# Optimizing the Web Mining technique using Heuristic Approach

**Gunjan Batra[1],Vijay Laxmi[2] and M. Afshar Alam[3]**
**[1, 2]Student, [3]Professor,**
**Hamdard University,**
**Hamdard Nagar**
**New Delhi**
**gunjan_batra27@yahoo.co.in, tyagi.vijaylaxmi@gmail.com, aalam@jamiahamdard.ac.in**

## Abstract

*With the exceptional growth of the Web, there is an escalating volume of data and information available in frequent Web pages. The swift extension of the web leads to several problems such as lacks of organization and structure. Moreover, the content is available in different dissimilar formats. Because of its hasty and muddled growth users are feeling sometimes disoriented, lost in that information overload that continues to expand. Web mining is a very extensive research area promising to solve the issues that arise due to the WWW phenomenon. So the usage of data mining methods and knowledge discovery on the web is now on the spotlight of a boosting number of researchers. The Web mining research overlaps substantially with other areas, such as Databases, IR and AI. In this work we propose research on various web mining concepts and different AI techniques used in Web Mining. We also survey A* searching algorithm that can facilitate the extraction process of web.*

Keywords: AI, IR, Web Mining, Heuristic Approach

## 1. INTRODUCTION

With the exceptional growth of the Web, there is an ever escalating volume of data and information available in Web pages. The research in Web mining aims to develop new techniques to effectively extract and mine useful knowledge or information from these Web pages. Etzioni [1] who first invented the term web mining which is concerned with extracting knowledge from web data. There has been huge interest of Researchers towards web mining. Three different research directions in the areas of web mining: web structure mining, web content mining and web usage mining.

1.1 Web structure mining is the technique to determine useful knowledge from the organization of hyperlinks.
1.2 Web content mining is the process to dig out useful information from Web page.
1.3 Web usage mining is the application of data mining techniques to large web data repositories [2].

## 2. THE TRIPLET OF RAW DATA, PATTERN & KNOWLEDGE

**Raw Data** that has been verified to be accurate and timely, is specific and organized for a purpose, is presented within a context that gives it meaning and relevance, and that can lead to an increase in understanding and decrease in uncertainty [3].
**Pattern** is a particular data behavior, arrangement or form that might be of a business interest.
**Knowledge** is the (iterative and interactive) nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. [Fayyad96] [4].

Raw data is the data input to processing.Although raw data has the potential to become "information," it requires selective extraction, organization, and sometimes analysis and formatting for presentation. As a result of processing, raw data sometimes ends up in a database, which enables the data to become accessible for further processing and analysis in a number of different ways.

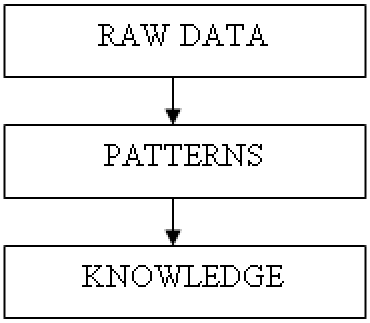

Figure 1. The Triplet

## 3. Web Mining

Web Mining is the technique used to crawl through various web resources to automatically discover and extract information from Web documents and services. Web mining can facilitate marketing patterns and tailor market to bring right products and services to right customers. It can help in making decisions in customer relationship management and also improve quality of mining. The general data mining process is shown in the figure 2.

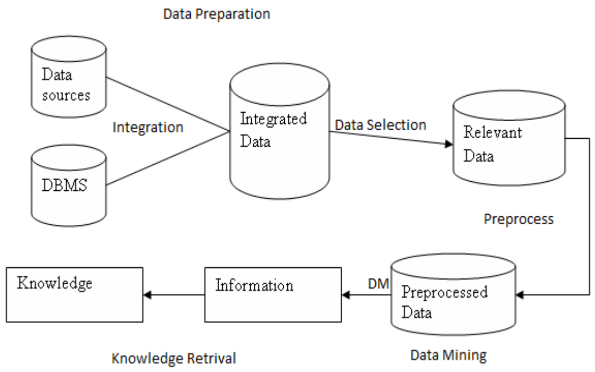

Figure 2. Data Mining Process

### 3.1 Web Mining's Subtask [5]:

- Mining Web Search Engine Data
- Analyzing the Web's Link structures
- Classifying web Documents automatically
- Mining Web page semantic structures and page contents
- Mining Web dynamics
- Building multilayered, multidimensional Web

### 3.3 Structure of web Mining

When a user views information, there are three basic factors that can influence the observation and evaluation process. They are:

• Web page Content
• Web page Design
• Website Design and Structure

On the basis of these factors web Mining is structured into three main categories. These are shown in figure 3 and explained below:

