International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-03, Issue 02, Apr 2017

Implementation of Semantic and Synaptic Web Mining- An Integrated Web Mining Algorithm

[1]Shubham P. Sharma, [2]Krishna M. Gupta, [3]Akshay D. Pabale, [4]Prof. Deepti Vijay Chandran
[1,2,3]UG Student, [4]Professor, [1,2,3,4]Smt. Indira Gandhi College of Engineering, Navi Mumbai, Maharashtra, India.
[1]sharma.shubham7@gmail.com, [2]Krishna20021994@gmail.com, [3]akshaypabale36@gmail.com, [4]dvcnr@yahoo.co.in

Abstract — There is always an ever growing information and the various sources that contain them however the lack of or limited means for improving processing capabilities require a smart method to discriminate between the sources of information to get to the exact information that we seek. The paper presents an integrated web mining approach that uses the three web mining approaches: - Web usage mining, Web content mining and Web structure mining to rank web pages according to their scores under criteria of the approaches given above to find the relevance of a particular source (web site) to our needs. In our project we put each source (website) under the three criteria for generating scores, Content Based, Link Based and Usage based. The three scores then are used for evaluation and a final score is used to determine the actual score for a website. This final score then is compared with the scores of other websites to generate the ranking index where top most result indicates most relevant result.

Key Words - content based score, inverse document frequency, page rank based score, term frequency, usage based score, web mining.

1. INTRODUCTION

The World Wide Web ("WWW" or simply the "Web") is a global information medium which users can read and write via computers connected to the Internet. In the last few decades there has been tremendous growth of data due to which retrieval of useful information for the user has become difficult. Hence arrived the need to deploy new methods for improving web mining to get the useful result that not only checks what is available on the internet but also what user wants.

Web mining is the application of data mining techniques to derive information and patterns from the World Wide Web. [1].The web involves three types of data: data on the web (content), web log data (usage) and web structure data. [3] Web mining approach can be categorized into three- Web usage mining, Web content mining and Web structure mining. [5]

1.1Problem Statement

The paper deals with the need to improve web mining techniques and approaches. In this paper we propose a web mining approach that uses semantic synaptic web model. [6][7][8] In addition we make use of usage details to determine better search results by using search engines.[4]

1.2 Proposed System

The main aim of this project is to create an integrated web mining approach where we use semantic-synaptic web mining model. Our system uses content based, interlinking between web pages and usage based scores to get appropriate information for web searches

Advantage of Proposed System

1. In this system search results that corresponding to a user's query is retrieve using the three basic approaches.
2. The first part deals with content present on the particular web page to determine its score.
3. Second part deals with the links the page retrieved for extracting information has with the others thus checking the interconnection with other pages and deriving some connection between similar pages thus checking the user's diversified interest in similar and related topics.
4. The third part deals with user's behavior as per its search and access or usage history to get certain idea about how to gather information example his/her favorite site.

2. PLANNING &FORMULATION

2.1Architecture

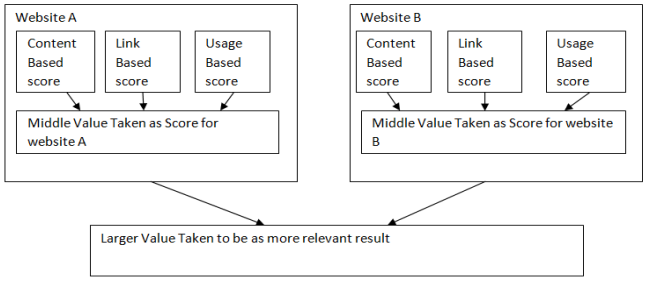

Fig 2.1 Main Architecture of integrated web mining technique.

The project works in three parallel stages.

The first stage is identifying the user which through login where a user can use the system as an individual or as a group of individuals with no certain converging interest.
The second stage is extracting information (web pages) from web server databases.

The third stage is implementation of three different algorithms in parallel to gain three different score content based, link based score and usage based score which are then compared for each website to get a certain final score for a website.

