Personalized Semantic Search Engine

Bhanu Teja Sudalagunta, Manikanta Pinabakala,
Hari Durga Kode and Damini Polamuri

March 18, 2024

PERSONALIZED SEMANTIC SEARCH ENGINE

Bhanu Teja Sudalagunta
Department of Computer Science Engineering and Technology
Parul Institute of Engineering and Technology
Vadodara, India
200303124479@paruluniversity.ac.in

Manikanta Pinabakala
Department of Computer Science Engineering and Technology
Parul Institute of Engineering and Technology
Vadodara, India
200303124419@paruluniversity.ac.in

Hari Durga Kode
Department of Computer Science Engineering and Technology
Parul Institute of Engineering and Technology
Vadodara, India
200303124285@paruluniversity.ac.in

Damini Polamuri
Department of Computer Science Engineering and Technology
Parul Institute of Engineering and Technology
Vadodara, India
200303124424@paruluniversity.ac.in

***Abstract*— The *Web 2.0,* which many folks use nowadays, serves as a vast collection of interconnected documents that gets transferred by computers and keeps being shown to individuals.**

A search engine is considered one of the most important tools to discover any information from WWW. In spite of having lots of development and novel research in current search engines techniques, they are still syntactic in nature and display search results on the basis of keyword matching without understanding the meaning of the query, resulting in the production of a list of Webpages containing a large number of irrelevant, and sometimes even unreliable documents as an output.

Simply, most of the current-day search engines are keyword-based search engines, meaning that they focus on each keyword of the search query they're fed with. Now, this methodology may seem ideal but most of the time, the search results of these search engines are irrelevant. And even the Web is mostly unstructured which makes it difficult to return a proper search result. Semantic Web (Web 3.0), the next version of the World Wide Web is being developed with the aim of mitigating the problems faced in Web 2.0 by representing data in a structured form, and for discovering such data from Semantic Web, Semantic Search Engines (SSEs) are being developed in many domains.

This paper provides a survey on some of the prevalent SSEs focusing on their architecture; their working and techniques; a practical work on the performance of an SSE and a normal keyword - based search engine; and presents a comparative study on the basis of techniques that different SSEs follow.

I. INTRODUCTION

Web and search engines have been one of the greatest inventions in human history, they didn't just make information finding easier, but the wideness and scalability of the web have truly been a great asset to humans. Information retrieval was once a very hectic task with the absence of resources and the level of technology that is available today. And then the internet brought a huge revolution in almost every field, with search engines like Google, Yahoo, Bing, etc. being developed at a rapid pace. But well, as always, human needs also evolve with the evolution of everything else. Similarly, in this case, humans managed to find out a problem that even though the search engines provide results quite easily at just a hit of the search button, we found out that the results we were shown on the search engines have most of the times been with a lot of noisy results. Users found out that searching a query solely does not necessarily give us relevant results. Most of the time, we have to scroll a lot through the web pages to find the documents that are relevant to us. The problem made us realize that the query alone will not be helpful in finding results, but it was also important for the machine to understand our query. Simply, instead of the keyword-based search on the web, making the machines understand the intent/context of our query will make them more efficient is what we figured out from years of research. That introduced the latest breakthrough technologies in web search - the semantic web, and semantic search engines. By combining the best of traditional keyword-based search engines with the power of the Semantic Web, semantic search engines are revolutionizing the way we search the web. The technique is simple, all the search engine has to do is understand the context the user is asking the query and give him an accurate result corresponding to that context instead of just showing up a large list of documents and results based on the keywords in the query.

