International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7948
***Article Received**: 02 May 2023 **Revised**: 26 June 2023 **Accepted**: 28 July 2023*

An Efficient Information Extraction Mechanism with Page Ranking and a Classification Strategy based on Similarity Learning of Web Text Documents

Sunil Kumar Thota[1], Dr. G V S Raj Kumar[2], Dr. B. Raja Koti[3], Dr. K. Naveen Kumar[4]
[1]Research Scholar, Department of Computer Science and Engineering, GST,
GITAM Deemed to be University, Visakhapatnam, India
sunilshivaji@gmail.com
[2]Professor, Department of Computer Science and Engineering, GST,
GITAM Deemed to be University, Visakhapatnam, India
gganapav@gitam.edu
[3]Assistant Professor, Department of Computer Science and Engineering, GST,
GITAM Deemed to be University, Visakhapatnam, India
rbadugu@gitam.edu
[4]Associate Professor, Department of Computer Science and Engineering, GST,
GITAM Deemed to be University, Visakhapatnam, India
nkuppili@gitam.edu,

Abstract: Users have recently had more access to information thanks to the growth of the www information system. In these situations, search engines have developed into an essential tool for consumers to find information in a big space. The difficulty of handling this wealth of knowledge grows more difficult every day. Although search engines are crucial for information gathering, many of the results they offer are not required by the user because they are ranked according on user string matches. As a result, there were semantic disparities between the terms used in the user inquiry and the importance of catch phrases in the results. The problem of grouping relevant information into categories of related topics hasn't been solved. A Ranking Based Similarity Learning Approach and SVM based classification frame work of web text to estimate the semantic comparison between words to improve extraction of information is proposed in the work. The results of the experiment suggest improvisation in order to obtain better results by retrieving more relevant results.

Keywords: information, ranking, classification, semantics.

I. INTRODUCTION

The acquisition of information has become an essential requirement for individuals in today's world. The definition of information, along with its transmission, has undergone significant transformations in recent decades. As per research [1], [2], search functionality has become one of the most popular applications on the internet. Conventionally, search systems have relied on metadata keywords to match user queries. However, these systems do not consider the semantic relationship between terms of the query and additional recognitions important to user. As a result, you can advance your search process by adding your own semantics. The proliferation of the www information system has prompted numerous scholars to design diverse strategies to manage and arrange extensive reservoirs of data. These methodologies encompass not only the efficacy of automated arrangement and categorization but also concerns related to scalability.

Web-based documents exhibit a wide range of diversity and are structured in various ways, many of which are inadequately organized. The categories of websites vary from basic personal homepages to extensive corporate sites. The information stored on the internet is used in semantic search applications for the Semantic Web [3], [4]. However, the meaning of terms can change between domains, as seen with the keyword "Apple" which can refer to different entities [5], [6]. Users searching for "apples" on the web may mean "apples" as fruits rather than “apples” as computers or mobiles. New terms are constantly being generated with new meanings being assigned to existing statements. Maintaining ontology to keep up with new terms and meanings is difficult and expensive. Semantic-Similarity is important for numerous tasks related to web databases [7], [12], and using a semantic similarity measure can help suggest or modify queries based on previous user queries. However, accurately measuring semantic-similarity between words remains a challenge in web-mining, information retrieval, and natural language processing [8], [9], [10], [11], and [19].

Previously, similarity learning has been employed for purposes such as community extraction, detection of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7948
***Article Received**: 02 May 2023 **Revised**: 26 June 2023 **Accepted**: 28 July 2023*

relationships, and removal of entity ambiguity [7], [12], [14], [25]. It has proven that evaluating semantic similarity accuracy between concepts and web texts has several limitations. Its goal in this effort is to improve the results of traditional search that rely on "information retrieval technology" by using data from the Semantic Web. It will improve the traditional search by allowing the core term meaning to be included [8], [26]. Traditional word frequency searches are improved by understanding the underlying semantics of retrieved documents and requests of the users [2]. Inadequate search information systems present a challenge as users often have difficulty expressing their information needs, and the available techniques to assist them are of low quality. [16].

To improve information extraction in web mining, a proposed mechanism uses Similarity Learning of Web Text Classification through Support Vector Machines (SVM) combined with ranking of the classified web pages. This mechanism aims to overcome the limitations of measuring precision and recall in semantic similarity. Because the majority of web applications receive content based on user input queries, these searches typically contain a small number of keywords, limiting the scope of semantic association for information retrieval. Due to the enormity of the internet, connecting the collection of documents acquired semantically with the user's specific search terms is a challenging task.

The suggested Similarity Learning of Web Text technique would describe a methodology for learning similarity in web text documents in terms of close collaboration and categorising pages in terms of keywords, as well as preparing classification patterns [22], [23]. The SVM classification algorithm will be used to generate classification patterns for each term. These patterns will then be used to determine the appropriate document class that matches the query. This paper endeavors to estimate semantic similarity automatically between keywords and retrieve documents, thereby enhancing information extraction and providing more precise outcomes by effectively evaluating the semantic similarity between the retrieved documents and keywords.

II. LITERATURE REVIEW

Even if English is considered a primary dominating language in a different kind of information business, the www database has become a genuine information source in other languages. Numerous methods and approaches for successful information retrieval have been developed, with search engines serving as notable examples [8], [20]. Information retrieval is widely used in various fields such as research, education, business, e-commerce, and entertainment. As the search has become increasingly competitive, some search engines have developed tailored search services. Users can describe the WebPages categories they're interested in with "Google's Custom Search" for example [17].

