Web Mining: Methodologies, Algorithms and Applications

Bussa V.R.R.Nagarjuna, Akula Ratna babu, Miriyala Markandeyulu, A.S.K.Ratnam

Abstract: ***The World Wide Web is a popular and interactive medium to disseminate information today. It is a system of interlinked hypertext documents accessed via the Internet. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia, and navigate between them via hyperlinks. With the recent explosive growth of the amount of content on the Internet, it has become increasingly difficult for users to find and utilize information and for content providers to classify and catalog documents on the World Wide Web. Traditional web search engines often return hundreds or thousands of results for a search, which is time consuming for users to browse. On-line libraries, search engines, and other large document repositories (e.g. customer support databases, product specification databases, press release archives, news story archives, etc.) are growing so rapidly that it is difficult and costly to categorize every document manually. To deal with these problems web mining is used. Web mining is the use of data mining techniques to automatically discover and extract information from the web documents and services. This paper presents an overview of web mining, its methodologies, algorithms and applications.***

Index Terms: Data mining, Methodologies, Web mining, World Wide Web.

I. INTRODUCTION

In a distributed information environment, documents or objects are usually linked together to facilitate interactive access. Examples for such information- providing environments include the World Wide Web (WWW) and online services such as America Online, where users, when seeking information of interest, travel from one object to another via facilities such as hyperlinks and URL addresses. The Web is an ever growing body of hypertext and multimedia documents. As the information offered in the Web grows daily, obtaining that information becomes more and more tedious. The main difficulty lies in the semi - structured or unstructured Web content that is not easy to regulate and where enforcing a structure or standards is difficult. A set of Web pages lacks a unifying structure and shows far more authoring styles and content variation than that seen in traditional print document collections. This level of complexity makes an " off - the - shelf " database - management and information – retrieval solution very complex and almost impossible to use. New methods and tools are necessary. Web mining [1][2] may be defined as the use of data - mining techniques to automatically discover and extract information from Web documents and services. It refers to the overall process of discovery, not just to the application of standard data - mining tools. Some authors suggest decomposing Web - mining task into four subtasks:

1. Resource Finding: T his is the process of retrieving data, which is either online or offline, from the multimedia sources on the Web, such as news articles, forums, blogs, and the text content of HTML documents obtained by removing the HTML tags.
2. Information Selection and Preprocessing*:* This is the process by which different kinds of original data retrieved in the previous subtask is transformed. These transformations could be either a kind of preprocessing such as removing stop words and stemming or a preprocessing aimed at obtaining the desired representation, such as finding phrases in the training corpus and representing the text in the first - order logic form.
3. Generalization: Generalization is the process of automatically discovering general patterns within individual Web sites as well as across multiple sites. Different general- purpose machine - learning techniques, data - mining techniques, and specific Web - oriented methods are used.
4. Analysis: This is a task in which validation and/or interpretation of the mined patterns is performed.

In this paper we present an overview of web mining. This paper is organized as follows: Section 2 presents the methodologies of web mining. Section 3 presents algorithms used for web mining. Section 4 presents the applications of web mining. And finally section 5 gives the conclusion.

II. WEB MINING METHODOLOGIES

Web mining the application of machine learning (data mining) techniques to web-based data for the purpose of learning or extracting knowledge. Web mining methodologies can generally be classified into one of three distinct categories:

- Web usage mining
- Web structure mining
- Web content mining

A. Web usage mining

Manuscript received on July, 2012.

Bussa V.R.R.Nagarjuna, M.Tech (CSE) Vignan's Lara Institute Of Technology And Science.Vadlamudi,Guntur,AP.,India.

Akula Ratnababu, M.Tech (CSE) Vignan's Lara Institute Of Technology And Science.Vadlamudi,Guntur,AP.,India.

Miriyala Markandeyulu, M.Tech (CSE) Vignan's Lara Institute Of Technology And Science.Vadlamudi,Guntur,AP.,India.

A.S.K.Ratnam,Head,Dept.of Computer Sceince Engineering Vignan's Lara Institute Of Technology And Science.Vadlamudi,Guntur,AP.,India.

