International Journal of Engineering & Technology, 7 (4) (2018) 4856-4861

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet. v7i418075
Research paper

A web text similarity learning and classification approach for efficient information extraction

Sunil Kumar Thota [1] *, Dr. Tummala Sita Mahalakshmi [2]

[1] Research Scholar, CSE Department, Gitam University, Vishakapatnam
[2] Professor, CSE Department, Gitam University, Vishakapatnam
**Corresponding author E-mail: sunilkumarthota.phd@gmail.com*

Abstract

Over the last few years, the explosion of the World Wide Web has allowed users to access more and more information. In this circumstance, search engines have become a necessary tool for users to uncover the information they require in a huge space. As a result, the task of organizing this rich information becomes more difficult every day. It plays an important function in accomplishing the information, but numerous of the returned results are not related to the user's necessitates, because they are ranked according to the string match of the user's query. This resulted in semantic differences involved in the meaning of the keywords in the retrieved documents and the terms used in the user's query. The problem of categorizing large sources of information into groups of similar topics has not yet been resolved. In this paper, it proposes a web-text similarity learning (WTSL) method and classification based on SVM mechanism. This proposal aims to automate the estimation of the semantic comparison among the words or article to enhance the information extraction. The experimental results suggest the improvisation towards retrieving more accurate results by retrieving more relevant documents.

Keywords: *Web Mining; Text Similarity, Classification; Information Extraction.*

1. Introduction

Today, information is a basic necessity for everyone. The concept of information, and consequently the transfer of information, has changed dramatically over the past few decades. Search is the most admired application on the web [1], [2]. Most of the traditional search systems typically use metadata keywords that match the query. However, these systems didn't take into the consideration of the semantic relations among "query terms" and additional "perceptions" that are important to the user. Consequently, the addition of unambiguous semantics is able to advance the retrieval process. The expansion of the "World Wide Web" has led many researchers to devise various methodologies for organizing huge information sources. It not only aims to the quality of automatic configuration and classification but also scalability issues.

Documents on the web are very diverse and structured differently, and most are not well structured. The nature of a website can range from very simple personal home pages to huge corporate websites. All contribute to the vast information store, and mostly perform the semantic Search-based application [3], [4] for the Semantic Web to search. However, semantic similarities among entities change eventually and transversely in the domains [5], [6]. For instance, searching for a keyword "apple" often connects to the "apple computer" or "apple mobile" on the web. Nevertheless, this "apple" sensation is not described in most "universal thesauri" or "dictionaries". A user who searches for "apple" on the Web might be significant in this intellect of "apple" and "not apple" as a fruit. New term expressions are continuously generating, and new senses are allocating to existing statements terms. Even sustaining ontology manually to confine these new terms and senses is expensive and complex.

The "Semantic similarity measurement" is significant in various web-related assignments [7], [12]. One way to uncover the suitable words to comprise in the query is to be evaluated with the old user queries utilizing the "semantic similarity measures". If the earlier query is semantically correlated to the present query, it will be capable of utilize it to suggest to the user or modify the original query from the underlying query engine. Regardless of the convenience of "semantic similarity measures" in these applications, accurate measurement of semantic similarity among two words or items left-over the challenges [8]. Even the accurate measurement of semantic similarities between words is an essential issue in "web mining", "information retrieval", and "natural language processing" [9], [10], [11], [19].

In the past, similarity learning is being used for the application of community extraction, relationship detection and entity ambiguity elimination [7], [12], [14], [25]. It has shown some limitation in measuring semantic similarity accuracy between concepts and web documents. Its aims to progress the conventional search results depend on the "information retrieval technology" by means of the data of the "Semantic Web" in this work. It will be improving the traditional search by allowing the search to incorporate the basic term meaning [8], [26]. Understanding the hidden semantics of retrieved documents and user queries improves traditional searches focused on word frequency [2]. The problem with poor search information systems is that users cannot convey their information requirements clearly, or there are low-ranking techniques [16] to estimate pages that are relevant to the query.

