iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XII **Month of publication:** December 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

A Survey on Semantic Page Ranking Techniques

Janisa J. Colaco[1], Shubha Tiwari[2]
[1,2] *Faculty, Department of Information Technology, Fr. Conceicao Rodrigues College of Engineering, Mumbai*

Abstract: The traditional search engines retrieve both relevant and irrelevant data. This results in wastage of user's time. Since the traditional ranking techniques are in the level of keyword matching [8]. They do not consider the semantics behind the user's query. To overcome the drawbacks of traditional ranking techniques there is semantic page ranking. In this paper we focus on the analysis of various semantic page ranking techniques and their comparative survey with respect to some similar factors among them.

Keywords: semantic, link analysis, Page Rank, HITS, relevance, semantic similarity.

I. INTRODUCTION

The World Wide Web is a vast resource of information. And with increase in the availability of information it is becoming more and more difficult to retrieve the appropriate information. As most of the traditional ranking techniques deal with only keyword matching and they do not consider the semantics behind the query. So the traditional ranking techniques either do not retrieve relevant results to the users or the relevant results are generally given a lower rank [8]. This normally results in the wastage of user's time as the user has to go through a number of irrelevant results before they can access their desired results. In case of the traditional ranking techniques the factors which influence the rank of a page include the following [4]

A. Number of matched terms in the query.
B. Frequency of terms.
C. Location of terms.
D. On the basis of link analysis.

The drawbacks associated with all such techniques are that they include only keyword matching or term frequency [8]. So the semantically similar pages that are desirable are often not retrieved. And the results are less relevant to the user's query.

With the semantic ranking techniques, the drawbacks of the traditional techniques could be improved. Semantic ranking techniques use the semantics behind the query to produce highly relevant search results. Some of the semantic measures used by them include:

E. Semantic similarity [9
F. Semantic intensity [10]
G. Semantic relationships [6] and
H. Semantic relevance [6].

Their main goal is to deliver the information which is semantically relevant to the users query. Most of the drawbacks of the traditional techniques are improved through semantic ranking.

An example of the most common traditional ranking technique is PageRank algorithm which depends on the out-links form the page to calculate the rank values [4]

I. Page Rank Algorithm [4]

1) It assumes that if a page has a link to another page then it votes for that page.
2) Each in-link to a page raises its importance.

Another is the HITS algorithm which considers both the in-links as well as the out links from the page to calculate the rank values.

II. HITS ALGORITHM [4]

A. Important pages are obtained on the basis of calculated authority and hubs value
B. Authority – Page that is pointed by many hubs

All such traditional ranking techniques do not include any kind of semantic measures while computing the rank values. So the main problem is of ranking the documents according to its semantic relevance to the query [8]. The flow of the paper is organized as follows: In Section II description of the semantic page ranking techniques is given which includes three semantic page ranking techniques. Section III includes the comparative survey on those techniques. Section IV gives the conclusion and future scope.

III.SEMANTIC RANKING TECHNIQUES

The semantic ranking algorithms usually can be generally categorized into certain types:

A. Based on Content Analysis [1]

They are based on ranking the pages based on relevancy of the page content with the user query.

B. Based on Link Analysis [2]

They consider the relevancy of the hyperlinks among web pages. And they are generally independent of the user's query.

C. Based on Ranking Semantic Associations among Entities [3]

The relevancy of the relationship among entities is calculated and thereby used to determine the rank of the pages.

The paper deals with analysis of three semantic page ranking techniques as described in below subsections. Subsection A includes description on a technique based on content analysis. Subsection B includes description of a technique based on Link analysis, While Subsection C includes description on semantic association ranking technique. Although there are many more techniques available, the described ones are some of the most commonly used.

