{C_J \in R^d | j = 1,2 \ldots, K\}$ with a user defined K << n as output depending on the similarity measure. For Vector Quantization (VQ), d = 16, K = 256 or 512 are the standard values assigned for execution. The following steps 1 to 8 are explained for grouping of similar web pages link.

Algorithm 1: LBG Clustering Algorithm

1. Input : TV = $\{x_1 \in R^d | i = 1,2, \ldots., n\}$.
2. Initiate: a CB C $\{C_j \in R^d | j = 1,2 \ldots., K\}$.
3. Set $D_0 = 0 and\ let k = 0$.
4. Group 'n' TV into K clusters according to $x_i = \in s_q if ||x_i - c_q||_p \leq ||x_i - c_j||_p for\ j \neq q$.
5. Update CC $c_j j = 1,2, \ldots, K\ by c_j = \frac{1}{s_J} \sum x_i \in s_j X_i$.
6. Set K ← + 1 and calculate the distortion $= \sum_{j=1}^{k} \sum x_j \in s_j ||x_j - c_j||_p$.
7. IF $(D_{k-1} - D_K)|\ D_k > \in$ (a repeat steps 4 ~6.
8. Output: the CB $= \{c_j \in R^d | j = 1,2, \ldots, k\}$. .

The LBG method is based on the initial codebook C, which is the distortion and the threshold; in implementation, we must provide the maximum number of iterations to ensure convergence.

RANKING WEB PAGE

Page rank has played an important part in the history and development of search engines. The page rank algorithm is one of the most important components of Google's ranking system; it is a link analysis method that assigns a number or rank to each hyperlinked web site on the World Wide Web. Stanford University created the page rank algorithm for Google in 1995. This algorithm allows websites to rank all web pages connected to them depending on the amount of links or visits. One important note is that this method ignores meta tags in order to prevent spam indexing misuse and other deceptive acts. The primary function of the page rank algorithm is to present a list of web sites in the order of most significant to least important, based on the result page provided by the search engine when a keyword search is performed.

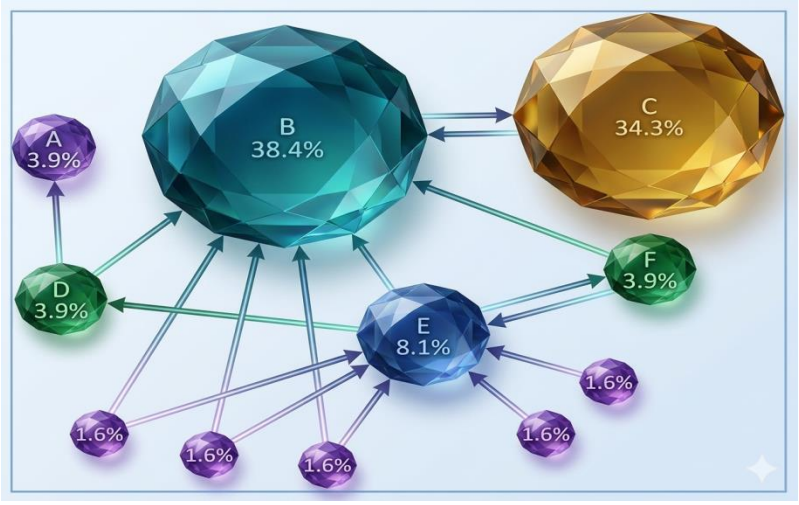

FIG. 11: RANKING THE WEB PAGES BASED ON GOOGLE'S PAGE RANK ALGORITHM

The primary operation in the page rank algorithm is determining the amount of links associated with a certain web site. If a highly rated website links to a web page, that web page is likewise highly ranked. Because highly rated websites may have useful links, a web page linking to highly ranked websites may contain important material. Thus, a web page with a big number of connections may be considered highly ranked under the democratic method of page rank algorithm. The basic essence of the page rank algorithm is to assess how frequently a person visits a specific online page after clicking one or two links on other websites, or by searching the web page based on text phrases using one or two keywords. Finally, the ranking was determined using the PR (page rank) algorithm, which took into account not only the number of links tied to the website, but also the validity of the data included in the linked webpage. Google's page rank algorithm is a mechanism for determining the significance and relevance of a web site among other web pages on the WWW.

Rank = $(P_S)/(N_S)$ (1)

Where s∈ {set of search engines},

PS=Each Page link from s,

NS=number of pages from s

Google re-indexes every month, therefore the page rank of all web pages may vary each month. Aside from rating, the site rank algorithm says nothing about the language of the information accessible on the web site, the amount of the data, or the string used in the hyperlink.