D. Zhou et al. [15] provide a unique strategy for structuring enhanced user profiles and developing tailored queries utilising the external corpus. The approach used in this model involves merging the topic models using two pseudo-alignment group of documets with a cutting-edge textual demonstration learning framework, which includes the utilization of "insertion of word". On basis of user profiles, it will develop two new query expansion technologies. The proposed approach employs two strategies based on topic relevance, which are achieved by using topic-weighted word insertion and search keywords derived from a user's profile. This approach has been thoroughly evaluated on two datasets and has outperformed current technologies, including both non-personalized and customized query expansion methods.

J. Hoxha et al. study effectively addresses recommending resources from multiple domains by combining the semantic content of resources with patterns in user browsing behaviour. The system suggests the most appropriate recommendations by using a support vector machine to determine relevance, with no overlap among retrieved domains. The study's effectiveness is demonstrated by analyzing a genuine dataset of user browsing activity and behavioral research logs that have been semantically enriched. The objective is to investigate the impact of structure on generating precise recommendations.

Xuan Wu et al. [20] studied semantic connections in social tagging systems, exploring links between tags, words, and both tags and words. Using tags and functions extracted from words, they generated three similarity graphs. Additionally, they incorporated feedback information from top-level documents related to physicians to standardize the strength of various relationships in the three graphs. The aim of this study is to use three similarity graphs as a new model for query expansion, with the goal of improving personalized search results. Empirical evaluations conducted on actual datasets validate the effectiveness of the proposed approach.

As per S. Lawrence et al. [21], 85 percent of internet users rely on search engines to obtain relevant information. Moreover, their study revealed that approximately 71 percent of website visitors tend to navigate to other websites while using search engines. Web search engines, on the other hand, are limited in terms of "exposure," "cost," "interface selection," "retrieving useful information," and "ranking the relevancy of the results." In summary, although search engines do have their limitations, they remain essential for conducting online searches. One distinctive feature of search engines is

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7948
***Article Received**: 02 May 2023 **Revised**: 26 June 2023 **Accepted**: 28 July 2023*

their employment of significantly advanced "IR approaches" that enable more effective web searching compared to traditional IR methods.

The development of web-based solutions to support a large number of concurrent users and pages searched in web search engines has created various problems and obstacles in information retrieval (IR) and classification. The number of concurrent users cannot be predicted in advance, which can lead to system overload. The quantity of pages available on the internet exceeds the scale associated with traditional data, and the number of suppliers of search engines, web users, and web pages is rapidly increasing, leading to greater memory space and computing power requirements.

In semantics, a semantic link refers to a relationship between two senses or concepts. Conventional taxonomic relationships, such as "synonymy" and "hyponymy," connect terms like "vehicle, automobile" and "bird, kiwi." "Classical relations" refer to such relationships, while composite relations are more complex and challenging to map. "Synonymy" and "hyponymy" are examples of semantic similarity, where "car and truck" share related characteristics and belong to the same category of vehicles. It can also be referred to as "semantic resemblance."

Any semantic relationship between two senses indicates the existence of a semantic link between them. Let's say the terms "vehicle, automobile" and "bird, kiwi" are used for an occasion. The two pairings are connected through conventional taxonomic relationships, namely "synonymy" and "hyponymy." The term "classical relations" describes such partnerships. Numerous terms, on the other hand, contribute to the formation of composite relations that are difficult to separate and map. Semantic similarity, which includes synonymy and hyponymy links, is the association between two concepts that have comparable qualities, compositions, or qualities. For instance, car and truck are semantically equivalent as they both refer to vehicles and share related attributes. Semantic similarity is a term used to describe the resemblance in meaning between two words

In many disciplines, the existing IR system is limited in its ability to find information using major search engines [7], [20], [17]. Several techniques have been suggested in the past, including [10], [15], [20], [21]. However, the challenge of accurately associating and categorizing user requests with the retrieved information necessitates semantic improvisation. An approach based on topic relevance and subject word weight [15] has shown improvement in query expansion, but it heavily relies on user profile personalization. The efficiency of support for IR and suggestion has also improved, according to a study and evaluation of website surfing patterns [10]. However, its classification and prediction are dependent on the semantic association of the blog. Through the suggested effort, the classification accuracy will be improved.

III. PROPOSED LEARNING AND CLASSIFICATION APPROACH

Figure 1 depicts the technique for learning and categorization for accurate information retrieval. The function is divided into three phases: "Information Retrieval", "Similarity Learning" and "Ranking and Classification." In the following sections, each of these phases is discussed separately.

3.1. Information Retrieval

The information retrieval process primarily focuses on answering user queries on the basis of a few keywords given by the users [14]. Users input of keywords are made up of user information requirements that must be prepared in the form of "keywords" from query input.

It usually goes through a tokenization process to extract each individual keyword from the query, resulting in a list of keywords, and then uses a query cleaning mechanism to extract generic phrases from query with the help of a pre-defined removal words dictionary.

The system performs a web page extraction procedure by utilizing the user's provided keywords to obtain relevant web documents. This process is repeated for every keyword set derived from the query input of user. The documents obtained through this process are then used for the subsequent phases of learning, classification, and ranking. Following the completion of learning and classification, the user is notified of the categorised results received in response to the input query.