Web usage mining [3] is a type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. CGI scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access.

Usage mining allows companies to produce productive information pertaining to the future of their business function ability. Some of this information can be derived from the collective information of lifetime user value, product cross marketing strategies and promotional campaign effectiveness. The usage data that is gathered provides the companies with the ability to produce results more effective to their businesses and increasing of sales. Usage data can also be useful for developing marketing skills that will out-sell the competitors and promote the company's services or product on a higher level.

Usage mining is valuable not only to businesses using online marketing, but also to e-businesses whose business is based solely on the traffic provided through search engines. The use of this type of web mining helps to gather the important information from customers visiting the site. This enables an in-depth log to complete analysis of a company's productivity flow. E-businesses depend on this information to direct the company to the most effective Web server for promotion of their product or service.

B. Web structure mining

Web structure mining [4][5] is a tool used to identify the relationship between Web pages linked by information or direct link connection. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages. This connection allows a search engine to pull data relating to a search query directly to the linking Web page from the Web site the content rests upon. This completion takes place through use of spiders scanning the Web sites, retrieving the home page, then, linking the information through reference links to bring forth the specific page containing the desired information.

Structure mining uses minimize two main problems of the World Wide Web due to its vast amount of information. The first of these problems is irrelevant search results. Relevance of search information become misconstrued due to the problem that search engines often only allow for low precision criteria. The second of these problems is the inability to index the vast amount if information provided on the Web. This causes a low amount of recall with content mining. This minimization comes in part with the function of discovering the model underlying the Web hyperlink structure provided by Web structure mining.

The main purpose for structure mining is to extract previously unknown relationships between Web pages. This structure data mining provides use for a business to link the information of its own Web site to enable navigation and cluster information into site maps. This allows its users the ability to access the desired information through keyword association and content mining. Hyperlink hierarchy is also determined to path the related information within the sites to the relationship of competitor links and connection through search engines and third party co-links. This enables clustering of connected Web pages to establish the relationship of these pages. On the WWW, the use of structure mining enables the determination of similar structure of Web pages by clustering through the identification of underlying structure. This information can be used to project the similarities of web content. The known similarities then provide ability to maintain or improve the information of a site to enable access of web spiders in a higher ratio. The larger the amount of Web crawlers, the more beneficial to the site because of related content to searches.

C. Web content mining

Web content mining[4][5] is also known as text mining. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query. With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query.

Text mining is directed toward specific information provided by the customer search information in search engines. This allows for the scanning of the entire Web to retrieve the cluster content triggering the scanning of specific Web pages within those clusters. The results are pages relayed to the search engines through the highest level of relevance to the lowest. Though, the search engines have the ability to provide links to Web pages by the thousands in relation to the search content, this type of web mining enables the reduction of irrelevant information.

Web text mining is very effective when used in relation to a content database dealing with specific topics. For example online universities use a library system to recall articles related to their general areas of study. This specific content database enables to pull only the information within those subjects, providing the most specific results of search queries in search engines. This allowance of only the most relevant information being provided gives a higher quality of results. This increase of productivity is due directly to use of content mining of text and visuals. The main uses for this type of data mining are to gather, categorize, organize and provide the best possible information available on the WWW to the user requesting the information. This tool is imperative to scanning the many HTML documents, images, and text provided on Web pages. The resulting information is provided to the search engines in order of relevance giving more productive results of each search.

III. WEB MINING ALGORITHMS

This section presents two most important algorithms for web mining.

A. HITS Algorithm

Before presenting HITS (Hypertext Induced Topic Search) [6] algorithm, we have to know

about two types of web pages: authorities, which provide the best source of information about a given topic and hubs, which provide a collection of links to authorities. Hub pages appear in a variety of forms, ranging from professionally assembled resource lists on commercial sites to lists of recommended links on individual home pages. These pages need not themselves be prominent, and working with hyperlink information in hubs can cause much difficulty. Although many links represent some kind of endorsement, some of the links are created for reasons that have nothing to do with conferring authority. Typical examples are navigation and paid advertisement hyperlinks. A hub's distinguishing feature is that they are potent conferrers of authority on a focused topic. We can define a good hub if it is a page that points to many good. authorities. At the same time, a good authority page is a page pointed to by many good hubs. This mutually reinforcing relationship between hubs and authorities serves as the central idea applied in the HITS algorithm that searches for good hubs and authorities.

