RESEARCHERID
THOMSON REUTERS

[ICRTCET-2018]

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

WEB PAGE RANKING BASED ON QUERY AND TITLE SEMANTIC SIMILARITY USING MODIFIED WU-PALMER METRIC (MWUP)

Harish Kumar B T[*1], Dr. Vibha L[2] & Dr. Venugopal K R[3]

[*1]Assistant Professor, Dept of CSE, Bangalore Institute of Technology
Bangalore, India

[2]Professor, Dept of CSE, BNMIT, Bangalore, India

[3]Principal, UVCE Bangalore, India

ABSTRACT

Web pages are the important modes for disseminating the information to the huge users. The number of web pages containing the information is increasing drastically every day. Web users tend to search the required information from the WWW (World Wide Web), which is the repository of all the websites consisting of web pages. Web user's search criteria consist of the user's input query string and the web user's result criteria consists of the ranked web pages list. Most of the existing works ranks the search results by counting the frequency of query keyword occurrences or by computing the semantic similarity only between the query keywords and URL (Uniform Resource Locator) keywords. The proposed work is based on the computation of the semantic similarity between the query string and the web page title tag string. The semantic similarity is computed by slightly modifying the WUP (Wu-Palmer) technique and renamed as MWUP (Modified Wu-Palmer) technique. The proposed work uses the Wordnik API for synonym extraction. The proposed work tries to improve the relevancy of the web page to the user context by raking them using the semantic similarity.

Keywords: *Web Page, Ranking, Wordnik, Semantic, Query, Title Tag.*

I. INTRODUCTION

Web page ranking is very important for presenting the users with the web pages highly relevant to their context. Web page relevancy with the user's context can be found by finding the semantic similarity between the user query text and the web page title text. Web page title defines the overall context of the web page and its relevancy to the user's query context. The proposed work uses the WordNik API to retrieve the synonym of the given word. The WordNik API produces the synonyms in the JSON (Javascript Object Notation) format. The JSON formatted output is processed to retrieve the synonym word list. Example 1 shows the WordNik API URI (Uniform Resource Identifier), its description and the sample output of the WorkNik API for the word "sports" in JSON format .

A. Example 1:Worknik API URI and its Description http://api.wordnik.com/v4/word.json/sports/related Words?useCanonical=true&relationshipTypes=synonym&li mitPerRelationshipType=10&api_key=a2a73e7b926c924fad 7001ca3111acd55af2ffabf50eb4ae5

http://api.wordnik.com/: is the domain name of the Wordnik API.

V4: is the version of the Wordnik API.

Word.json: represents the format in which the output is required.
Education: is the keyword for which the synonym set is required.
RelatedWords: defines the type of word set the API should output.

UseCanonical: If this attribute is set to true then the API will auto perform the typos correction and stemming of keywords?

ISSN 2348 – 8034
Impact Factor- 5.070

RelationshipTypes: Defines the type of word that is required.

LimitPerRelationshipType: Defines the maximum number of synonym words to be present in the output.

API_KEY: is the key unique that is passed to make the API operate.

JJ. JSON Output for Word Sports [{"relationshipType":"synonym","words":["moc

k","mirth","diversion","frolic","play","jeer","game","mocker y","pastime","amusement"]}]

C. Motivation
Web users have to spend the considerable amount of time to search the required information from the web. Ranking of the web pages according to the users query string semantic similarity to the web page is highly beneficial to the web users. Web page ranking is motivating. in terms of retrieving highly context relevant web pages and reducing the search time.

D. Contribution
The proposed works has four major parts. First part involves algorithm for preprocessing the URL data set and representing each URL in the canonical form. URL keyword extraction and title tag keyword extraction and preprocessing algorithm in second part. Part three consists of algorithm to extract the synonym keyword set for URL and TITLE tag keywords and building the synonym database. Last part contains algorithm for preprocessing the query string and finding the similarity between the query and the URL and TITLE tag keywords using MWUP.