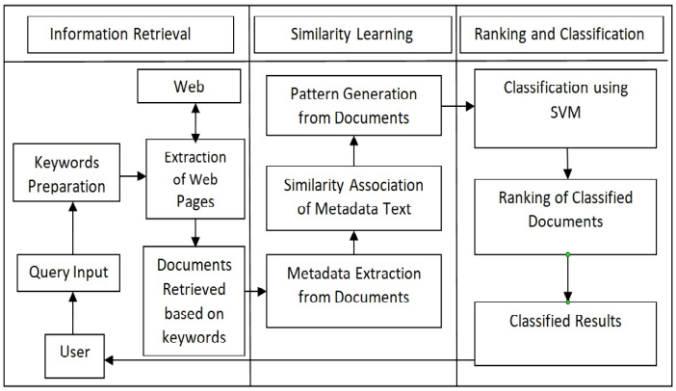


Fig 1: Information Extraction System Architecture

3.2 Similarity Learning Mechanism

Similarity Learning of Web Text is a mechanism for generating relevant patterns of knowledge with respect to an input query by separation of relevant documents which are positive and negative [13], [18]. The documents extracted may

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7948
***Article Received**: 02 May 2023 **Revised**: 26 June 2023 **Accepted**: 28 July 2023*

contain a mix of semantically relevant and unrelated phrases. In this scenario, learning unrelated and related phrases is critical in order to return the classified documents as query results. The web text metadata is pre-processed to eliminate stop words before relevance calculation.

3.2.1 Web Text Terms Extraction

For each document Dk web text is extracted as Ed, on which text cleaning is applied which creates a vector, Vd, comprising words. Ed's terms go through a purification process to get rid of any unnecessary stop words.

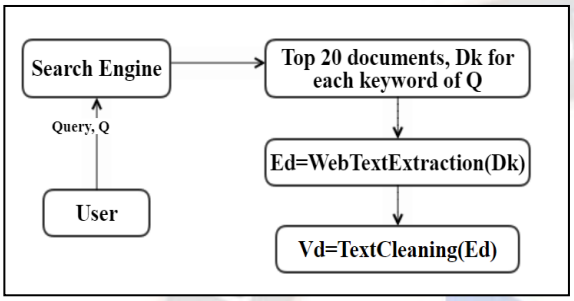


Fig 2: Web Text Extraction Procedure

Figure 2 depicts the web text extraction procedure. The system utilizes a web search engine to find relevant documents based on the user's query input and generate keywords. For each keyword, the top twenty documents are gathered, and the WebTextExtraction method is used to extract phrases. The extracted phrases are cleaned using the TextCleaning method to remove stop words and construct the final set of terms, Vd. The similarity association index in Vd is then computed.

3.2.2 Similarity Association and Pattern Generation

The system selects the most frequent keywords from the user input query and calculates their similarity with other keywords present in the extracted document terms Vd. This process helps to establish a pattern for classification. This mechanism's approach is described in Algorithm-1.

Algörithm. 1: Similarity_Association of keywords.

Input: Set of Keywords as, K[ ].

Set of Documents terms , V[R],

(The total no. of documents retrieved is R).

Output: Keywords similarity associated value array, SA_V [ ].

Method: Similarity_Association (K[ ], V[R])

Step 1. for every keyword of the set K[], do the following:

Step 2. Retrieve the keyword and store it in Wk.

Step 3. Assign a frequency count of 0 to fc.

Step 4. For each document in the set Vt[R], do the following:

Step 5. Retrieve the terms of the document and store them in Dj[].

Step 6. For each term in the document retrieved from Vt[R], do the following:

Step 7. Retrieve the term and store it in Dt.

Step 8. If the term Wk is equal to Dt, then do the following:

Step 9. Increment fc by 1.

Step 10. End the loop for term.

Step 11. End the loop for document.

Step 12. If the frequency count fc is greater than 0, then do the following:

Step 13. Calculate the term frequency s_a_v as ((fc * 100) / R).

Step 14. Store the term frequency s_a_v in the array SA_V[i].

Step 15. End the loop for keyword.

The result of Similarity Association of Keywords generates an array of similarity-related values for keywords, SA_V[]. It processes and generates the patterns required for document classification using the value of SA_V[]. Algorithm-2 depicts the pattern creation procedure.

Algorithm-2: Generation of Patterns

Input: Keywords Set as, K [ ].

Keywords similarity associated values array, SA_V[ ]

Output: Keywords Patterns Array, PV[ ]

Method: Pattern-Generation(K[ ], SA_V[ ])

Step 1. For each keyword in the set K[], do the following:

Step 2. Retrieve the key value and store it in H1.

Step 3. Retrieve the corresponding similarity value from SA_V[] and store it in K_S1.

Step 4. Store the value of H1 in H_V.

Step 5. For each remaining keyword in the set K[], do the following:

Step 6. Retrieve the key value and store it in H2.

Step 7. Retrieve the corresponding similarity value from SA_V[] and store it in K_S2.

Step 8. If K_S2 is greater than K_S1, then do the following:

Step 9. Store H2 in H_V.

Step 10. Store K_S2 in W_V.

Step 11. End the loop for remaining keywords.

Step 12. Store the value of H_V in PK_L[].

Step 13. Store the value of W_V in PK_V[].

Step 14. Remove the value H_V from Key[].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7948
***Article Received**: 02 May 2023 **Revised**: 26 June 2023 **Accepted**: 28 July 2023*

Step 15. Remove the value W_V from SA_V[].

Step 16. End the loop for the keyword set.

Step 17. Retrieve the highest value from PK_V[] and store it in Phigh.

Step 18. Assign the value of Midx to 0.

Step 19. While the value of Phigh is greater than 0, do the following:

Step 20. Store a space in P_V.

Step 21. For each term present in PK_L[], do the following:

Step 22. Retrieve the PK_L value and store it in H_V.

Step 23. Retrieve the PK_V value and store it in K_V.

