ISSN : 2229-4333(Print) | ISSN : 0976-8491(Online)

Survey of Web usage Mining

[1]Rakesh Kumar Malviya, [2]Mahesh Chandra Malviya,[3]Vinay Kumar Soni,
[4]Ritesh Joshi, [5]Preetesh Purohit
[1,2]Dept. of CSE, Jawaharlal Institute of Technology, Borawan, India
[3]Dept. of MCA, Jawaharlal Institute of Technology, Borawan, India
[4,5]Dept. of MCA, Medicaps Institute of Technology & Management, Indore, M.P., India

Abstract

Web mining, i.e. the application of data mining techniques to extract knowledge from Web content, structure, and usage, is the collection of technologies to fulfil the potential of extracting valuable knowledge from the World Wide Web and its usage pattern. Web mining techniques seek to extract knowledge from Web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. This article provides an overview of past and current work in the three main areas of Web mining research—content, structure, and usage. In this paper we define Web mining and, in particular, present an overview of the various research issues, techniques, and development efforts in Web content mining and Web usage mining.

Keywords

World Wide Web, web mining, web content mining, web structure mining, web usage mining.

I. Introduction

The advent of the World-Wide Web (WWW) has overwhelmed the typical home computer user with an enormous flood of information. To be able to cope with the abundance of available information, users of the WWW need to rely on intelligent tools that assist them in finding, sorting, and filtering the available information. Just as data mining aims at discovering valuable information that is hidden in conventional databases, the emerging field of Web mining aims at finding and extracting relevant information that is hidden in Web-related data, in particular in text documents that are published on the web. Depending on the nature of the data, one can distinguish three main areas of research within the Web mining community:

A. Web Content Mining

application of data mining techniques to unstructured or semi-structured data, usually HTML-documents

B. Web Structure Mining

Use of the hyperlink structure of the Web as an (additional) information source

C. Web Usage Mining

analysis of user interactions with a Web server (e.g., click-stream analysis) i.e. collecting data from web log records.

II. Web Content Mining

In recent years the growth of the World Wide Web exceeded all expectations. Today there are several billions of HTML documents, pictures and other multimedia files available via internet and the number is still rising. But considering the impressive variety of the web, retrieving interesting content has become a very difficult task. Web Content Mining uses the ideas and principles of data mining and knowledge discovery to screen more specific data. The use of the web as a provider of information is unfortunately more complex than working with static databases. Another important aspect is the presentation of query results. Due to its enormous size, a web query can retrieve thousands of resulting WebPages. Thus meaningful methods for presenting these large results are necessary to help a user to select the most interesting content. The Web document usually contains several types of data, such as text, image, audio, video, metadata and hyperlinks. Some of them are semi-structured such as HTML documents, or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data. The unstructured characteristic of Web data force the Web content mining towards a more complicated approach.

A. Mining by developing a knowledge-base repository of the domain

When information is given a well-defined meaning by defining the relationship between web pages and their contents, we are said to be creating an ontology. However this definition of relationship is very difficult to identify for a number of reasons, some of which are, i. The identification of vocabulary that is used to describe the relevant concepts within the document. ii. Finding definitions, for the vocabulary identified, that best describes the term. iii. Identifying correct relations between the above two, where one term may be linked to many definitions and one definition may be for more than one term. Analysis of mined knowledge.

B. Interpretation of Mined Knowledge

One of the open issues in data mining, in general, and Web mining, in particular, is the creation of intelligent tools that can assist in the interpretation of mined knowledge. Clearly, these tools need to have specific knowledge about the particular problem domain to do any more than filtering based on statistical attributes of the discovered rules or patterns. In Web mining, for example, intelligent agents could be developed that based on discovered access patterns, the topology of the Web locality, and certain heuristics derived from user behaviour models, could give recommendations about changing the physical link structure of a particular site. As a simple example, consider an agent that (among other things) looks at the difference between the visit frequency for a particular page and the number of frequent user paths ending in that page. This difference could be used to determine if the page constitutes an entry point. This may suggest the other navigational links should be placed on that page to increase traffic to other clusters of pages.