The rest of the paper is organized as depicted below. Section II presents a brief literature review related to the proposed research work. Brief description of the problem statement, aims and objectives of the proposed work are covered in section III. Section IV shows the main modules involved in the proposed work in the form of general architecture. Working model of the proposed methodology is presented with an example in section V. Section VI describes the algorithms proposed. Experimental organization and performance interpretation is covered in section VII. Final conclusion on the proposed research work is presented in section VIII.

II. RELATED WORK

Many researchers have already worked on the specified research area and have presented their ideas and solutions to the problems that they were able identify. This section focuses on providing the brief description of the techniques and the methodology used in the existing works related to this research area.

Juhi Agarwal et al., in [1] have proposed an algorithm to rank the web pages based on frequency of keywords and semantic words in the web page. Frequency means the number of times the keywords and the semantic words occurring in the web page. Authors used WordNet database for finding the synonyms. In [2] S. Ramana Murthy et al., has proposed genetic algorithm for ranking the web pages. This algorithm has used the concepts of mutation and crossover parameters in WordNet. Mutation is considered as synonym and crossover is considered as the topic based ranking, in which both the frequency count and also hyperlink count of number of times a given keyword and their synonym words appear in the web page.

Syntactic classification of web pages using fuzzy C-means algorithm and neural network classification based page ranking algorithm is proposed by Debajyothi Mukhopadhyay et al., in [3] where a single query may generate different ranking to the web pages depending on the web page category. George Tsatsaronis et al., in [4] has introduced semantic rank algorithm. Ranking of web pages is done by keyword and text semantic similarity by combining the WordNet and Wikipedia.

In [5] Marius-Gabriel Gutu et al., has compared the semantic similarity methodologies like WordNet based technique, explicit semantic analysis trained on Wikipedia and latent semantic analysis trained on Wikipedia. Dr.

ISSN 2348 – 8034
Impact Factor- 5.070

Daya Gupta et al., in [6] have presented a page ranking algorithm based on user preferences. This algorithm UPBR (User Preference Based Ranking) has employed structure, usage and content mining techniques. B. Prasanthi et al., in [7] have presented a methodology for image retrieval from web pages by re-ranking the images using the query specific semantic signatures.

Rekha Singhal et al., in [8] have proposed an algorithm for search engine optimization by page ranking using the in-linked weight-age of the web pages. Yuan Ziqian in [9] has proposed an improved ranking algorithm using cheating similarity and cheating relevance. This work is based on the white-list and the black -list pages. Web pages with lower black-list similarity and higher white-list similarity are more likely to be white-list page and vice versa.

In [10] Ali I El-Dsouky et al., using Analytical Hierarchy Process (AHP) and semantic relations has proposed a methodology for ranking of web pages. AHP is a tool for decision making and semantic relations are found using the WordNet ontology. Joeran Beel et al., in [11] has identified several factors and provided an introductory overview on each factors to rank the articles in the Google scholar.

III. PROBLEM STATEMENT

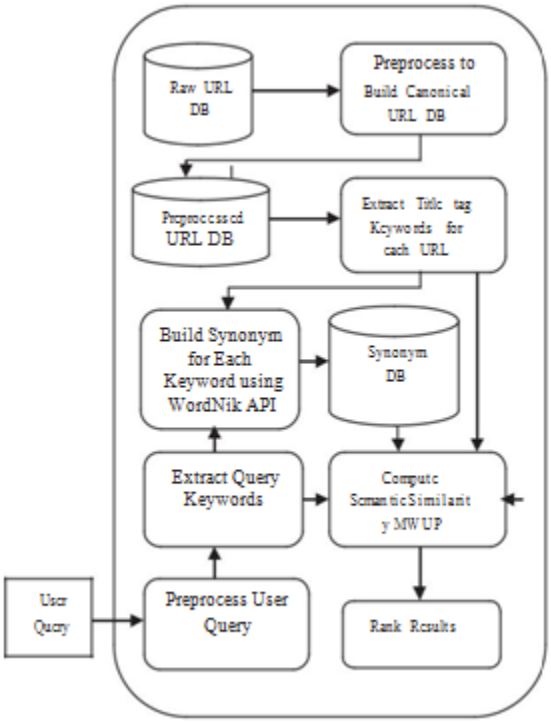
Web search engines searches the WWW to retrieve the web pages Pi that matches the given query *Qi*. Web search engines produces the huge list of search result web pages *SR={P1,P2,...Pn}* that matches the given query *Qi*. These search result web pages must be ranked based on the user query *Qi* and web page title tag text *Ti* semantic similarity, to get the web pages *Pi* that is more context semantic similar to the query *Qi*.

