Power System Technology
ISSN:1000-3673

Received: 16-06-2025 *Revised: 05-07-2025* *Accepted: 20-08-2025*

Relevance Feature Discovery for Text Mining

Dr. Kumar P K[1], Renukaradya V[2], Shreyas M S[3], Suryakant[4], Nazia Sultana[5]

[1,2,3,4,5]Dept of CSE, VTU Center for PG Studies, Mysuru, Visvesvaraya Technological University, Belagavi-590018, Karnataka, India

[1]pandralli@gmail.com,[2]renukaradhyanishu@gmail.com, [3]shreyasuresh1999@gmail.com, [4]suryakantbidemani@gmail.com, [5]naziaimedoh@gmail.com

Abstract: - With so many phrases, patterns, and noise in text, ensuring the quality of features identified for defining user performance that are relevant is a difficult task. Most of the popular text mining and categorization algorithms now in use utilize term-based techniques. Polysemy has afflicted all of them, though. Most Internet search results are based on patterns rather than meaning. when user searches for a word, only specific keywords will be considered for results and includes noise. Instead of looking for patterns, the focus of this study is to determine the meaning of provided words and then provide comparable material because of that inquiry and that can be implemented using Sematic search algorithm. In the end, the user receives the accurate information they requested, along with context. Here, it gives around 80% - 90% of accuracy in results with cancelled noise. Positive, negative, and neutral facts are all readily apparent in this search.

Keywords: *Noise in text, Text mining, Sematic search algorithm, Crawl Module, search strategies.*

1. Introduction

The use of a search engine is now essential while conducting online research. There would be little use in storing data on a website, blog, etc., if users could not easily find it using a search engine, as it would be extremely inconvenient to visit each individual site to get certain information. Semantics is the study of meaning. Emphasis is placed on meaningful relationships between words, phrases, signs, and symbols. Web 2.0 is "a data web that machines can process either directly or indirectly," according to Tim Berners-Lee, the man who created the World Wide Web. The need for more effective ways to retrieve and organize online content has grown in importance due to the rapid advancements in technology and the steadily increasing number of internet users. Current demands for speed, data volume, and information quality are frequently beyond the scope of traditional search techniques.

A method known as text mining has surfaced. This procedure entails gleaning valuable information, trends, and insights from sizable datasets of either structured or unstructured text data. Text is now the most common type of data in the modern digital world, especially with the growth of user-generated content on social media and other platforms. Nowadays, people frequently share written content, photos, and thoughts online, producing enormous volumes of data every day.

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 20-08-2025

When it comes to analysing this data, text mining is essential, particularly when it comes to figuring out trends, gauging public opinion about brands, goods, or services, and comprehending consumer behaviour. Additionally, it has uses in several fields, such as organizational research, personal analytics, and social science. Text mining needs to combine and analyse data from several sources to find more d Because the data we are gathering from various sources is, by definition, unstructured, it is recommended that we format the data once we have collected it. A straight application of the method to such texts presents several challenges.

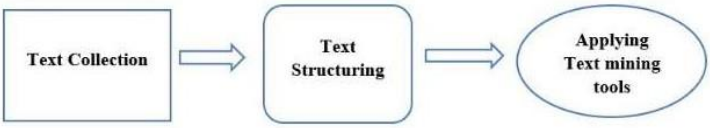


Figure 1: Process of Text Mining

When we talk about semantics, we're essentially referring to the meaning or essence behind words and expressions. In the context of search technologies, semantics goes beyond simple keyword matching it involves analysing the deeper meaning and intent behind a user's query. Rather than just scanning for specific terms, semantic search tries to understand what the user is looking for and the context in which they are searching.

This approach allows for far more accurate and relevant search results. By interpreting the query, grasping its context, and connecting it with data spread across various sources, semantic search delivers insights that traditional search methods often miss. It relies on the principles of language understanding to draw logical connections between words and concepts.

Unlike conventional search engines that match exact words, semantic search is a key component of text mining that considers the entire context of the search query including surrounding words, sentence structure, and the underlying intent. It factors in elements like user location, related phrases, trending topics, synonyms, and even different word forms to enhance the search experience.

