Grenze International Journal of Engineering and Technology, June Issue

Advancements in Search Engines

Bharath Kumar[1], M Kavya Sree[2] Pavandath Reddy[3] and Dr.T.Prathima[4]
[1-4]Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, India
Email: bk521878@gmail.com, 160120737066kavyasree@gmail.com,160120737097pavandath@gmail.com, and prathimareddy.t@gmail.com

***Abstract*—In the never-ending digital realm, searching for particular information is a vital aspect of the Internet and is carried out by millions of users daily. More than half of all website visitors arrive there from a search engine, rather than actually typing in the URL. Search engines stand as the gateways to an ocean of information. Gaining insight into the evolutionary progression and operational mechanisms of these search engines is imperative for the facilitation of effective information retrieval processes. The domain of Information Retrieval (IR) is dedicated to structuring, storing, and effectively retrieving information from diverse sources. At the core of IR lies the formidable challenge of bridging the gap between the user's information needs and the wealth of data available an effort which is made easier by search engines. Despite their remarkable ability to access vast information, challenges persist, including information over load, potential misinformation propagation, privacy concerns, contextual understanding limitations, and filter bubble creation. Understanding these drawbacks highlights the imperative for ongoing research and innovation to mitigate these shortcomings and optimize the user search experience. While search engines have revolutionized information retrieval, it is important to acknowledge and address the challenges they encounter. Through dedicated research and comprehensive analysis of diverse sources, this project is intended to uncover innovative solutions that can enhance search engine capabilities. By confronting issues such as information overload, misinformation, and privacy concerns head-on, we aspire to pave the way for more efficient, reliable, and user-centric search engine technologies in the future.**

***Index Terms*— Information Retrieval, Crawling, Indexing, Bidirectional Encoder Representation from Transformer, Natural Language Processing.**

I. INTRODUCTION

In the contemporary digital landscape, the ability to effi ciently access and retrieve pertinent information from the vast expanse of the internet is fundamental. With millions of users relying on search engines daily, these tools serve as crucial gateways, facilitating the exploration of the vast realm of information available. This research delves into the domain of Information Retrieval (IR) and its essential role in shaping the landscape of digital knowledge dissemination. Research in the f ield of search engines has advanced across various domains, encompassing user behavior analysis, extraction mechanisms, location-based effects, XML keyword search efficiency, di versification algorithms, reinforcement learning for result re ranking, association between click rates and relevance, influ ence of search engine marketing, metasearch engines, web search user interfaces, image-based search engines, page rank ing improvement strategies, metadata-driven search engine selection, category - based PageRank methods, and performance evaluation in code retrieval tasks. These contributions

Grenze ID: 01.GIJET.10.2.535

collec tively reflect the multifaceted progress in understanding and optimizing search engine functionalities.