D. Ranking of searched documents using semantic technology[1]

In this paper the semantic results are first extracted using term frequency and synonyms and then the retrieved results are filtered based on user attention time. Word net is used for getting synonyms of the documents

E. Flow of the Ranking process

1) *User Profile [1]:* The first step is the user profile wherein the user has to register on the website. He is then given a passage for reading and when he starts reading he will click on the start button and after finishing reading he will click the stop button so that record the time taken by the user to read the passage. And the length of the given document (docLen) is calculated in total terms contained in that particular document. Based on which the words per minute of the user (wpm) can be calculated as:wpm= docLen/ time taken to read the passage (1)
2) *Search[1] :*In the searching step, the query is processed to get the individual query terms. And Links to all the documents containing the query terms is returned. Then the total term frequency for each document is calculated.

$t_f[i][j]$=freq(docs[i], terms[j]) (2)

totalFreq[i]= total Freq[i]+$t_f[i][j]$ (3)

Where $t_f[i][j]$ denoted the term frequency of terms j in document i. The documents are then sorted based on the decreasing order of term frequency.

3) *Filter Result[1]:* Input to this will be the sorted list of documents according to term frequency. And when the user clicks on a particular document the length of that document will be calculated and the time taken by the user to read that document is calculated.

F. The flow of the filtering algorithm is as follows [1]

1) User clicks on a document
2) Calculate the length of the document = docLen
3) Calculate threshold time t = docLen / wpm
4) t_{ci} = time spent by the user on document i
5) If t_{ci} = t
6) Set user Interest as true
7) Increase the PR of Di by 1.
8) Move similar documents up in the docs[] List

While the user is reading a document if the threshold limit is reached it means that the user is interested in that document so the rank of that document will be increased by one unit, and all other documents which are similar to that document will be moved up in the document list.

G. Advantages

1) The algorithm combines term frequency along with synonyms.
2) Also considers user's interest for ranking.

H. Disadvantages

1) Every User has to register on the website.

2) User interest is based only on the time spent on the document.

I. ST Rank: A Site Rank Algorithm using Semantic Relevance and Time Frequency[2]

Almost all users tend to click those hyperlinks which have high semantic relevance with the content of the web page. So the similarity between anchor text and body of the web page has to be considered for computing the rank values [5].Also pages which update more frequently have a high probability of being visited. So the updating frequency of websites should also be considered while computing the rank values.

*1) Semantic Relevance [2][5]:*For computing the relevance between anchor text and web page the probability of websites has to be considered. The probability of visiting different websites is given by the transition probability matrix.The page transition probability is calculated as follows:

$$p_{ij} = \alpha \times s(i, j) + (1-\alpha) \times 1/n \; ; \quad L(i, j) \neq 0 \qquad (4)$$

$$= (1-\alpha) \times 1/n; \quad L(i, j)=0 \qquad (5)$$

Where s(i,j) denotes the jumping probability; L(i, j) is the number of hyperlinks between page i and page j; α is the damping factor, usually set to 0.85.

$$s(i,j) = \beta \times 1/ d_i + (1- \beta) \times Sim(at_j, con_i) \qquad (6)$$

d_i denotes the out-degree of page i; β is another damping factor set between 0.4 and 0.6. $Sim(at_j, con_i)$ is the similarity between the anchor text which points to page j and the body of page i. which is computed using the vector space model.

$$Sim (D_1, D_2) = Cos \theta = \frac{\sum_{k=1}^{N} W_{1k} W_{2k}}{\sqrt{(\sum_{k-1}^{n} W_{1k}^2)(\sum_{k-1}^{n} W_{2k}^2)}} \qquad (7)$$

In the vector space model each term t_k in a document D (t_1, …t_k..t_n) is given a weight W_k.

*2) Time Frequency[2]:*The updating frequency of web site is computed as follows:

$$Freq(s) = \partial \times N_a/D + (1- \partial) \times N_{na}/D \qquad (8)$$

Where N_a denotes the count of updated thematic pages in a website; N_{na} denotes the count of updated non-topic pages. D is the updating time interval for calculating updated pages. ∂ is a damping factor, usually set to 0.85.In cases where a number of non topic pages are been updated to cheat the updating frequency, the cheating can be prevented as the count of updated thematic and non-thematic pages are considered.

