Web Miining: Summary

[1,]Sonia Gupta , [2,]Neha Singh
[1,]*Computer Science Department Iftm University*
Moradabad(India)

Abstract

World Wide Web is a very fertile area for data mining research, with huge amount of information available on it. From its very beginning, the potential of extracting valuable knowledge from the Web has been quite evident. The term Web mining has been used in two different ways. The first, called Web content mining and the second, called Web usage mining. The web content mining is the process of information discovery from sources across the World Wide Web. Web usage mining is the process of mining for user browsing and access patterns. Interest in Web mining has grown rapidly in its short existence, both in the research and practitioner communities.

Keywords: *Web mining, information retrieval, information extraction*

I. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from Web data - including Web documents, hyperlinks between documents, usage logs of web sites, etc. With more than two billion pages created by millions of Web page authors and organizations, the World Wide Web is a tremendously rich knowledge base. The knowledge comes not only from the content of the pages themselves, but also from the unique characteristics of the Web, such as its hyperlink structure and its diversity of content and languages.

Web mining research overlaps substantially with other areas, including data mining, text mining, information retrieval, and Web retrieval. The classification is based on two aspects: the purpose and the data sources. ***Retrieval*** research focuses on retrieving relevant, existing data or documents from a large database or document repository, while *mining* research focuses on discovering new information or knowledge in the data. Web retrieval and Web mining share many similarities. Web document clustering has been studied both in the context of Web retrieval and of Web mining. Web mining is not simply the application of information retrieval and text mining techniques to Web pages; it also involves non textual data such as Web server logs and other transaction-based data.

Table 1.1 A classification of retrieval and mining techniques and applications

Purpose/Data information sources	Any data	Textual data	Web-based data
Retrieving known data or documents efficiently and effectively	Data Retrieval	Information Retrieval	Web Retrieval
Finding new patterns or knowledge previously Unknown	Data Mining	Text Mining	Web Mining

It is also interesting to note that, although Web mining relies heavily on data mining and text mining techniques, not all techniques applied to Web mining are based on data mining or text mining.

Web mining is the application of data mining techniques to extract knowledge from Web data, where **at least one of structure (hyperlink) or usage (Web log) data is used in the mining process** (with or without other types of Web data).

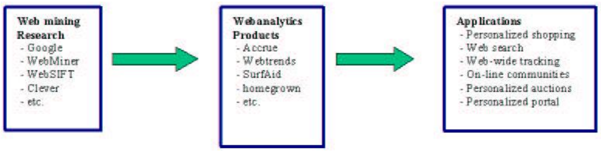


Figure 1.1: Web mining research & applications

Web mining technique could be used to solve the information overload problems above directly or indirectly. However, we could not claim that Web mining techniques are only tools to solve these problem. Other techniques and works from different research areas such as database(DB), information retrieval(IR), natural language processing(NLP), and the web document community, could also be used.

II. WEB MINING TAXONOMY

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined:

[1] **Web Content Mining:** Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. While there exists a significant body of work in extracting knowledge from images - in the fields of image processing and computer vision - the application of these techniques to Web content mining has not been very rapid.

[2] **Web Structure Mining:** The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be further divided into two kinds based on the kind of structural data used.

- **Hyperlinks:** A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to a different Web page. A hyperlink that connects to a different part of the same page is called an *Intra-Document Hyperlink*, and a hyperlink that connects two different pages is called an *Inter-Document Hyperlink.*

- **Document Structure:** In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page.

Web Usage Mining: Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications.

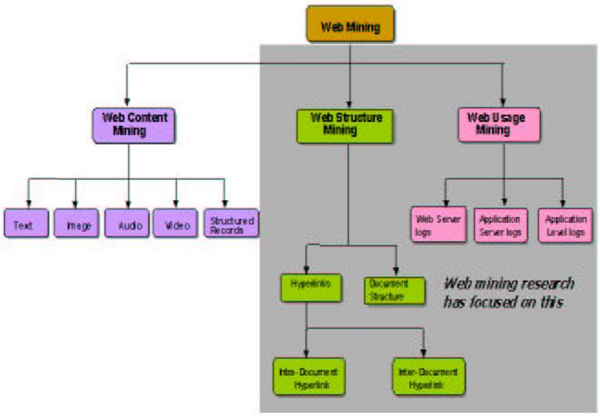


Figure 2.1: Web mining Taxonomy

III. WEB MINING

3.1 OVERVIEW

Web mining is the use of data mining technique to automatically discover and extract information from Web documents and services. Web mining is decomposed into subtasks such as:

[1] **Resource finding**: the task of retrieving intended Web document.
[2] **Information selection and pre-processing:** automatically 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.