C. Iterative refinement of user queries for personalised search

This is a very interesting approach, where the initial results of the user's query are re-ordered, by observing user preferences, and by asking some questions to assist the user in the search process. One might argue with the need for this approach with the existence of powerful search engines available to us today like Yahoo! © Well just try and look up a keyword on it. If one compares the total number of hits resulted by it with the actual desired ones, he will realize the importance of query refinement search engines, such as this technique proposes. Imagine the uses of such a technique just in shopping and the time it would save a conventional customer!

D. Process of Iterative Query Refinement

The queries in this method are accepted through an interface, passed to any general-purpose search engine available. The search engine would retrieve the entire list of web pages that according to the search criteria of the search engine appear the best. This list is passed on to the user interface once again, where by the user can choose the best desired links at random from the initial list. This list is passed on to the Refinement Algorithm, which adjusts the weights of the pages resulted from the search. The Ranker finally assigns new ranks to the pages and the new list is displayed before the user. The user may choose to further fine-tune the results displayed, or if the results are rather to his liking he may accept the results and continue with the traversal of the pages to carry on with the search.

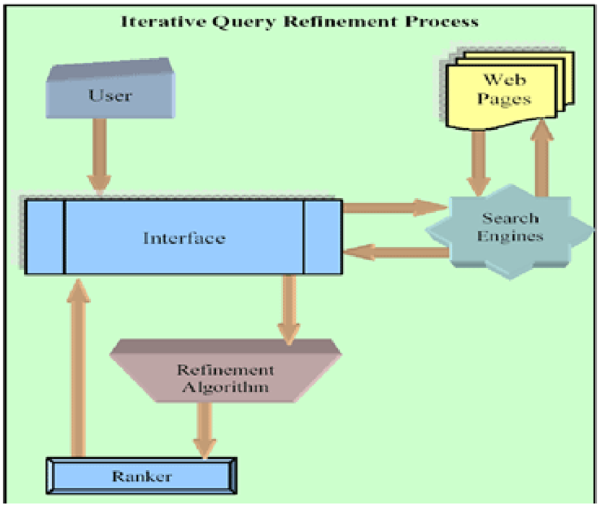

Fig. 1 : Iterative Query Refinement Process

III. Web Structure Mining

The goal of Web structure mining is to generate structural summary about the Web site and Web page. Technically, Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Web structure mining describes the connectivity in the Web subset, based on the given collection of interconnected web documents. The structural information generated from the Web structure mining includes the follows:

1. The information measuring the frequency of the local links in the Web tuples in a Web table.
2. The information measuring the frequency of Web tuples in a Web table containing links that are interior and the links that are within the same document
3. Information measuring the frequency of Web tuples in a Web table that contains links that are global and the links that span different Web sites
4. The information measuring the frequency of identical Web tuples that appear in the Web table or among the Web tables.

A. Web Graph Structure

The structure of a typical Web (directed) graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. While conventional information retrieval focuses primarily on information that is provided by the text of Web documents, the Web provides additional information through the way in which different documents are connected to each other via hyperlink.

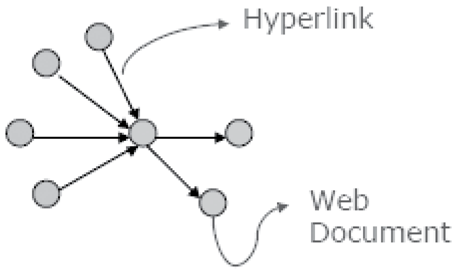

Fig. 2 : Web Graph Structure

B. Web Structure Terminology

Web-graph: A directed graph that represents the Web. Node: Each Web page is a node of the Web-graph. **Link:** Each hyperlink on the Web is a directed edge of the Web-graph.

In-degree: The in-degree of a node, p, is the number of distinct links that point to p.

Out-degree: The out-degree of a node, p, is the number of distinct links originating at p that point to other node

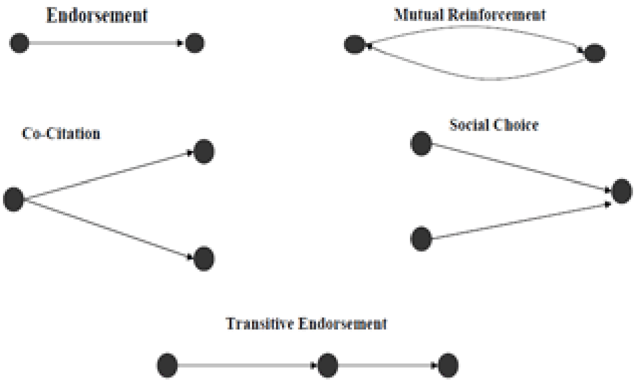

Fig. 3 :

Directed Path: A sequence of links, starting from p that can be followed to reach q.