II. LITERATURE SURVEY

Ranjna Jain paper discusses the evolution of the web and search engines. It compares Web 2.0 (unstructured data) and Web 3.0 (structured data) and how this impacts search results. Semantic search engines (Swoogle, Falcon, Hakia, etc.) are introduced as a way to provide more meaningful search results using structured data, but the paper criticizes the choice of outdated engines for comparison. Overall, the paper highlights the potential of semantic search for better web searches[1]. G.Madhu and Dr.A.Govardha survey paper examines different Semantic Web Search Engines (SWSEs) and their techniques. It highlights the paper's strengths: clear descriptions of SWSE techniques and the variety of SWSEs covered. However, the paper's brevity is a weakness, lacking detailed analysis and technique comparisons. Despite this, the paper is a good starting point for newcomers due to its clear explanations. The paper also explores challenges in SWSE development, including precision/recall issues, user intention, and generalizing user patterns. These challenges are valuable insights for researchers. Overall, the paper is a valuable resource for those interested in semantic web and semantic search engines[2]. Junaid Rashid paper offers a comprehensive overview of semantic search engines (SSEs) for researchers and developers. It covers key areas including basic search engines, different SSEs, their functionalities, and comparisons between them. The paper is praised for its clear explanations, diverse engine selection, and analysis of both strengths and limitations[3]. Dinesh Jagtap paper explores semantic web, a future web with user-understandable databases. It uses RDF (Resource Description Framework) as the standard for this web. The paper also introduces Ontology, a building block of the semantic web, and how it's created using RDF and OWL. The paper highlights issues with traditional search engines and how semantic engines can address them. These issues include, Low precision, high recall (lots of results, but not very relevant),Difficulty understanding user intent, Inaccurate queries, Crawler inefficiency. The paper argues that semantic search engines can improve upon traditional ones. The authors provide a practical experiment where their proposed semantic engine outperforms existing ones. They acknowledge that semantic search isn't a perfect solution, but it has the potential to significantly improve traditional search[4]. Upasana Sinha &Vikas Dubey paper examines semantic search engines (SSEs) for both theory and practice. It explains how SSEs work and differ from traditional search engines. The paper compares existing SSEs and then analyzes the performance of DuckDuckGo (a semantic search engine) vs Google on real queries. This practical test shows that semantic search engines have the potential to significantly improve web search[5]. Johannes Darms research explores using semantic search for COVID-19 preprint research. Traditional search methods struggle with the large volume of COVID-19 preprints. Semantic search offers a more accurate way to find relevant preprints by considering the meaning and context of the search query and preprints themselves. The authors built a semantic search engine for COVID-19 preprints and showed it outperforms traditional search methods. This highlights the potential of semantic search for this area and provides a valuable tool for researchers[6].

Semantics can improve agriculture by, **Standardizing data:** Creating a common language for crops, livestock, etc. This allows easier information sharing between regions and organizations. **Sharing data:** Semantic web technologies allow data sharing in a format understandable by machines. This streamlines data exchange and decision-making. **Analyzing data:** Semantic analysis helps find patterns in large datasets like weather, soil, and crop yields. This informs decisions on crop management and other practices. Overall, semantics in agriculture leads to better data management, collaboration, and decision-making for farmers and others in the food industry[7]. Sindhu Nair , Jheel Somaiya & Dharmish Shah paper proposes a fuzzy semantic search engine to improve search accuracy. Traditional search engines struggle with variations in wording. This fuzzy engine uses fuzzy matching and semantic technologies to understand the meaning behind a search query and return more relevant results, even if the wording isn't exact. The system was tested on research articles and showed improvement over traditional search methods[8]. Poonam Gupta article discusses question answering (QA) systems specifically designed for COVID-19 information. These systems are crucial for delivering accurate and timely answers to the public and healthcare workers. The paper explores different QA system approaches and highlights challenges and opportunities in developing them. Finally, it explores promising future directions for this technology[9]. Sen Ma paper proposes a new semantic search engine focused on computer hardware and gaming. This system uses advanced techniques to understand the meaning of user queries and provide more accurate and relevant search results compared to existing search engines. The authors evaluated their system using real user queries and found it to outperform existing options[10]. Yufei Li, Yuan Wang, and Xiaotao Huang paper argues that current search engines struggle with meaning and proposes the Semantic Web as a solution. The Semantic Web uses RDF triples to embed semantic information (metadata) in web pages. The paper introduces OntoLook, a search engine that leverages this metadata to improve search results. OntoLook uses algorithms to process this semantic information and improve the accuracy of search results [11].