Step 24. If K_V is equal to Phigh, then do the following:

Step 25. Add W_V to P_V.

Step 26. Store the value of P_V in F_P[Midx].

Step 27. Add 1 to Midx.

Step 28. Subtract 1 from Phigh.

Step 29. End the loop while.

F_P[], the final created pattern, which will be used in the document classification and ranking process and to generate the classed results required.

3.3. Classification & Ranking

During web text learning process, F_P[] is obtained and used as input for SVM-based classification which is a reliable classifier for both linear and non-linear data. In order to generate results, the SVM evaluates the significance of each retrieved online document in comparison to the pre-established patterns.

The Support-Vector-Machines mechanism distinguishes between positive and negative features based on the distance between them on a hyperplane boundary. The most significant pattern indicates the highest positive relevancy, while the least significant pattern indicates the lowest positive relevancy regarding the connection with the query. Each pattern serves as a segmentation the border.

In Algorithm-3, the mechanism of the classification process is depicted.

Algorithm-3: Retrieved Documents Classification

Input: Keywords Set as, K[ ].

Patterns Generated , F_P[ ].

Documents Set Retrieved, R[D].

Output : Classified-Results , CR[ ].

Method: **Documents Classification** (K[ ], F_P[ ], R[D])

Step 1. Initialize widx to 0.
Step 2. Iterate over each pattern in F_P[].
Step 3. Set the value of fp to the current pattern in F_P[].
Step 4. For each document retrieved in Ret[], denoted as Ret[d], do the following:
Step 5. Set ps to false.
Step 6. Call the function getDocumentTerms on Ret[d], denoted as a, and store the result in Rj[].
Step 7. Call the function getDocumentPatterns on Ret[d] and Key[], and store the result in D_P[].
Step 8. Call the function comparePatternSimilarity on D_P[] and fp, and store the result in ps.
Step 9. If ps is true, do the following:
Step 10. Store the value of Ret[d] in CR[widx].
Step 11. Increment widx by 1.
Step 12. Set ps to false.
Step 13. Call the function removeDoc to remove Ret[d].
Step 14. End the inner loop.
Step 15. End the outer loop.

The technique comparePatternSimilarity (...) tries to find a positive or negative relationship between two patterns by comparing them. It's likely that the document has nothing to do with the high positive pattern and instead has something to do with the mid or low relevant pattern. In relation to the user question, highly positive classification results will be considered highly accurate. It uses a few real-time online documents retrieved from a different domain to assess this approach. On the classified documents a ranking algorithm is applied to order the documents.

A new algorithm named Weighted Page Content Rank (WPCR) has been developed to address the shortcomings of conventional search engine algorithms like PageRank and Weighted PageRank. The WPCR approach incorporates techniques such as web content mining and web structure mining to enhance the ordering of pages in search results, which improves user experience and makes it easier for users to locate relevant and important pages.

The Weighted Page Content Rank (WPCR) algorithm orders pages in search results based on a numerical value assigned to each page. The algorithm uses web structure mining to assess a page's importance by analyzing its inlinks and outlinks, while web content mining evaluates its relevance to the user's query. When a web page becomes more relevant to a query, its WPCR score increases. This leads to improved accuracy and efficiency of search results and enhances users' overall search experience.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7948
***Article Received**: 02 May 2023 **Revised**: 26 June 2023 **Accepted**: 28 July 2023*

Algorithm: Web Page Content Ranking Algorithm

Input: Classified Pages CR[],

Weights of Inlinks and Oütlinks of CR[],

Query terms of Q, d the damping factor.

Output: Ranks of Pages

a) Relevance Calculation

Step 1. Get a page from CR[] and put it in P.
Step 2. Count all important word groups that make up Q (also known as N).
Step 3. Check to see if the N sequences are contained in P.
Step 4. Store sum of occurrances of all N sequences into Z
Step 5. The longest sequences that can exist in P are grouped and stored in S and the total number of times a string appears in S is stored in X .
Step 6. The ContentWeight (cw) is computed by dividing X by Z and store the value cw into the array CW.
Step 7. The number of search phrases found in P are stored in C and the number of valid query terms are stored in D.
Step 8. The ProbabilityWeight(pw) is computed by dividing C by D and stored in the array PW.
Step 9. For each page of CR, repeat Step 1 to Step 8.

b) Rank calculation

Step 1.Retrieve a page P from the array CR
Step 2.Identify all of P's backlinks (B).
Step 3. $PR(P) = (1-d) + d[\sum_{V \epsilon B} PR(V)W^{in}(P,V)W^{out}(P,V)](CW[P] + PW[P])$
Step 4. Store the ranking value obtained from PR(P) into Rank[P].
Step 5. For each page of CR, repeat Step 1 to Step 5

IV. EXPERIMENT ANALYSIS

4.1 Description of dataset and Metrics

The World-Wide-Web database contains a vast number of web pages that reference diverse semantic relationships. When utilising a keyword-based Web search engine to look for an entity in a specific semantic connection, the user must create a query that includes some keywords linked to the entity and the relation. As a result, we gathered a collection of papers from diverse domains that are used informally to compute usage patterns or distributional models in order to conduct the evaluation.

Using the Google search engine, 100 web data records from a domain "Shopping online" and “Booking Online” were obtained for review.