Objectives:

i. Preprocessing the URL database to build the canonical form for each URL.
ii. Extracting the URL Keywords and the Title tag text from each URL.
iii. Building the synonym database for each URL and Title tag Keyword.
iv. Preprocessing user query and extracting the query keywords and finding the synonyms.
v. Computing the semantic similarity between the user query and the web page title tag text.
vi. Ranking the web pages based on the semantic similarity.

ISSN 2348 – 8034
Impact Factor- 5.070

IV. ARCHITECTURE AND MODELING



General architecture of the proposed work is as depicted in Fig. 1. The raw URLs are preprocessed and represented in the canonical form in the preprocessed URL database. Title tag string from each URL is extracted, preprocessed and synonyms are extracted for each title tag keyword using WordNik API and stored in synonym database. User query string is preprocessed to remove the stop words and stemming is performed. Synonyms are extracted for each query keywords and synonym database is updated. The preprocessed URL title tag string semantic similarity with the preprocessed user query is computed using the MWUP and the results are ranked.

V. PROPOSED METHODOLOGY

The proposed methodology consists of three major components.

A. Representing URL in Canonical Form:

URL is the major means to identify the resource uniquely in the WWW. Representing all the URLs, those maps to the same resource in the canonical form, is highly beneficial to filter the duplicate URLs pointing to the same resource. Give two URLs *P* and *Q* with *N1* and *N2* characters respectively, the substitution process for forming

ISSN 2348 – 8034
Impact Factor- 5.070

the canonical form of *P* and *Q* is obtained using the matrix *M* of size *(N1 + 1) X (N2 + 1)* and the elements of the matrix M are filled using equation (1).

□□

ے

ā □ □ ā □□ (1)

. □ ā

ہ ā □

Table 1 show the sample URLs and their canonical form obtained using the substitution process as depicted in equation (1). Table 2 demonstrates the actual substitution process involved between the URLs with the ID 1 and 2 in table I.

Table I: Sample URLs and their Canonical form

URL-ID	URL
1	www.sbi.co.in/portal/web
2	https://sbi.co.in/portal/web
3	https://sbi.co.in/web
	Canonical form of URLs 1, 2 and 3
	https://www.sbi.co.in/portal/web/

Table II: Example of Substitution Process
Assume A↶https://, B↶sbi.co.in, C↶Portal, D↶web, E↶www.

		URL-2 (https://sbi.co.in/portal/web)						
			A	B	/	C	/	D
	0	0	0	0	0	0	0	0
E	0	←0	0	0	0	0	0	0
B	0	0	↖1	1	1	1	1	1
/	0	0	1	↖2	2	2	2	2
C	0	0	1	2	↖3	3	3	3
/	0	0	1	2	3	↖4	4	4
D	0	0	1	2	3	4	↖5	5

If *M(i,j) = M(i-1,j-1) +1* the no substitution is required. Diagonal arrow indicates no substitution is required. If *M(i,j) = M(i-1,j)* then substring of URL-2 must be inserted into URL-1 String. Left arrow indicates that substring in row must be substituted to string in column and if *M(i,j) = M(i,j-1)* then substring of URL-1 must be inserted into URL-2 string, up arrow indicates sub-string in column must be substituted to string in row. The net canonical from of URL-1, URL-2 and URL-3 is as shown in table I.

