International Journal of Computer Science and Informatics

Volume 2 | Issue 4 Article 1

April 2013

A Lime Light on the Emerging Trends of Web Mining

Udayasri. B
Dept. of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysore, Karnataka, India, udaya.sri84@gmail.com

Sushmitha. N
Dept. of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysore, Karnataka, India, sush2007n@gmail.com

Padmavathi. S
Dept. of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysore, Karnataka, India, padmavathi.030@gmail.com

Recommended Citation
B, Udayasri.; N, Sushmitha.; and S, Padmavathi. (2013) "A Lime Light on the Emerging Trends of Web Mining," *International Journal of Computer Science and Informatics*: Vol. 2 : Iss. 4 , Article 1.
DOI: 10.47893/IJCSI.2013.1096
Available at: https://www.interscience.in/ijcsi/vol2/iss4/1

A Lime Light on the Emerging Trends of Web Mining

Udayasri.B, Sushmitha.N, Padmavathi.S

Dept. of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysore, Karnataka, India
E-mail : udaya.sri84@gmail.com, sush2007n@gmail.com, padmavathi.030@gmail.com

Abstract **-** The World Wide Web is a huge, information center for a variety of applications. Web contains a dynamic and rich collection of hyperlink information. It allows Web page access, usage of information and provides numerous sources for data mining. The goal of Web mining is to discover the pattern of access and hidden information from huge collections of documents.

In this paper we are presenting the various emerging web mining techniques that are effectively efficient in overcoming the demerits of existing technologies and also give the superficial knowledge and comparison about data mining. This paper describes the past, current and future of web mining. Web mining attempts to determine useful knowledge from secondary data obtained from the interactions of the users with the web. We have also described the personalization on web which is used to manipulate the information presented to the users through the various personalization strategies.

Key words **-** *Web Mining; Web Content Mining; Web Structure Mining; Web Usage Mining.*

I. INTRODUCTION

In recent years, internet is expanding very rapidly in the world. It has penetrated into every areas of society and has also become a huge, pervasive distribution and global information service center. Existing commercial systems seek to do some minimal personalization based on declarative information directly provided by the user, such as their zip code or keywords describing their interests or specific URLs or even particular pieces of information they are interested in (e.g. price for a particular stock). Our research aims at creating systems that automatically tailors the content delivered to the user from a website. We do so by mining the web - both the contents, as well as the user's interaction. In the World Wide Web, documents of all sorts of formats, contents and descriptions are collected and interconnected with hyperlinks making it the largest repository of data. Despite of its dynamic and unstructured nature, its heterogeneous characteristic and it's very often redundancy and inconsistency, the World Wide Web is the most important data collection regularly used for reference because of the broad variety of topics covered and the infinite contributions of resources and publishers. The Web is becoming a central part of social, cultural, political, education, academic and commercial life and contains a wide range of information and applications. Competitive trends in modern society require a large number of Internet appearances and produce real-time information with in-depth analysis. Combine the traditional Data mining with the Web to carry out Web mining. The Web documents and Web activities are used to extract the information that the users are interested in, hidden useful patterns and hidden information. Web mining techniques are the results of long process of research and product development. This paper spills the lime light on the benefits the web mining technology offered to the present day world.

II. DATA MINING

The process of extracting previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions can take on different approaches depending on the type of data involved and the desired objectives. Data Mining consists of three components: the captured data, which must be integrated into organization-wide views, often in a Data Warehouse; the mining of this warehouse; and the organization and presentation of this mined information to enable understanding.

1. Data selection :

Domain experts understand the meaning of the metadata. They collect, describe and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital. In the data exploration phase, traditional data analysis tools, for example, statistics are used to explore the data.

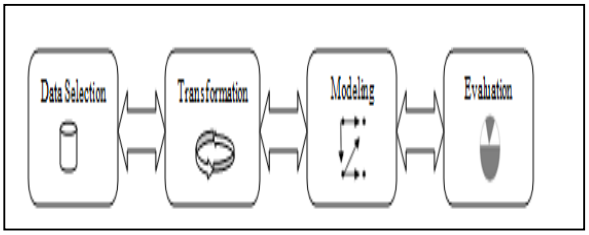

Fig. 1 : Data Mining Process

2. *Transformation:*

It can also be known as data consolidation and it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

3. *Modeling:*

Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model.

4. *Evaluation:*

Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions:

1. Does the model achieve the business objective?
2. Have all business issues been considered?

At the end of the evaluation phase, the data-mining experts decide how to use the data mining results.

Drawbacks in the existing approaches:

1. The response time perceived by the user is too long.
2. The explosive growth of the Web has imposed a heavy demand on networking
3. Hence, an obvious solution in order to improve the quality of Web services would be to increase the bandwidth, but such a choice involves increasing economic cost.
4. Web caching scheme has three significant drawbacks: If the proxy is not properly updated, a user might receive stale data and as the number of users grow, the original servers typically become bottlenecks.
5. Main drawback of systems which have enhanced pre-fetching policies is that some pre-fetched objects may not be eventually requested by the users. In such a case, the pre-fetching scheme increases the network traffic as well as the Web server's load.

III. WEB MINING

Web Mining is based on the knowledge discovery from web. It will extract the knowledge framework and represents it in a proper way. Web mining is useful to extract the information, image, text, audio, video, documents and multimedia. Before the dawn of web mining it was difficult o extract information in proper way from web. But with the advent of web mining it became easy to extract all the features and information about multimedia. Web mining is the application of data mining techniques to discover patterns from the Web.

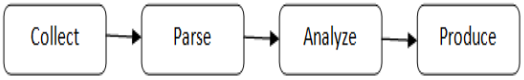

Fig. 2 : Steps of Web Mining

Steps in Web Mining:

- Collect - fetch content from web
- Parse - extract data from formats
- Analyze - tokenize, rate, classify, cluster
- Produce - useful data

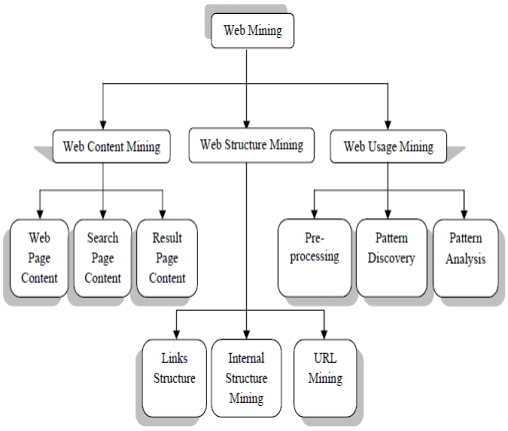

Fig. 3 : Classification of Web Mining

Web mining can be categorized into three areas of interest based on which part of the web to mine:

a) Web Content Mining

b) Web Structure Mining

c) Web Usage Mining

a) Web Content Mining:

Web content mining is related but different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in Web content mining. It is related to text mining because much of the web contents are texts. However, it is also quite different from data mining because Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data. Web content mining is also different from text mining because of the semi-structure nature of the Web, while text mining focuses on unstructured texts. Web content mining thus requires creative applications of data mining and/or text mining techniques and also its own unique approaches.

The various contents of Web Content Mining are

- Web page
- Search page
- Result page

Web Page: A Web page typically contains a mixture of many kinds of information, e.g., main content, advertisements, navigation panels, copyright notices, etc. For a particular application only some part of the information is useful and the rest are noises.

Search Page: A search page is typically used to search a particular Web page of the site, to be accessed numerous times in relevance to search queries. The clustering and organization of Web content in a content database enables effective navigation of the pages by the customer and search engines.

Result page: A result page typically contains the results, the web pages visited and the definition of last accurate result in the result pages of content mining.

b) Web Structure Mining:

It derives information and knowledge mainly from the Web and the links between the organizational structures. Based on scientific citation analysis theory, the interconnection between the data in the document contains a wealth of useful information. The usual search engines consider only the Web as a flat collection of documents because of taking into account the complexity of the structure, ignoring the structure of information. Mining of structure and Web page structure, guides the classification and clustering of pages to find authoritative pages, center pages, to improve retrieval performance. Web page also can be used to guide the collection work to improve collection efficiency.

The various contents of Web structure mining are

- Links Structure Mining
- Internal Structure Mining
- URL Mining

Links Structure: Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts have resulted in a newly emerging research area called Link Mining. It consists Link-based Classification, Link-based Cluster Analysis, Link Type, Link Strength and Link Cardinality.