These datasets will be subjected to a learning and classification mechanism to evaluate precision, recall, and accuracy percentages, using the Eq. 3.1, 3.2 and 3.3 as given below.

$$\text{Precision}=\frac{\text{Total number of documents correctly classified (True Positives)}}{\text{Total number of documents correctly classified (True Positives)+Incorrectly Classified (False Positives)}}*100 \quad (3.1)$$

$$\text{Recall}=\frac{\text{Total number of documents correctly classified (True Positives)}}{\text{Total number of documents correctly classified (True Positives)+Incorrectly Classified (False Negatives)}}*100 \quad (3.2)$$

$$\text{Accuracy}=\frac{\text{Total number of documents correctly classified}}{\text{Total number of documents retrieved}}*100 \quad (3.3)$$

In the context of text classification, accuracy is calculated by dividing the number of documents that were classified correctly by the total number of documents that were retrieved, using the input query as a basis for retrieval. The research results were compared to established text classification methods, including Multivariate-Bernoulli-Naive-Bayes (MVNB) and Multinomial-Naive-Bayes (MNB).

4.2 Result Analysis

A query is run to generate the required learning patterns, which were then categorised using these patterns to measure the precision, recall, and accuracy. Tables 3.1, 3.2 illustrates the generated keyword patterns for the various domain searches.

A boundary level was selected for each generated pattern based on the most favourable and negative associations. The highest level of positive association is considered the most positive, while the lowest level is considered the least favourable. We judged patterns with more than two values to be high, patterns with two values to be mid, and patterns with one value to be low. A total of top 25 results are gathered for each query in order to perform classification based on the produced pattern. We compute the classified result precision, recall and accuracy % for each iteration against the created pattern, as shown in the Table.

The obtained results reveal that the percentage of precision, recall and accuracy is excellent at the high and mid boundary levels. The results of the comparison are displayed below.

It demonstrates that in the case of a high boundary level, more accuracy and relevancy may be attained in the support of the created pattern. However, it may vary depending on the length of the query keyword. The lengthier the query, it was argued, the more specific and accurate the results would be.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7948
Article Received: 02 May 2023 Revised: 26 June 2023 Accepted: 28 July 2023

Table-3.1: shows the outcome for the query "*online airline booking*"

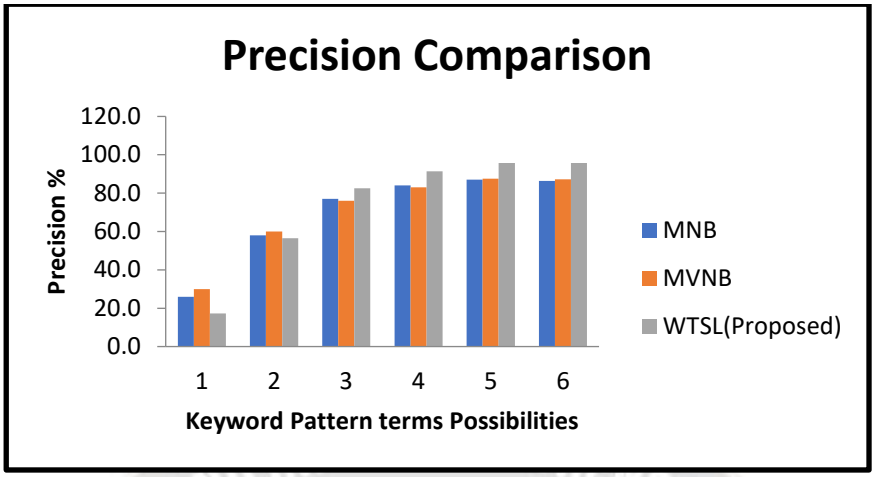
Keyword Pattern Terms Possibilities	Keyword Patterns	Boundary Level	Total Classified	Correctly Classified (True-Positives)	Correctly Classified (True-Negatives)	Incorrectly Classified (False-Positives)	Incorrectly Classified (False-Negatives)
1	Booking	Low	25	4	0	19	2
2	Online	Low	25	13	0	10	2
3	Booking, online	Mid	25	19	0	4	2
4	Online, booking	Mid	25	21	0	2	2
5	Booking, airline, online	High	25	22	0	1	2
6	Online, airline, booking	High	25	22	0	1	2

Precision Values		
MNB	MVNB	WTSL (Proposed)
26.0	30.0	17.3
58.0	60.0	56.5
77.0	76.0	82.6
84.0	83.0	91.3
87.0	87.5	95.7
87.0	87.5	95.7

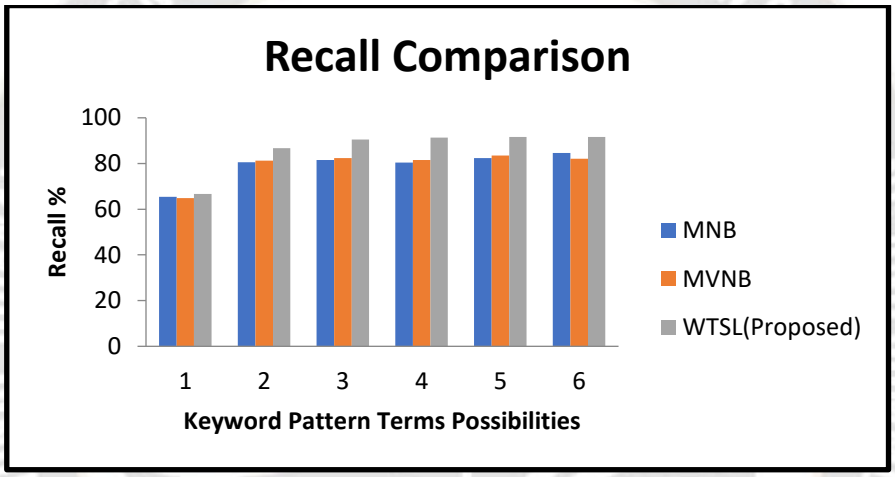
Recall Values		
MNB	MVNB	WTSL (Proposed)
65.4	64.8	66.67
80.6	81.2	86.67
81.5	82.3	90.48
80.4	81.5	91.30
82.3	83.5	91.67
84.6	82.1	91.67