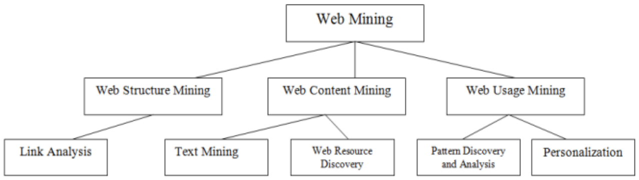

Figure 3 Structure of Web Mining

#### 3.3.1 Web Structure mining

This process involves discovering the data on the basis if underlying link structure on the web. This form of mining is based on the topology of the hyperlinks with or without a description of the links.
Web Structure is a useful source for extracting information such as

- Quality of Web Page
- Interesting Web Structures
- Web Page Classification

#### 3.3.2 Web Content Mining

A lot of the knowledge in the Web is inside documents,
i.e., in their content/data/documents. The discovery process of this useful information from that content is called Web Content Mining.
Web Content Mining is useful to

- Identify the topics represented by a Web Documents
- Categorize Web Documents
- Find Web Pages across different servers that are similar
- Find the relevant data according to user query and/or task based relevance
- Recommendations –List of top "n" relevant documents in a collection or portion of a collection.
- Filters –Show/Hide documents based on relevance score

#### 3.3.3 Web Usage Mining

This process involves analysis of the data generated by user behavior and browsing history. Web content mining and web structure mining relies on primary data but web usage mining relies on secondary data like data from the server logs, browser logs, user profiles etc.
Web Usage mining is useful to

- determine the best way to structure the Web site
- identify "weak links" for elimination or enhancement
- A "site-specific" web design agent
- Pre-fetch files that is most likely to be accessed
- Intra-Organizational Applications
- enhance workgroup management & communication
- evaluate Intranet effectiveness and identify structural needs & requirements

## 4. AI techniques used in Web mining :

### 4.1 ARTIFICIAL NEURAL NETWORKS:

An **artificial neural network** (ANN), often just called a "neural network" (NN), is a mathematical model based on biological neural networks, in other words, is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. Neural network method is used for classification, clustering, feature mining, prediction and pattern recognition.

### 4.2 Genetic Algorithm

Genetic Algorithm [7] is a heuristic function for optimization, where the extreme of the function (i.e., minimal or maximal) cannot be established analytically. A population of potential solutions is polished iteratively by employing an approach inspired by Darwinist evolution or natural selection. Genetic Algorithms promote "survival of the fittest".

### 4.3. Bayesian networks

Bayesian networks are directed acyclic graphs whose nodes represent variables, and whose arcs encode conditional independencies between the variables. The graph provides an intuitive description of the dependency model and defines a simple factorization of the joint probability distribution leading to a tractable model which is compatible with the encoded dependencies.

### 4.4 Decision Trees

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. Decision tree algorithms tend to automate the entire process of hypothesis generation and then validation much more completely and in a much more integrated way than any other data mining techniques.

### 4.5. Rule Induction

Rule induction is one of the major forms of data mining and is perhaps the most common form of knowledge discovery in unsupervised learning systems. Rule induction on a data base can be a massive undertaking where all possible patterns are systematically pulled out of the data and then an accuracy and significance are added to them that tell the user how strong the pattern is and how likely it is to occur again.

The bane of rule induction systems is also its strength - that it retrieves all possible interesting patterns in the database.

## 5. Algorithms for web mining

**5.1 HITS search algorithm** - HITS (acronym for Hyperlink-Induced Topic Search) is a search algorithm developed by Jon Kleinberg to find the most authoritative pages for a wide search query. It is based on the scoring pages either as "hubs" in the sense of pages that link to numerous authoritative Web pages for a given topic, or authoritative pages, which are Web pages that are pointed to by the hub pages.

Kleinberg reasoned that there were two types of search queries, specific and broad. For more specific queries, the problem is to find any pages that match the criteria. For broad queries, there is an abundance problem: there are too many pages that match the criteria and therefore the problem is to find the most relevant ones for that query [8].

### 5.2 Page Rank

Page Rank is an objective measure of citation significance that corresponds with people's skewed idea of importance. PageRank was proposed by Brin and Page [13] as a possible model of user surfing behavior. The PageRank of a page represents the probability that a random surfer (one who follows links randomly from page to page) will be on that page at any given time. A page's score depends recursively upon the scores of the pages that point to it. Source pages distribute their PageRank across all of their out links.

The PageRank algorithm and implementation details are described in [9, 10]. The PageRank algorithm represents the structure of the Web as a matrix, and PageRank values as a vector.

## 6. Problem Statement

**HITS-**

- Disadvantage of this algorithm is that because it is executed at query time, it may have very poor performance, depending on the number of iterations required to rank the Hubs and authorities.
- Another disadvantage is that it may be making assumptions about the structure of the Web that no longer hold true.

**PageRank-**

- The main disadvantage is that it favors older pages, because a new page, even a very good one, will not have many links unless it is part of an existing site.
- Another disadvantage is easy-to-produce 'link-farms' schemes designed to artificially push up the PageRank.

## 7. Solution

These problems can be avoided by using heuristic approach.

*Heuristic* is a rule of thumb that probably leads to a solution. Heuristics play a major role in search strategies because of exponential nature of the most problems. Heuristics help to reduce the number of alternatives from an exponential number to a polynomial number.

In order to solve larger problems, domain-specific knowledge must be added to improve search efficiency. Information about the problem includes the nature of states, cost of transforming from one state to another, and characteristics of the goals. This information can often be expressed in the form of heuristic evaluation function, say f(n,g), a function of the nodes n and/or the goals g.[11]

In this paper, we proposed a modified A* algorithm to improve the use of importance as well as relevance.

## 8. Modified A* Algorithm

A* is a computer algorithm that is widely used in path finding and graph traversal, the process of plotting an efficiently traversable path between points, called nodes. A*-search algorithm treats Internet as a directed graph, webpage as node and hyperlink as edge, so the search operation could be abstracted as a process of traversing directed graph. This is the optimum approach to traverse web pages and getting the desired result in minimum time. Here, crawling is on the basis of relevancy, which is calculated in terms of fuzzy logics which lie in the range of 0 to 1.

0 – no similarity/ match

1 – Perfect match

Relevance of a page is measured in terms of frequency and location of searching phrase in the page and connectivity of that page.

$f(n) = g(n) + h(n)$

g(n):- cost calculated by phrase frequency and location. Term appearing near the top of a web page, such as in the title or in the first few paragraphs of text, it is assumed that the page is more relevant than if the term is used at the bottom of the page. Pages where the words appear more frequently in relation to the other words on the page also qualifies the page as being more relevant than other web pages. Here, crawler represents a fetched Web page as a vector of words weighted by occurrence frequency and location. The crawler then computes the cosine similarity of the page to the query or description provided by the user, and scores the unvisited URLs on the page by this similarity value.

h(n):– cost of the page is analyzed by incoming and outgoing links. If a page is linked to from a large number of other pages, then it is ranked more highly.

More the value of f(n) better will be the page to traverse. Algorithm maintains a list, which keeps URL of page to be searched. The URLs have different priority, the URL with more superior priority will be located at the front of the list, and will be searched sooner than others.
To implement this approach Priority queue is maintained as a data structure. After each iteration, the link with the highest f(n) function value is picked from the frontier.

### 7.1 Features:-

- This algorithm makes use of both best first search and Page rank algorithm and has advantages of both the approaches.
- The use of importance as well as relevance.
- It favors older pages, because a new page, even a very good one, will not have many links unless it is part of an existing site.
- Poor retrieval effectiveness (relevance) as little vocabulary control is exercised by web site developers and the index engines

## 8. CONCLUSION and future Scope

The World Wide Web is huge, universal, heterogeneous and unstructured. Web mining is a very broad research area trying to solve issues that arise due to the WWW phenomenon. In this work, we analyzed the three separate categories and subtasks of Web mining. After that, we briefly surveyed various AI techniques used in web mining. In next section, we briefly explained PageRank and HITS algorithms with their disadvantages. We proposed a heuristic approach to overcome the problems of these algorithms. One simple improvement to enhance efficiency is inclusion of heuristic approach in Web Mining. The WWW will keep growing, even in a somewhat different form than how we know it today. Therefore the need for discovering new methods and techniques to handle the amounts of data existing in this universal framework will always exist.

## 9. REFERENCES

[1] O.Etzioni. The World Wide Web: Quagmire or Gold Mine. Communications of the ACM, 39 CII): 65-68, 1996.

[2] Robert Cooly, Bamshad Mobasher, Jaideep Srivastava (1999) : Data Preparation for Mining World Wide Web browsing Pattern.

[3] http://www.businessdictionary.com/definition/information.html

[4] http://www.kmining.com/info_definitions.html

[5] Jaiwei Han,Kevin Chen-Chaun Chang, University of Illinois at Urbna Champaign Data Mining for Web Intelligence.

[6] Navin Kumar Tyagi , A.K. Solanki & Sanjay Tyagi; AN "ALGORITHMIC APPROACH TO DATA PREPROCESSING IN WEB USAGE MINING" International Journal of Information Technology and Knowledge Management July-December 2010.

[7] David. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Publication Addison-Wesley Professional

[8] Kleinberg, J. Authoritative sources in a hyperlinked environment" *Journal of the ACM* 46:5:604-632, , 1999.

[9] T. H. Haveliwala: Efficient Computation of PageRank, unpublished manuscript, Stanford University (1999)

[10] A. Y. Ng, A. X. Zheng, and M. I. Jordan: Stable Algorithms for Link Analysis, Proceedings of the 24th ACM SIGIR Conference (2001), 258-266

[11] Justin heyes A* algorithm tutorial http://www.heyes-jones.com/astar.html

[12]Artificial intelligence book by Elaine rich and Kevin Knight

[13] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", http://infolab.stanford.edu/~backrub/google.html