Internal Structure Mining: It can provide information about page ranking or authoritativeness and enhance search results through filtering i.e., tries to discover the model underlying the link structures of the web. This model is used to analyze the similarity and relationship between different web sites.

URL Mining: It gives a hyperlink which is a structural unit that connects a web page to different location, either within the same web page (intra_document hyperlink) or to a different web page (inter_document) hyperlink.

c) Web Usage Mining:

It focuses on techniques that could predict user behavior while the user interacts with the web and also it discovers the meaningful pattern from data generated by client server transaction on one or more web localities. A web is a collection of interrelated files on one or more web servers. Web usage mining aims at utilization of data mining techniques to discover the usage patterns from web based application. It automatically generates the data for the server access logs, refers logs, agent logs, client sides cookies, user profiles, metadata, page attributes, page contents & site structures. It is a technique to predict user behavior when the user interacts with the web.

Web usage mining is categorized into three phases:

- Preprocessing
- Pattern Discovery
- Pattern Analysis

Preprocessing : According to the client, server and proxy server, the preprocessing is the first approach to retrieve the raw data from web resources and process the data. It automatically transforms the original raw data to the next process.

Pattern Discovery : According to the data preprocessing, the raw data is used to discover the knowledge and to implement the techniques which will be used for machine learning. This makes use of data mining procedures.

Pattern Analysis : It is the process after pattern discovery. It checks whether the pattern is correct on the web and guides the process of extraction of the information/ knowledge from the web.

IV. BENEFITS, LIABILITIES OF WEB MINING AND ITS APPLICATIONS

There are many benefits that can be obtained through the applications of web mining technology.

Application areas of Web Mining:

- E-Commerce
- Search Engines
- Personalization
- Website Design

How Web mining is different from classical data mining?

Web mining:

- The web is a collection of interrelated files, even though it is not a relation.
- Web mining is the discovery of knowledge from the web.
- Usage data is huge and growing rapidly.
- Ability to react in real-time usage of patterns.

Data mining:

- Textual information and linkage structure.
- Google's usage logs are bigger than their web crawl.
- Data generated per day is comparable to largest conventional data warehouse.
- No human in the loop.

The attention paid to Web mining, in research, software industry and Web-based organizations, has led to the accumulation of lot of experiences. It is our attempt in this paper to capture them in a systematic manner and identify the directions for future research.

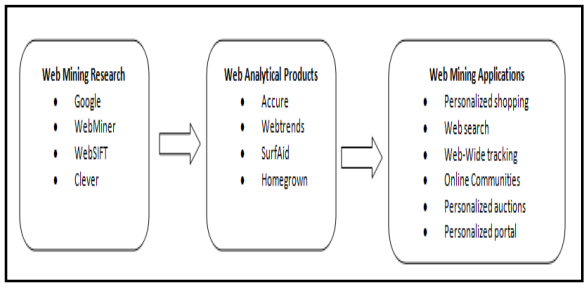

Fig. 4 : Web mining research & applications

V. PERSONALIZATON ON THE WEB

Web personalization is a strategy, a marketing tool, and an art. Personalization requires implicit or explicit collection of visitor information and leveraging that knowledge in the content delivery framework to manipulate what information has been presented to the users and how is it presented. Web personalization can be seen as an interdisciplinary field that includes several research domains from user modeling, social networks, web data mining, human-machine interactions to Web usage mining. Web usage mining is an example of approach to extract log files containing information on user navigation in order to classify users. Personalization process has been enriched at the semantic level, based on user modeling and on log files analysis.

i. Personalization Strategies:

Personalization falls into four basic categories, ordered from the simplest to the most advanced:

Memorization: It is the simplest and most widespread form of personalization. The user information such as name and browsing history is stored (e.g. using cookies), which will be used later to recognize and greet the returning user. It is usually implemented on the Web server.

Customization: This form of personalization takes the user's preferences as input from the registration forms in order to customize the content and structure of a web page.

Guidance or Recommender Systems: A guidance based system tries to automatically recommend hyperlinks that are deemed to be relevant to the user's interests, in order to facilitate access to the needed information on a large website.

Task Performance Support: In these client side personalization systems, a personal assistant executes actions on behalf of the user, in order to facilitate access to relevant information.

ii. The Web personalization process can be divided into four distinct phases:

Collection of Web data: Implicit data includes past activities/click streams as recorded in Web server logs and/or via cookies or session tracking modules. Explicit data usually comes from the registration forms and rating questionnaires.

