SEMANTIC WEB RANKING USING HITS METHODOLOGY

John Vaseekaran.S[1]
Research Scholar,Sathyabhama Institute of Science and Technology, India
vaseekaranjohn168@gmail.com
Dr. N. Srinivasan[2]
Professor,Department of Computer Science and Engineering,
Sathyabhama Institute of Science And Technology, India,
nsrinivasan.cse@sathyabama.ac.in

Abstract

Mining or searching the World Wide Web (WWW) seems to be a hard task as the search engines sometimes search or display the websites or pages not needed by the user. The content that are displayed may sometimes seems to be unrelated.

In this paper we would try to solve the major problems faced in HITS methodology like topic drift, Mutual reinforcing relation between host problematic and irrelevant authorities. We try to find a solution by using the Web Structure Mining Technology.

1. INTRODUCTION

The main aim is to discover the structural summary about the web site and web pages. There are two types of mining namely [3]

1. Web Content Mining
2. Web Structure Mining

Web Content Mining :

Basically, web content mining focuses on intra – document structure (within the document).

Web Structure Mining:

It gives an idea of structure in the web pages.

Web structure mining tries to focus on inter – document structure (within the web) that is, to discover the link structure of hyperlinks.

Web structure mining[8], one of categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection.

2. ALGORITHMS

The two Link based algorithms are[2],

- Page Rank
- HITS (Hyperlink Induced Topic Search)

PAGE RANK

The page rank algorithm is given below[4][7]

$$PageRank\ of\ site = \sum \frac{PageRank\ of\ inbound\ link}{Number\ of\ links\ on\ that\ page}$$

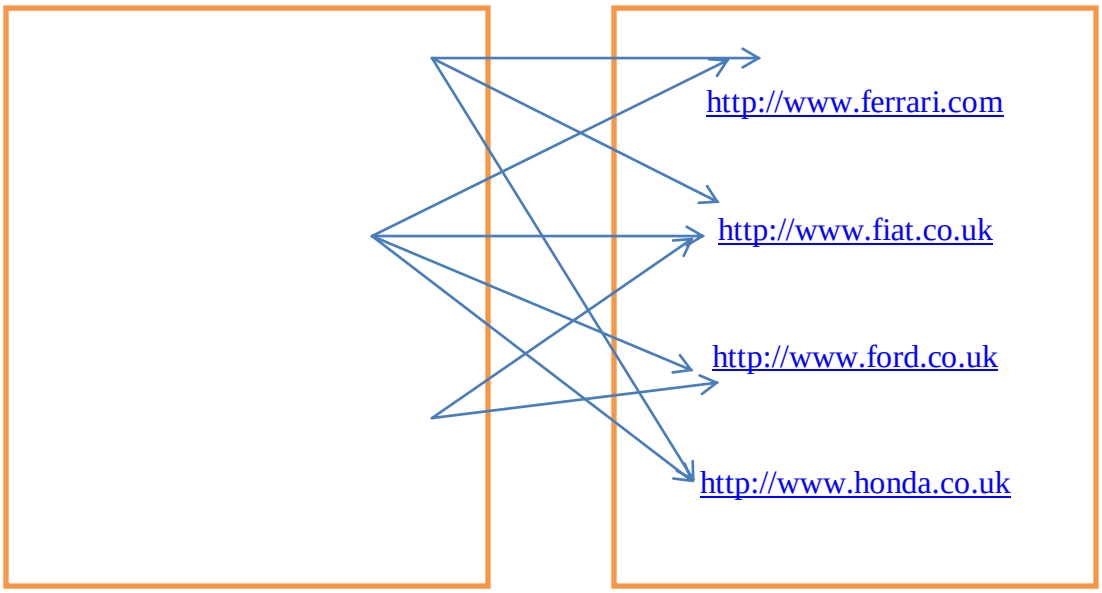
OR

$$PR(u) = (1-d) + d \times \sum \frac{PR(v)}{N(v)}$$

HITS ALGORITHM

In the HITS algorithm [5], the first step is to retrieve the most relevant pages to the search query. This set is called the *root set* and can be obtained by taking the top pages returned by a text-based search algorithm. A *base set* is generated by augmenting the root set with all the web pages that are linked from it and some of the pages that link to it.