II. RELATED WORK

Working Of Search Engine:
Search engines play a crucial role in today's digital world, enabling users to navigate the vast expanse of the internet with ease and efficiency. Understanding the inner workings of these powerful tools unveils the intricate algorithms and processes that underpin their ability to deliver accurate and timely results. At the heart of search engine operation lies a three-step process: crawling, indexing, and ranking. Crawl ing involves systematically exploring the internet, gathering information from web pages, images, and videos. This meticulously extracted data forms the foundation of the index, a comprehensive repository of indexed documents. When a user poses a query, the search engine meticulously analyses it, ensuring clarity and consistency. The query is then matched against the index, identifying documents that align with the user's intent. Sophisticated algorithms evaluate the relevance of each matched document, considering factors like term frequency, document importance, and freshness. The most relevant documents are then presented on the search results page, guiding users to the information they seek. This intricate process, powered by advanced algorithms and continuous user feedback, enables search engines to connect users with the information they seek, transforming the way we interact with the vast repository of knowledge available on the internet. Everyone of us had thought at least once how search engines work so fast. The short answer is distributed computing. Search engines like Google and Bing employ a distributed computing architecture to handle the vast volume of search queries and deliver results quickly. Instead of relying on a single server, they utilize a network of hundreds, thousands, or even hundreds of thousands of servers distributed globally. When a user initiates a search, the query is distributed among these servers, each responsible for retrieving a portion of the results. This parallel processing approach eliminates potential bottlenecks and ensures that users receive their results promptly, even during peak traffic periods.
B. Evolution of Search Engines Before computers had screen and keyboard how did people store data? The credit goes to Joseph Jacquard, a French weaver and merchant. His invention involved the use of punched cards to control the pattern being woven by a loom. The punched cards contained the instructions for the pattern, and by placing these cards in a specific sequence, weavers could produce intricate patterns in textiles. Charles Babbage, a 19th-century mathematician and inventor, introduced the concept of punched cards for use in his proposed analytical engine, a mechanical general-purpose computer. Using this punchcard technique, computers took instruction, did math, solved equations. How did this work?. Simple , holes punched in the card represent data to be placed in the computer Then computers got screens and keyboards.But we still couldn't talk to it like you'd talk to a human. We had to write it in code(C colon, slash carat smartdrv dot exe.)So how do people access this data. Then came search engines. These are gateways (like a door) standing between user and internet.
1) Influence of World Wide Web:WWW is like a digital library consisting of tons and tons of information.So users needed a way to find relevant information quickly as traditional methods of file directories were insufficient for the dynamic and expansive nature of the WWW.This gave birth to search engines.Platforms like WebCrawler (1994) and Lycos (1994) were among the early attempts to index and search web content.But what mattered was how accurately the user is being directed to these webpages. To make content search able, search engines began using keyword-based indexing. In this process, automated crawlers(robots also know as spider) explore web pages, extracting textual content and identifying significant keywords that encapsulate the main themes of the content. These keywords are then cataloged in an index, forming an associative map between keywords and the web pages containing them. When users input search queries, the search engine's algorithm scans the index to identify and rank relevant pages based on factors like keyword frequency and context. The ranked results are then displayed to users, allowing them to click on links for access.The earliest search engines, like Archie (1990) and Gopher (1991), were primarily based on keyword matching. Users entered queries, and these systems retrieved pages containing the specified keywords. The disadvantage was it may not accurately capture the context or intent behind a user's query, leading to instances of irrelevant or overly broad results.
2) PageRank Algorithm: Co-founded by Larry Page and Sergey Brin, PageRank algorithm introduce by Google in 1998 revolutionized the landscape of search engines by introducing a novel approach to result ranking. Moving from traditional keyword-based indexing, PageRank focused on evaluating the hyperlink structure of the web. Google's algorithm treated each hyperlink as a ”vote” for a page's significance, considering not only the presence of keywords but also the quality and quantity of links pointing to a particular page. This innovative method aimed to surface pages that were not just keyword-relevant but also perceived as authoritative within the vast expanse of the web. By prioritizing content with a higher number of quality inbound links, PageRank

significantly improved the relevance and reliability of search results, setting a new standard for search engine algorithms and influencing subsequent develop- ments in the field.

3) Crawling and Indexing: With the proliferation of web content, search engines faced the challenge of efficiently cataloging and retrieving information. Crawling, the process of discovering and accessing web pages, became more intricate due to the vastness of the web. The WWW's influence led to the development of sophisticated crawling algorithms that could navigate the complex structure of interconnected pages. Additionally, indexing techniques evolved to handle the diverse and dynamic content available on the web. Traditional keyword based indexing expanded to include more advanced methods that took into account factors like relevance, authority, and user experience. The interconnected nature of the WWW also prompted the introduction of hyperlink analysis, where the relationships between web pages, measured through links, played a crucial role in determining a page's significance.