Overall Algorithm of STRank [2] [11]

Calculate the page transition probability p_{ij} . Then get the new n × n transition probability matrix Q(α)

(By rearranging the values of p_{ij} as every page will finally belong to some website the transition probability matrix is obtained) and n denotes the number of web pages.

Using matrix transforming and the theory of stochastic complement: Rank values are computed $\|\phi_i(\alpha)\|_1$

Those rank values are combined with the updating frequency of websites.

$$SR (s_i) = \lambda \times \|\phi_i(\alpha)\|_1 + (1- \lambda) \times Freq(s_i) \qquad (9)$$

λ is a damping factor set between 0.2 and 0.7. This gives the overall rank of the website (SR (s_i)).

3) Advantages

a) Considers updating frequency of websites.

b) Cheating of websites.

c) Quality of updated pages is considered.

4) Disadvantages Computing page rank for whole web graph is time consuming and costly.

J. Ranking Documents Semantically Using Ontological Relationships [3] [6]

Using user defined criteria it first identifies interesting semantic associations among entities in ontology.

And then the documents are ranked using the relevance measure of relationships

1) *Ranking Complex Relationships [3] [7]:* It assesses the overall relevance of associations among entities in the ontology. And users are enabled to browse through the ontology and mark their region of interest. So the Associations passing through these regions are considered relevant Certain ranking criteria are used for computing the relevance of semantic associations
2) *Ranking criteria [3] [7]*

a) Context definition (C_A) – identify important classes or properties
b) Subsumption (S_A) – classes lower in the hierarchy are considered to be more relevant.
c) Trust (T_A) – if the source of the entity can be trusted.
d) Rarity (R_A) – based on situations either the frequent association or rare associations can be relevant.
e) Popularity (P_A) – Either popular or non – popular associations can be considered relevant.
f) Association Length (L_A) – Sometimes short and direct connections are preferred whereas sometimes long associations are preferred.Also all these criteria's are assigned certain weight values (K): For E.g.: In some cases, Popularity might be given more weight than association length.

Then the overall weight of the association is computed as follows:-

Overall Ranking Criterion [3] [6]

$$W_A = K_1 \times C_A + K_2 \times S_A + K_3 \times T_A + K_4 \times R_A + K_5 \times P_A + K_6 \times L_A \quad (10)$$

3) *Ranking Documents using Relationships [3]:*A relevance measure is used for ranking the documents based on relationships. It determines how relevant an entity is with respect to the neighbouring entities in the document.
4) *Relevance Measure Algorithm [3]:* It takes as input the match-entity, other entities in the document and list of sequences with their importance levels which is assigned by the domain expert [6].
5) The total score is initialized to be zero.
6) Each sequence

a) Based on the importance levels it is identified which of the neighboring entities are important
b) The Neighboring entities are then added to either of these sets: *lowSet, mediumSet and highSet.*

7) By taking each entity in the "other entities set"

a) If it is in lowSet, then the corresponding low-score is added to the total score.
b) If it is in mediumSet, then the corresponding medium-score is added to the total score.
c) If it is in highSet, then the corresponding high-score is added to the total score.

The total score is computed based on how of the annotations of the document belong to a particular set. The total score determines the relevance of a entity with respect to the other entities occurring in the same document. The score of a document d is basically a function of three things: the entity e that does match the user input, the set A of other annotations in the document, and the ontology O

$$\text{Score } d = r(e, A, O)$$

8) *Advantages*

a) Considers user defined criteria for identifying important semantic associations.
b) Robust for multiple entity matches.

9) *Disadvantages*

a) Domain experts manually assign importance level.
b) Requires complete ontology with named entities.