Creating an effective semantic search system involves various advanced techniques and algorithms. These may include keyword-to-concept mapping, graph-based pattern detection, and fuzzy logic, all of which help in refining the accuracy and relevance of search results based on how humans use language.

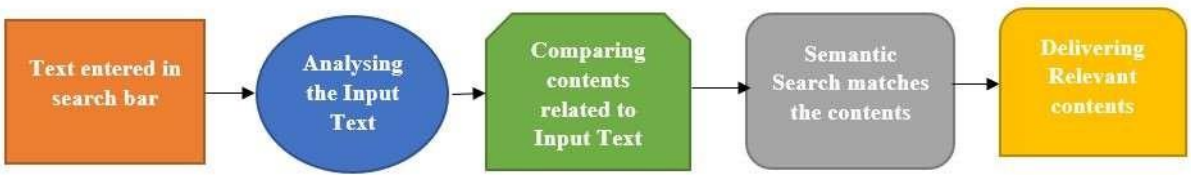


Figure 2: Sequence of Semantic search technique

Power System Technology
ISSN:1000-3673

Received: 16-06-2025 *Revised: 05-07-2025* *Accepted: 20-08-2025*

Existing search engines have a few flaws. As we narrowed in on the problems plaguing these search engines, we came across a wide range of challenges, including but not limited to: ambiguity; subjective content; high volatility; rapid technical development; reliance on results; monetary influences; and many others. There are some occasions when the user's query leads us to a huge number of pages that have nothing to do with what the user was looking for. These are search engines that make no claims about the accuracy of the information they return. Rather of focusing on relevance, search engines may priorities marketing, spamming, or self-promotion in their rankings. Unspoken or unmentioned factors that may be just as significant as those that receive more attention. When it comes to fast developing topics, search engines just cannot keep up with the newest in peer-reviewed scholarly study like printed journals, papers, and books can. The search engines can support many languages; however, the English translation may not be perfect.

As part of our study, we put forth a methodology for ensuring that users only encounter the most pertinent results for their queries. Text mining based semantic search algorithm will be combined in this approach. The good, the bad, and the most up-to-date information that can be found on the web are all brought to you. Our main aim is machine to understand the meaning of the query given by the user and deliver a greater number of related contents to the user. It should provide all positive, negative, and neutral contents which are available in the internet. For example, if a user is searching for Four stroke Engine. Our method is to find the meaning and collect the data from the internet relevant to that in cases like Positive, Negative data using our proposed method. And it should be capable of giving more information in short period of time.

2. Related Study:

In the modern-day work, the internet has become an irreplaceable tool that can be used in various ways. This process of searching and surfing the Internet has overtaken even the seriousness of emailing in the recent past years. The level of trust that these electronic searches attract however, is widely different. The amount of content that is acquired is quite significant, and much of it is in shady, indeed evil places. The more established systems include Google, Yahoo, Bing and Ask. They both have kept the central purpose of a search engine in mind, that is, to find relevant content but both in different ways than the other, through different operating methods. Some search engines utilize advanced semantic search by relying upon the semantic context of the query. During the relatively brief history of the World Wide Web (WWW) itself several generations of search technologies have arisen and then died out, most notably those based on keyword-oriented approaches or more traditional database-driven architectures. Higher order algorithms have been proposed by a greater sophistication of

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 20-08-2025

researchers, and the current debate shows that when compared head-to-head, the results made by semantic-related search engines are more prone to achieving results of high level of accuracy. To conclude on the relative merit of Google, Bing, Yahoo and Ask, the report will compare the effectiveness of each search engine as it relates to search precision. Survey data is presented and analyzed in this report to provide an examination of various search engine types as they relate to semantic web browsing. This study compares the results of some recent searches on the four most popular search engines to find out how well they work [1].