By resource finding we mean retrieving the data that is either online or offline from the text sources that is available on the Web such as electronic newsletter, electronic newswires, newsgroups, the text content of HTML document obtained by removing HTML tags and also manual selection of Web resources. The information selection and pre-processing step is any kind of transformation process of any kind of original data retrieving in IR process. Thus, Web mining refers to overall process of discovering potentially useful and previously unknown information or knowledge from the Web data.

3.2 WEB MINING AND INFORMATION RETREIVAL

Information Retrieval is automatic retrieval of all relevant documents while at same time retrieving as few of the non-relevant as possible. Information retrieval has the primary goals of indexing the text and searching for useful documents in a collection and nowadays in research. Information retrieval includes modeling, document classification and categorization, user interfaces, data visualization, filtering etc. Web mining is the part of Information Retrieval process.

3.3 WEB MINING AND INFORMATION EXTRACTION

Information Extraction has the goal of transforming a collection of documents, usually with the help of Information Retrieval system into information that is more readily digested and analyzed. Building Information Extraction system manually is not feasible and salable for such a dynamic and diverse medium such as Web contents. Due to this nature of Web, most information Extraction system focuses on the specific Websites to extract. Others use machine learning or data mining techniques to learn extraction patterns or rules for Web document semi-automatically or automatically. Web mining is the part of Information Extraction.There are basically two types of Information Extraction: Information Extraction from unstructured text and Information Extraction from semi-structured data.

3.4 Web Mining and Machine Learning Applied on the Web

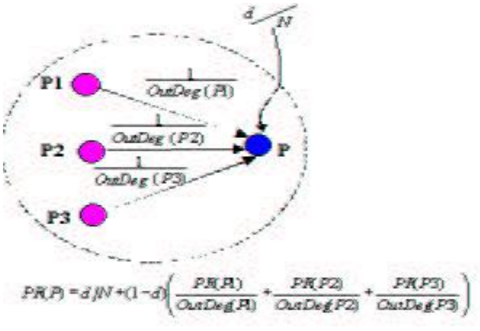
Web mining is not the same as learning from the Web or machine learning techniques applied from the Web.

IV. KEY ACCOMPLISHMENTS

We briefly describe the key new concepts introduced by the Web mining research community.

4.1 Page ranking metrics - Google's Page Rank function

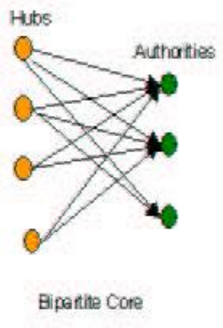
Page Rank is a metric for ranking hypertext documents that determines the quality of these documents. The key idea is that a page has high rank if it is pointed to by many highly ranked pages.


d/N
P1
$\frac{1}{OutDeg(P1)}$
P2
$\frac{1}{OutDeg(P2)}$
$\frac{1}{OutDeg(P3)}$
P3
P
$PR(P) = d/N + (1-d)\left(\frac{PR(P1)}{OutDeg(P1)} + \frac{PR(P2)}{OutDeg(P2)} + \frac{PR(P3)}{OutDeg(P3)}\right)$

Page rank

4.2 Hubs and Authorities - Identifying significant pages in the Web

Hubs and Authorities can be viewed as ‘fans’ and ‘centers’ in a bipartite core of a Web graph. A Core (i, j) is a complete directed bipartite sub-graph with at least i nodes from F and at least j nodes from C. With reference to the Web graph, i pages that contain the links are referred to as ‘fans’ and the j pages that are referenced are the ‘centers’. From a conceptual point of view ‘fans’ and ‘centers’ in a Bipartite Core are basically the Hubs and Authorities. The hub and authority scores computed for each Web page indicate the extent to which the Web page serves as a “hub” pointing to good “authority” pages or as an “authority” on a topic pointed to by good hubs. First a query is submitted to a search engine and a set of relevant documents is retrieved. This set, called the ‘root set’, is then expanded by including Web pages that point to those in the ‘root set’ and are pointed by those in the ‘root set’.



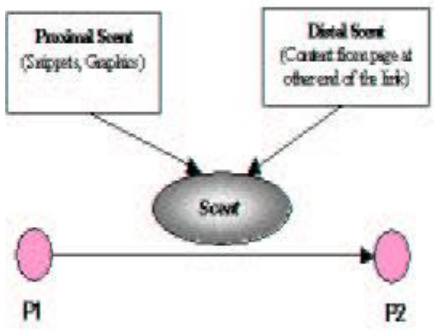
Hubs and authorities

4.3 Robot Detection and Filtering - Separating human and non human Web behavior

Web robots are software programs that automatically traverse the hyperlink structure of the World Wide Web in order to locate and retrieve information. The importance of separating robot behavior from human behavior prior to extracting user behavior knowledge from usage data. First of all, e-commerce retailers are particularly concerned about the unauthorized deployment of robots for gathering business intelligence at their Web sites. In addition, Web robots tend to consume considerable network bandwidth at the expense of other users. Sessions due to Web robots also make it more difficult to perform click-stream analysis effectively on the Web data. Conventional techniques for detecting Web robots are often based on identifying the IP address and user agent of the Web clients. While these techniques are applicable to many well-known robots, they may not be sufficient to detect camouflaging and previously unknown robots.

4.4 Information scent - Applying foraging theory to browsing behavior

Information scent is a concept that uses the snippets and information presented around the links in a page as a “scent” to evaluate the quality of content of the page it points to and the cost to access such a page. The key idea is a user at a given page “ foraging” for information would follow a link with a stronger “scent”. The “scent” of the pages will decrease along a path and is determined by network flow algorithm called spreading activation.



Information scent

4.5 User profiles - Understanding how users behave

The Web has taken user profiling to completely new levels. For example, in a 'brickand-mortar' store, data collection happens only at the checkout counter, usually called the 'point-of-sale'. This provides information only about the final outcome of a complex human decision making process, with no direct information about the process itself. In an on-line store, the complete click-stream is recorded, which provides a detailed record of every single action taken by the user - which can provide much more detailed insight into the decision making process.

4.6 Interestingness measures

When multiple sources provide conflicting evidence One of the significant impacts of publishing on the Web has been the close interaction now possible between authors and their readers. In the pre-Web era, a reader's level of interest in published material had to be inferred from indirect measures such as buying/borrowing, library checkout/ renewal, opinion surveys, and in rare cases feedback on the content. For material published on the Web it is possible to track the precise click-stream of a reader to observe the exact path taken through on-line published material, with exact times spent on each page, the specific link taken to arrive at a page and to leave it, etc. Much more accurate inferences about readers' interest about published content can be drawn from these observations. Mining the user click-stream for user behavior, and use it to adapt the 'look-and-feel' of a site to a reader's needs.

V. FUTURE DIRECTIONS

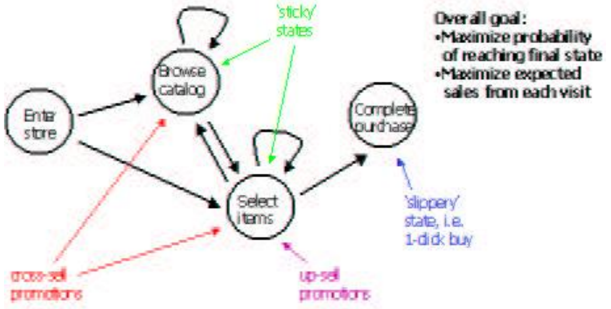
As the Web and its usage grows, it will continue to generate evermore content, structure, and usage data, and the value of Web mining will keep increasing. Outlined here are some research directions that must be pursued to ensure that we continue to develop Web mining technologies that will enable this value to be realized.

5.1 Process mining

Mining of 'market basket' data, collected at the point-of-sale in any store, has been one of the visible successes of data mining. However, this data provides only the end result of the process, and that too decisions that ended up in product purchase. Research needs to be carried out in (i) extracting process models from usage data, (ii) understanding how different parts of the process model impact various Web metrics of interest, and (iii) how the process models change in response to various changes that are made - changing stimuli to the user.