Zhi Quan Zhou study explores using metamorphic testing to check the quality of web search engines. Traditional testing methods struggle with search engines because there's no single "correct" answer. Metamorphic testing creates test cases based on expected relationships between searches and their results. The researchers applied this technique to Google, Bing, and Baidu and found it effective in identifying issues. This benefits both developers (who can improve their engines) and users (who can learn how to use them better). The study also suggests that search engine quality can vary depending on factors like language and query type. Overall, metamorphic testing shows promise as a way to improve web search engine quality[12]. SCORE is a powerful semantic engine that bridges information retrieval, AI, and knowledge representation. It allows for semantic search, which focuses on meaning over just keywords. This lets users find and analyze data more effectively. SCORE also helps build knowledge bases for organizations to store and manage large amounts of data. The paper highlights that SCORE is a big step in semantic technology, but the field is still evolving. The Semantic Web is being developed to connect data with meaning, allowing for even more powerful information retrieval systems in the future. Overall, SCORE shows the promise of semantic technologies for data analysis and the Semantic Web suggests a future where information systems can understand the meaning of data[13]. Semantic Web search engines aim to improve web searches by understanding the meaning of information, not just keywords. This paper proposes a new algorithm to improve these engines. The algorithm simplifies ranking search results by focusing on the user's query and the relationships between data. This reduces the need to process massive amounts of data and improves search accuracy. Overall, this research is a step towards more effective and user-friendly web search experiences[14]. The Semantic Web aims to improve web search by using ontologies to categorize information.

Ontologies define the relationships between concepts, allowing for more context-aware searching. However, there can be mismatches between how people search and how information is categorized. This paper describes a search engine that uses approximate query processing to bridge this gap and improve search accuracy[15]. Rudi L. Cilibrasi and Paul M.B. Vitanyi paper proposes a new method to measure similarity between words and phrases. It uses Google search results and information distance to calculate a score called "Google similarity distance." This method leverages the vast amount of information on the web to understand how similar words are based on how often they appear together in search results. The paper shows this method can be used for various tasks like clustering, classification, and translation, and performs well compared to existing methods[16]. Pimwadee Chaovalit paper compares two methods for analyzing movie reviews: machine learning and semantic orientation. Both methods aim to categorize reviews as positive or negative. The study shows these methods are effective for sentiment analysis of movie reviews, but challenges exist due to the mix of fact and opinion, and the use of irony in reviews. Future work should focus on improving techniques to handle these complexities[17].

Benjamins paper proposes a new method for automatically combining web services on the Semantic Web. This method uses Golog, a type of programming language, to simplify the process and make it more user-friendly. The method also allows for customization and can be verified to ensure it works as expected. This new method has potential applications in areas like cognitive robotics[18]. Pakistan can improve socio-economic development through ICT by making technology accessible in local languages. Most people in Pakistan don't speak English fluently, so ICT education and resources need to be offered in local languages to bridge the digital divide and empower more people. This will allow for more innovation, economic growth, and social development[19]. The Semantic Web aims to make web searches more powerful by adding meaning to webpages. This is done with machine-readable labels (metadata) that describe the content and how it relates to other things. Search engines aren't very good at using this information yet, but it has the potential to create a new generation of smarter search engines and personalized web experiences[20].

III. METHODOLOGY

The workflow and methodology of building a "Personalised Semantic Search Engine" can be broken into several steps. Here are the high-level steps involved in achieving the project:

• **Collect Data:** Gather data on the user's interests and preferences. This can be done through
surveys, user feedback, or data from previous search queries.

• **Build a User Profile:** Based on the data collected, build a user profile that includes information
such as the user's preferred topics, search history, and search behavior.

• **Data Preprocessing:** Preprocess the data by cleaning, normalizing, and transforming it into
a suitable format for analysis. This might possibly involve eliminate stop words, stemming or lemmatizing words, and converting text into a vector representation.

• **Train a Machine Learning Model:** Train a machine learning model to identify and prioritize
relevant search results based on the user's preferences. This can very possibly be done using some techniques such as supervised learning or maybe even unsupervised learning; it all kind of depends on the available data.