4) Ranking Algorithms:The evolution of search engine algorithms beyond keyword-centric approaches represents a pivotal shift in the quest for more refined and contextually meaningful search results. Google, at the forefront of this transformation, introduced algorithms such as Florida, Hummingbird. Penguin, Panda and many more. The Florida update in 2003 marked a pivotal moment in search engine history. It targeted over-optimization tactics that sought to manipulate rankings through excessive keyword optimization. This shift in approach led to significant fluctuations in website rankings, prompting a re evaluation of SEO strategies to align with Google's evolving algorithms. Another significant algorithm, Panda (2011), focused on content quality as a determinant of search rankings. It assessed the overall quality and uniqueness of content on web pages, penalizing sites with thin, duplicate, or poorly written content. User experience considerations, such as page layout and ad-to-content ratio, were also factored into Panda's evaluation, emphasizing the importance of providing valuable and engaging content. In 2012, Google introduced the Penguin algorithm to tackle webspam and manipulative link building practices. Penguin scrutinized the quality of inbound links, penalizing websites engaged in tactics like link schemes and keyword stuffing. Notably, this algorithm operated in real-time, allowing for quicker adjustments and recoveries for websites that promptly addressed identified issues. Hummingbird in 2013 to delve into the intricacies of user intent. This algorithm went beyond mere keyword matching, incorporating semantic search technology to comprehend the context and meaning of words within queries. Hummingbird excelled in recognizing synonyms, processing conversational queries, and delivering more nuanced search results reflective of user intent. In essence, these algorithms were introduced to enhance the quality and relevance of search results, addressing issues ranging from webspam and link manipulation to content quality and user experience. The iterative nature of these updates underscores Google's commitment to refining its algorithms, ensuring that search results align with user expectations and intent.

5) Natural Language Processing : Natural Language Processing (NLP) has revolutionized the functionality of search engines by introducing a more sophisticated and context aware approach to user queries. Unlike traditional keyword based indexing, NLP enables search engines to understand the intricacies of language, including synonyms, contextual meanings, and user intent. This transformation is particularly evident in voice search, where NLP allows for more conversational interactions. BERT, an NLP algorithm introduced by Google has transformed search queries by enhancing the understanding of context and semantics.. BERT stands for Bidirectional Encoder Representations. It allows search engines to comprehend the nuances of human language more effectively. Unlike traditional keyword-based searches, BERT considers the context of words in a query, enabling it to understand the intent behind a user's search more accurately. Before BERT, search engines primarily focused on individual keywords, often leading to results that were contextually inaccurate. With BERT, the search algorithm takes into account the entire context of a query, including the relationships between words and the overall structure of the sentence. This bidirectional approach enables the algorithm to understand the meaning of words in relation to their context, making search queries more akin to how people naturally speak and write. BERT has particularly improved the handling of conversational and long-tail queries. It recognizes the importance of prepositions, conjunctions, and other connecting words, ensuring that the search results align with the user's intent.

6) Comparison between keyword search and BERT: For example, let us consider a query, ”Can i get a medicine for someone from pharmacy”. A traditional keyword search only identifies keywords ”medicine” , ”someone” and ”pharmacy”. So the machine returns us webpages which contain keywords of this. But it does not really understand the true meaning. Ofcourse, it's a machine. How can it understand our intentions?. This field in Information Retreival where the algorithms works even on words that are prefix and postfix to the keywords is a very very complex task. But BERT algorithm understands the true meaning of the query and it identifies the words ”for”, ”someone” along with keywords and picks out the best results for our query. By making search queries more natural, BERT has enhanced the user experience, providing more relevant and contextually

appropriate results. This trans formation reflects a shift from mere keyword matching to a deeper understanding of the intricacies of language, making search engines more adept at interpreting the diverse ways people express their information needs.