The Basic idea is to identify a small sub graph of web and apply link analysis on this sub graph and to locate the authorities and hubs for the given query [3].



3. EXISTING PROBLEMS OF HITS

- Topic drift [2]
- Mutual reinforcing relation between host problematic
- Irrelevant authorities[6]

4. THE SEARCHING PROCESS

The searching process of a website is more complex than just compare the query against a list of documents and return the matches. For eg- car search – the best automobile maker in last 4 years.

5. THE DIFFERENCE IN SEARCHING

The list of Pages that we get back while algorithmically correct might be different . Here the given word in a given set of documents will not do intelligent rephrasing for us.

For Eg- Needed Pages have less rank and unwanted or not needed websites have more ranks or the words will not match.

The main problem is that the displayed webpages may be different or not needed by the user. The reason is that the top companies such as Hyundai, Toyota might not even use the term ”Automobile Makers” on their website. Instead they might use “ Car Manufacture”[2].

When this question is asked to a human being we want the humans to understand that Automobile means Car, Vehicle etc. But when the computer is asked to search, the searching is different. That is the HITS methodology might not use the words that are used by the Hundai or Toyota Company.

6. THE SOLUTION – WHAT TO BE DONE

It can be useful if the companies have a Dictionary such that for any query, they could figure out Synonyms– Equivalent meaning of Phrase. This might improve the Quality.

But we would be left with the initial problem of sorting the huge number of pages that are relevant to the different meanings of the query phase. That is the web pages in the web structure might be more and some pages might not even sometimes match with the needed word.

7. THE CONCLUSION

1. If trying to find the pages that contain the query words should be the starting point. The ranking algorithm must be based on Authoritative for a given query. Page “i” is called an Authority(Main Page) for a given query – “Automobile Maker” if it contains valuable information on the given topic or subject. For Eg – The official websites for car manufactures such as www.bmw.com, www.hyundai.usa.com[3], would be authorities for this search.

2. The hubs contain useful links towards the authoritative pages. In other words hubs points to the search engine in the right direction.

For Eg – In real life when we buy a car, we are more inclined to purchasing it from a certain dealer that your friend recommends.

Following the analogy, the authority in this case would be a car dealer and the hub would be the Friend.

REFERENCES

1. IUP Journal of Information Technology. Sep2012, Vol. 8 Issue 3, p64-72. 9p. 2 Diagrams, Author(s) : Deepa, Sankaranarayanan; Hariharan, Shanmugasundaram
2. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 4 (2016) pp 2552-2556 © Research India Publications. http://www.ripublication.com 2552 Analytical Implementation of Web Structure Mining Using Data Analysis in Educational Domain Dr. S. P. Victor Professor CS, St. Xaviers College, Tirunelveli, Tamil Nadu, India.
3. Mr. M. Xavier Rex Research Scholar, M. S. University, Tirunelveli, Tamil Nadu, India.
4. Musa, H. H., & Noureldien, N. A. (2018). Comparing the Ranking Performance of Page Rank Algorithm and Weighted Page Rank Algorithm. Advanced Science Letters, 24(1), 750–753. doi:10.1166/asl.2018.11807
5. Tian, X. (2014). Improvements of HITS Algorithm Based on Triadic Closure. Journal of Information and Computational Science, 11(6), 1861–1868. doi:10.12733/jics20103219
6. Bennett, M., Stone, J., & Zhang, C. (2006). A Scalable Parallel HITS Algorithm for Page Ranking. First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06)
7. Sen, T., Chaudhary, D. K., & Choudhury, T. (2017). Modified Page Rank Algorithm: Efficient Version of Simple Page Rank with Time, Navigation and Synonym Factor. 2017 3rd International Conference on Computational Intelligence and Networks (CINE). doi:10.1109/cine.2017.24
8. Wen, J.-R. (n.d.). Enhancing Web Search through Web Structure Mining. Encyclopedia of Data Warehousing and Mining. doi:10.4018/9781591405573.ch084