B. Web Page Title Tag Extraction and Pre-processing.
Web page title string is extracted using the HTTP.Get( ) method of visual studio. Stop words removal and stemming is performed on the title string and URL title database is built. Synonym database is built for every word of the title tag using WordNik API. User search query is preprocessed and synonyms are found for each word and the synonym database is updated.

C. Semantic Similarity Computation between User Query and Title Tag String
Semantic similarity is computed by constructing the Semantic Similarity Matrix *SSM(i, j)* of size *L1 X L2*. Where *L1* and *L2* are the number of words in user query *Q* and Title string *T*. The elements of *SSM(i,j)* are filled by computing the semantic similarity between the user query word *Q(Wi)* and title tag string word *T(Wj)* using the equation (2).

(2)

Where,

SSW[Q(Wi), T(Wj)] is the semantic similarity between user query word *Q(Wi)* and title tag string word

T(Wj).

Q(Wi) is the i^{th} word of user query *Q*.
T(Wj) is the j^{th} word of title tag string *T*.
SC(Q(Wi)) is the synonym words count for the

word *Q(Wi)*.

SC(T(Wi)) is the synonym words count for the word *T(Wj)*.
C is the count of matching synonym words between synonyms for *Q(Wi)* and *T(Wi)*.

Table IV shows the semantic similarity computation between the user query *Q* = {Artifice training in India} and the URL title tag string *T* = {Arts and Culture education in India}. The raw user query and title tag string are preprocessed. Preprocessed user query *Q*= {Artifice training India} of length *L1* = 3 and *T* = {Arts culture education India} of length *L2* = 4. Table III shows the synonyms for the words of user query *Q* and title tag string *T*. Semantic Similarity between the *Q* and *T* represented as *SS(Q, T)* is computed using equation (3).

(3)

Table III: Synonyms

ISSN 2348 – 8034
Impact Factor- 5.070

User Query Q = {Artifice training India}	
Word	**Synonym**
Artifice	Handicraft, Trade, Workmanship
Training	Education, Discipline, Drill, Drilling, School, Manage, Making
India	India
Title tag String T = {Arts culture education India}	
Word	**Synonym**
Arts	Duplicity, Artifice, Trade, Profession, Literature, Calling, Readiness, Contrivance, Dexterity, Business
Culture	Civilization, Cultivate, education
Education	Training, Breeding, Teaching, Instruction
India	India

Table IV: Semantic Similarity Computation

	Q = {Artifice training India}			
T = {Arts		**Artifice**	**Training**	**India**
	Arts	0.15	0	0
	Culture	0	0.2	0
	Education	0	0.16	0
	India	0	0	1
SS(Q, T) = 1.51				

VI. ALGORITHMS

This Section presents the major algorithm involved in the proposed research work. Table V shows the algorithm for building the canonical form of the URL and Table VI presents the algorithm for computing the semantic similarity between the user query and title tag string.

ISSN 2348 – 8034
Impact Factor- 5.070

Table V: Algorithm for Building Canonical URL

```
Input: P, Q // Raw URLs
Output: Canonical Form URL //Normalized
URL Variables:
i, j, A, B, C Integers //Looping Variables P, Q
String // For Storing raw URLs N1, N2 Integers //
Lengths of P and Q M(N1+1,N2+1) Matrix //
Substitution Matrix

Steps:
For i = 0 to N1+1
M(i,0) = 0 // Initialize first row to zeros
End for
    For j = 0 to N2+1
      M(0,j) = 0 // Initialize first column to
    zeros End for
      /* Split P and Q using 'https://' or 'http://'
      and '/' as the delimiters */
    Tokens(P) = Split (P into Tokens)
    Tokens (Q) = Split (Q into Tokens)
For i = 1 to Length(Tokens(P))
      For j = 1 to Length(Tokens(Q))
        If Token(Pi) = Token(Qj) then
          A= M(i-1,j-1) + 1
        End if
        B = M(i-1, j)
        C = M(i, j-1)
        M(i, j) = Max{A, B, C}
      End for
End for
For i = Length(Tokens(P)) to 1
      For j = Length(Tokens(Q)) to 1
        If M(i, j) = M(i-1, j-1) + 1
         Skip //No Substitution
        End if
        If M(i, j) = M(i, j-1)
         Append Token(Qj) in P
        End if
        If M(i, j) = M(i-1, j)
         Append Token(Pi) in Q
        End if
      End for
End for
```