The two main steps of the HITS algorithm are

1. The sampling component, which constructs a focused collection of Web pages likely to be rich in relevant information, and
2. The weight - propagation component, which determines the estimates of hubs and authorities by an iterative procedure and obtains the subset of the most relevant and authoritative Web pages.

In the sampling phase, we view the Web as a directed graph of pages. The HITS algorithm starts by constructing the subgraph in which we will search for hubs and authorities. Our goal is a subgraph rich in relevant, authoritative pages. To construct such a subgraph, we first use query terms to collect a root set of pages from an index - based search engine. Since many of these pages are relevant to the search topic, we expect that at least some of them are authorities or that they have links to most of the prominent authorities. We therefore expand the root set into a base set by including all the pages that the root - set pages link to, up to a designated cutoff size. This base set V typically contains from 1000 to 5000 pages with corresponding links, and it is a final result of the first phase of HITS.

In the weight - propagation phase, we extract good hubs and authorities from the base set V by giving a concrete numeric interpretation to all of them. We associate a nonnegative authority weight *a* p and a nonnegative hub weight *h* p with each page p ∈ V. We are interested only in the relative values of these weights; therefore, normalization is applied so that their total sum remains bounded. Since we do not impose any prior estimates, we set all *a* and *h* values to a uniform constant initially. The final weights are unaffected by this initialization.

We now update the authority and hub weights as follows. If a page is pointed to by many good hubs, we would like to increase its authority weight. Thus, we update the value of a p for the page p to be the sum of h q over all pages q that link top:

$$a_p = \Sigma h_q, \forall q \text{ such that q -> p}$$

where the notation q → p indicates that page q links to page p. In a strictly dual fashion, if a page points to many good authorities, we increase its hub weight

$$h_p = \Sigma a_q, \forall \text{ q such that p ->q}$$

B. Page rank Algorithm

PageRank[7] is a method for rating Web pages objectively and mechanically, paying attention to human interest. Web search engines have to arrange with inexperienced users and pages manipulating conventional ranking functions. Any evaluation strategy which counts replicable features of Web pages is unimmunized to manipulation. The task is to take advantage of the hyperlink structure of the Web to produce a global importance ranking of every Web page. This ranking is called PageRank.

The structure of the Web is based on a graph with about 150 million nodes (Web pages) and 1.7 billion edges (hyperlinks). If Web pages A and B link to a page C, A and B are called the backlinks of C. This circumstance is illustrated in Figure 1. In general, highly linked pages are more important. Thus they have more backlinks. But the important backlinks are often less in quantity. For example a Web page with a single backlink from Yahoo has to be ranked higher than a page with a couple of backlinks from unknown or private sites. A Web page has a high rank, if the sum of the ranks of its backlinks is also high.

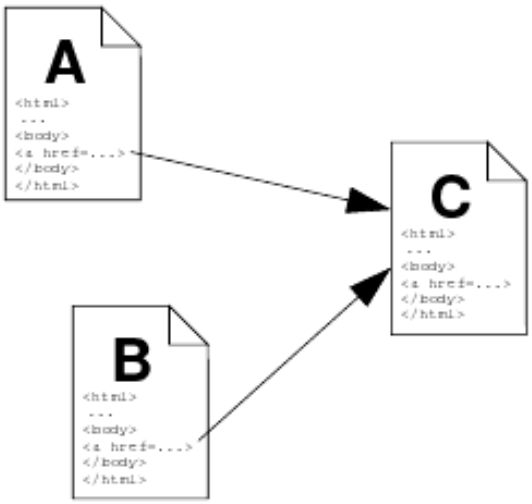

Figure 1: A and B are backlinks of C

The following is the simplified version of PageRank:

Let u, v be Web pages. Then let Bu be the set of pages that point to u. Further let Nv be the number of links from v. Let c < 1 be a factor for normalization. We define a simple ranking R, which is a simplified version of PageRank:

$$R(u) = c \sum_{u \in Bu} \frac{R(v)}{N_v}$$

The rank of a page is divided among its forward links evenly to contribute to the ranks of the pages they point to. The equation is recursive. But there is a problem with this simplified function. If there where two Web pages that point to each other but to no other page while some other Web page points to one of them, a loop will be generated during the iteration. This loop will

accumulate the rank but will never distribute any ranks. This trap formed by loops in a graph without outedges are called rank sinks.