In the modern digital research, it is possible to note that almost every individual with online presence frequently makes use of the internet in a very wide range of activities- including thorough research investigations and light recreation, trade and communal interactions. Searching and web surfing have become even more common than checking inbox directories, and this switches up digital practices to great extent. However, the quality of such search results is subject to a lot of differentiation. Much of the content that is obtained in answering a query is produced by sites that are not credible or there is also the threat they are harmful to the user. The top dogs, Google, Yahoo, Bing, and Ask are already capturing the market with each major difference in the approach of its implementation and provision of similar core functions. The normal engines work by matching the actual key word phrases and the semantic engine takes it a notch higher with the lexical and syntactic analysis of user intent.

The history of the World Wide Web has passed through several eras in the development of search engines where the primitive architecture used keywords-based search-engines developed into database-driven search-engines. In the current situation, the increased interest in semantics and user intention pushes researches to create the algorithms that refine the search. The part below critiques the difference between old and semantic searching and, by studying the results of such large-scale engines as Google, Bing, Yahoo, and Ask, exposes that semantic methods are much more accurate and provide results in a contextually acceptable manner. Survey data is presented and analyzed in this report to provide an examination of various search engine types as they relate to semantic web browsing. This study compares the results of some recent searches on the four most popular search engines to find out how well they work [2].

As members of the digital humanities community, most of us, who teach, find one main job in front of us: to teach students the abilities needed to overcome the modern information environment. Our initial difficulty is to explain the reason behind the search engines; they are created to make the users find a piece of information on the huge territories of the World Wide Web. Although there are several search engines, Google, Yahoo and Bing turn out to be the most frequently selected. When applied wisely, these tools can provide relevant information with a high rate and extreme accuracy (all depends on the questions posed and the documents that they execute), linking user questions to the information in the corpus to an extent where

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 20-08-2025

until recently, no other tool could possibly achieve. However, the future of web searching is pointing to the use of semantic search engines, which goes beyond mere matches of single words, and obtains the meaning and context of the message when a query is typed. A continued disadvantage of these earlier systems was the time-consuming length, and by extension, manual labor the user would have to expend to mine useful information out of databases. This inefficiency was bound to affect the performance as well as the user experience of the current traditional search tools. The appearance of semantic search engines therefore marks a more clever and sophisticated reaction. These engines not only interpret the lexical meaning of words but also place them in the context of the hidden environment making the results more quickly and probably with a greater extent of accuracy. Mostly when used effectively, semantic search technology significantly improves the speed of information search and accuracy of finding information. Correctly implemented, a semantic search engine may render very specific results within a relatively short period of time and with a significant level of efficiency.

Results from blogs and other websites are more common in search engine results. Since the information found on blogs and websites is not always reliable, the user cannot rely on the outcomes. As such, we employ XML meta-tags and their associated capabilities. There will be both predefined and customizable tags on the xml page. The RDF format can be used to import information from this XML about the metadata of a page [3].

There is now a massive amount of data because of developments in digital data collection methods. More than 80% of the data we deal with today is in some form of non-structure. It is very hard to find the right patterns and trends in the data to evaluate the text documents. The goal of text mining is to find meaningful connections between large amounts of textual data. There are a variety of text mining methods and technologies available for gleaning useful insights for use in planning for the future. To increase efficiency and reduce the time and effort needed to extract useful information, choosing the most suitable and appropriate text mining approach is crucial. This paper provides a concise discussion and analysis of text mining methods and their usefulness in a variety of contexts. As a bonus, the problems that text mining has and how they affect how reliable and useful the results are also explained [4].

Most traditional search engines retrieve web pages based on the presence of exact keywords within their content. However, this keyword-based approach has several limitations. One major issue is that these engines often ignore synonyms or related phrases, meaning users must try multiple similar terms manually to find what they are looking for. Additionally, conventional search systems treat all keywords equally, failing to recognize that some terms might carry more importance than others in a query. Another shortcoming is the lack of intelligent data categorization. Without the ability to group or filter information meaningfully, the search process can become overwhelming and less effective.