Table VI: Algorithm for Computing Semantic Similarity

```
Input      :User Query Q, Title Tag String T
Output     :Semantic Similarity Score Matrix
             SSM[ ]
Variables  :
i, j, K, L Integer // for looping
L1, L2 Integer // Storing lengths of Q and T
Q(W), T(W) List // Stores words of Q and T
SSM[L1, L2] Matrix // Stores the scores

Steps:
For i = 0 to Length(Q(W))
      For j = 0 to Length(T(W))
          Symlist1 = Synonym(Q(Wi))
          Symlist2 = Synonym(T(Wj))
        For K = 0 to Length(Symlist1)
          For L = 0 to Length(Symlist2)
            If Symlist1K= Symlist2L
              Symcount = Symcount + 1
            End if
          End for
        End for
SSM[i,j] =

      End For
End For
```

ISSN 2348 – 8034
Impact Factor- 5.070

VII. EXPERIMENTAL SETUP AND PERFORMANCE EVALUATION

The proposed methodology is implemented using visual studio VB.NET. To evaluate the performance of the proposed research work URLs from the various categories like education, sports, media etc are collected. The URL database size from all the categories combined is 4,67,074 URLs. Preprocessing to remove all the URLs other than .html the number reduced to 4,00,100. Applying canonical form to URLs the size further reduced to 2,80,000 records. For better accuracy the user query must contain minimum of three to five distinct keywords after stop word removal and stemming. The first twenty highest score URLs in the Semantic Similarity Matrix are selected and displayed to the user. The efficiency of the proposed work is compared with the existing works and the performance of the proposed work is found better. The graph shown in Fig. 2 shows the precision comparison between the proposed approach (MWUP) and the Semantic Relations Using Analytical Hierarchy Process (SRAHP) using the equation (4).

(4)

Where,

NWc Number of Correct Web Pages Returned NWt Total Number of Web Pages Returned

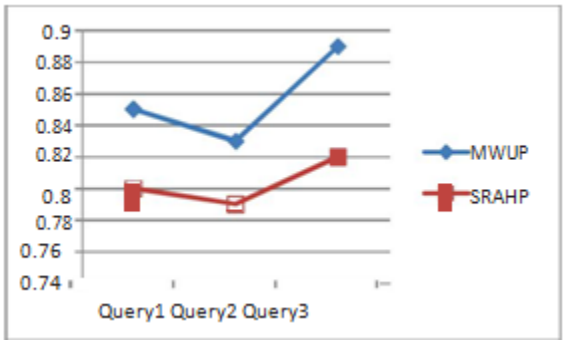


Fig. 2 Precision Comparison

Fig. 3 shows the Recall comparison between the proposed work (MWUP) and SRAHP using equation (5). Fig. 4 shows the False-Measure comparison using equation (6).

(5)

Where,

ISSN 2348 – 8034
Impact Factor- 5.070

NWr☞ is the total number of related web pages, but they are not necessarily correct pages.