This paper proposes a Web Text Similarity Learning (WTSL) and Classification based on SVM mechanism to overcome the limitation of measuring semantic similarity accuracy and support more accurate information mining for the various web mining application. As most of the web application retrieve documents based on the user input query, these queries consist of limited keywords which limit the scale of semantic association for the information retrieval. As the vastness of the web is big, it is difficult to relate the set of the document retrieved semantically in associating to the user's limited input search keywords.

The proposed WTSL method will define a methodology to learn the web text similarity in relates of terms close association and categorize the documents in relates to the keywords and prepare the classification patterns [22], [23]. The prepared classification patterns for each keyword will be used for the most accurate class document is relevant to the query utilizing an SVM classification method. This work aims to automatically estimate semantic similarities between keywords terms and retrieve documents to improve information extraction and also to provides more accurate results by efficiently estimating semantic similarity between keywords and retrieved documents.

2. Related works

2.1. Existing problem and solution

The web has become the authentic source of information in different languages, even though English is considered as a major dominating language at present in a different kind of information service. Many methods and approached are proposed for effective information retrieval and search engines are leading examples of these techniques [8], [20]. It is mostly used for information retrieval in "research work", "education", "business", "e-commerce", and "entertainment sectors". As competition in the search market has accelerated, some search engines have launched customized search services. For example, "Google's Custom Search" allows users to specify the categories of WebPages they're interested in [17].

D. Zhou et al. [15] suggest an innovative model for organizing enhanced user profiles using the external corpus for personalized query development. This model combines the contemporary "state-of-the-art" textual demonstration learning framework, such as "word insertion", with the topic models in the two pseudo-alignment document groups. It will build two new query expansion technologies based on user profiles. These two techniques are based on topic relevance between topic-weighted word insertion and search terms within a user's profile. An in-depth experimental evaluation of two actual data sets using different outer corpus showed that our approach is superior to existing technologies, including traditional "non-personalized" and "customized query expansion" methods.

J. Hoxha et al. [10] solves the problem of suggesting resources from different domains by combining the semantic content of these resources with user browsing behavior patterns. It suggests, not to have domain overlap among obtained domains with newly developed associations supports on investigated semantic content of web resources. If the user is currently viewing a particular page, it will suggest an approach to applying a "support vector machine (SVM)" to learn the relevance of the resource and to predict what is best for the user to recommend. It learns the impact of the structure on generating accurate recommendations from the actual data set of semantically rich logs of user browsing performance across multiple Web sites and behavioral experiments to exhibit the effectiveness of our approach.

Xuan Wu et al. [20] investigate "multiple semantic relationships in social tagging systems", including among "tags", between words, and between tags and words. Three similarity graphs are created based on the tags and functions obtained from the word. It also standardizes the effortlessness of multiple relationships through three similar graphs, incorporating physician-related "feedback information" from top-level documents. The objective of this work is to improve the customized search results by considering the three similarity graphs above as a new query expansion model. Experiments performed on real data sets validate the proposed approach.

S. Lawrence et al. [21] estimated that "85 %" of web users utilize search engines to uncover their information requires. They designated that "71 %" of web users accomplish other websites during accessing the search engine. They also condition that the most significant action accomplished on the Internet user's rate searching. However, "web search engines" are restricted in terms of "exposure", "cost", "interface selection", how fine they "retrieve relevant information" and how fine they "rank the relevance of the results". In short, while the search engines restrictions, they are essential for searching the web. There is a little abnormal characteristic of search engines which exploit a diversity of comparatively superior "IR techniques" to improve the searching on the web in comparison to the conventional IR methods.

The development in web-based solutions has raised many problems and challenges in related to IR and its classification to support the number of concurrent users in admired search engines and the number of documents being accessed [1], [8], [24]. Further, in particular, to the number of concurrent users of a search engine at any specified time cannot be forecast in advance and it can overload a system. The number of documents widely available on the Internet surpasses the magnitude associated with the foundations of "classical data" by numerous guidelines of magnitude [27], [28]. Moreover, the quantity of providers of "Internet search engines", "Web users", and "Web pages" is mounting at a incredible rapidity, with every page of the media that occupies additional "memory space" and include unusual category of "multimedia information" such as "images", "graphics", "audio", and "video".