Shortest Path: Of all the paths between nodes p and q, which has the shortest length, i.e. number of links on it. **Diameter:** The maximum of all the shortest paths between a pair of nodes p and q, for all pairs of nodes p and q in the Web-graph.

C. Hubs and Authorities

Hyperlink-induced topic search (HITS) is an iterative algorithm for mining the Web graph to identify topic hubs (pages with good sources of content) and authorities (pages with good sources of links). “Authorities” are highly ranked pages for a given topic; “hubs” are pages with links to authorities. The algorithm takes as input search results returned by traditional text indexing techniques, and filters these results to identify hubs and authorities. The number and weight of hubs pointing to a page determine the page's authority. The algorithm assigns weight to a hub based on the authoritativeness of the pages it points to. For example, a page containing links to all authoritative news servers (CNN, CNBC, and so on) is a powerful news hub.

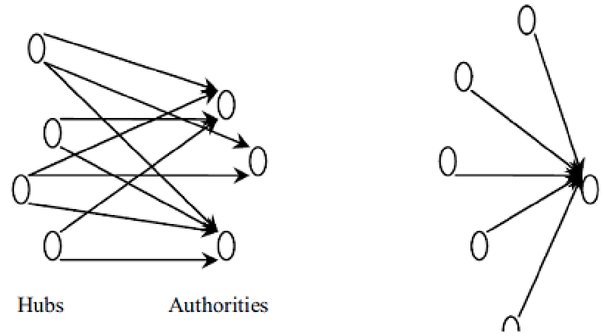

Fig. 4

According to Kleinberg, "Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs".

IV. Web Usage Mining

Web usage mining is the type of Web mining activity that involves the automatic discovery of user access patterns from one or more Web servers. As more organizations rely on the Internet and the World Wide Web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context. Organizations often generate and collect large volumes of data in their daily operations. Most of this information is usually generated automatically by Web servers and collected in server access logs. Other sources of user information include referrer logs, which contain information about the referring pages for each page reference, and user registration or survey data gathered via tools such as CGI scripts. Analyzing such data can help these organizations to determine the life time value of customers, cross marketing strategies across products, and effectiveness of promotional campaigns, among other things. Analysis of server access logs and user registration data can also provide valuable information on how to better structure a Web site in order to create a more effective presence for the organization. In organizations using intranet technologies, such analysis can shed light on more effective management of workgroup communication and organizational infrastructure. Finally, for organizations that sell advertising on the World Wide Web, analyzing user access patterns helps in targeting ads to specific groups of users. Most of the existing Web analysis tools provide mechanisms for reporting user activity in the servers and various forms of data filtering. Using such tools, for example, it is possible to determine the number of accesses to the server and the individual files within the organization's Web space, the times or time intervals of visits, and domain names and the URLs of users of the Web server. However, in general, these tools are designed to deal handle low to moderate traffic servers, and furthermore, they usually provide little or no analysis of data relationships among the accessed files and directories within the Web space.

A. Web Usage Mining Architecture

While extracting simple information from web logs is easy, mining complex structural information is very challenging. Data cleaning and preparation constitute a very significant effort before mining can even be applied. The relevant data challenges include: elimination of irrelevant information such as image files and cgi scripts, user identification, user session formation, and incorporating temporal windows in the user modeling. After all this pre-processing, one is ready to mine the resulting database. We have developed a general architecture for Web usage mining. The WEBMINER is a system that implements parts of this general architecture. The architecture divides the Web usage mining process into two main parts. The first part includes the domain dependent processes of transforming the Web data into suitable transaction form. This includes pre-processing, transaction identification, and data integration components. The second part includes the largely domain independent application of generic data mining and pattern matching techniques (such as the discovery of association rule and sequential patterns) as part of the system's data mining engine. The overall architecture for the Web mining process is depicted in fig. 5 below.