Accuracy Values		
MNB	MVNB	WTSL (Proposed)
24.0	25.0	16.0
50.2	52.3	52.0
70.8	71.4	76.0
74.3	76.9	84.0
78.9	80.4	88.0
81.2	81.4	88.0

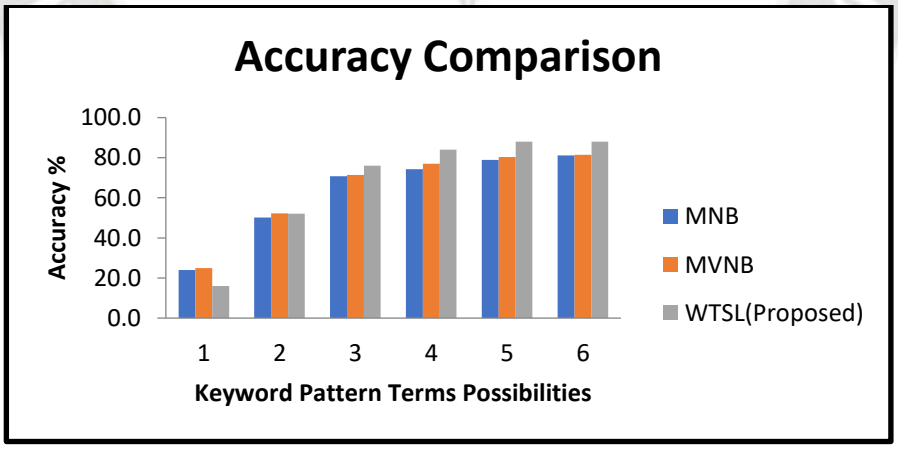
International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7948
***Article Received**: 02 May 2023 **Revised**: 26 June 2023 **Accepted**: 28 July 2023*



3.3(a)



3.3(b)



3.3(c)

Fig. 3.3(a)Precision, 3.3(b) Recall and 3.3(c) Accuracy Comparison for the query online airline booking

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7948
***Article Received**: 02 May 2023 **Revised**: 26 June 2023 **Accepted**: 28 July 2023*

Table 3.2 Result for the search term "*online electronic shopping*"

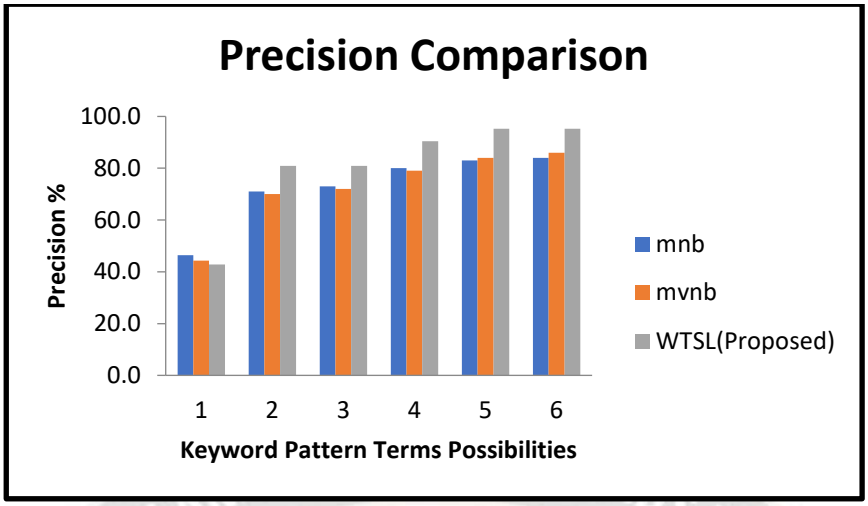
Keyword Pattern Terms Possibilities	Keyword Patterns	Boundary Level	Total Classified	Correctly Classified (True-Positives)	Correctly Classified (True-Negatives)	Incorrectly Classified (False- Positives)	Incorrectly Classified (False-Negatives)
1	Shopping	Low	25	4	0	19	2
2	Online	Low	25	13	0	10	2
3	Shopping, online	Mid	25	19	0	4	2
4	Online, Shopping	Mid	25	21	0	2	2
5	Shopping, elelectric, online	High	25	22	0	1	2
6	Online,electronic, shopping	High	25	22	0	1	2

Precision Values		
MNB	MVNB	WTSL (Proposed)
46.4	44.3	42.9
71.0	70.0	81.0
73.0	72.0	81.0
80.0	79.0	90.5
83.0	84.0	95.2
84.0	86.0	95.2