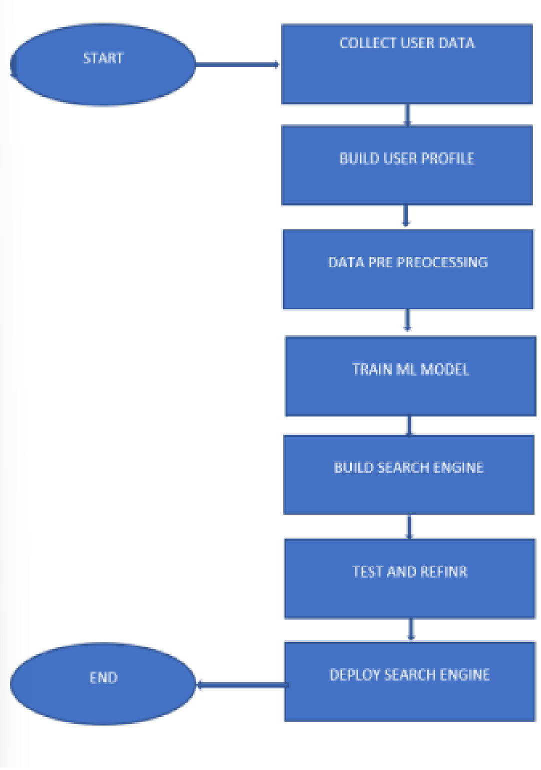
• **Build a Search Engine:** Develop a search engine that integrates the machine learning model
to provide personalized search results to the user. The search engine should be able to take in a user
query, match it with the relevant content, and rank the results based on the user's preferences.

• **Test and Refine:** Test the search engine with a set of test queries and refine the model as
necessary. This may involve tweaking the model parameters or adjusting the user profile based on
user feedback.

• **Deploy the Search Engine:** When the search engine is prepared, put it online or in a mobile application so that users may access it. Improvising search engine such as bug fixes is also essential in the coming future.

Overall, building a personalized semantic search engine requires a combination of data collection, data preprocessing, machine learning, and search engine development skills. It is important to ensure that the search engine is accurate, scalable, and user-friendly to provide a high-quality user experience.

FLOW CHART:-



IV. IMPLEMENTATION

Fig: Home page of search engine

Fig: New user registering the details to get access of search engine

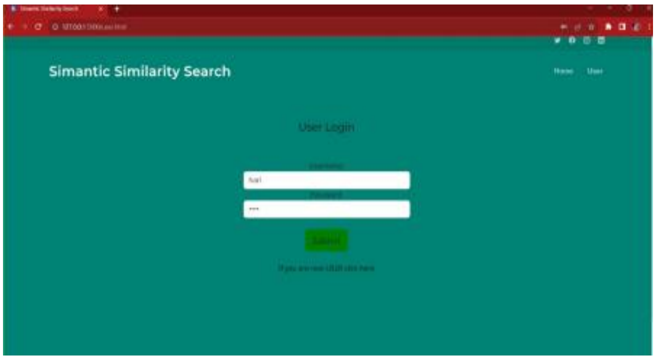


Fig : Login with valid credentials

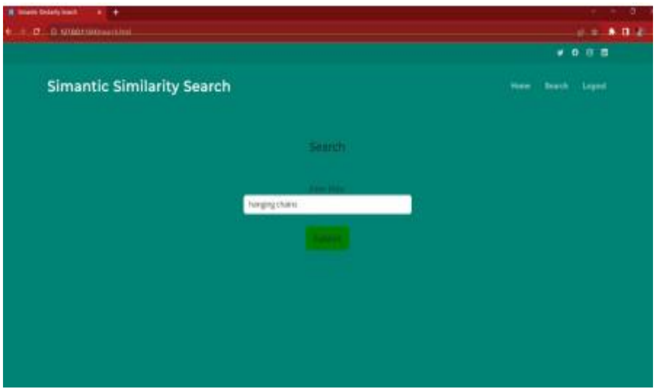


Fig: Finding the results according to given dataset

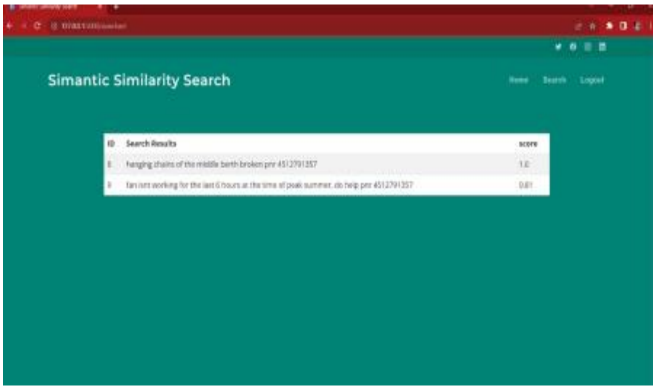


Fig : Output with score

CONCLUSION

In conclusion, our proposed project on developing a personalized semantic search engine has the potential to greatly benefit users by providing a more accurate and personalized search experience. By building a user profile and using machine learning algorithms, we aim to identify and prioritize search results that are relevant to the user's interests and preferences.