A semantic connection between any two perceptions designates the existence of a semantic relation linking them. Consider, for an occasion let the term couples as "car, automobile" and "bird, kiwi". The terms in these two pairs are associated through "classical taxonomic relations" such as "synonymy", ("car and automobile are synonyms") and "hyponymy" ("kiwi is a bird"). Such relationships are called "classical relations". However, numerous terms contribute to further composite relations which cannot be effortlessly distinct and mapped. Semantic similarity consists of semantic associations among two concepts that have "similar nature", "composition or attributes". Examples of semantic similarity are "synonymy" and "hyponymy" relations. For a case, "car and truck" are semantically comparable because together are vehicles, and allocates a set of related features. A "semantic similarity" is a particular feature of "semantic relatedness". Semantic similarity wraps up nearly every one of the earlier talk about classical associations.

The current IR system facing limitation in search of information utilizing major search engine in many domains [7], [20], [17]. As many solutions are designed in the past [10], [15], [20], [21], but the inability of accurately association and classification between the user request and the retrieved information are though semantically demands a need for the improvisation. A topic relevance based approach [15] based on topic word weight shows an improvisation towards query expansion but it mostly depends on user profile personalization. A study and evaluation of websites browsing pattern also show an improvement towards the effectiveness of support for IR and recommendation [10]. But, its prediction and classification depend on the weblog semantically association. This proposed work will enhance the accuracy of classification through a new WTS Learning method to overcome the limitation of information retrieval in various domains.

3. Proposed learning and classification approach

The mechanism for learning and classification for accurate information retrieval is presented in Fig. 1. It segments the function in three phases as, "Information Retrieval", "WTS Learning", and "Classification". It discusses all these phases individually in the following sections.

3.1. Information retrieval

The mechanism of information retrieval mostly focused to answer the user questions based on a few keywords as input from the user [14]. The input keywords by the users consist of user information requirement concept, which needs to prepare in the form of "keywords" from the query input.

It generally undergoes a process of tokenization of keywords to extract each individual keywords from the query to build a set of keywords as K and applies a query cleaning method to remove the generic terms from the query in support of a pre-defined removal words dictionary.

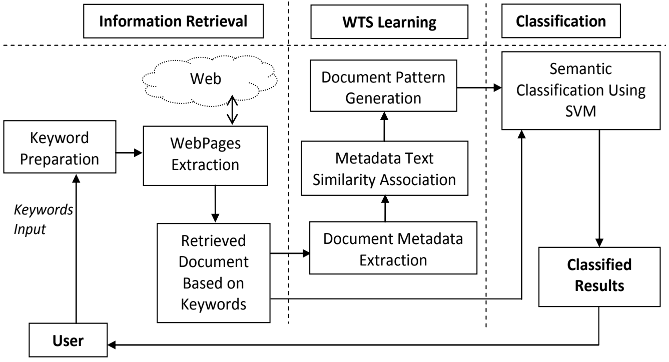
Information Retrieval
WTS Learning
Classification
Web
Keyword Preparation
WebPages Extraction
Keywords Input
Retrieved Document Based on Keywords
User
Document Pattern Generation
Metadata Text Similarity Association
Document Metadata Extraction
Semantic Classification Using SVM
Classified Results

Fig. 1: System Architecture.

Based on the prepared keywords it implements a webpage extraction mechanism to retrieve the keywords relevant web documents. The process of web documents extraction are repeated for each number of keywords prepared from the user query input and the obtained sets of documents are stored for the next phase learning and classification mechanism. On completion of learning and classification, the classified results obtained are replied to the user against the input query.