IV. COMPARISON

TABLE I
COMPARATIVE ANALYSIS

Technique/ Parameters	1.Ranking documents using semantic technology	2. STRank	3.Ranking documents semantically using ontological relationships
Specific for	Hidden Web	Surface Web	Hidden Web
Mining Technique	Web content mining	Web structure mining	Web content and web structure mining

Input parameters	-Words per minute of the user -synonyms	-Anchor text -hyperlinks	Importance level of relationship sequences.
Semantic measure	Semantic similarity	Semantic relevance	Semantic relationships and relevance.
Advantages	-Combining term frequency and synonyms. -Considers user's interest also.	-Considers updating frequency of web sites. -Cheating of updating frequency is handled. -Quality of updated pages.	-Considers user defined criteria for ranking associations. -Robust for multiple entity match. -Interlinking not required.
Limitations	Every user has to register on the web site.	Computationally expensive -Newly added pages are not considered.	Importance level assigned manually by domain experts. -Requires complete ontology with named entities.
Query Dependency	More	Less	More
Correlation	High	Low	High
Relevancy	More(uses semantic similarity among terms)	Less than other two	More (uses relationships among entities)

V. CONCLUSION AND FUTURE SCOPE

A. To summarize the analysis of the techniques described above

1) Ranking documents using Semantic Technology [1] has benefit of combining user's interest with semantic approach of words.So it has better relevancy to user's interest. And there is high correlation with user's query.

2) STRank algorithm [2] has advantage of dealing with cheating of higher site updating frequency and the quality of pages. It provides high quality of ranked results. It has better quality, better crawling time and computationally efficient.

3) Ranking documents using ontological relationship [3] is robust for multiple entity matches. And it is highly correlated to user human ranking.

All these techniques employ some form of semantic measure while determining the ranking order. Every technique has its own pros and cons. Comparative survey analyses them with respect to some common characteristics. So they can be considered to be a complement to the traditional ranking techniques and the search process.

They thereby help in finding semantically relevant documents which the traditional techniques cannot find. And thus improve the overall ranking process.

In Future these techniques can be implemented and compared so that their performance can be measured in terms of precision, recall, accuracy in order to find the best one among them; which will provide us with an in-depth analysis of the techniques. Also a new approach may be designed by combining the common aspects of these approaches.

REFERENCES

[1] Juhi Agrawala, Nishkarsh Sharmab, Pratik Kumarc, Vishesh Parshavd, R H Goudare, "Ranking of Searched Documents using Semantic Technology" International Conference On Design and Manufacturing, Elsevier Procedia Engineering, 2013.

[2] Hongzhi Guo, Qingcai Chen, Xiaolong Wang, Zhiyong Wang, Yonghui Wu , "STRank: A SiteRank Algorithm using Semantic Relevance and Time Frequency" Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 2009.

[3] Boanerges Aleman-Meza, I. Budak Arpinar, Mustafa V. Nural and Amit P. Sheth , "Ranking Documents Semantically Using Ontological Relationships" IEEE Fourth International Conference on Semantic Computing, 2010.