III. Literature Survey

1) Algorithms: The evolution of search engine algorithms has progressed through distinct eras, beginning with the reliance on keywords in the early stages of search engines such as Archie and Veronica. [1]The late 1990s saw a paradigm shift with the introduction of PageRank by Larry Page and Sergey Brin, revolutionizing result rankings by considering link structures. This paper deals with PageRank algorithm which is based on the idea of hyperlink analysis, where the significance of a web page is decided by how many links there are to it and how good those links are. The computation of PageRank involves solving a system of equations, and for large-scale web graphs, this can be computationally expensive. TF-IDF became a cornerstone for information retrieval during the same period, assessing term importance in documents. The early 2000s witnessed the advent of algorithms like HITS, emphasizing hyperlink structures, and LSI, introducing semantic indexing. BM25, a probabilistic extension of TF-IDF, gained prominence in the mid-2000s. [2]2003 was a revolutionary year where Google updated its algorithm. Florida was introduced and then suddenly non-spam sites lost rankings. These non-spam sites that lost rankings were labelled as "false positives." The blowback was so massive that Google promised to try not to ever roll out a significant update before the holidays again. [3] Panda and Penguin updates in 2011 and 2012 released by Google targeted content quality and link spam. The sub sequent years brought innovations like RankBrain's machine learning for ambiguous queries (2015), BERT's contextual understanding (2018),and ongoing developments in personalized search algorithms such as Facebook's DeepRank. Continuous advancements in learning-to-rank algorithms demonstrate the field's dynamic nature, combining historical foundations with cutting-edge machine learning techniques.
2) Metasearch Engines: Paper [20] introduces Meta Crawler, a search service designed to address issues with existing web search services such as incomplete results, irrelevant responses, and tedious querying processes. The key premises of Meta-Crawler include the recognition that no single search service is sufficient, many references from services are irrelevant and can be filtered with more expressive queries, low-quality references can be quickly detected and removed, and these features will be widely used by the web's population. Meta-Crawler aims to enhance the search experience by allowing users to query multiple services simultaneously, removing irrelevant results, and quickly identifying low-quality references. The paper highlights the popularity of Meta-Crawler, receiving over 7000 queries per week, and discusses the variety of information users search for, emphasizing that explicit content queries constitute less than 4percent of total queries. Rest of the paper deals about the design and implementation of Meta-Crawler, experiments validating its premises, discussions on related work, and considerations for future work and potential impact. The paper concludes with a summary of its key findings and contributions. The system aims to balance the goals of maximizing result quality and minimizing resource consumption, demonstrating success with over 20,000 daily queries. The learning algorithm, refined over almost two years, has proven effective with minimal query term uses. The paper envisions the system's future evolution, emphasizing personalizing in meta-search and its role as an information-gathering tool supporting higher-level goals. Overall, SAVVYSEARCH represents a significant step towards efficient and responsible use of web resources in today's world. The evolution of metasearch engines includes the implementation of sophisticated algorithms for result aggregation and relevance determination, enabling a more accurate synthesis of information from diverse sources. Additionally, advancements in real-time querying and user customization features have been introduced, providing users with more current and personalized search experiences. Despite these strides, challenges persist, such as effectively managing diverse search engine APIs, ensuring result freshness, and navigating differences in ranking algorithms. Nevertheless, metasearch engines remain pivotal in offering users a consolidated and varied view of search results, contributing to a more robust and user-centric information retrieval landscape.
3) Dataset Search Engines: The evolution of dataset search engines signifies a pivotal advancement in the realm of search engine functionality. Traditionally, mainstream commercial search engines were predominantly geared toward indexing and retrieving textual content, lacking specialized features for dataset search. The in- creasing significance of data across various domains, spanning academia to industry, prompted a realization of the inadequacies in conventional search engines for handling structured datasets effectively. This evolution unfolded through several phases, starting with the acknowledgement of the growing importance of data for decision making and innovation. Subsequent research initiatives delved into acquiring datasets, profiling their characteristics, and augmenting them to enhance user comprehension. Information extraction frameworks

