Analysis of Link Algorithms for Web Mining

Monica Sehgal
<monica.sehgal326@gmail.com>

Abstract- As the use of Web is increasing more day by day, the web users get easily lost in the web's rich hyper structure. The main aim of the owner of the website is to provide the relevant information to the users to fulfill their needs. Web mining technique is used to categorize users and pages by analyzing users behavior, the content of pages and order of URLs accessed. Web Structure Mining plays an important role in this approach. In this paper we discuss and compare the commonly used algorithms i.e. PageRank, Weighted PageRank and HITS.

Index Terms- Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, PageRank, Weighted PageRank and HITS.

I. Introduction

The World Wide Web (WWW) is rapidly growing on all aspects and is a massive, explosive, diverse, dynamic and mostly unstructured repository of data. Till now, WWW is the huge information repository referenced for Knowledge. Generally, user faces a lot of challenges in the Web:

1) Huge amount of information on the web
2) Web information coverage is very wide and diverse
3) All types information/data must exist on the web
4) Much of web information is semi structured
5) Much of the web information is linked
6) Availability of redundant web information.
7) The web is noisy
8) The web is also about services
9) The web is dynamic
10) The web is a virtual Society.

This paper is organized as follows: Web Mining is introduced in Section II. The Taxonomy of Web Miningnamely Web Content Mining, Web Structure Mining, Web usage Mining are discussed in Section III. Section IV describes the various Link analysis algorithms. Section IV(A) defines Page Rank Algorithm, IV(B) defines Weighted Page Rank Algorithm and IV(C) defines Hyperlink Induced Topic Search Algorithm . Section V provides the comparison of various Link Analysis Algorithms.

II. Web Mining

Web Mining is the use of data mining techniques to automatically discover and extract information from Web documents and services.

Web Mining Process : The Complete process of extracting knowledge from Web data [5][8] is as follows in Figure-1:

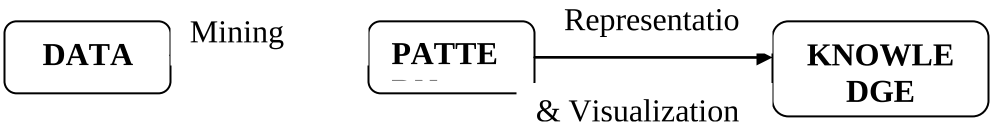

Figure 1: Web Mining Process

Web Mining can be decomposed into several tasks, namely:

1. **Resource Finding** refers to the task of getting/ reading intended web credentials.
2. **Information selection and pre-processing refers to** Robotically selecting and pre-processing definite from information retrieved Web resources.
3. **Generalization** refers to Robotically ascertains certain general patterns at individual or multiple Web Situates .
4. **Analysis**: Mined Pattern Rationale and interpretation.

III. Web Mining Categories

Web Mining areas according to the Web data used as input as follows:

Web Content Mining (WCM) is responsible for exploring the proper and relevant information from the contents of web. It focuses mainly inner document level.

Web Structure Mining (WSM) is the process by which we discover the model of link structure of the web pages. We catalog the links; generate information like the similarity and relations among them by taking advantage of Hyperlink topology. Its Goal is to generate structured summary about the website and web page.

Web usage Mining (WUM) is responsible for recording the user profile and user behavior inside the log file of the web.

Table 1 gives an overview of above Web Mining categories.

	Web Mining			
	Web Content Mining		**Web Structure Mining**	**Web usage Mining**
	IR View	**DB View**		
View of Data	- UnStructured - Structured	- Semi – Structured - Web Site as DB	- Link Structure	- Interactivity
Main Data	- Text Documents - HyperText Documents	- Hypertext Documents	- Link Structure	- Server Logs - Browser Logs
Representation	- Bag of Words, n-gram Terms, - Phrases, Concepts or ontology - Relational	- Edge labeled Graph, - Relational	- Graph	- Relational Table - Graph
Method	- Machine Learning - Statistical (including NLP)	- Proprietary Algorithms - Association Rules	- Proprietary algorithms	- Machine learning - Statistical - Association rules
Application Categories	- Categorization - Clustering - Finding extract rules - Finding patterns in Text	- Finding frequent Sub Structures - Web Site Schema Discovery	- Categorization - Clustering	- Site Construction - Adaptation and management - Marketing - User Modeling

IV. Link Analysis Algorithms

Web mining techniques provides the additional information through hyperlinks where different documents are connected.[2] We can view the web as a directed labeled graph whose nodes are the documents or pages and edges are the hyperlinks between them. This directed graph structure is known as web graph. There are number of algorithms proposed based on link analysis. Three important algorithms Page rank[5], Weighted Page rank[6] and HITS[7] are discussed below:

A. Page Rank

Page Rank, an algorithm developed by Brin and Page at Stanford University, is a numeric value that represents how important a page is on the web. Page Rank is the method used by Google for measuring a page's "Significance". After considering all other factors like Title Tag and Keywords , Google uses Page rank to adjust results so that more "Significant" pages move up in the results page of a user's search result display.

The Page rank value for a page is calculated based on the number of backlinks to a page. Page Rank is displayed if you've installed the Google toolbar (http://toolbar.google.com/). And its rank only goes from 0 – 10 and seems to be like a logarithmic scale:

Toolbar PageRank (Log base 10)	**Real PageRank**
0	0 - 100
1	100 – 1,000
2	1,000 – 10,000
3	10,000 – 100,000
4	And so on….

Following are some of the terms used:

(1) **PageRank (PR)** : the actual, real, page rank for each page as calculated by Google.

(2) **TOOLBAR PR** : The Page rank displayed in the Google toolbar in your browser ranges from 0 to 10.

(3) **BACKLINK** : If page A links out to page B, then page B is said to have a "back link" from page A.

We can't know the exact details of the scale because the maximum PR of all pages on the web changes every month when Google does its re-indexing! If we presume the scale is logarithmic then Google could simply give the highest actual PR page a toolbar of 10 and scale the rest appropriately.

The another definition given by Google is as follows: We assume page A has pages T1…Tn which points to it. The parameter d (damping factor), can be set between 0 and 1. But usually set to 0.85. C(A) refers to the number of links going out of page A. The Page rank of a page A is given as follows:

$$PR(A) = (1 – d) + d\ (PR(T1)/C(T1) + … + PR(Tn)/C(Tn))$$

Note that Page Ranks form a probability distribution over web pages, so the sum of all web pages' Page Ranks will be one. Page rank or PR(A) has the following sections:

1. **PR (Tn)** – Each page has a notion of its own self-importance. That's "PR(T1)" for the first page in the web all the way up to "PR(Tn)" for the last page.
2. **C (Tn)** – Each page spreads its vote out evenly amongst all of it's outgoing links. The count , or the number , of outgoing links for page 1 is "C (T1)" , "C (Tn)" for page n, and so on for all pages.
3. **PR (Tn)/C (Tn)** – so if out page (page A) has a backlink from page "n" the share of the vote page A will get is "PR(Tn)/C(Tn)".
4. **d(**… - All these fractions of votes are added together but, to stop the other pages having too much influence, this total vote is "drown" by multiplying it by 0.85 (the factor "d").

B. Weighted Page Rank

The more popular web pages are the more linkages that other web pages tend to have to them or are linked to by them. The proposed extended Page Rank algorithm – a Weighted Page Rank Algorithm[10] – assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its out link pages. Each out link page gets a value proportional to its popularity (its number of in links and out links). The popularity from the number of in links and out links is recorded as $W^{in}(v,u)$ and $W^{out}(v,u)$, *respectively*. $W^{in}(v,u)$ is the weight of link(v,u) calculated based on the number of in links of page u and the number of in links of all reference pages of page v.

$$W^{in}(v,u) = \frac{Iu}{\sum p\in R(v) Ip}$$

Where Iu and Ip represent the number of in links of page u and page p, respectively. $R(v)$ denotes the reference page list of page v.

$W^{out}(v,u)$ is the weight of link(v,u) calculated based on the number of out links of page u and the number of out links of all reference pages of page v.

$$W^{out}(v,u) = \frac{Ou}{\sum p \in R(v) Op}$$

Where Iu and Ip represent the number of out links of page u and page p, respectively. $R(v)$ denotes the reference page list of page v. Figure-1 shows an example of some links of a hypothetical website.

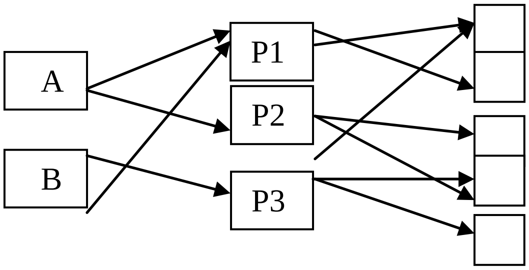

Figure 1: Links of a Web Site

PageRank VS Weighted PageRank

In order to compare the WPR with the PageRank, the resultant pages of a query are categorized into four categories based on their relevancy to the given query. [8] They are

1. Very Relevant Pages (VR): These are the pages that contain very important information related to a given query.
2. Relevant Pages (R): These Pages are relevant but not having important information about a given query.
3. Weakly relevant Pages (WR): These Pages may have the query Keywords but they do not have the relevant information.
4. Irrelevant Pages (IR) : These Pages are not having any relevant information and query Keywords.

The PageRank and WPR algorithms both provide ranked pages in the sorting order to users based on the given query. So, in the resultant list, the number of relevant pages and their order are very important for users. Relevance Rule is used to calculate the relevancy value of each page in the list of pages. That makes WPR different from PageRank.

Relevancy Rule: The Relevancy of page to a given query depends on its category and its position in the page –list. The larger the relevancy value, the better is the result.

$$K = \sum_{i \in R(p)} (n - i) * W i$$

Where,
i = ith page in the result page –list
R(p), n = the first n pages chosen from the list R(p)
Wi = weight of ith page as given below

$$W\ i = (v1, v2, v3, v4)$$

Where, v1,v2,v3 and v4 are the values assigned to a page if the page is VR,R,WR,IR respectively. These values are always v1>v2>v3>v4. Experimental studies show that WPR produces larger relevancy values than the PageRank.

C. HITS (Hyperlink Induced Topic Search)

Kleinberg developed a WSM based algorithm named Hyperlink Induced Topic Search (HITS) in 1988 which identifies two different forms of web pages called hubs and authorities. Authorities are pages having important contents. Hubs are pages that act as resource lists, guiding users to authorities. Thus, a good hub page for a subject points to many authoritative pages on that content, and a good authority page is pointed by many fine hub pages on the same subject.

Kleinberg states that a page may be a good hub and a good authority at the same time [8, 9]. This spherical relationship leads to the definition of an iterative algorithm called HITS.

The HITS algorithm treats WWW as a directed graph G (V, E), where V is a set of vertices representing pages and E is a set of edges that match up to links. Figure 2 shows the hubs and authorities in web [2].

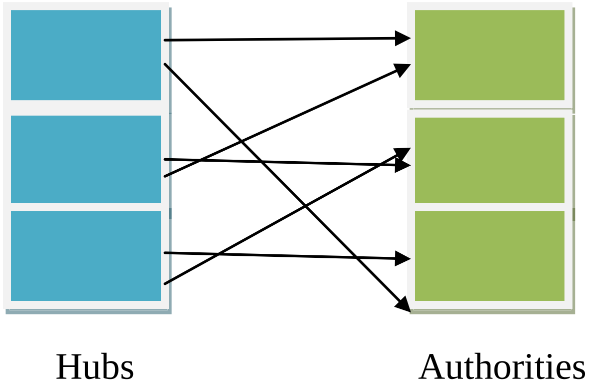

Figure 2: Hubs and Authorities

It has two steps,

1. Sampling Step – In this step a set of relevant pages for the given query are collected.
2. Iterative Step – In this step Hubs and Authorities are found using the output of sampling step.

In this HITS Algorithm, the weights of the Hub (Hp) and the weights of the Authority (Ap) are calculated using following algorithm.

HITS Algorithm

```
1. Initialize all weights to 1
2. Repeat Until the weights converge:
3. For every hub p ϵ H
4.
   $H\ p = \sum_{q \in I(p)} Aq$
5. For every authority p ϵ A
6.
   $A\ p = \sum_{q \in B(p)} Hq$
7. Normalize
```

Where Hq is Hub Score of a page, Aq is authority score of a page, I(p) is set of reference pages pf page p and B(p) is set of referrer pages pf page p. The authority weight of a page is proportional to the sum of hub weights of pages that link to it.