This score is then compared with that of others to gain the final rank. This ranking is dynamic in nature since the score based on the three criteria content based, link based and usage based are subject to changes in the future. Hence every user can get different personalised results as per its usage and identity.

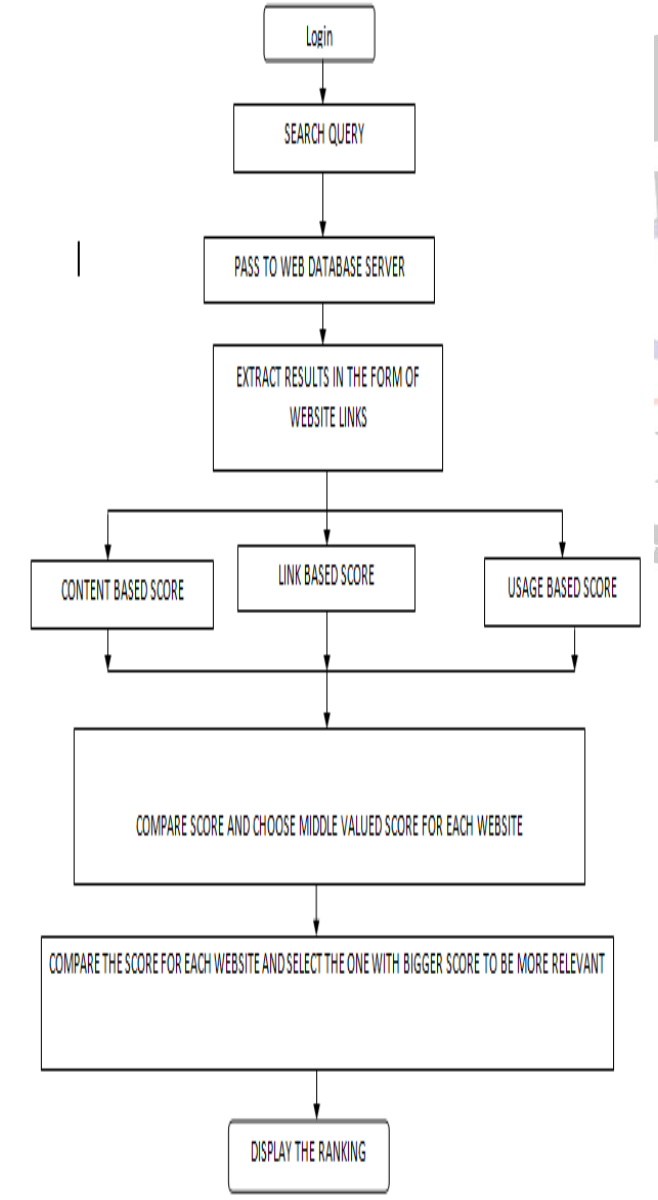

Fig 2.2 Flow of integrated web mining technique

2.2Modules in the technique

The three modules used here are:
1. Content Based Score for semantic analysis.
2. Link Based Score for synaptic based analysis.
3. Usage based score

Content based score involves searching for terms(query) in metadata(web site description and title).

Content based score used TF-IDF (Term Frequency and Inverse Document Frequency) where Tf uses word count (query) while Idf calculates term density in the document
Link based Score using pagerank for analysis over hyperlink structure.
The Pagerank uses hyperlinks network among the pages to get the pagerank score.

Usage based score involves using user's search history to understand basic usage.

This usage is divided as Personal and Group based usage where one is based on individual specific based usage and other calculates overall usage on anonymity. While first one presents the personalized view of Web to the user while the other indicates overall usage pattern on the machine.

III. ALGORITHM DEVELOPMENT

Page Rank

PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."

A hyperlink to a page counts as a vote of support. The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it ("incoming links"). A page that is linked by many pages with high rank receives a high rank itself. If there are no links to a web page there is no support of this specific page. The Google Toolbar Page Rank goes from 0 to 10. It seems to be a logarithmic scale. The exact details of this scale are unknown.

PageRank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page.

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + +PR(Tn)/C(Tn))

Where:
- PR(A) is the PageRank of page A,
- PR(Ti) is the PageRank of pages Ti which link to page A,

C(Ti) is the number of outbound links on page Ti d is a damping factor which can be set between 0 and 1.