□□ □כ כ□ (6)
□ □□ □□□□□ □□□□

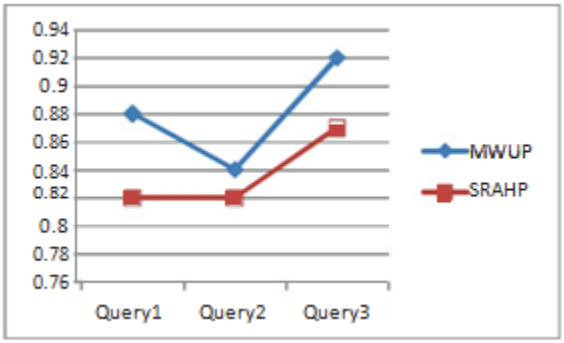


Fig. 3: Recall Comparison

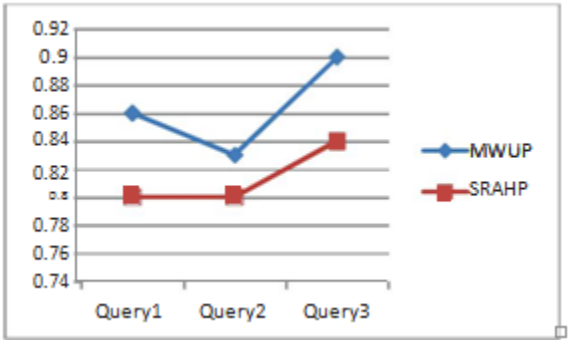


Fig. 4: F-Measure Comparison

VIII. CONCLUSION

The proposed work ranks the web pages based on semantic similarity using Modified Wu-Palmer Metric. The graphs shows performance of the proposed work is better than the existing work. In this work a new algorithm for semantic similarity measure is proposed which is based on the WordNik API and an algorithm for building the canonical URL.

REFERENCES

1. *JuhiAgarwal, Nishkarsh Sharma, Pratik Kumar, VisheshParshav and R H Goudar, "Ranking of Searched Documents Using Semantic Technology", In ElsevierInternational Conference on Design and Manufacturing, IconDM 2013.*
2. *S Ramana Murthy and G Anuradha, "Implementation of Page Ranking Using Genetic Algorithm", In Proceedings of Seventh IRF International Conference, 12th October 2014, Goa, India, ISBN: 978-93-84209-577-5.*
3. *DebajyothiMukhopadhyay, PradiptaBiswas and Young Chou Kim, "A Syntactic Classification based Web Page Ranking Algorithm", In Sixth International Workshop on MSPT Proceedings (MSPT-2006).*
4. *George T Satsaronis, IraklisVarlamis&KjetilNorvag, "SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs", In Proceedings of the Twenty Third International Conference on Computational Linguistics (Coling 2010) Pages 1074-1082, Beijing, August-2010.*

5. *Marius-Gabriel Gutu, TraianRebedea and Stefan Trausan-Matu, "A Comparison of Semantic Similarity Techniques for a Corpus of CSCL Chats", In IEEE Fourteenth RoEduNet International Conference – Networking in Education and Research (RoEduNet NER)-2015, Craiova, Romania.*
6. *Dr. Daya Gupta and Devika Singh, "User Preference Based Page Ranking Algorithm", In IEEE International Conference on Computing, Communication and Automation (ICCCA-2016).*
7. *B Prasanthi, Suresh Pabbaju and D Vasumathi, "Specific Query Semantic Signatures in Web Page Re-ranked Image Retrieval", In IEEE International Conference on Computational Intelligence and Computing Research-2016.*
8. *RekhaSinghal and SaurabhRanjanSrivastava, "Enhancing the Page Ranking for Search Engine Optimization Based on Weightage of In-Linked Web pages", In IEEE International Conference on Recent Advances in Innovations in Engineering (ICRAIE) Dec 23rd – 25th -2016, Jaipur, India.*
9. *Yuan Ziqian, Zhang Wenhui, Fu Huijuan and Tuzhixiao, "A PageRank- Improved Ranking Algorithm Based on Cheating Similarity and Cheating Relevance", In IEEE/ACIS Sixteenth International Conference on Computer and Information Science (ICIS)-2017, Wuhan, China.*
10. *Ali I El. Dsouky, Hesham A. Ali and Rabab S Rashed, "Ranking Documents Based on the Semantic Relations Using Analytical Hierarchy Process", In International Journal of Advanced Computer Science and Applications (IJACSA) Vol. 7, Page No. 2, 2016.*
11. *JoeranBeel and BelaGipp, "Google Scholar's Ranking Algorithm: An Introductory Overview", In Proceedings of the 12th International Conference on Scientometircs and Informatics (ISSI), Vol 1, Page No. 230-241, Rio De Janeiro (Brazil)-2009, ISSN: 2175-1935.*