EXPERIMENTAL RESULTS

The proposed strategy is tested using a variety of search engines. The assessment is performed utilizing the Java platform and search engines such as GYBA. To assess the performance of the suggested technique, the following parameters are utilized and defined.

PRECISION

Within the context of information retrieval, precision is defined as the portion of retrieved documents that are relevant to the query. Precision evaluates all retrieved documents and asks how near they are to the goal concept, but it may also be evaluated at a specific cut-off rank, taking into account just the highest results generated by the system, as indicated in Equation 2.

$$Precision = \frac{|Relevant\ Document \cap Retrieved\ Documents|}{|\{Retrieved\ Document\}|} \quad (2)$$

RECALL

In information retrieval (IR), recall refers to the portion of documents that are successfully retrieved that are relevant to the query. Equation 3 shows how successfully a search identifies every conceivable document that could be of interest to the searcher.

International Journal in IT & Engineering (IJITE)
Volume 13 Issue 02, February 2026 ISSN: 2321-1776 Impact Factor: 8.256
Journal Homepage: http://ijmr.net.in, Email: irjmss@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal

$$Recall = \frac{|\{Relevant\ Document\ \cap Retrieved\ Documents\}|}{|\{Retrieved\ Document\}|} \quad (3)$$

Relevant documents are webpages that have a rank of less than 0.5 across all search engines. The number of papers retrieved equals the number of web links displayed at the end.

TABLE 1 : RANK BASED COMPARISON FOR THE PROPOSED AND OTHER WEB SEARCH ENGINES

Top Links For the search "Data Mining"	Google	Yahoo	Bing	Ask	Proposed
http://en.wikipedia.org/wiki/Data_mining	1	1	1	1	1
http://www.britannica.com/technology/data-mining	8	6	5	4	2
http://www.statsoff.com/textbook/Data-Mining-Techniques	10	4	4	8	3
http://www.investopedia.com/terms/d/datamining.asp	5	7	6	7	4
http://www.techpedia.com/defeinition/1181/data-mining	6	8	9	6	5
http://thearling.com	7	5	7	9	6
http://economintimes.indiatimes.com/definition/data-mining	9	9	10	5	7

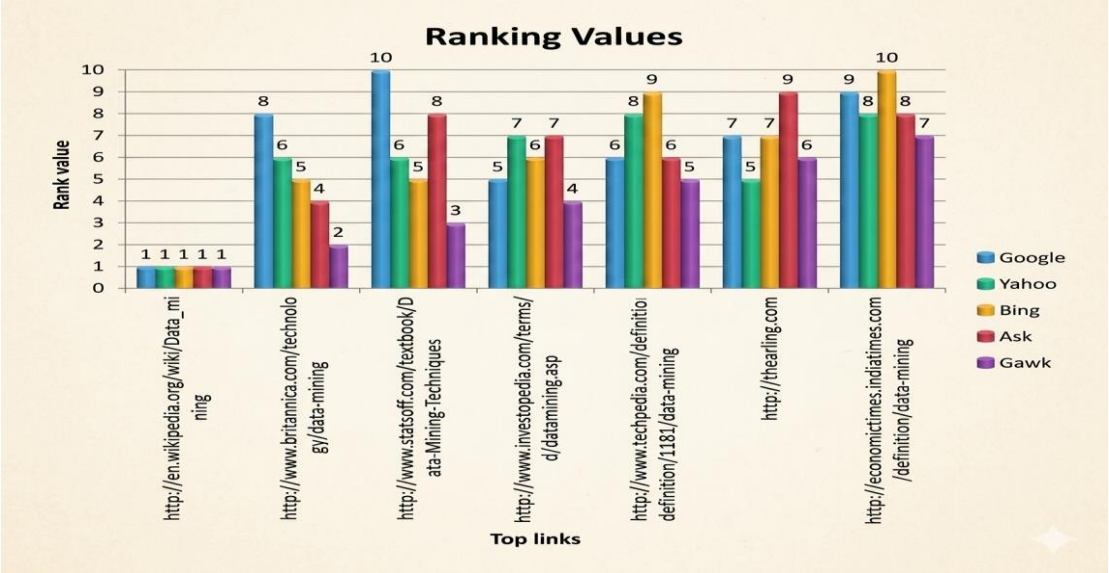

FIG. 12: RANK BASED COMPARISON FOR THE PROPOSED AND OTHER WEB SEARCH ENGINES

Figure 12 and Table 1 describe the Multi Keyword Page Rank search Algorithm comparison between the present SE and the proposed Search Engine based on Rank value. Among other things, the suggested framework provides appropriate rankings for web sites when compared to other search engines.