Example:

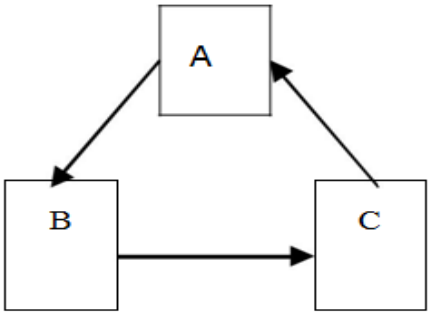

Fig 3.1 Simple connection of links between websites.

The number of web pages $N = 3$
The damping parameter $d = 0.7$
The damping parameter d = 0.7

PR(A) = (1 – d) × (1 / N) + d × (PR(C) / 1)
PR(B) = (1 – d) × (1 / N) + d × (PR(A) / 1)
PR(C) = (1 – d) × (1 / N) + d × (PR(B) / 1)
So
$PR(A) = 0.1 + 0.7 \times PR(C)$
$PR(B) = 0.1 + 0.7 \times PR(A)$
$PR(C) = 0.1 + 0.7 \times PR(B)$

By solving the above system of linear equations, we get
$PR(A) = 1/3 = 0.33$
$PR(B) = 1/3 = 0.33$
$PR(C) = 1/3 = 0.33$

Term Frequency-Inverse Document Frequency

Tf-idf stands for *term frequency-inverse document frequency*, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

One of the simplest ranking functions is computed by summing the tf-idf for each query term; many more sophisticated ranking functions are variants of this simple model.

Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

How to Compute

Typically, the tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones.

Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

Step 1: TF (Term Frequency)
Step 2: Check no of times terms repeated in document.
Step 3: Check total no of words in the document.
Step 4: IDF (Inverse document frequency)

$$TF(term) = \frac{(\text{no of times terms required repeated in the document})}{(\text{total no of words in the document})}$$

Step 5: Total number of documents in database (m)
Step 6: Total number of documents with required term (n)
Step 7: IDF (t) = Log_e (m/n)
Step 8: TF-IDF weight = TF(t)×IDF(t)

Example:
Consider a document containing 100 words wherein the word *bat* appears 3 times. The term frequency (i.e., tf) for *bat* is then (3 / 100) = 0.03. Now, assume we have 10 million documents and the word *cat* appears in one thousands of these. Then, the inverse document frequency (i.e., idf) is calculated as log(10,000,000 / 1,000) = 4. Thus, the Tf-idf weight is the product of these quantities: 0.03 * 4 = 0.12.

IV. RESULTS AND DISCUSSION

The given screenshot displays the outcome of the search query .The result generated is the search result based on the three scores

- Content Based score
- Pagerank Based score
- Usage Based score

Fig 5.1 Indexing of scores based on scores

The second screenshot (red rectangle portion) displays the score generated for each web page that has been used for indexing the web pages.

Fig 5.2 Scores of websites shown (red markings for indication purposes only)

In given figure (5.3), score highlights the working and dynamic nature in which rank adjust themselves as per usage here the first and the second link are not initially ever visited by the user hence the ranking does not make much use of usage based score and ranks are given as such.

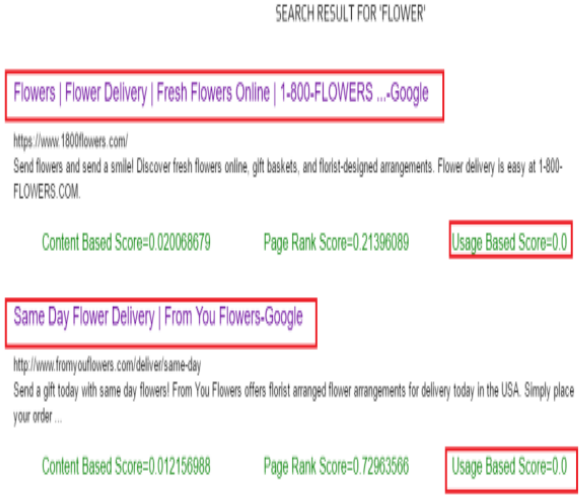

Fig 5.3 Scores of websites and comparison with initial usage score. (Red markings for indication purposes only)

Now in the next snapshot, the second ranked link was visited by the user leading to its increased score and hence the rank of the page also changed improving its ranking and placing it at first position from the second.