3.2. WTS learning mechanism

The mechanism of Web Text Similarity Learning (WTSL) generate knowledge patterns relevant to the input query to segregates the positive and negative relevant documents [13], [18]. The extracted documents might have few semantically relevant terms and many unrelated terms. In such case, it is important to learn the unrelated and related terms to quantify the related documents as query results. The extracted metadata as web text from the web document is initially pre-processed to remove the stop words and implements a similarity calculation to determine the relevancy as explained in the following section.

1) Extracting the web text terms

It constructs a vector, V_d of terms from each document web text extracted as E_d. The terms in E_d undergoes a cleansing process to remove the unwanted stop words.

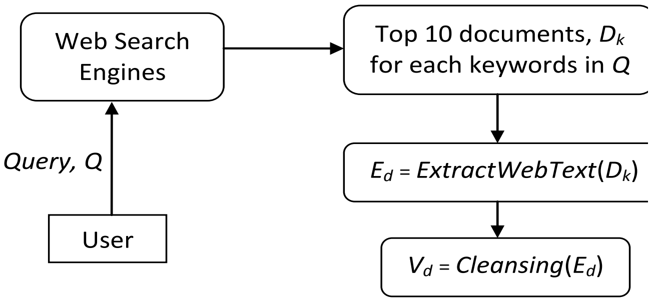
Web Search Engines
Top 10 documents, D_k for each keywords in Q
Query, Q
E_d = $ExtractWebText(D_k)$
User
V_d = $Cleansing(E_d)$

Fig. 2: Web Text Extraction Process.

The process of extraction of web text is shown in Fig.2. Based on the user Query, Q input and the generated keywords from Q, it retrieves the keyword relevant documents as D_k from a web search engine. It collects the top 10 results documents for each keyword and an ExtractWebText (D_k)method is processed to extract the terms from the document as E_d. To remove the stop words a Cleansing (E_d) method is processed to build the final V_d for a keyword retrieve documents. These terms in V_d(are)further utilized for computing (computing)the similarity association index.

2) Similarity Association and Pattern Generation

To compute the similarity association between the keywords, K, and the terms of the document, V_d it initially identifies the most frequent keywords among K and then identifies the similarity association between the most frequent with others keywords to generate the required pattern for the classification. The Algorithm-1 describes the procedure of the mechanism.

Algorithm-1: Similarity Association of Keywords.
Input:
Set of keywords as, K[ ].
Sets of documents terms, V[R],
(where R is the no. of documents retrieved).

Output: Array of keywords similarity associated value, SA_Value [ ].

```
Method: Similarity_Association (K[ ], V[R])
//-- Similarity Association for each keywords in K --
for (k=0; k <size of K; k++ )
{
        $W_k$ = K [k];
        $f_{count}$ = 0;
        //-- For each document retrieved as, V[R] --
        for ( r = 0; r < R; r++ )
        {
// -- Getting each documents terms --
$D_r$[ ] = V [r];

//-- For each terms of a document --
for ( d = 0; d<size of $D_r$; d++ )
{
$D_{term}$= $D_r$[d];
if ($W_k$ == $D_{term}$ )
{
        $f_{count}$ ++;
}
}
        }
        //-- Compute Similarity Association --
        if ($f_{count}$> 0)
        {
        sim_asoc_value=(($f_{count}$*100)/ R);
        }
        //-- Array of Similarity Association Values --
        SA_Value [ k] = sim_asoc_value ;
}
```

The outcome of the Similarity Association of Keywords generate an array of keywords similarity associated values, SA_Value [ ]. Utilizing the value of SA_Value [ ], it process and generate the patterns required for the document classification. The method for the generation of the pattern is illustrated in Algorithm-2.