5.2 Temporal evolution of the Web

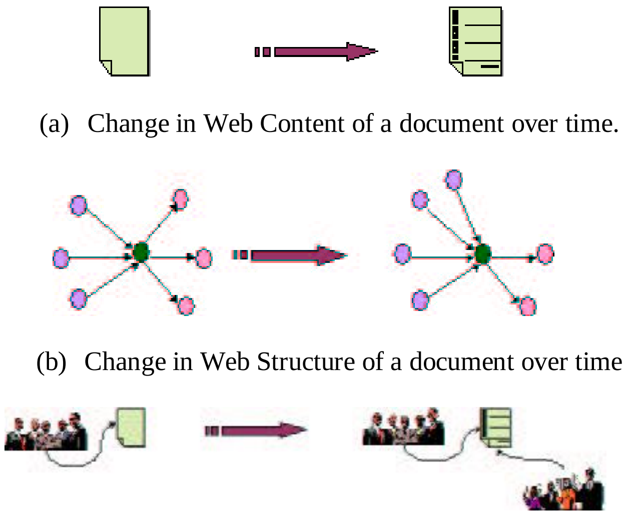
While storing the history of all of this interaction in one place is clearly too staggering a task, at least the changes to the Web are being recorded by the pioneering. Research needs to be carried out in extracting temporal models of how Web content, Web structures, Web communities, authorities, hubs, etc. are evolving. Large organizations generally archive (at least portions of) usage data from there Web sites. The temporal behavior of the three kinds of Web data: Web Content, Web Structure and Web Usage.



Shopping Pipeline modeled as State Transition Diagram

5.3 Fraud and threat analysis

The anonymity provided by theWeb has led to a significant increase in attempted fraud, from unauthorized use of individual credit cards to hacking into credit card databases for blackmail purposes. Since all these frauds are being perpetrated through the Internet, Web mining is the perfect analysis technique for detecting and preventing them. Research issues include developing techniques to recognize known frauds, and characterize and then recognize unknown or novel frauds, etc. The issues in cyber threat analysis and intrusion detection are quite similar in nature.



(a) Change in Web Content of a document over time.

(b) Change in Web Structure of a document over time

(c) Change in Web Usage of a document over time

VI. CONCLUSION

As the Web and its usage continues to grow, so 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 have briefly described the key computer science contributions made by the field, the prominent successful applications, and outlined some promising areas of future research. Our hope is that this overview provides a starting point for fruitful discussion.

REFRENCES

[1] J. Borges and M. Levene. Mining Association Rules in Hypertext Databases. InKnowledge Discovery and Data Mining, pages 149–153, 1998.

[2] K. Bollacker, S. Lawrence, and C.L. Giles. CiteSeer: An autonomous webagent for automatic retrieval and identification of interesting publications. In Katia P. Sycara and Michael Wooldridge, editors, Proceedings of the Second International Conference on Autonomous Agents, pages 116–123, New York, 1998. ACM Press.

[3] E.H. Chi, P. Pirolli, K. Chen, and J.E. Pitkow. Using Information Scent to model user information needs and actions and the Web. In Proceedings of CHI 2001, pages 490–497, 2001.

[4] S.Abiteboul. Querying semi- structured data. In F. N. Afrati and P. Kolaitis , editors database theory-ICDT '97, 6th International Conference, Delphi Greece, January 8-10, 1997, proceedings, volume 1186 of lecture note in computer science, pages 1-18. Springer, 1997.

[5] S.Abiteboul, D.Quass, J. Mchugh, J. Widom and J. L. Wiener. The lorel Query language for semi-structured data. Int . J. on Digital Libraries 1(1):68-88, 1997.

[6] E.H. Chi, J. Pitkow, J. Mackinlay, P. Pirolli, R. Gossweiler, and S.K. Card. Visualizingthe evolution of web ecologies. In Proceedings of the Conference on Human Factors in Computing Systems CHI'98, 1998.

[7] E. Colet. Using Data Mining to Detect Fraud in Auctions, 2002.

[8] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. PhD thesis, University of Minnesota, 2000.

[9] K.L. Ong and W. Keong. Mining Relationship Graphs for Eective Business Objectives.

[10] B. Prasetyo, I. Pramudiono, K. Takahashi, M. Toyoda, and M. Kitsuregawa. Naviz user behavior visualization of dynamic page.