Fig 5.4 Scores of websites and comparison with usage score incremented.(Red markings for indication purposes only)

V. CONCLUSION

The paper presents the model where the websites are classified on basis of their relevancy to the user which is not only based on content and internal structure of links between web pages but also the user's behavior as an personalized individual as well as group of anonymous user.

In the project results are used to generate not only the relevance score but also to create a feel for personalized web for the user. Since the requirement of each user is different from the other. The project allows multiple scores for website that is based on the usage and that too by different user in a sense that two users searching for the same thing but in different detail or focus can end up having different rank index due to difference in usage pattern of results also it provides them more control over what, how and from where they like to get information rather than depending on the computerized scores and search activity of other people.

Apart from the focused searches the project also has provided option for group search that allows search results that are based on group activity and can provide to the user an insight into general searching patterns. Thus this project classifies websites on the relevance of their information both in terms of technicalities such as content and link structures as well as factoring the user's interests.

REFERENCES

[1] Oren Etzioni. *"The world wide web: Quagmire or gold mine"*.Communications of the ACM, Vol.39(11), Pp.65-68 (1996)

[2] Ding, c., Chi, c.-H., and Luo, T., "*An Improved Usage-Based Ranking*", W AIM '02: Proceedings of the 3rd International Conference on Advances in Web-Age Information Management, London, UK: Springer-Verlag, p.p. 346- 353, 2002. .

[3] S. K. Madria, S. S. Bhow mick, E. P. Lim etal. "*Research issues in web data mining".* In proceeding conference, Dawak,99,Pp.303-3012,(1999).

[4] R. Cooley. "*The web usage mining: Discovery and Application of Interesting patterns from web data*", Phd thesis,Dept.of computer science, university of Minnesota, May 2000.

[5] M. Spiliopoulou, *"Data mining for the web"*. In proceeding of principles of data mining and knowledge Discovery,Third European conference, PKDD 99, Pp.588-589.

[6] F. Sebastini, "*Machine Learning in Automated Text Categorization.Tech."* report B4-31, Istituto di Elaborazione dell'Informatione,Consiglio Nazionale delle Ricerche, pisa, (1999).

[7] S. Chakarabarti, "*Data Mining for Hypertext: A Tutorial Survey*",ACM SIGKDD Explorations, Vol. 1, no. 2, pp. 1-11, 2000.

[8] J. Fumkranz, "*Web Structure Mining: Exploiting the graph Structure of the World Wide Web*", Osterreichische Gesellschaft fur Artificial Intelligence (OGAI), vol. 21, no.2, Pp. 17-26 (2002).

[9] Hofgesang, P., "*Relevance of Time Spent on Web Pages, in 'Workshop on Web Mining and Web Usage Analysis*", the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), 2006.

[10] Kellar, M., Watters, c., Duffy, 1., and Shepherd, M., "*Eflect of Task on Time Spent Reading as an Implicit Measure of Interest*". ASIST 2004 Annual Meeting, p.p. 168- 175,2004.

[11] Kritikopoulos A., Sideri M. and Varlamis I. , "*Success index: measuring the efficiency of search engines using implicit user feedback*", in: Proceedings of the 11th Pan-Hellenic Conference on Informatics, Special Session on Web Search and Mining, 2007.

[12] Liu Y., Liu T., Gao B., Ma Z. and Li H., "*A framework to compute page importance based on user behaviors*", Information Retrieval, Vol. 13, pp. 22-45,2010.

[13] Jain R. and Dr. Purohit, "*Page Ranking Algorithms for Web Mining*, International Journal of Computer Applications", VoU3, No. 5, pp. 22-25,2011.

[14] Rekha c., Usharani J. and Iyakutti K., "*Improving the Information Retrieval System through Eflective Evaluation of Web Page in Client*.