Power System Technology
ISSN:1000-3673

Received: 16-06-2025 *Revised: 05-07-2025* *Accepted: 20-08-2025*

Suppose then I put the matter this way: present search engines are inadequate in being able to provide the user with an answer to the search that will be precise and subtle rather than based on severe keyword matching. The current paper suggests the Fuzzy-Go model, a more advanced model, where fuzzy-logic reasoning is incorporated at the level of semantic implementation. When those high standards of lexical literalism are replaced by relaxed ones Fuzzy-Go will produce a fuzzy ontology a controlled system of relationships and family resemblance between concepts represented as fuzzy-logic variables. Due to this ontological infrastructure, the system finds and retrieves web document containing not only the specifically typed keywords, but also similar and contextually compatible ones. In addition, Fuzzy-Go could apply different weights to different search keywords, and therefore the user can dial in their request with greater precision. Combining fuzzy-logic inference with the robust semantics analysis, Fuzzy-Go not only makes its search-results more user-intent-aware and subtly context-sensitive, but has also effectively increased the substantive level of the search-experience [5].

Upon consideration of the material in focus, it becomes evident that the pattern-searching engine (PSE) method becomes a topical solution to the domains where the patterns mining is relevant and the relationships between patterns must be explored systematically such as the case with the emotions and affective phrases. The model uses affinity identification of documents in which vocabularies show similarities in terms of the patterns of word use. By so doing, it forecasts the exemplification of its prototypical representation with eminent productivity, and efficacy by synergizing a semantic natural search structure, which openly mediates between terms which have comparative implications and at the same time acknowledges the usage of words, which have disparate connotative implications. [6].

3. Methodology:

In the contemporary field of information science, diverse body of text-mining techniques has risen to critically investigate and analyze patterns contained within the textual collections of arbitral cases. They are frequently used to carry out functions that include text classification, document classification, keyword search and query indexing. The propositional project under consideration attempts to promote the area by proposing a new paradigm of semantic searching, whose aim is retrieving more precise and salient data distributed on the internet. Through this framework, both the researcher and the practitioner can isolate coherent context sensitive facts and present it to the end user.

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 20-08-2025

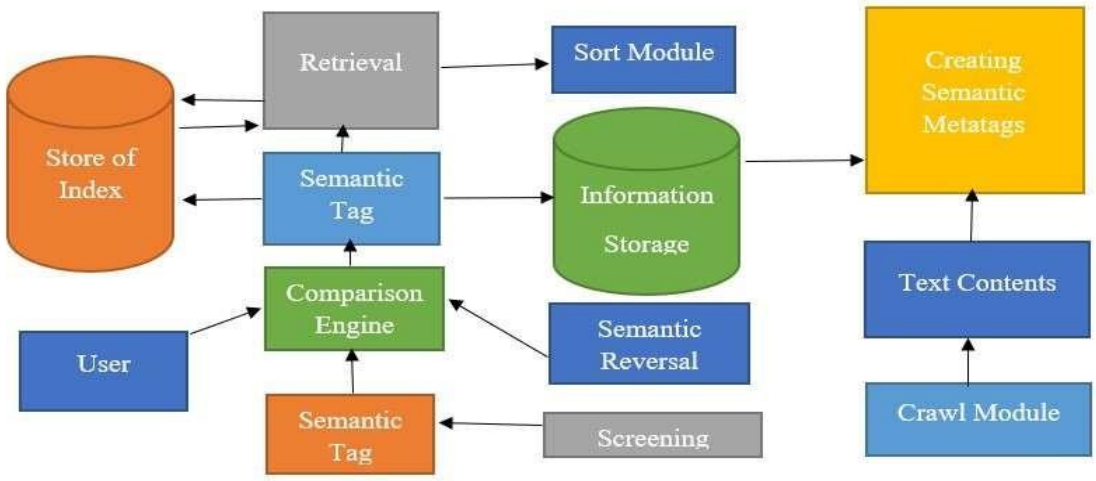


Figure 3: System Architecture

3.1 Crawl Module

Everything that happens during a crawl is managed by the Crawl module. The static web pages collected through crawling actions are processed in one of two ways by the crawl module: either all of the URLs from the web pages are collected and added to the URL sets, or the web pages are added to the text document data sets (if the sets exist) or stored in the cache for a long enough period to complete notes and index.

For future crawling, the URL sets offer the URLs of static web pages. For each set of URLs, crawling selects one to crawl again. The Crawl module can crawl as many pages as the available storage space and human judgement allow.