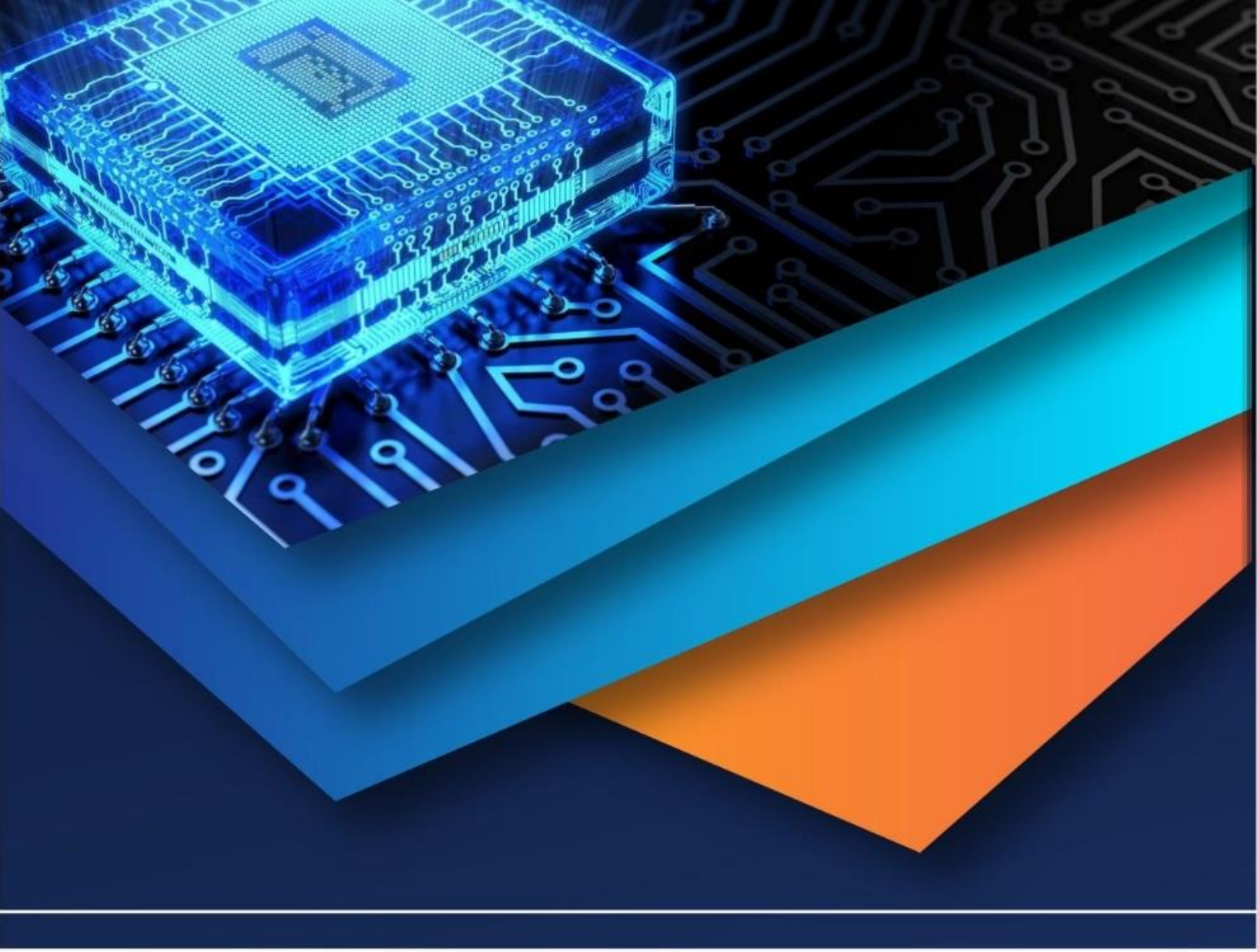
[4] Pooja Devi, Ashlesha Gupta, Ashutosh Dixit, **"**Comparative Study of HITS and PageRank Link based Ranking Algorithms" , International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 2, 2014.

[5] Guang Feng, Tieyan Liu, Ying Wang, et al. "AggregateRank: Bringing Order to Websites", In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, USA. Pages: 75 - 82. 2006.

[6] B. Aleman-Meza, "Ranking Documents based on Relevance of Semantic Relationships", PhD Dissertation, University of Georgia, 2007.

[7] Lakshmana Phaneendra Maguluri, M Vamsi Krishna, P S S Sridhar, "A novel approach for discovering relevant semantic associations on social Web mining", Conference on IT in Business, Industry and Government (CSIBIG), IEEE, 2014.

[8] Sanjay, Dharmender Kumar, "A Review Paper on Page Ranking Algorithms", International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), Vol. 4, Issue 6, 2015.
[9] Nisha Bansal, Dr. Paramjeet Singh, "Improved Web Page Ranking Algorithm Using Semantic Similarity and HITS algorithm", International Journal of Emerging Trends & Technology in Computer Science (IJETICS), Vol. 3, Issue 4, 2014.
[10] Nida Aslam, Irfan Ullah, Jonathan Loo, RoohUllah, Martin Loomes, "SemRank: ranking refinement strategy by using the semantic intensity", World Conference on Information Technology, Procedia Elseveir, 2011.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY