# An Intelligent Model for Redesigning Websites using Web Mining Techniques

Gauri Jain
ITM University,
Gurgaon,
India

Dr. Varun Kumar
HOD, Computer Science
Department, ITM University,
Gurgaon, India

## ABSTRACT
The main objective of this paper is to evolve an Intelligent Model which can help organizations to restructure their website so that website structure can be refined and it becomes more efficient and user friendly. In order to reach the main goal web mining process, web mining algorithm are applied. These can help to suggest the possible changes in the design of the wesite so that a common user feels much more comfortable in browsing the website. The data for this paper has been collected from a university website using a web crawler. Web mining algorithm are applied on the data so that possible changes can be suggested.

## General Terms
Web Mining, Web Structure Mining, Page Rank, Page Rank Algorithm, Crawlers, Websites.

## Keywords
Web Mining, Web Mining Process, Web Mining Techniques, Page Rank algorithm.

## 1. INTRODUCTION
With a huge amount of data available on the World Wide Web, we are drowning in the pool of data. We can use this data to our advantage if we can extract useful information from it. But, this is not possible manually, since human capacity does not allow processing of this huge amount of data. This paper is organized in four sections. The first section throws some light on the web mining, its process, and types of web mining technique. The second section is about web structure mining with an emphasis on page rank algorithm. Then, we discuss the current trends in website development. After that we suggest how web structure mining can help in refining the structural design of a website and is supported with the help of implementation example.

## 2. WEB MINING
### 2.1 Overview
Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services [1].This is of utmost importance research these days due to vast amount of information available on WWW. However, web mining technique requires a process to be followed which is given by [10]:

1. Resource finding: the task of retrieving intended Web documents.
2. Information selection & preprocessing: Data is frequently pre-processed to put it into a format that is compatible with the analysis technique to be used in the next step. Preprocessing may include cleaning data of inconsistencies, filtering out irrelevant information according to the goal of analysis
3. Generalization: automatically discovers general patterns at individual web sites as well as across multiple sites
4. Analysis: validation and /or interpretation of mined patterns.

### 2.2 Web Mining Process
To clarify the confusion of what forms Web mining. Kosala and Blockeel [2] had suggested a decomposition of Web mining in the following tasks:

1. *Resource finding*: The task of retrieving intended Web documents.
2. *Information selection and pre-processing*: *A*utomatically selecting and pre-processing specific information from retrieved Web resources.
3. *Generalization:* automatically discovers general patterns at individual Web sites as well as across multiple sites.
4. *Analysis*: validation and/or interpretation of the mined patterns.

### 2.3 Categories of Web Mining
There are three categories of web mining as shown in figure 1:

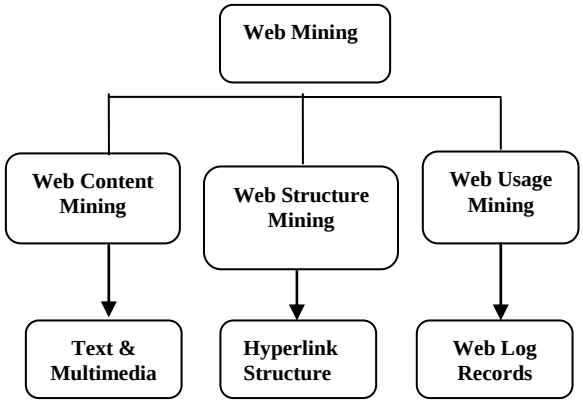


**Figure 1. Web mining Categories**

1. **Web Content Mining**

   Web Content mining describes the discovery of useful information from the web contents/data/documents.

2. **Web Structure Mining**

   Web Structure mining tries to discover the model underlying the link structure of the web.

3. **Web Usage Mining**

   Web Usage mining tries to make sense of the data generated by web surfer's sessions or behaviors. [2]

## 3. Web Structure Mining

The challenge for Web Structure mining is to deal with the structure of the hyperlinks within the web itself [3]. Web structure mining tries to discover the model underlying the link structures of the Web. The model is based on the topology of the hyperlink with or without the link description. This model can be used to categorize the Web pages and is useful to generate information such as similarity and relationships between Web sites [1]. This link structure between the different web pages on the WWW helps in deriving important information about the web pages like the importance and authoritativeness of a web page. For Example, if there are two pages A and B, A pointing towards B as shown in figure 1. Then, it means that B is an important page which is referenced by A. The more important page is referenced by more number of pages.

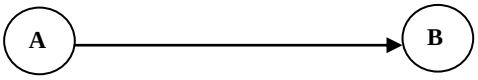


**Figure 2.** Page A referencing Page B

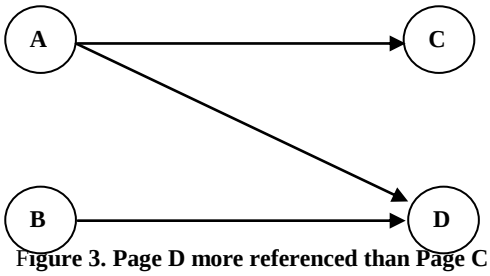


**Figure 3. Page D more referenced than Page C**

In the Figure 2, page D is referenced by page A and page B. We can say that page D is more important that C.

### 3.1 Algorithms

Since the amount of data on the Web is very huge, we can't calculate the importance for each page manually. There are two major link-based search algorithms, HITS (Hypertext Induced Topic Search) [12] and PageRank[5, 11]. The basic idea of the HITS algorithm is to identify a small sub-graph of the Web and apply link analysis on this sub-graph to locate the authorities and hubs for the given query. The sub-graph that is chosen depends on the user query. The selections of a small sub-graph (typically a few thousand pages), not only focus the link analysis on the most relevant part of the Web, but also reduce the amount of work for the next phase. The main weaknesses of HITS are known to non-uniqueness and nil-weighting [4]. Page Rank algorithm calculates the importance of web pages using the link structure of the web. In their approach Brin and Page extends the idea of simply counting in-links equally, by normalizing by the number of links on a page. The Page Rank algorithm is defined as:

We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor, which can be set between 0 and 1. We usually set d to 0.85.Also C (A) is defined as the number of links going out of page A. The Page Rank of a page A is calculated as shown in eq (1):

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) \quad \text{Eq(1)}$$

### 3.2 Organizational Website

Most of the organizations are having their websites developed. This helps in increasing their popularity, their business and their presence in the competitive world. Current website design scenario is shown in Figure 3. Some of these organizations have the least idea about website design so they are dependent on the development team for the designing. The development team themselves may be having the design prototype of the website already developed. If they don't they design the website according to their own preferences.

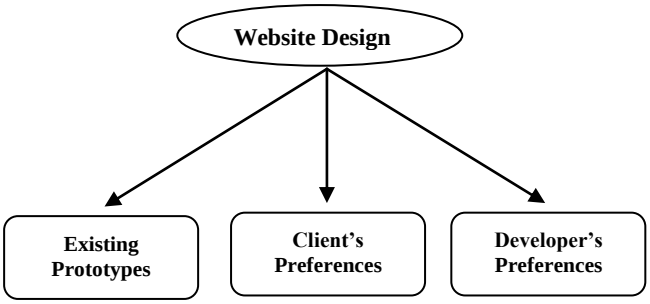


**Figure 4. Current website design trends**

Therefore, we need a strong technical ground on which the websites should be redesigned or rearranged so that they are more users friendly and efficient. This redesigning is done on the basis of importance of pages. More important pages should be much easier to locate in comparison to pages having less importance. The importance is calculated in terms of how many other pages refer to that particular page. The more referred, the more important it is. Location of a page is related to the effort required in reaching the page i.e. the number of hits required to reach a particular page. In this research, we have calculated the importance with the Page Rank Algorithm.

## 3.3 Experiment

We have collected the research data with the help of a web crawler from the website of ITM University. There are in total 355 pages. The data consists of incoming links and the number of outgoing links for each page. We have calculated the page rank for each page. The pages are arranged in decreasing order of page rank and accordingly changes are made as shown in Table 1 below.

**Table 1. Calculated Page Rank of some pages**

| Page No | Page Rank |
|---|---|
| 328 | 8.651913549 |
| 319 | 8.651913485 |
| 310 | 8.651913331 |
| 309 | 8.651913295 |
| 303 | 8.651913201 |
| 299 | 8.651913129 |
| 297 | 8.651913088 |
| 288 | 8.651912962 |
| 284 | 8.651912894 |
| 280 | 8.651912826 |
| 278 | 8.651912758 |
| 150 | 8.65191252 |
| 289 | 7.917159699 |
| 191 | 7.159706445 |
| 192 | 7.159706445 |
| 193 | 7.159706445 |
| 311 | 7.037048627 |
| 312 | 2.70850765 |
| 308 | 2.70850758 |
| 301 | 2.708507567 |
| 170 | 2.708507551 |
| 159 | 2.708507543 |
| 151 | 2.708507534 |
| 146 | 2.708507481 |
| 129 | 2.708507472 |
| 105 | 2.708507461 |
| 104 | 2.708507453 |
| 94 | 2.708507445 |
| 85 | 2.708507436 |
| 320 | 1.965124274 |
| 318 | 1.797985006 |

## 3.4 Observations

Few observations were prominent from the above experiment.

1. Most of the pages were according to increasing order of page rank as shown in Figure 5.

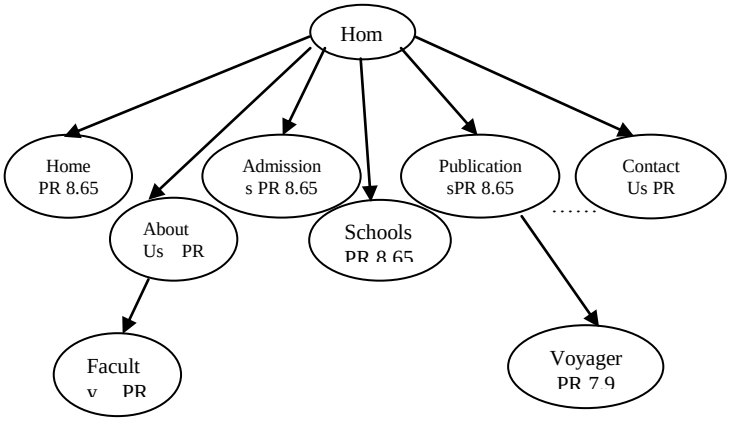


**Figure 5. Increasing order of Page Rank**

2. There were links of the pages with lower rank on the home page.

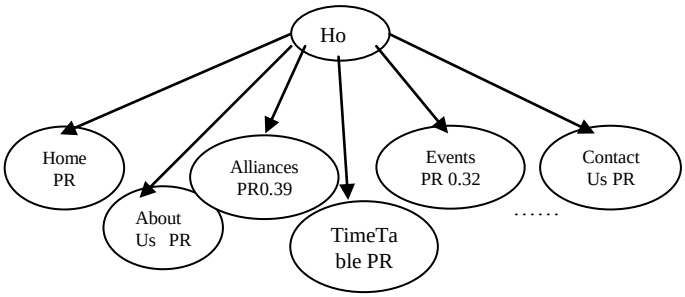


**Figure 6. Low Page Rank Links on the Home Page**

3. Important links like faculty having higher page rank are below in the hierarchy. It can be brought up in the website structure.

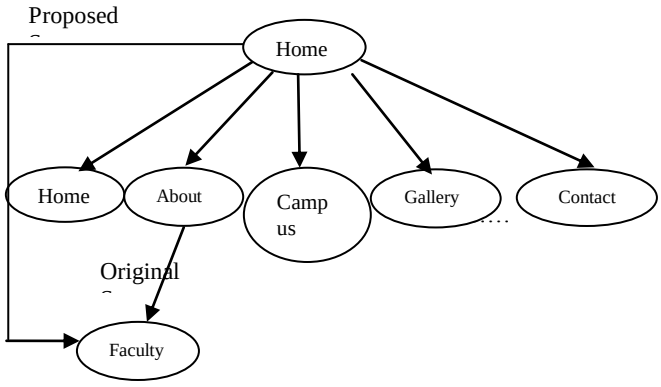


**Figure 7. Relocating Faculty Page**

There were some general observations about the website like

- The website doesn't have any directory structure.
- There was presence of several dead links on the website.
- There was duplicacy of webpage. These were having same information but the page numbers were different.

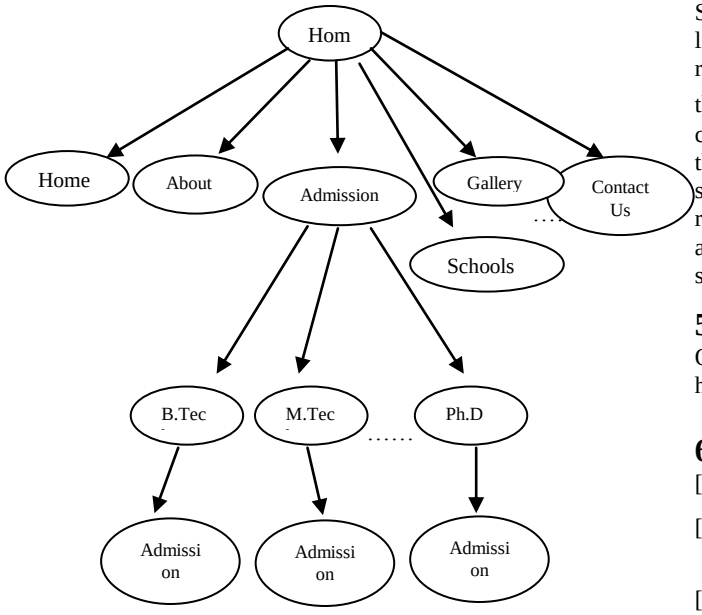


**Figure 8. Duplicate Admission Notice Page**

These can be redirected to one common page having consistent information.