The Page Rank algorithm starts with the conversion of each URL from the database into an integer. The next step is to store each hyperlink in a database using the integer IDs to identify the Web pages. The iteration is initiated after sorting the link structure by the parent ID and removing dangling links. A good initial assignment has to be chosen to speed up convergence. The weights from the current time step are kept in memory and the previous weights are accessed on disk in linear time. After the weights have converged the dangling links are added back and the rankings are recomputed. The calculation performs well but could be made faster by easing the convergence criteria and using more efficient optimization strategies.

IV APPLICATIONS OF WEB MINING

The following are some of the applications of web mining:

1. Web mining is used to discover how users navigate a web site and the results can help in improving the site design and making it more visible on the web.
2. In customer relationship management (CRM), Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. Web mining is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign.
3. The popularity of digital images is rapidly increasing due to improving digital imaging technologies and convenient availability facilitated by the Internet. However, how to find user-intended images from the Internet is non-trivial. The main reason is that the web images are usually not annotated using semantic descriptors. To retrieve web images from the internet, web mining is used.
4. Web mining is used for keyphrase extraction. Keyphrases are useful for a variety of purposes, including summarizing, indexing, labeling, categorizing, clustering, highlighting, browsing, and searching. The task of automatic keyphrase extraction is to select keyphrases from within the text of a given document. Automatic keyphrase extraction makes it feasible to generate keyphrases for the huge number of documents that do not have manually assigned keyphrases.
5. Web mining is used for social network analysis. Social network is the study of social entities (people in an organization, called actors), and their interactions and relationships. Social network analysis is useful for the Web because the Web is essentially a virtual society, and thus a virtual social network, where each page can be regarded as a social actor and each hyperlink as a relationship. Many of the results from social networks can be adapted and extended foruse in the Web context. The ideas from social network analysis are indeed instrumental to the success of Web search engines.

V.CONCLUSION

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. Web mining can be broadly defined as the discovery and analysis of useful information from the WWW. Web mining methodologies are of three types: In web usage mining the goal is to examine web page usage patterns in order to learn about a web system's users or the relationships between the documents. In web structure mining, we examine only the relationships between web documents by utilizing the information conveyed by each document's hyperlinks. In web content mining we examine the actual content of web pages (most often the text contained in the pages) and then perform some knowledge discovery procedure to learn about the pages themselves and their relationships.

This paper explains two most popular algorithms of web mining: HITS and PageRank. By using web mining algorithms, significant patterns about the user behavior on the web can be extracted and thus improve the relationship between the website and its users.

REFERENCES

1. Chang , G. , M. J. Haeley , J. A. M. McHugh , J. T. L. Wang , *Mining the World Wide Web: An Information Search Approach* , Kluwer Academic Publishers , Boston, MA , 2001 .
2. S. Chakrabarti. mining the Web. Morgan Kaufmann, San Francisco, CA, 2003.
3. R.W. Cooley .Web usage mining: Discovery and application of Interesting patterns from Web data. PhD thesis , dept of computer science, university of Minnesota, May 2000.
4. R. Kosala and H. Blockeel. Web mining research: A survey. SIGKDD Explorations, 2(1), 2000.
5. Osmar R. Za¨ane. From resource discovery to knowledge discovery on the internet. Technical Report TR 1998-13, Simon Fraser University, 1998.
6. http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/ Lecture4/lecture4.html
7. Brin, S.; Motwani, R.; Page, L.; Winograd, T.: The PageRank Citation
8. Ranking: Bringing Order to the Web. Technical Report, 1998.