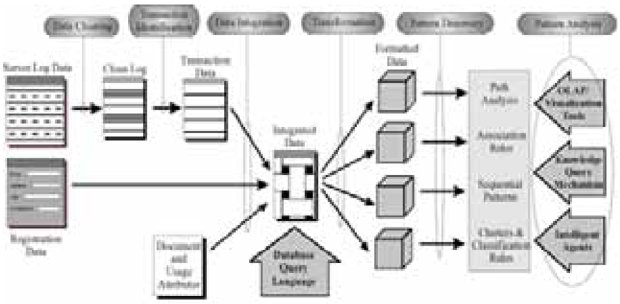

Fig. 5 :

Data cleaning is the first step performed in the Web usage mining process. Any of the cleaning techniques can be used to pre-process a given Web server log. Currently, the WEBMINER system uses the simplistic method of checking filename suffixes. Some low-level data integration tasks may also be performed at this stage, such as combining multiple logs, incorporating referrer logs, etc. After the data cleaning, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules. The clean server log can be thought of in two ways; either as a single transaction of many page references, or a set of many transactions each consisting of a single page reference. The goal of transaction identification is to create meaningful clusters of references for each user. Therefore, the task of identifying transactions is one of either dividing a large transaction into multiple smaller ones or merging small transactions into fewer larger ones. This process can be extended into multiple steps of merge or divide in order to create transactions appropriate for a given data mining task. A transaction identification module can be defined as either a merge or a divide module. Both types of modules take a transaction list and possibly some parameters as input, and output a transaction list that has been operated on by the function in the module in the same format as the input. The requirement that the input and output transaction format match allows any number of modules to be combined in any order, as the data analyst sees fit. The WEBMINER system currently has reference length, maximal forward reference, and time window divide modules, and a time window merge module. Access log data may not be the only source of data for the Web mining process. User registration data, for example, is playing an increasingly important role, particularly as more security and privacy conscious client-side applications restrict server access to a variety of information, such as the client user IDs. The data collected through user registration must then be integrated with the access log data. There are also known or discovered attributes of references pages that could be integrated into a higher level database schema. Such attributes could include page types, classification, usage frequency, page meta information, and link structures. While WEBMINER currently does not incorporate user registration data, various data integration issues are being explored in the context of Web usage mining. Once the domain-dependent data transformation phase is completed, the resulting transaction data must be formatted to conform to the data model of the appropriate data-mining task. For instance, the format of the data for the association rule discovery task may be different than the format necessary for mining sequential patterns.

Finally, a query mechanism will allow the user (analyst) to provide more control over the discovery process by specifying various constraints. The emerging data mining tools and systems lead naturally to the demand for a powerful data mining query language, on top of which many interactive and flexible graphical user interfaces can be developed. Such a query mechanism can provide user control over the data mining process and allow the user to extract only relevant and useful rules. In WEBMINER, a simple Query mechanism has been implemented by adding some primitives to an SQL-like language. This allows the user to provide guidance to the mining engine by specifying the patterns of interest.

V. Conclusion

As the web and its usage continues to grow, so too grows the opportunity to analyze web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper we survey the research area of Web mining, focusing on the category of Web mining. Since this is a huge area, and there a lot of work to do, we hope this paper could be a useful starting point for identifying opportunities for further research

References

[1] M. Agosti, G.M. Di Nunzio, A. Niero, “From Web Log Analysis to Web User Profiling”, In DELOS Conference 2007. Working Notes. Pisa, Italy, 2007, pp 121–132.

[2] B. Berendt, B. Mobasher, M. Nakagawa, M. Spiliopoulou, “The Impact of Site Structure and User Environment on Session Re- construction in Web Usage Analysis”, WEBKDD 2002, LNAI 2703, pp 159-179, 2003.

[3] C. W. Cleverdon, “The Cranfield Tests on Index Languages Devices”. In Readings in Information Retrieval, Morgan Kaufmann Publisher, Inc., San Francisco, California, pp.4760, 1997.

[4] F.M. Facca, P.L. Lanzi, “Mining interesting knowledge from Weblogs: a survey”, Data and Knowledge Engineering Vol. 53, No. 3, June 2005, pp 225-241.

[5] P.M. Hallam-Baker, B. Behlendorf, “Extended Log File Format, W3C Working Draft WD-logfile-960323”, [Online] Available : http://www.w3.org/TR/WD-logfile.html.

[6] D. Nicholas, P. Huntington, A. Watkinson, “Scholarly journal usage: the results of deep log analysis”, Journal of Documentation Vol 61 no2 2005