Algorithm-2: Pattern Generation

Input:
Set of keywords as, K [ ].
An array of keywords similarity associated value, SA_Value [ ]

Output: Array ofKeywords Patterns, PValue [ ]

```
Method:PatternGeneration(K[ ], SA_Value [ ])
//-- Order the obtained Keywords based on its SA_Value --
for (k = 0; k <size of K; k++ )
{
$W_1$ = K [k];
Key_$SA_1$ =SA_Value [k];

//-- For each value on SA_Value --
H_Val=$W_1$;
for (j = 1; j<size of K; j++)
{
        $W_2$ = K [j];
        $Key\_value_2$ =SA_Value [j];
        if($Key\_SA_2$>$Key\_SA_1$)
        {
        W_Val=$W_2$;
        H_Val=$Key\_SA_2$;
}
}
PKey_List [k] = W_Val;
PKey_Value [k] = H_Val;

RemoveKey ( W_Val, K[ ] );
RemoveValue( H_Val, SA_Value [ ] );
}

//-- Pattern Generation --
$P_{high}$ = getHighestValue (PKey_Value [ ] );
$f_{idx}$ = 0;
while ($P_{high}$> 0)
{
P_Val = "";
for (p = 0; p<size of PKey_List; p++)
{
        W_Value = PKey_List [p] ;
        K_Value = PKey_Value [p];

if(K_Value==$P_{high}$)
{
P_Val + = W_Value ;
}
}
F_Pattern [$f_{idx}$] = P_Val;

$f_{idx}$ = $f_{idx}$ + 1;
$P_{high}$= $P_{high}$- 1;
}
```

The final generated pattern, F_Pattern[ ]will be utilized the further classification process of the documents and to generate the required classified results.

3.3. Classification based on SVM

The obtained pattern, F_Pattern [ ] through the web text learning process is an input to the classification based on SVM process. As SVM is considered as a promising classifier for the linear and non-linear data. This classification approach measures the relevancy of each retrieved web document against the generated pattern to produce the results.

As SVM works on a hyperplane boundary based on the feature distance separating the features positively and negatively. Here, it considered each pattern is a boundary of segregation and the highest pattern is considered as highest positive relevancy, whereas the lowest level pattern is considered the lowest positive relevancy in terms of associating with the query. The mechanism of the classification process is illustrated in the Algorithm-3.

Algorithm-3: Classification of Retrieved Document

```
Input:
Set of keywords as, K[ ].
Generated pattern, F_Pattern[ ].
Set of Retrieved Documents, R[D].

Output : Classified Results , CResult[ ].
Method:Document_Classification (K[ ], F_Pattern[ ], R[D ])
$c_{idx}$= 0;
for( k=0;k< size of F_Pattern;k++)
{
$F_{patt}$ = F_Pattern[ k ];
//-- For each document retrieved based on the keywords
for( d=0; d <size of R; d++ )
{
patt_sim = false;
// -- Getting each documents terms --
$R_d$ [ ] = getDocumentTerms( R [d] );
D_Pattern = getDocumentPattern( $R_d$ [ ], K[ ] );
patt_sim=comparePatternSimilarity (D_Pattern, $F_{patt}$);
        if (patt_sim = = true )
        {
CResult[$c_{idx}$ ]= R [d];
$c_{idx}$++;
patt_sim = false;
removeDoc (R [d]);
        }
}
}
```

Here, the method comparePatternSimilarity (...) try to compare the find the positive or negative relationship with the comparing pattern. It is possible that the document might be not relevant to the high positive pattern but may relate to the mid or low relevant pattern. The high positive classify results will be considered as highly accurate results in relates to the user query. To evaluate this mechanism it implements this against few real-time web document retrieved from a different domain. It discusses it more briefly in the next section.