3.2 Semantic Notes

With the use of ontologies and a knowledge warehouse, meaningful Web tags may be applied to regular text documents obtained via crawling to derive semantic comments. That is, to create the entity-attribute-relationship graphs of the required texts' semantic metadata tags. Semantic Notes is predicated on extracting relevant information from texts, such as keywords and concepts. This key assumption is critical for classifying the various relationships and producing meaningful labels.
Using the ontology classes and the words that serve as attributes, the system offers the access operations of internal form results and transforms them into the appropriate ontology language, such as the RDF triples model of OWL syntax (semantic tags).

3.3 Screening and Comparison

Power System Technology
ISSN:1000-3673

Received: 16-06-2025 *Revised: 05-07-2025* *Accepted: 20-08-2025*

As a result of this procedure, only a subset of the semantic tags is useful It will suffice to start with the fact that some of the tags of entity annotations may have been transplantation fertilized, before the process of ontological graft by operation. Perhaps, the source of donor was not good, or some of the duplicates could be unwittingly put in place. In any case, provided we thoroughly weed out these constituent tags--eliminating the inappropriate, muzzling the redundant, and cutting out the out-of-the-way--we shall erect a more reliable stream to build on, an inferential crawl space on which to reconstruct new inferential systems. It is then possible to combine such filtered sets into definitive sets of rules and inject them into the emerging ontology and knowledge warehouse. Within that enhanced schema, the palette of semantic relations will become expanded to provide to the mechanism of automated reasoning, and the search IR engine will record a positive increase of search effectiveness: plus, wider search scope and plus more relevant results. The incumbent of such an undertaking should interrogate the trust model which is the basis of the Semantic Web first to bring about a responsible governance of the same. Index documents include RDF Triples groups and the wildcard of RDF Triples groups, giving users more leeway in how they describe their queries. Better matches between the retrieval formula and the relevant phrases can be provided thanks to semantic tags' ability to capture the semantic linkages between words. Semantic tags will boost information retrieval efficiency in a document or index expression.

3.4 Retrieval

Although keyword searches form the basis of current retrieval technology, some users may want to organize their searches according to broader semantic ideas that incorporate more commonly used terms and a deeper understanding of a topic. Semantic search engines, websites that are customized for each user, and smart information services will all benefit from improvements in retrieval technology that can pull out the semantic links between words.

3.5 Analysis

Standard Web text documents exist alongside their parallel counterparts on the Semantic Web. The semantic web documents' annotations, on the one hand, encapsulate the meaning of the assertion made in the online text document by way of metadata and machine interpretation. On the other hand, machines will be able to grasp and deal with the knowledge in online text documents if semantic descriptions are added to the web text documents using semantic web documents.

Power System Technology
ISSN:1000-3673

Received: 16-06-2025 *Revised: 05-07-2025* *Accepted: 20-08-2025*

3.6 Algorithm

Input: E → List of Contents

Al → List of concepts t → New Sentence in E x → New word in E
Ev → visited document

tu → Semantics for Documents for each concept d_j

do

compute $frq_d$of d_j compute $frq_{st}$of d_j compute $frq_x$of d_j
for each Ev

for each Sv

insert d_j in Al

end

end

for each concept $d_j = d_v$ do

update frq_E of d_j add new concept → Cl end
end

output: concepts list Cl

4. Results and Discussion:

This part shows the results obtained in the comparison test conducted for our proposed semantic search-based technique with the existing methods. For this examination we have taken 10 queries which needs to be searched in our Proposed method as well as with other existing methods also. Here we compare with Google, Duck Duck Go, Bing. So, the queries are given in the Table 4.1. By giving these queries in every search engine, we will observe the quantity of data it delivers per search and time consumed to return the results respectively. Here we are considering our Semantic search method as Proposed method.