emerged as a critical aspect, facilitating the extraction of meaningful information from sources like research articles. Specialized search algorithms tailored for tabular dataset searches be- came a focal point, leveraging techniques such as generated schema labels, pretrained language models, and graph neural networks to optimize search outcomes. The concept of universal dataset encoders, treating datasets as point sets for efficient retrieval of similar datasets, further contributed to this transformative journey.
4) Feature Snippet Extraction: Featured snippets are concise, information-rich excerpts displayed prominently at the SERPs, providing users with quick and direct answers to their queries. These snippets are extracted from web pages deemed most relevant to the user's search intent. Consequently, understanding and optimizing for featured snippet extraction has become a focal point for webmasters, content creators, and digital marketers alike. This involves deciphering the algorithms used by search engines to select and present these snippets, tailoring content to meet snippet criteria, and strategically positioning a website as an authoritative source to increase the likelihood of being featured. In this context, featured snippet extraction plays a pivotal role in the dynamic landscape of SEO, influencing how information is not only discovered but also presented to a vast online audience. Key contributors include the authors of BERT (Bidirectional Encoder Representations from Transformers) The introduced language representation model, BERT (Bidirectional Encoder Representations from Transformers), offers a unique approach to pretraining deep bidirectional representations from unlabelled text. In contrast to previous models, BERT conditions on both sides covering the context at all layers, making it to achieve NLP tasks with minimal task- specific architecture modifications during fine-tuning. BERT's conceptual simplicity and empirical power are evident in its performance across eleven NLP tasks, demonstrating substantial improvements in tasks like Q and A and language inference. The model's unified architecture, relying on a hidden language model known as MLM and a (NSP) task, contributes to its effectiveness in capturing bidirectional context. The fine- tuning process is streamlined, leveraging BERT's self-attention mechanism for various downstream tasks, making it a versatile and powerful tool in the NLP landscape.This paper discusses "featured snippet" where it displays answer before even clicking on the Google search button. The algorithm starts with a list of keywords. These keywords can come from two sources: suggestions from search engines and user autocompletion. The list is then expanded using a technique called snowballing." The next step is to identify patterns in how search engines answer questions directly. These patterns have four parts: the answer itself, where it was found on the search results page, the full URL of the answer, and the type of answer. From the full URL, we can extract the domain name and main domain name. Finally, the algorithm collects information about the snippets that appear alongside the answers. These algorithms have significantly improved contextual analysis, enabling a deeper understanding of user intent and enhancing the extraction of relevant information for featured snippets. User engagement feedback is increasingly utilized to refine extraction techniques, ensuring the continual adaptation of algorithms to evolving user behaviour and information needs ensuring to deliver more accurate, relevant, and contextually rich featured snippets.

IV. Conclusion

In the contemporary digital landscape, search engines are integral to efficiently accessing information on the internet. This research explores the evolution of Information Retrieval (IR) within search engines across various domains. From the foundational algorithms of Larry Page and Sergey Brin till today's advancements, the journey encompasses user behaviour analysis, data retrieval techniques, audio-enabled search engines, image-based search engines, metasearch engines, and more. The literature review highlights the dynamic evolution of search engine algorithms, data retrieval techniques, and specialized search functionalities like audio-enabled and image based searches. The progression of metasearch engines is acknowledged, with examples like Erik Selberg's Meta Crawler representing pioneering innovations. User behaviour analysis reflects advancements in understanding and optimizing search engine functionalities. Evolution in location-based ranking, featured snippet extraction, and structured data searching is detailed, showcasing adaptations for more accurate, con textually relevant, and personalized results. SEM and SEO illustrate how advertising and optimization strategies have evolved alongside search engine advancements. In conclusion, collaborative efforts have shaped the intricate landscape of search engines, reflecting continuous progress in understanding, adapting to, and optimizing for the ever-changing digital information ecosystem.

References

[1] Utkarsh Jain, Ayush Mishra, B. Jaganathan, and Pankaj Shukla. 2021. Study and analysis of category based PageRank method. Wirel. Netw. 27, 8 (Nov 2021), 5461–5476. https://doi.org/10.1007/s11276-021 02617-y