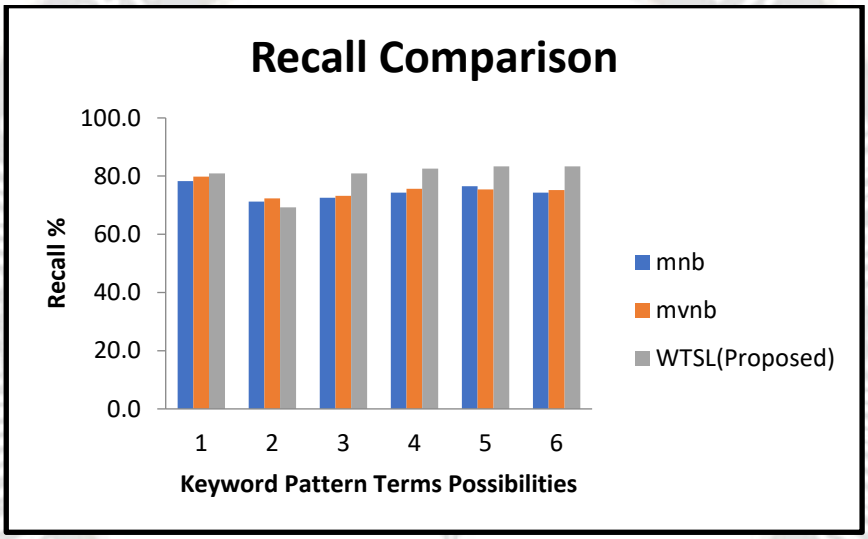
Recall Values		
MNB	MVNB	WTSL (Proposed)
78.3	79.8	80.95
71.2	72.4	69.23
72.6	73.2	80.95
74.3	75.6	82.61
76.5	75.4	83.33
74.3	75.2	83.33

Accuracy Values		
MNB	MVNB	WTSL (Proposed)
38.0	40.0	36.0
60.2	61.3	68.0
61.4	62.3	68.0
68.3	69.5	76.0
71.2	70.8	80.0
73.4	72.6	80.0

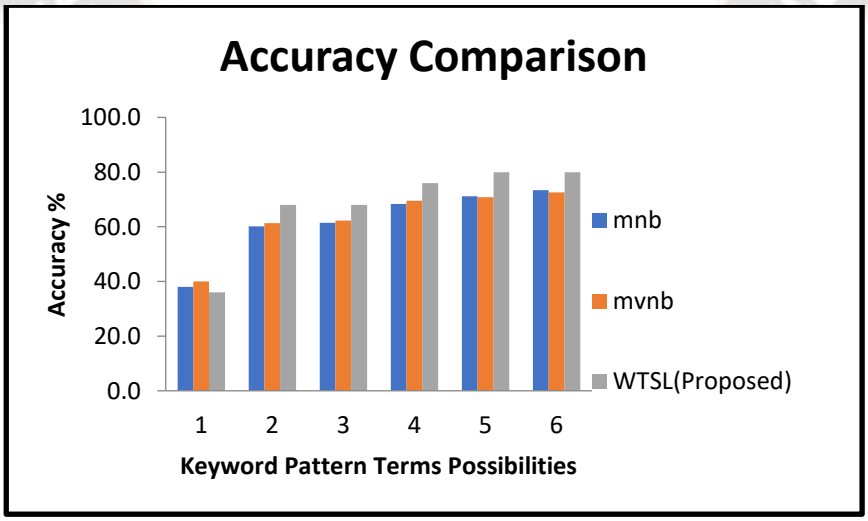
International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7948
***Article Received**: 02 May 2023 **Revised**: 26 June 2023 **Accepted**: 28 July 2023*



3.4(a)



3.4(b)



3.4(c)

Fig. 3.4(a) Precision, 3.4(b) Recall and 3.4(c) Accuracy Comparison for the query online electronic shopping

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7948
***Article Received**: 02 May 2023 **Revised**: 26 June 2023 **Accepted**: 28 July 2023*

V. CONCLUSION

The vast growth of the www information system has led to an enormous quantity of information available to users. To navigate this vast space, search engines have become an essential tool to retrieve pertinent information. However, the current ranking system, which relies heavily on keyword matching, frequently produces irrelevant outcomes. This has led to the need for a semantic-based approach to improve the information retrieval accuracy. The paper, proposes a Ranking Based Similarity Learning Approach and SVM-based classification framework to estimate the comparison of the meanings of two or more words to determine their level of similarity in order to improve the extraction of the information. Based on the experimental outcomes, it appears that this technique can enhance the precision of information retrieval by retrieving a greater number of pertinent results. However, further research is required to address the issue of classifying information sources into groups of similar subjects, which remains unresolved. Ultimately, this research highlights the importance of developing more sophisticated methods to manage the ever-expanding knowledge available on the web.

REFERENCES:

[1] J. Shen, E. Zheng, Z. Cheng, C. Deng, "Assisting Attraction Classi-fication by Harvesting Web Data", IEEE Access Volume: 5 Pages: 1600 - 1608, 2017.

[2] Tzu-Yi Chan, Yue-Shan Chang, "Enhancing Classification Effec-tiveness of Chinese News Based on Term Frequency", IEEE 7th In-ternational Symposium on Cloud and Service Computing (SC2), Pages: 124- 131,2017.

[3] C. Chen, X. Meng, Z. Xu, T. Lukasiewicz, "Location-Aware Person-alized News Recommendation with Deep Semantic Analysis", IEEE Access, Volume: 5 Pages: 1624 - 1638,2017. https://doi.org/10.1109/ACCESS.2017.2655150.

[4] J. Gracia, E.Mena, "Web-Based Measure of Semantic Relatedness", In Proceedings of 9th International Conference On Web Information Systems Engineering (Wise '08), Vol. 5175, Pp. 136-150, 2008.

[5] J. Ruohonen, "Classifying Web Exploits with Topic Modeling", 28th International Workshop on Database and Expert Systems Applica-tions (DEXA) Pages: 93 - 97, 2017.

[6] U. Kumaresan, K. Ramanujam, "Web Dat a Extraction from Scien-tific Publishers' Website Using Heuristic Algorithm", International Journal of Intelligent Systems and Applications (IJISA), Vol.9, No.10, pp. 31 - 39,

[7] R. L. Cilibrasi, P.M.B. Vitanyi, "The Google Similarity Distance", IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No 3, 370-383, 2007.