Table 1: List of Queries taken to test the speed and accuracy

S.No	Queries raised by users
1	Best vegetarian Hotels
2	Creative 2.1 speakers
3	Fastest Internet broadband network in Mumbai
4	Photostable sunscreen spf55
5	Who is the President of India
6	Covid-19 cases today

Power System Technology
ISSN:1000-3673
Received: 16-06-2025 Revised: 05-07-2025 Accepted: 20-08-2025

7	Indian banks with highest interest percentages

As shown in the Table 4.1, These are the queries taken to search in different search engines. Here the comparison we are doing with Google, Bing and Duck Duck Go. Here we are obtaining the accuracy of content-based results to be delivered to the user. As per our proposed method it gives more accuracy about content. It collects the data from throughout the internet and gives the user. It provides Positive, Negative as well as neutral Information about the content what user search for.

Table 2: List of Queries Searched comparatively with other search engines

S. No	Proposed Method	Google	Duck Duck Go	Bing
Q1	7	5	6	4
Q2	9	6	5	7
Q3	6	4	5	6
Q4	8	5	6	7
Q5	7	4	5	3
Q6	8	3	7	4
Q7	9	7	6	3
Average	**83%**	**78%**	**72%**	**68%**

As per the Average value gathered after the implementation and comparison of content-based searches. Our Proposed method shows the highest accuracy 83% and Google gives 78% of accuracy and Duck Duck Go provides 72% of accuracy and finally Bing gives 68% of Accuracy.

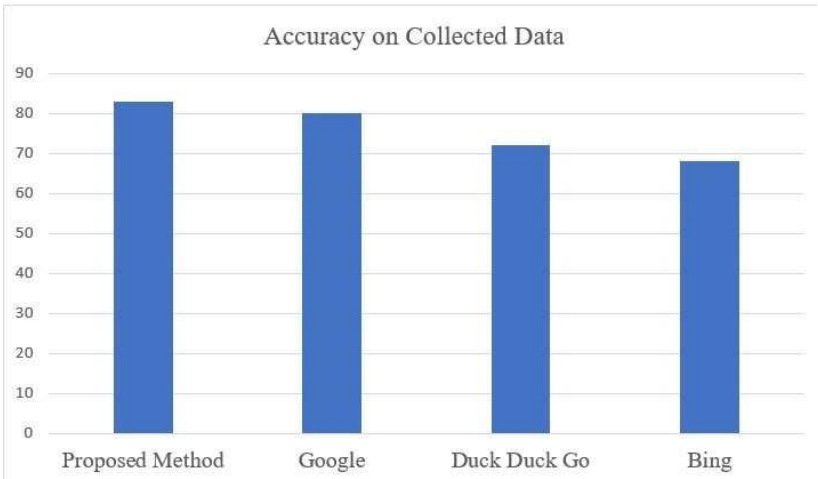
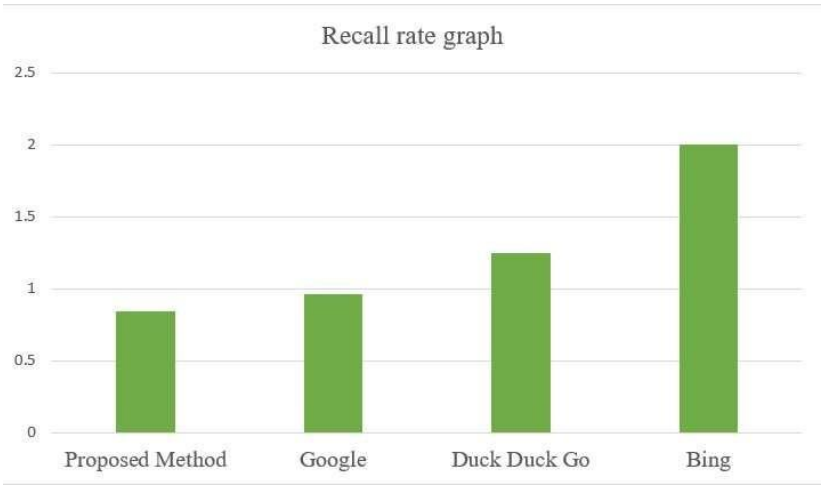


Figure 4: Accuracy based Comparison

Power System Technology
ISSN:1000-3673

Received: 16-06-2025 *Revised: 05-07-2025* *Accepted: 20-08-2025*

As like the comparison based on accuracy, there is a comparison called Recall rate which is about how much time is taken for a Search engine to process the user required data. So, for



the comparison in that case, we have compared our Prosed Method with Google, Duck Duck Go and Bing.