[2] https://www.searchenginejournal.com/google-algorithm-history/florida update/close
[3] https://www.searchenginejournal.com/google-algorithm-history/alg-list
[4] G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. Commun. ACM 18, 11 (Nov. 1975), 613–620. https://doi.org/10.1145/361219.361220
[5] G. Salton ,Automatic Text 343(1970).DOI:10.1126/science.168.3929.335 Analysis.Science168,335
[6] CLEVERDON, C. (1970), ”Evaluation Tests of Information Retrieval Systems”, Journal of Documentation, Vol. 26 No. 1, pp. 55-67. https://doi.org/10.1108/eb026487
[7] Cleverdon, C.W., 1991, September. The significance of the Cranfield tests on index languages. In Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 3-12).
[8] Z. Bao, J. Lu, T. W. Ling and B. Chen, ”Towards an Effective XML Keyword Search,” in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 8, pp. 1077-1092, Aug. 2010, doi: 10.1109/TKDE.2010.63.
[9] https://www.searchenginejournal.com/what-is-voice search/463344/close
[10] Divy Patel, Dhairya Shah, Param Modi, Mohendra Roy,Artificial intelligence powered material search engine,Materials Today: Proceedings,Volume 57, Part 1,2022, Pages 11-13,ISSN 2214 7853,https://doi.org/10.1016/j.matpr.2022.01.120.
[11] M. Van Thong, P. J. Moreno, B. Logan, B. Fidler, K. Maffey and M. Moores, ”Speechbot: an experimental speech-based search engine for multimedia content on the web,” in IEEE Transactions on Multimedia, vol. 4, no. 1, pp. 88-96, March 2002, doi: 10.1109/6046.985557.
[12] Faris Elghaish, Jatin Kumar Chauhan, Sandra Matarneh, Farzad Pour Rahimian, M. Reza Hosseini,Artificial intelligence based voice assistant for BIM data management,Automation in Construction,Volume 140,2022,104320,ISSN 5805,https://doi.org/10.1016/j.autcon.2022.104320. 0926
[13] Tiago F. Pereira, Arthur Matta, Carlos M. Mayea, Frederico Pereira, Nelson Monroy, Jo˜ao Jorge, Tiago Rosa, Carlos E. Salgado, Ana Lima, Ricardo J. Machado, Lu´ ıs Magalh˜aes, Telmo Ad˜ ao, Miguel ´ Angel Guevara L´opez, Dibet Garcia Gonzalez, A web-based Voice Interaction framework proposal for enhancing Information Systems user experi ence,Procedia Computer Science,Volume 196,2022,Pages 235-244,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2021.12.010.
[14] https://blog.scaleflex.com/voice-search-optimization/
[15] C. Geetha and P. Geetha, ”Diversifying Image search results,” 2016 International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC), Melmaruvathur, India, 2016, pp. 102 107, doi: 10.1109/ICCPEIC.2016.7557232.
[16] Kevin Corace and Ken Kocon. 2005. Image based search engine. J. Comput. Sci. Coll. 20, 5 (May 2005), 106–107.
[17] Jeff Ubois, Jamie Davidson, Marko Grobelnik, Paul Over, and Hans Westerhof. 2010. Video search: are algorithms all we need? In Proceedings of the 19th international conference on World wide web (WWW ’10). Association for Computing Machinery, New York, NY, USA, 1339–1340. https://doi.org/10.1145/1772690.1772918
[18] Yu Zheng, Heng Tian, Minxin Du, and Chong Fu. 2022. Encrypted video search: scalable, modular, and content-similar. In Proceedings of the 13th ACM Multimedia Systems Conference (MMSys ’22). Association for Computing Machinery, New York, NY, USA, 177–190. https://doi.org/10.1145/3524273.3528190
[19] Li Zhang, Lei Shi, Jiashu Zhao, Juan Yang, Tianshu Lyu, Dawei Yin, and Haiping Lu. 2022. A GNN-based Multi-task Learning Framework for Personalized Video Search. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM ’22). Association for Computing Machinery, New York, NY, USA, 1386–1394. https://doi.org/10.1145/3488560.3498507
[20] Selberg, E. and Etzioni, O., 1995, December. Multi-service search and comparison using the MetaCrawler. In Proceedings of the Fourth Int’l WWWConference, Boston.