4. Experiment evaluation

The "World Wide Web" enclose an enormous number of "Web pages", which relates to numerous semantic associations. When a user needs to search for an article in a precise "semantic relation" by means of a "keyword-based Web search engine", the user has to prepare a query through several keywords associated to the individual and the relation. So, to perform the evaluation it collected a set of various domains documents that are utilized in a familiar approach to computing "distributional models" or "usage patterns". An evaluation datasets of 100 web data records are collected using Google search engine from the different domain as, "Tours and-Travel", "Treatment and Health Care", and "Online e-shopping". Over this datasets, it will perform the learning and classification mechanism to measure the percentage classification precision, recall, and accuracy using the equation-1, 2 and 3.

$$Precision = \left(\frac{\sum No.of\ Correctly\ Associated}{Total\ Number\ of\ Document\ Classified}\right) \times 100 \tag{1}$$

$$Recall = \left(\frac{\sum No.of\ Incorrectly\ Associated}{Total\ Number\ of\ Document\ Classified}\right) \times 100 \tag{2}$$

$$Accuracy = \left(\frac{\sum(No.of\ Classified\ \cap No.of\ Incorrect\ Associated)}{Total\ No.of\ Classified\ Document}\right) \times 100 \tag{3}$$

In general, a classifier estimates the accuracy based on the number of the document is classified, and the number of documents is being correctly associated and the number of document incorrectly classified in relevant to the input query. A difference between the number of a document classified and a number of incorrectly associated with the total number of document record retrieved measure the accuracy percentage. It compares the output of our result with the popular Google search engine results for the different query.

4.1. Result analysis

To measure the result it executed four different queries to generate the required learning patterns and using these patterns classification was performed to compute the precision, recall, and accuracy.

Table-1, 2 and 3 show the generated keyword patterns for the different domain query.

Based on each generated pattern it set a boundary level based on most positive and negative association. High is considered most positively associated and low is least associated. It considered pattern having > 2 values as high, having 2 value will be mid and, having 1 value is low. As per each query, a total of top 25 results are retrieved to perform the classification based on the generated pattern. On each iteration against the generated pattern, it computes the classified result measures as precision, recall and accuracy percentage as shown in the Tables-1, 2 and 3.

The obtained results show that with high and mid boundary level the percentage of precision and accuracy is impressive. So, it compares the computed results of high level against the popular Google's search result. The comparison results are shown below.

Table 1: Outcome for the Query "Online Airline Booking"

Keyword Patterns	Boundary Level	Total Retrieved	Total Classified	Correctly Associated	Incorrectly Associated	PRECISION	RECALL	AC
booking, online, airline	High	25	23	22	1	95.65217391	4.347826087	
booking, airline, online	High	25	23	22	1	95.65217391	4.347826087	
online, booking	Mid	25	23	21	2	91.30434783	8.695652174	
booking, online	Mid	25	23	19	4	82.60869565	17.39130435	
online	Low	25	23	13	10	56.52173913	43.47826087	
booking	Low	25	23	4	19	17.39130435	82.60869565	
Google Results		**25**	**25**	**20**	**5**	**80**	**20**	

Table 2: Outcome for the Query "Online Electronic Shopping"

Keyword Patterns	Boundary Level	Total Retrieved	Total Classified	Correctly Associated	Incorrectly Associated	PREC	RECALL	AC
online, shopping, gifts	High	25	21	20	1	95.23809524	4.761904762	
electronic, online	Mid	25	21	19	2	90.47619048	9.523809524	
online, shopping	Mid	25	21	17	4	80.95238095	19.04761905	
online	Low	25	21	9	12	42.85714286	57.14285714	
electronic	Low	25	21	17	4	80.95238095	19.04761905	
Google Results		**25**	**25**	**19**	**7**	**76**	**28**	

Table 3: Outcome for the Query " Eye and Cancer Care Hospitals"

Keyword Patterns	Boundary Level	Total Retrieved	Total Classified	Correctly Associated	Incorrectly Associated	PREC	RECALL	AC
cancer, care, hospitals	High	25	23	22	1	95.65217391	4.347826087	
eye, care, hospitals	High	25	23	21	2	91.30434783	8.695652174	
hospitals, care	Mid	25	23	19	4	82.60869565	17.39130435	
cancer	Low	25	23	13	10	56.52173913	43.47826087	
eye	Low	25	23	16	7	69.56521739	30.43478261	
Google Results		**25**	**25**	**23**	**2**	**92**	**8**	

Fig. 3 and Fig. 4 Show the Comparison Results of Precision and Recall Percentage with A Different Query. It Shows That with Learning the Keyword Similarity in Respect to the Retrieved Document Web Text Can Improve the Result Precision and Minimize the Rate of Recall. in Comparing to Google Results the High-Level Patterns Results Show More Précised, As It Has High Similarity Association to the Web Text and Minimize Recall Rates.