Figure 5: Recall rate graph

As shown in the Figure 4.2, It shows that our Proposed Method takes less time to process when compared with other search engines.

5. Conclusion

In this work, we provide a high-level overview of the search strategies used by a few popular semantic web search engines. We also talked about how semantic web search engines work and what they can tell us about things like high recall but low accuracy, figuring out what the user wants, wrong searches, and crawler efficiency.

In this paper, we introduce a semantic web tool for doing site-wide searches. The goal of our future work is to create a semantic web search system capable of providing complete results to even the most intricate queries. Here we achieved with 83% of accuracy and 1.75 recall rate. This study also provides a quick summary of the finest semantic search engines, each of which takes a somewhat different approach to providing a personalized search experience. Searching the internet nowadays is difficult, and it is estimated that roughly half of the hard inquiries remain unanswered. Semantic search may be used to improve upon standard online searches. But it's still not clear whether a search engine can give results that meet all of these criteria.

Power System Technology
ISSN:1000-3673

Received: 16-06-2025 *Revised: 05-07-2025* *Accepted: 20-08-2025*

Refrences

[1] L. Lai, C. Wu, P. Lin and L. Huang, "Developing A Fuzzy Search Engine Based on Fuzzy Ontology and Semantic Search," 2011 Ieee International Conference on Fuzzy Systems (Fuzz-Ieee 2011), 2011, Pp. 2684-2689, Doi: 10.1109/Fuzzy.2011.6007378.

[2] Sanjib Kumar Sahu, D.P. Mahapatra2, R.C. Balabantaray,"Comparative Study of Search Engines in Context of Features and Semantics", 20th June 2016. Vol.88. No.2

[3] Dinesh Jagtap, Nilesh Argade, Shivaji Date, Sainath Hole, Mahendra Salunke,"Implementation of Intelligent Semantic Web Search Engine", Vol. 4 Issue 04, April-2015

[4] Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha, Fakeeha Fatima,"Text Mining: Techniques, Applications and Issues", Vol. 7 No. 11, 2016.

[5] Lien- Lai Chao-Chin Wu Pei-Ying Lin, Liang-Tsung Huang,"Developing A Fuzzy Search Engine Based on Fuzzy Ontology and Semantic Search", Une 27-30, 2011, Taipei, Taiwan.

[6] Wei-Dong Fang, Ling Zhang, Yan-Xuan Wang, Shou-Bin Dong,"Toward A Semantic Search Engine Based on Ontologies", Proceedings of The Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005

[7] Hossein Hassani, Christina Beneki, Stephan Unger, Maedeh Taj Mazinani and Mohammad Reza Yeganeg,"Text Mining in Big Data Analytics", Ig Data and Cognitive Computing.

[8] Berners-Lee, T., Hendler, J. And Lassila, O. "The Semantic Web", Scientific American, May 2001.

[9] Deborah L. Mcguinness. "Ontologies Come of Age". In Dieter Fensel, Jim Hendler, Henry Lieberman, And Wolfgang Wahlster, Editors. Spinning The Semantic Web: Bringing The World Wide Web To Its Full Potential. Mit Press, 2002.

[10] G. Sudeepthi1, G. Anuradha, Prof. M. Surendra Prasad Babu, "A Survey on Semantic Web Search Engine", International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.

[11] Ramprakash Et Al "Role of Search Engines in Intelligent Information Retrieval on Web", Proceedings of the 2nd National Conference; Indiacom.

[12] R. Bekkerman and A. Mccallum, "Disambiguating Web Appearances of People in A Social Network," Proc. Int "L World Wide Web Conf. (Www "05), Pp. 463-470, 2005.

[13] G. Salton and M. Mcgill, Introduction to Modern Information Retrieval. Mcgraw-Hill Inc., 1986.

[14] F. Manola, E. Miller, And B. Mcbride, Rdf Primer. W3c Recommendation, Vol. 10, no., 2004.