[8] Tchiegue, R. Li, S. Ma, "A web text classification technique for un-labeled training samples", 6th IEEE International Conference on Software Engineering and Service Science (ICSESS) Pages: 437 - 440, 2015.

[9] T. M. Veeragangadhara Swamy, G. T. Raju, "A Novel Prefetching Technique through Frequent Sequential Patterns from Web Usage Data", An International Journal of Advanced Computer Technology, Vol. 4, No. 6, June 2015.

[10] J. Hoxha, P. Mika, R. Blanco, "Learning Relevance of Web resources across Domains to make recommendations", 12th international con-ference on Machine Learning and Applications, vol. 2, pp. 325-330, 2013.

[11] Y. Li, A. Algarni, M. Albathan, Y. Shen, and M.A. Bijaksana, "Rel-evance Feature Discovery for Text Mining", In IEEE Trans. Knowl. Data Eng., vol. 26, Jan. 2015.

[12] P. Li, H. Wang, K. Q. Zhu, Z. Wang, and X. Wu, "Computing term similarity by large probabilistic is a knowledge", In Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management, ser. CIKM '13, New York, NY, USA, pp. 1401-1410, 2013.

[13] Kalaivani, A. ., Karpagavalli, S. ., & Gulati, K. . (2023). Expert Automated System for Prediction of Multi-Type Dermatology Sicknesses Using Deep Neural Network Feature Extraction Approach. International Journal of Intelligent Systems and Applications in Engineering, 11(3s), 170–178. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2557

[14] Y. Li, A. Algarni, and N. Zhong. "Mining positive and negative pat-terns for relevance feature discovery", In KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge dis-covery and data mining, pages 753-762, New York, NY, USA, 2010.

[15] C. Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, and Khaled F. Shaalan. "A survey of Web information extraction sys-tems", IEEE Transactions on Knowledge and Data Engineering, 18(10):1411-1428, 2006.

[16] D. Zhou, X. Wu, W. Zhao, S. Lawless, J. Liu, "Query Expansion with Enriched User Profiles for Personalized Search Utilizing Folksonomy Data", IEEE Transactions on Knowledge and Data Engg Volume: 29, Issue:7, Pages: 1536 - 1548, 2017.

[17] M. A. Siddiqui,"Mining Wikipedia to Rank Rock Guitarists", Inter-national Journal of Intelligent Systems and Applications (IJISA) , vol.7, no.12, pp.50 - 56,.

[18] X. He, C.H.Q. Ding, H. Zha, H.D. Simon, "Automatic topic identifi-cation using webpage clustering", In Proceedings of IEEE Interna-tional Conference on Data Mining, pp.195-202, 2001.

[19] W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge", IEEE Transactions on Knowledge and Data Engineering, 1041-4347, 2016.

[20] A. Ashari, M. Riasetiawan, "Document Summarization using Tex-tRank and Semantic Network", International Journal of In telligent Systems and Applications (IJISA), Vol.9, No.11, pp. 26 - 33,

[21] X. Wu, Dong Zhou, Yu Xu, S. Lawless, "Personalized query expan-sion utilizing multi-relational social data", 12th International Work-shop on Semantic and Social Media Adaptation and Personalization (SMAP) Pages: 65-70, 2017.

[22] S. Lawrence, L. Giles, A. Spink, "Inquirus Web metasearch tool: A user evaluation", In Proceedings of WebNet, PP. 819-820, 2000.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7948
Article Received: 02 May 2023 Revised: 26 June 2023 Accepted: 28 July 2023

[23] S. T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern re-finement in text mining", In Proc. IEEE Conf. Data Mining, pp. 1157-1161, 2006.
[24] N. Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", Vol. 24, No.1,Jan 2012.
[25] A. Anagnostopoulos, A. Broder, and K. Punera, "Effective and Effi-cient Classification on a Search-Engine Model, Knowledge and In-formation Systems, 2007.
[26] Matti Virtanen, Jan de Vries, Thomas Müller, Daniel Müller, Giovanni Rossi. Machine Learning for Intelligent Feedback Generation in Online Courses . Kuwait Journal of Machine Learning, 2(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/188
[27] Z. Zhang, Q. Li, and D. Zeng, "Mining evolutionary topic patterns in community question answering systems", IEEE Trans. Syst., Man, Cybern. Vol. 41, no. 5, pp. 828-833, 2011. https://doi.org/10.1109/TSMCA.2011.2157131.
[28] J. Zhu, Member, K. Wang, Y. Wu, Zhongyi Hu, and H. Wang, "Min-ing User-Aware Rare Sequential Topic Patterns in Document Streams", IEEE Transactions on Knowledge and Data Engineering, 2016.
[29] Thota, D. S. ., Sangeetha, D. M., & Raj , R. . (2022). Breast Cancer Detection by Feature Extraction and Classification Using Deep Learning Architectures. Research Journal of Computer Systems and Engineering, 3(1), 90–94. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/48
[30] M. Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broad-head, and Oren Etzioni. "Open information extraction from the Web". In Proceedings of the 20th International Joint Conference on Artificial Intelligence, pages 2670-2676, 2007.
[31] M. S. Kamel, "An Efficient Concept Based Mining Model for En-hancing Text Clustering", IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10, October 2010.
[32] Nidhi Grover, Ritika Wason, “Comparative Analysis Of Pagerank And HITS Algorithms”, Vol. 1 Issue 8, October – 2012.