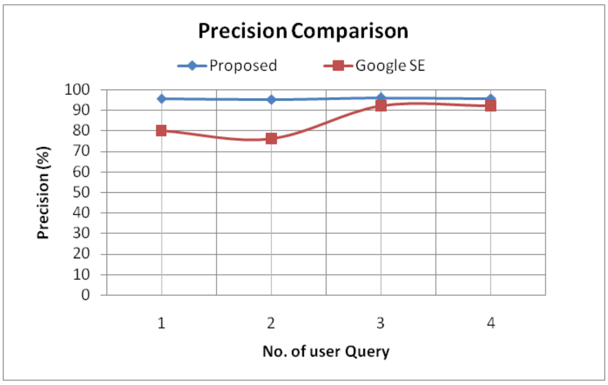

Fig. 3: Precision Comparison.

Fig.5 shows the accuracy assessment of the proposed and Google results outcomes. It shows that in support of generated pattern more accurate and relevant can be achieved in case of high boundary level. But it might have variance depends on the query keyword length. It suggested that the longer the query more precise and accurate the results will be, but in case of the search engine it's results are more associated but it attained more recalls.

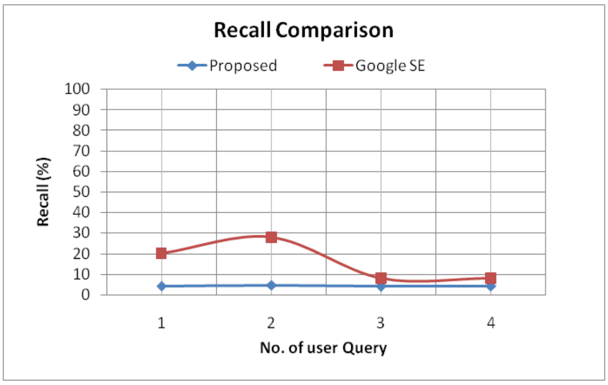

Fig. 4: Recall Comparison.

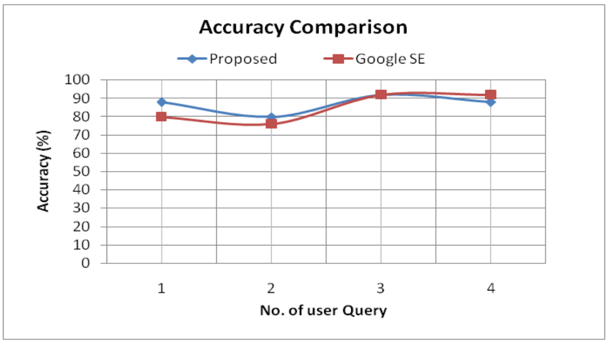

Fig. 5: Accuracy Comparison.

5. Conclusion

This paper presented a web-text similarity (WTS) learning and classification approach to improvising the accuracy of search results. As user mostly submits different variance and context query it is important to relate the retrieve documents to categorized accurately to reduce the recall. It has implemented a WTS learning mechanism to learn the close association between the query and web text document and generate a keywords pattern. The generated patterns are utilized to perform the classification of the retrieved documents to suggest the best accurate results. An experiment was performed over a dataset constructed from online web documents. Utilizing the generated patterns it performs the classification of these datasets posing a different query. The results analysis between the Google and proposed results shows an improvisation the accuracy and low recall rate. A low recall rate shows more precise and might be more relevant and satisfactory results to a user query.

References

[1] J. Shen, E. Zheng, Z. Cheng, C. Deng, "Assisting Attraction Classification by Harvesting Web Data", IEEE Access Volume: 5 Pages: 1600 - 1608, 2017. https://doi.org/10.1109/ACCESS.2017.2656878.