Similarly a hub of a page is proportional to the sum of authority weights of pages that it links to.

Following are some constraints of HITS algorithm

- Hubs and authorities: It is not easy to distinguish between hubs and authorities because many sites are hubs as well as authorities.
- Topic drift: Sometime HITS may not produce the most relevant documents to the user queries because of equivalent weights.
- Automatically generated links: HITS gives equal importance for automatically generated links which may not have relevant topics for the user query.
- Efficiency: HITS Algorithm is not efficient in real time.

A HIT was used in a prototype search engine called Clever for an IBM research project. Because of the above constraints HITS could not be implemented in real time search engine.

V. Comparison

Table 2 shows comparison of all the three algorithms discussed above [11].

Table 2: Comparison of Algorithms

Algorithm	**Page Rank Algo**	**WPR Algo**	**HITS Algo**
Mining Technique Used	WSM	WSM	WSM and WCM
Working	Computes scores at indexing time. Results are sorted according to importance of pages.	Computes scores at indexing time. Results are sorted according to Page importance.	Computes Hub and Authority scores of n highly relevant pages on the fly.
I/P Parameters	Back Links	Back Links, Forward Links	Back Links, Forward Links & Content
Complexity	O(log N)	< O(log N)	< O(log N)
Limitations	Query independent	Query independent	Topic Drift and Efficiency Problem
Search Engine	Google	Research Model	Clever

VI. Conclusion

Web Mining is powerful technique used to extract the information from past behavior of users. Various Algorithms are used to rank the relevant pages. PageRank, Weighted PageRank and HITS treat all links equally when distributing the rank score. PageRank and weighted PageRank are used in WSM. HITS are used in both WSM and WCM. PageRank and Weighted PageRank calculates the score at indexing time and sort them according to importance pf page where as HITS calculates the hub and authority score of n highly relevant pages. The input parameters used in Page Rank are BackLinks, Weighted PageRank uses BackLinks and Forward Links as Input Parameter, HITS uses BackLinks, Forward Link and Content as Input Parameters. Complexity of PageRank Algorithm is O(log N) where as complexity of Weighted PageRank and HITS algorithms are <O(log N).

After going through the analysis of algorithms for ranking of web pages against the various parameters such as methodology, input parameters, relevancy of results and importance of the results, it is concluded that existing techniques have limitations particularly in terms of time response, accuracy of results, importance of the results and relevancy of results. An efficient web page algorithms should meet out these challenges efficiently with compatibility with global standards of web Technology.

References

[1] T.Munibalaji,C.Balamurugan, "Analysis of Link Algorithms for Web Mining",IJEIT Vol 1, Issue 2, Feb 2012.

[2] Raymond Kosala, Hendrik Blockee, "Web Mining Research : A survey",ACM Sigkdd Explorations Newsletter, June 2000, Vol 2.

[3] Laxmi Choudhary and Bhawani Shankar Burdele, "Role of Ranking Algorithms for Information Retrieval".

[4] Neelam Tyagi,Simple Sharma, "Comparative study of various Page Ranking Algorithms in Web Structure Mining (WSM)",International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol 1,Issue 1, June 2012.

[5] Neelam Duhan , A.K.Sharma , Komal Kumar Bhatia, "Page Ranking Algorithms: A Survey", Proceedings of the IEEE International Conference on Advance Computing, 2009.

[6] S. Brin and L.Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Networks and ISDN Systems, Vol 30, Issue 1-7, 1998.

[7] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithms, Proceedings of the second Annual Conference on Communication Networks and Services Research (CNSR) IIEEE, 2004.

[8] Tamanna Bhatia, "Link Analysis Algorithms for Web Mining", IJCST Vol 2, Issue 2, June 2011.

[9] Dilip Kumar Sharma, A.K.Sharma, "A Comparative Analysis of Web Page Ranking Algorithms",International Journal on Computer Science and Engineering, Vol 2,No.08, 2010

[10] Rekha Jain, Dr. G.N.Purohit, "Page Ranking Algorithms for Web Mining ",International Journal of Computer Applications (0975 – 8887), Vol 13, Jan 2011.