The project involves several key steps, including data collection, user profiling, data preprocessing, machine learning, and search engine development. These steps will be implemented using a combination of programming languages, libraries, and tools, such as Python, TensorFlow, and Elasticsearch(in necessary cases). While there are existing search engines that offer some degree of personalization, we believe that our proposed approach will offer a more advanced and effective solution. By leveraging the power of machine learning, we aim to create a search engine that not only provides relevant results but also adapts and learns from the user's search behavior over time.. However, we are confident that with careful planning and execution, we can overcome these challenges and deliver a high-quality personalized semantic search engine.

In summary, our proposed project represents an exciting opportunity to innovate in the field of search engine development and improve the search experience for users. We look forward to implementing the project and sharing our results with the broader community.

REFERENCES

[1] Ranjna Jain,Neelam Duhan,A.K. Sharma;Comparative Study on Semantic Search Engines;Volume 131 – No.14, December2015.

[2] G.Madhu and Dr.A.Govardha,Dr.T.V.Rajinikanth;Intelligent Semantic Web Search Engines;) Vol.2, No.1, January 2011.

[3] Junaid Rashid,Muhammad Wasif Nisar;A Study on Semantic Searching, Semantic Search Engines and Technologies Used for Semantic Search Engines;DOI: 10.5815/ijitcs.2016.10.10.

[4] Dinesh Jagtap*1, Nilesh Argade1, Shivaji Date1, Sainath Hole1, Mahendra Salunke2;Implementation of Intelligent Semantic Web Search Engine;Vol. 4 Issue 04, April-2015

[5] Upasana Sinha, Vikas Dubey;The Technique of Different Semantic Search Engines;ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020.

[6] Semantic Search to Explore COVID-19 Research Preprints Authors: Johannes Darms, Roman Baum, Lisa Langnicke.

[7] survey of semantic web technology for agriculture ; Hitzler , pascal.

[8] Fuzzy Semantic Search Engine Authors: Sindhu Nair , Jheel Somaiya , Dharmish Shah.

[9] Question Answering Systems for Covid-19 Authors: Poonam Gupta, Ruchi Garg.

[10] Semantic search engine wiith rule reasoning Authors: Sen Ma, Wen Zhao, Shi-kun Zhang, Hang Zhang.

[11] Yufei Li, Yuan Wang, and Xiaotao Huang based on,A Relation-Based Search Engine in Semantic Web,IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19,NO. 2,FEBRUARY 2007.

[12] Zhi Quan Zhou, Shaowen Xiang, and Tsong Yueh Chen,based on Metamorphic Testing for Software Quality Assessment,in IEEE Transactions on Software Engineering · January 2015,DOI: 10.1109/TSE.2015.2478001.

[13] Amit P. Sheth, Clemens Bertram, David Avant;Managing Semantic Content for the Web;IEEE Internet Computing, 6 (4), 80-87.

[14] Lamberti F., Sanna A., Demartini C., “A relation-based page rank algorithm for semantic web search.

[15] Olivier Corby, Rose Dieng-Kuntz, Catherine Faron Zucker, Fabien Gandon. Searching the Semantic Web: Approximate Query Processing based on Ontologies. IEEE Intelligent Systems, 2006, IEEE Intelligent Systems, 21 (1), pp.8. 10.1109/MIS.2006.16. hal-01150215.

[16] Rudi L. Cilibrasi and Paul M.B. Vitanyi ,based on The Google Similarity Distance ,VOL. 19, ´ NO 3, MARCH 2007, 370–383.

[17] Pimwadee Chaovalit,Lina Zhou,on Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches,January 2005 DOI: 10.1109/HICSS.2005.445.

[18] V. Benjamins et al. IBROW3: An Intelligent Brokering Service for Knowledge-Component Reuse on the World WideWeb. In KAW’98, Banff, Canada.

[19] Sheila McIlraith,Tran Cao Son,Adapting Golog for Composition of Semantic Web Services; item Dr. Muddassar Farooq,Research Impact: A Labyrinth of Wrong Myths.

[20] JAROSLAV POKORN Y,WEB SEARCHING AND INFORMATION RETRIEVAL;1521- ´ 9615/04/.