A Detailed Survey on Web Mining

A.Ashikali[1], B.Loganathan[2]
M.Phil. Research Scholar[1]
PG & Research Department of Computer Science
Government Arts College (Autonomous), Coimbatore-18.
Associate Professor & Head[2]
PG & Research Department of Information Technology
Government Arts College (Autonomous), Coimbatore-18.

Abstract: **In this paper the survey is conducted regarding World Wide Web which is a fertile area for data mining and research. The web mining is one of the data mining domains where data mining techniques are used for extracting information from the web servers. Web mining is universal set of web content mining, web structure mining and web usage mining. This paper summarises various techniques of web mining like feature extraction, transformation and representation of different application domains.**

Keywords: **Web mining, web usage mining, web structure mining, and web content mining.**

INTRODUCTION

The expansion of the World Wide Web has result in enormous data that is now in General freely available for user access.[4] Web Mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web.[7] The emerging field of web mining aims at finding and extracting relevant information that is hidden. The related data in particular is text document published on the web. Web data is continuously updating at every moment. So it is observable that the data or the web page which is gained by the user will be gained another time in the other structure and disorder. Analysis and Discovery of useful information from World Wide Web poses a phenomenal challenge to the researcher.

WEB MINING

Overview: Web mining process is divided into four steps namely resource finding, selection of data, performing pre-processing, generalization and analysis [8] [2]. Data mining is commonly defined as the process of discovering useful pattern or Knowledge from data source e.g., databases, texts, images, the web, etc...

Web mining and Information Retrieval: The resource or document discovery (IR) on the web is an instance of web Content mining and web mining is also associated with intelligent Information Retrieval. Actually Information Retrieval is the automatic retrieval of all applicable documents while reducing the retrieval of non-relevant document as less (few) as possible [13].

The web document classification used for indexing is a task which is taken as an instance of web mining.

Web Mining and Information Extractions: Information Extraction has the goal of transforming a collection of documents, usually with the help of an Information Retrieval system into information that is more readily digested and analyzed [5]. Information Extraction aims to extract relevant facts from the document while IR aims to select relevant document [11].

While IE is interested in the structure or representation of a document, Information Retrieval views the text in a document just as a case of unordered words [14].

Thus in general IE works at a finer granularity level than Information Retrieval does on the document.

Web Mining and Machine Learning Applied on the web page: Web mining is not the same as learning beginning the web or machine learning technique applied on the web. On the one hand, there are some applications of machine learning applied on the web that are not instance of web mining.

Some examples are some proprietary algorithms that are used for mining the hubs and authorities, data guides, and web schema discovery.

Machine Learning techniques support with help web mining as they could be applied to the processor in web mining.

Web mining intersects with the application of machine learning on the web.

WEB MINING CATEGORIES

There are three types are performed in web mining.

I. Web Content Mining
II. Web Structure Mining
III. Web Usage Mining

1) Web Content Mining

Web Content Mining refers to the invention of helpful info from web contents as well as text, image, audio, video, etc.

Web Content Mining are classified into two categories one is web page mining and another one is search results. Web page mining different class of web data (text, html, xml and multimedia) used web content of web pages.

Web search mining is intended to extract pattern from web search engines.

Web content mining identify the useful information from the web consists of structure data such as data in the tables, unstructured data for instance free texts, also semi structured data such as HTML documents. Web content mining needs to select helpful information before analysis.

Web content mining has the follow approached to mine data 1.unstructured text mining 2.structure mining 3.semi-structured text mining 4.multimedia mining [9].

i. Unstructured text mining
Most of the web content is of unstructured text data content mining requires application of data mining and text mining techniques [15].

- Information extraction
- Topics Tracking
- Summarization
- Categorization
- Clustering
- Information Visualization

ii. Structured data mining
The Structure data on the web represent their host pages. Structured data is easy to extract when compared to unstructured text the technique used for mining structured data are.

- ✓ Web crawler
- ✓ Wrapper generation
- ✓ Page content mining

iii. Semi-structured data mining
Data evolving from rigidly structured relational tables through numbers and strings to enable the natural representation of complex real world objects with no sending the application writer into contortions.

iv. Multimedia data mining
The multimedia data mining techniques are SKICAT, colour histogram matching, multimedia miner and shot boundary detection.

2) Web Structure Mining
Web Structure Mining studies the model underlying the link structure of the web. Based on links used it is classified into two types. One is internal and other is external.

Web Structure Mining intended to expose the structure of web sites and how they are connected.

Web information retrieval tools make use of only the text on pages ignoring precious information contained in links. The focus of structure mining is therefore on time information. This is an important aspect of web data.

The goal of web structure mining is to generate structured review about websites and web pages. It uses tree – like structure to analyze and describe HTML or XML.

This type of mining can be additional divided into two kinds based on the kind of structured date used.