TABLE 2: COMPARISON OF SEARCH ENGINES USING PRECISION VALUE

Search Engines	Precision Value
Google	0.5
Yahoo	0.55
Bing	0.4
Ask	0.5
Proposed	**0.7**

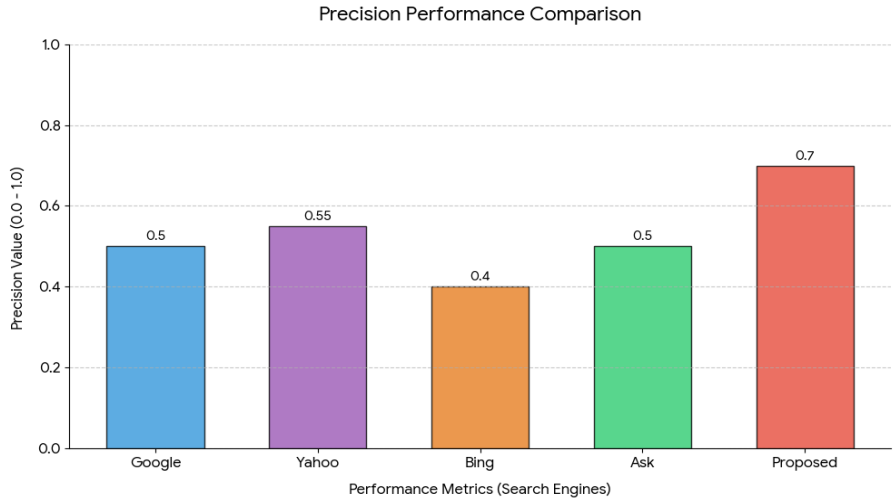

FIG. 13: COMPARISON OF SEARCH ENGINES USING PRECISION VALUE

Figure 13 and Table 2 compare the Multi keyword Page Rank search algorithm to the present SE and the new framework based on precision value. It is observed that the new framework has a high accuracy value of 0.7 when compared to other search engines.

TABLE 3: COMPARISON OF SEARCH ENGINES USING RECALL VALUE

Search Engines	Precision Value
Google	0.5
Yahoo	0.55
Bing	0.4
Ask	0.5
Proposed	**0.7**

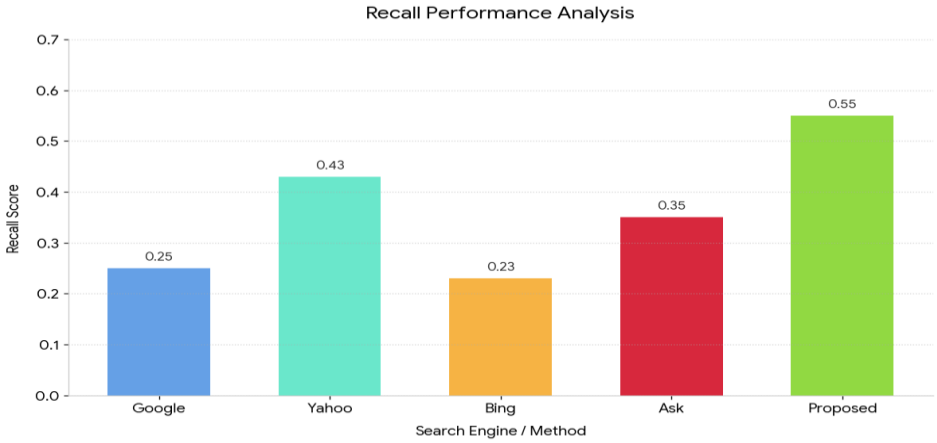

FIG. 14: COMPARISON OF SEARCH ENGINES USING RECALL VALUE

Figure 14 and Table 3 explain the performance of the Multi Keyword Page Rank search algorithm compared to existing search engines and the suggested search engine based on the recall value. It should be noted that the proposed search engine has a high recall value of 0.55 when compared to existing search engines.

CONCLUSION

The search engine returns URLs as search results, and the user must select the most appropriate link. Obtaining the information that the user requires takes time and effort. As a consequence, customizing search results on the web is the greatest answer to the above-mentioned difficulty that consumers face on the online. In this research various search engines and its methodologies are discussed. This paper also illustrates the text based webpage ranking system for search engine. A combination of machine learning methods and text mining techniques were applied to obtain a search engine using multi keyword and produces accurate web pages matching user requirement based on the rank value of webpage generated by novel web page ranking algorithm. The novel method provides better precision and recall value than existing search engines like Ask, Google, Bing and Yahoo to rank the web pages.