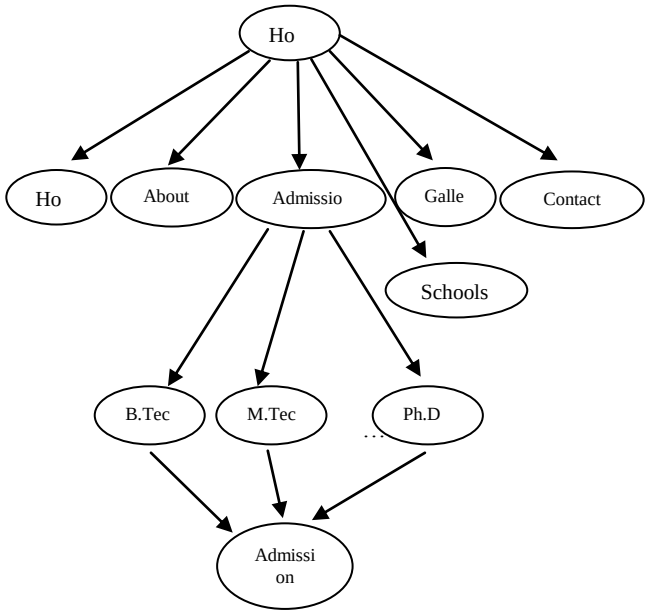


**Figure 9. Same Admission Notice Page**

## 4. Conclusion

With the results of the experiment we can say that since most of the important pages are designed according to their calculated page rank. So, we can say that website structure follow the technical aspect of website design. But, we can further improve it by following the design changes driven by experiments and observations. We have suggested relocating the pages having low page rank from the home page since there importance is less as compared to pages having high page rank. Also we have suggested a separate tab for the faculty page on the home page. Some general observations were also of significant importance like following the directory structure for the entire website. This results in not only easy searching of web pages but also provides the proper hierarchical design to the website. The website contained few duplicate pages i.e. their content being same but they are linked from different sources in the hierarchy. Another suggestion was there are many nonexistent pages which needs to redirected or removed. At the end we conclude that the efficiency and usability of website can be improved by following the above suggestions.

## 5. ACKNOWLEDGMENTS

Our thanks to ITM University and experts whose contributions has been very significant in the completion of this research.

## 6. REFERENCES

[1] Wookey Lee. Hierarchical Web Structure Mining.

[2] Raymond Kosala & Hendrik Blockeel. Web Mining Research: A Survey. ACM SIGKDD, July 2000.

[3] Da Gomes Jr., M.G. and Z. Gong, 2005. Web structure mining: An introduction. Proceeding of the IEEE International Conference on Information Acquisition, June 27-July 3, IEEE Xplore Press, Hong Kong and Macau, China, pp: 6. DOI: 10.1109/ICIA.2005.1635156

[4] Haveliwala, T. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE TKDE. 15(4), pp 784-796, 2003.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. Proceedings of the Seventh International World Wide WebConference, 1998.

[6] A.Senthil Kumar & N Palanisamy, Challenges for Web Mining 2008, Proceedings of the 2008 International Conference on Computing, Communication and Networking (ICCCN 2008) 978-1-4244-3595-1/08/©2008 IEEE

[7] P. Ravi Kumar and Ashutosh Kumar Singh, Web Structure Mining-Exploring Hyperlinks and algorithm for information Retrieval, 2010, American Journal of Applied Sciences 7 (6): 840-845

[8] Data Mining Techniques, Arun K. Pujari, Universities Press 2001

[9] L. Getoor, Link Mining: A New Data Mining Challenge. *SIGKDD Explorations*, vol. 4, issue 2, 2003.

[10] O. Etzioni. The world wide web: Quagmire or goldmine. Communications of the ACM, 39(11):65-68,1996.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bring order to the web. Technical report, Stanford University, 1998.

[12] Kleinberg, J.M., Authoritative sources in a hyperlinked environment. In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1998, pages 668-677 – 1998.

[13] Sona Bairamzadeh & Alireza Bolhari. Investigating Factors Affecting Students' Satisfaction of University Websites. 978 -1-4244-5540-9/10 IEEE 2010.