- Hyper Links
- Document Structure.

i. Link based classification: Link based classification is the most recent upgrade of a classic data mining task to linked domains [1]. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links among pages, anchor text, html tags and other possible attributes found on the web page.

ii. Link based cluster analysis: The goal in cluster analysis is to find naturally occurring sub-classes.the data is segmented into graphs, where similar object are grouped together and dissimilar objects are grouped into different graphs. Different than the previous task, link based cluster analysis in unsupervised and can be used to discover hidden pattern from data.

iii. Link Type: There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities (or) predicting the purpose of a link.

iv. Link Strength: Links can be associated with weights.

v. Link Cardinality: The main task here is to predict the number of links between objects. There is many ways to use the link structure of the web to create notions of authority. The main goal in early applications for link mining is to made good use of the understanding of their intrinsic social organization of the web.

3) Web Usage Mining
Web usage mining focuses on victimisation data processing techniques to investigate search logs or different activity logs to search out interesting patterns. Web usage mining goes through server log files to extract pattern that reveals usage of website by the users. Web usage mining is the application of data mining techniques for discovering motivating usage pattern from web usage data, in order to understand and better serve the needs of web based application. Web usage mining is also known as web log mining is used to analysis the behaviour of website users.

The important phases of web usage mining are explained here

- ❖ Data collection
- ❖ Data Pre-processing
- ❖ Pattern discovery
- ❖ Pattern analysis
- ❖

✓ **Data collection**

The data collection is the discovery of secret information and usage pattern trends, which could aid the web managers for improving the management, performance and controlling of the web servers.

✓ **Data Pre-processing**

Pre-processing of web log data is the fore most important step before doing analysis on any kind of data.web log pre-processing is a process of applying pre-processing step on web log.

✓ **Pattern discovery**

Once after the web data is pre processing step the required interesting patterns or rules can be discovered out of web data by apply statistical methods and as well as data mining methods like clustering association rules, classification rules, path analysis, etc.

✓ **Pattern Analysis**

In this phase the uninteresting patterns from the patterns discovered in preceding pattern discovery phase are removed. And also the discovered patterns are analyzed by making use of some of OLAP tools or by SQL query mechanism.

✓ **Data Clustering**

The method of clustering is broadly used in different projects by researches for finding the usage pattern or user profiles. The clustering algorithms become the most mining method in websites and the clustering objects include user groups and web pages.

CONCLUSION: In this paper we survey the research in the area of web mining. The web presents new challenges to the conventional data mining algorithm that work on flat data. The mining of web data still be present as a challenging research problem in the future. There are many implemented concepts available in web mining. Though various algorithm and techniques have been proposed still work has to be done in discovering new tools to used mine the web.

REFERENCES

[1] Arvind Kumar Sharma, P.C. Gupta, ―Exploration of efficient methodologies for the improvement in web mining techniques-A survey, International Journal of Research in IT & Management (ISSN 2231-4334)Vol.1, Issue 3, July 2011.
[2] Ashish Kumar Garg, Mohammad Amir, Jarrar Ahmed, Man Singh, Sham Bansa," Implementation ofa Search Engine" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064
[3] Chang G, Healey MJ, McHugh JAM, Wang JTL. Web minig. In Mining the*World Wide Web—An Information Search Approach*, Dordetch: Kluwer; 2001.
[4] Cooley.R, J.Srivastava, and B.Mobasher. Web mining: Information and pattern discovery on the world wide web. In *Proceedings ofthe 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, 1997.
[5] Cowie.J and W.Lehnert. Information Extraction Communications of the ACM,39 (1):80-91,1996.
[6] Etzioni.O, The World-WideWeb: Quagmire or Gold Mine? *Communications of the ACM*, 39(11):65–68, 1996.
[7] Han, J., Kamber, M. Kamber. "Data mining: concepts and tech-niques". Morgan Kaufmann Publishers, 2000.
[8] Jaideep Srivastava, "Web Mining: Accomplishments & Future Directions", University of Minnesota USA, srivasta@cs.umn.edu, http://www.cs.umn.edu/faculty/srivasta.html
[9] Johnson,F., Gupta,S.K., *Web Content Minings Techniques: A Survey*,International Journal of Computer Application. Volume 47 – No.11, p44, June(2012).
[10] Kosala.R, H.Blockeel, ―Web Mining Research: A Survey, SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.
[11] Pazienza.M.T, editor. Information Extraction. A multidisciplinary Approach to an Emerging Information Technology, Volume 1299 of LectureNotes in Computer Science. International Summer School, SCIE-97,Frascati (rome), Springer, 1997.
[12] Srivastava.G, K. Sharma, V. Kumar," Web Mining: Today and Tomorrow", in the Proceedings of 2011 3rd International Conference on Electronics Computer Technology (ICECT), pp.399-403, April 2011.
[13] Van RijsbegenC.J, Information Retrieval Butter Worths,1979.
[14] Wilks.Y, Information Extraction as a care language technology, volume 1299 of lecture Notes in Computer science, chapter In M-T.Pazienza(ed.), Information Extraction,page 1-9.Springer,1997.
[15] Zhang,Q., Segall,R.S., Web Mining: A Survey of Current Research, Techniques, and Software, International Journal of Information Technology & Decision Making. Vol.7, No. 4, pp. 683-720. World Scientific Publishing Company (2008).