REFERENCES

1. Taher H. Haveliwala (2002),” Topic-sensitive PageRank”, Proceeding WWW '02 Proceedings of the 11th international conference on World Wide Web, Hawaii, USA,ACM,pp. 517-526, 2002.
2. G. Jeh, J. Widom (2003), ”Scaling personalized web search”, Proceedings of the 12th international conference on world wide web, WWW ’03, New York, NY, USA, ACM , pp. 271-279, 2003.
3. Phyo Thu Thu Khine, Htwe Pa Pa Win, Khin New Ni Tun,” Indexing Relational Databases for Efficient Keyword Search”, International Journal of Scientific & Engineering Research, Volume 2, Issue 10, ISSN 2229-5518, 2011.
4. Sumit Bhatia, Cliff Brunk, PrasenjitMitra,” Analysis and Automatic Classification of Web Search Queries for Diversification Requirements”, ASIST 2012, October 28-31, 2012.

5. Zhou Hui, Qin Shigang, Liu Jinhua, Chen Jianli, "Study on Website Search Engine Optimization", International Conference on Computer Science and Service System IEEE, 2012.
6. Robin Sharma, Ankita Kandpa and Priyanka Bhakuni, "Web Page Indexing through Page Ranking for Effective Semantic Search", Proceedings of7'h International Conference on Intelligent Systems and Control IEEE, ISBN: 978-1-4673-4603-0, 2012.
7. John B. Killoran, "How to use Search Engine optimization techniques to Increase website visibility", Transactions Professional Communications IEEE, ISBN: 0361-1434, 2013.
8. S.G.Choudhary, S.R.Kalmegh, Dr. S. N. Deshmukh, "Semantic Search Algorithms based on Page Rank and Ontology: A Review", 3rd International Conference on Intelligent Computational Systems (ICICS'2013) January 26- 27, Hong Kong (China), pages 17-20, 2013.
9. Shikha Goel ,Sunita Yadav, "Search Engine Evaluation Based on Page Level Keywords", 3rd IEEE International Advance Computing Conference (IACC), 2013.
10. Arooj Fatima, Cristina Luca, George Wilson, "User Experience and Efficiency for Semantic Search Engine", IEEE, 2014.
11. Sachin Pardeshi, "Literature Survey on Web Personalization International Journal of Scientific Research in Science, Engineering and Technology (ijsrset.com), 2015.
12. Lei Duan, Guanting Tang, Jian Pei, James Bailey, Akiko Campbell, Changjie Tang , "Mining Outlying Aspects on Numeric Data", Data Mining and Knowledge Discovery, vol. 29,no. 5, pp.1116-1151, 2015.
13. Vajenti Mala, D. K. Lobiyal,"Semantic and keyword based web techniques in information retrieval", IEEE 2016 International Conference on Computing, Communication and Automation (ICCCA), ISBN: 978-1-5090- 1666-2, 2016.
14. Luigi Laura, Gianluigi Me,"Searching the Web for illegal content: the anatomy of a semantic search engine", methodologies and application, Soft Computing, pp.1245–1252, 2017.
15. Hikmat A, M. Abdeljaber," Profile-Based Semantic Method using Heuristics for Web Search Personalization",Digital Object Identifier, (DOI) : 10.14569/IJACSA.2018.090926, 2018.
16. Amelec Viloriaa , Omar Bonerge Pineda Lezama," An intelligent approach for the design and development of a personalized system of knowledge representation", International Workshop on Web Search and Data Mining (WSDM) April 29 - May 2, , Leuven, Belgium, Procedia Computer Science 151 (2019) 1225–1230, 2019.
17. Jesus Silvaa, Lucelys del Carmen Vidal Pachecob , Kevin Parra Negretec Johana Cómbita Niñod , Omar Bonerge Pineda Lezamae , Noel Varelaf, "Design and Development of a Custom System of Technology Surveillance and Competitive Intelligence in SMEs", International Workshop on Web Search and Data Mining (WSDM) April 29 - May 2, Leuven, Belgium, Procedia Computer Science 151 (2019) 1231–1236, 2019.
18. Mhawi, D. N., Oleiwi, H. W., Saeed, N. H., & Al-Taie, H. L. (2022). An efficient information retrieval system using evolutionary algorithms. Network, 2(4), 583-605.