Preprocessing of Web data: 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. It includes cleaning of data inconsistencies, filtering out irrelevant information according to the goal of analysis.

Analysis of Web data: This step applies machine learning or Data Mining techniques to discover interesting usage patterns and statistical correlations between web pages and user groups.

Decision making/Final Recommendation Phase: This phase makes use of the results of the previous analysis step to deliver recommendations to the users. The recommendation process typically involves generating dynamic Web content on the fly, such as adding hyperlinks to the last web page requested by the users.

VI. CONCLUSION & FUTURE WORK

Web data is growing at a significant rate. Web mining is a fertile area of research with many successful applications. As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract useful knowledge from it. To extract the specific data from web warehouse, the three categories (Web Content Mining, Web Structure Mining and Web Usage Mining) of web mining play a major role. Web mining is one of the most important applications of data mining. It is having its own benefits and successful applications with which we can overcome the problems or difficulties faced in data mining. In this paper a clear picture of how web mining is efficiently different from data mining has been highlighted. As the usage of the internet in the present day is growing in faster rate, the personalization process of the web mining provides us a great opportunity of maximizing the efficient usage of the internet.

Cloud mining is a new approach to apply data mining to the customer data by using web mining process. By a cloud, we mean an infrastructure that provides resources and/or services over the Internet. In fact among all the potential use of web mining in future, the growing online shopping activities, e-services industry and e-commerce are important domains. Hence the future of the web mining can be seen in the field of Cloud Mining.

ACKNOWLEDGEMENT

The authors gratefully acknowledge Dr. Mahesh Rao, HOD, Dept of Computer Science and Engineering, Vidyavardhaka College of Engineering Mysore, Karnataka and Dr. B. Sadashivegowda, Principal, Vidyavardhaka College of Engineering, Mysore, Karnataka for their invaluable and continuous support.

REFERENCES

[1] F. Masseglia, P. Poncelet, and R. Cicchetti, "An Efficient Algorithm for Web Usage Mining", Networking and Information Systems Journal (NIS), vol.2, no. 5-6, pp. 571-603, 1999.

[2] J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, vol. I, no. 2, pp. 12-23, 2000.

[3] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, SIGKDD Explorations, Copyright 2000 ACM SIGKDD, July 2000.

[4] Michael Jennings," What are the major comparisons or differences between Web mining and data mining?" Information Management Online, June 25, 2002.

[5] Ying Zhang, "The study of web data mining in EB," the excellent collection of magisterial and doctoral thesis, May 2007.

[6] Sravan Kumar, D. and Naveena Devi, B. "Learner's Centric Approach for Web Mining " et al. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1(2), 2010.

[7] Kavita Sharma, Gulshan Shrivastava, Vikas Kumar, "Web Mining: Today and Tomorrow".

[8] B.N.Laxmi, G.H Raghunandhan, "A C o n c e ptual Overview of Data Mining", Proceedings of the National Conference on Innovations in Emerging Technology-2011 Kongu Engineering College, Perundurai, Erode, Tamilnadu, India.17 & 18 February, 2011.pp.27-*32*.

[9] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining - Concepts, Applications & Research Directions and Web Mining" – Accomplishments & Future Directions

[10] A.Jebaraj Ratnakumar, "An Implementation Of Web Personalization Using Web Mining Techniques", Journal of Theoretical and Applied Information Technology © 2005 - 2010 JATIT. All rights reserved.

[11] Manoj Pandia, Subhendu Kumar Pani, Sanjay Kumar Padhi, Lingaraj Panigrahy, R.Ramakrishna, "A Review Of Trends In Research On Web Mining", A Review Of Trends In Research On Web Minin, International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011.

[12] Tan, Steinbach, Kumar, "Introduction to Data Mining"

[13] Introduction to Data Mining and Knowledge Discovery, Third Edition by Two Crows Corporation.

[14] Michael J. A. Berry, Gordon S. Linoff, Mining the Web: Transforming Customer Data, J. Wiley, 2002.

[15] Eirinaki, M., Vazirgiannis, M. (2003) "Web Mining for Web Personalization", ACM Transactions on Internet Technology, Vol.3, No.1, February 2003.

◈◈◈