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

[3] C. Chen, X. Meng, Z. Xu, T. Lukasiewicz, "Location-Aware Personalized 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. https://doi.org/10.1007/978-3-540-85481-4_12.

[5] J. Ruohonen, "Classifying Web Exploits with Topic Modeling", 28th International Workshop on Database and Expert Systems Applications (DEXA) Pages: 93 - 97, 2017. https://doi.org/10.1109/DEXA.2017.35.

[6] U. Kumaresan, K. Ramanujam, "Web Dat a Extraction from Scientific Publishers' Website Using Heuristic Algorithm", International Journal of Intelligent Systems and Applications (IJISA), Vol.9, No.10, pp. 31 - 39, https://doi.org/10.5815/ijisa.2017.10.04.

[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. https://doi.org/10.1109/TKDE.2007.48.

[8] Tchiegue, R. Li, S. Ma, "A web text classification technique for unlabeled 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 conference on Machine Learning and Applications, vol. 2, pp. 325-330, 2013. https://doi.org/10.1109/ICMLA.2013.144.

[11] Y. Li, A. Algarni, M. Albathan, Y. Shen, and M.A. Bijaksana, "Relevance Feature Discovery for Text Mining", In IEEE Trans. Knowl. Data Eng., vol. 26, no. 6, pp., 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 &#38; Knowledge Management, ser. CIKM '13, New York, NY, USA, pp. 1401-1410, 2013. https://doi.org/10.1145/2505515.2505567.

[13] Y. Li, A. Algarni, and N. Zhong. "Mining positive and negative patterns for relevance feature discovery", In KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 753-762, New York, NY, USA, 2010. https://doi.org/10.1145/1835804.1835900.

[14] C. Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, and Khaled F. Shaalan. "A survey of Web information extraction systems", IEEE Transactions on Knowledge and Data Engineering, 18(10):1411-1428, 2006. https://doi.org/10.1109/TKDE.2006.152.

[15] 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 Engineering Volume: 29, Issue: 7, Pages: 1536 - 1548, 2017. https://doi.org/10.1109/TKDE.2017.2668419.

[16] M. A. Siddiqui,"Mining Wikipedia to Rank Rock Guitarists", International Journal of Intelligent Systems and Applications (IJISA) , vol.7, no.12, pp.50 - 56, https://doi.org/10.5815/ijisa.2015.12.05.

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

[18] 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.

[19] A. Ashari, M. Riasetiawan, "Document Summarization using TextRank and Semantic Network", International Journal of In telligent Systems and Applications (IJISA), Vol.9, No.11, pp. 26 - 33, https://doi.org/10.5815/ijisa.2017.11.04.

[20] X. Wu, Dong Zhou, Yu Xu, S. Lawless, "Personalized query expansion utilizing multi-relational social data", 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP) Pages: 65 - 70, 2017. https://doi.org/10.1109/SMAP.2017.8022669.

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

[22] S. T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining", In Proc. IEEE Conf. Data Mining, pp. 1157-1161, 2006. https://doi.org/10.1109/ICDM.2006.50.

[23] N. Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", Vol. 24, NO. 1, January 2012.

[24] A. Anagnostopoulos, A. Broder, and K. Punera, "Effective and Efficient Classification on a Search-Engine Model, Knowledge and Information Systems, 2007.

[25] 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.

[26] J. Zhu, Member, K. Wang, Y. Wu, Zhongyi Hu, and H. Wang, "Mining User-Aware Rare Sequential Topic Patterns in Document Streams", IEEE Transactions on Knowledge and Data Engineering, 2016. https://doi.org/10.1109/TKDE.2016.2541149.

[27] M. Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. "Open information extraction from the Web". In Proceedings of the 20th International Joint Conference on Artificial Intelligence, pages 2670-2676, 2007.

[28] M. S. Kamel, "An Efficient Concept Based Mining Model for Enhancing Text Clustering", IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10, October 2010.