International Journal of Recent Trends in Engineering & Research
ISSN (ONLINE) : 2455 – 1457
IMPACT FACTOR : 4.101

A SURVEY ON WEB MINING TECHNIQUES

K. Velkumar[1], Dr. P. Thendral[2]
[1]Assistant Professor, Department of Computer Science and Engineering, Theni Kammavar Sangar College of Technology, Theni
Associate Professor Department of Computer Science and Engineering, Kalasalingam Academy o Research and Education, Krishnankovil

Abstract— WWW is the most generally known and important source of information for data mining research. So, it becomes a challenging task to retrieve useful and novel information and knowledge from this huge, dynamic structurally complex and ever-growing World Wide Web. Web Mining is the procedure of Data Mining techniques to automatically ascertain and extract knowledge from Web documents and services. The main motive of web mining is ascertaining useful information from the WWW and its usage patterns. Web mining can be broadly divided into three different types of mining techniques: they are Web Content mining, Web Structure mining and Web Log Mining. Web Content mining retrieve knowledge from the content of web documents. Web structure mining is retrieve the structure information from the web. Web usage mining is identifying or discovering interesting usage patterns from large amount of data.
Keywords— *Data Mining Web Mining, Web Mining Algorithms, Data Mining Tools*

I. INTRODUCTION

In Today's world, people rely on World Wide Web to send and receive information, to distribute their knowledge, for conducting online businesses, to express their opinions and views, to discuss with the people all around the world, for entertainment etc. Thus World Wide Web has become the largest source of information for mining tasks. The information present on the web is useful depending upon the choices and preferences of people searching for information. For example, for a person who is engaged in steel industry business, the information related to other domains may seem less worthy. Even though almost every kind of information is present on the web, it becomes a challenging task to extract the information based on user's preferences and interests. Web Data mining is the application of data mining which uses advanced data mining algorithms and techniques to extract interesting and potentially useful patterns from the web data.

The Difference between the Data mining and Web mining

Table 1: difference between data mining and web mining

S.No	Data Mining	Web Mining
1.	Data Mining is the process that attempt to retrieve the pattern and hidden information in large data sets in any system.	Web Mining is the process of data mining techniques to automatically retrieve and extract knowledge from web documents.

2.	Data Mining is used for web page analysis.	Web Mining is used for a particular website and electronic services.

II Web Mining

Web Mining is the procedure of Data Mining techniques to automatically ascertain and extract knowledge from Web documents and services. The web mining is classified as web content mining, web structure and web log mining.

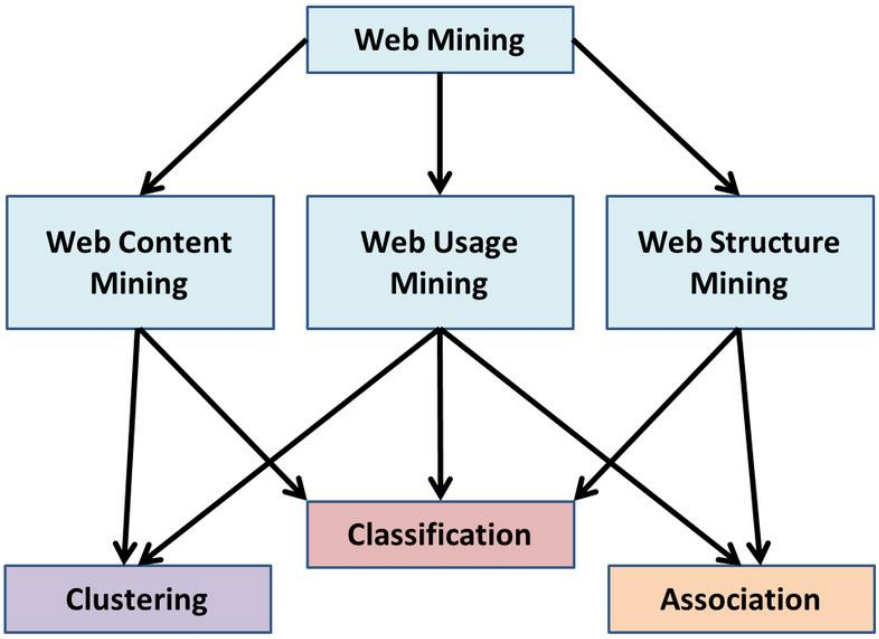

Fig 1Taxonomy of web Data mining

a) Web Content Mining

Web content is retrieve the content from the web page. The web content like text, images, audios and videos. This kind of mining achieves scanning and mining of the text, images and groups of web pages according to the content of the input. This has prompted many researchers to develop new and more intelligent algorithms and techniques for information retrieval and organizing and interpreting semi structured and unstructured data. It is also called agent bases approach. This approach is used automatically retrieve the significant information from the world wide web. There are three types of agents. Intelligent search engine, information filtering and personalized web agents.

b) Web structure mining

The web structure mining is retrieve the structural information from the world wide web. The structure of the web graph comprises of web pages as nodes, and hyperlinks as

edges connecting associated pages. Structure mining fundamentally displays the structured summary of a specific website. Web structure Mining goals to create structural summary about websites and web pages.

c) Web Usage Mining

Web usage mining is also named web log mining. The web log mining is retrieve the interesting pattern from the web log server. The user browsing is captured in this web log mining. The web log mining contains three types; preprocessing, pattern discovery and pattern analysis.

Applications of Web Mining**:**

1. Web mining helps to refine the control of web search engine by classifying the web documents and recognizing the web pages.
2. It is used for Web Penetrating e.g., Google, Yahoo etc and Vertical Searching e.g., FatLens, Become etc.
3. Web mining is used to predict user performance.
4. Web mining is used for specific Website and electronic service e.g., landing page optimization.

II. Web mining algorithms

There are many algorithms are used to extract the content from the world wide web.

The web contents algorithms are Decision Tree, Naive Bayes, Support Vector Machine, Neural Network algorithms.

a) Decision Tree Algorithm

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, individual branch signifies the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

The following decision tree is for the concept buy_computer that indicates whether a customer at a company is likely to buy a computer or not.

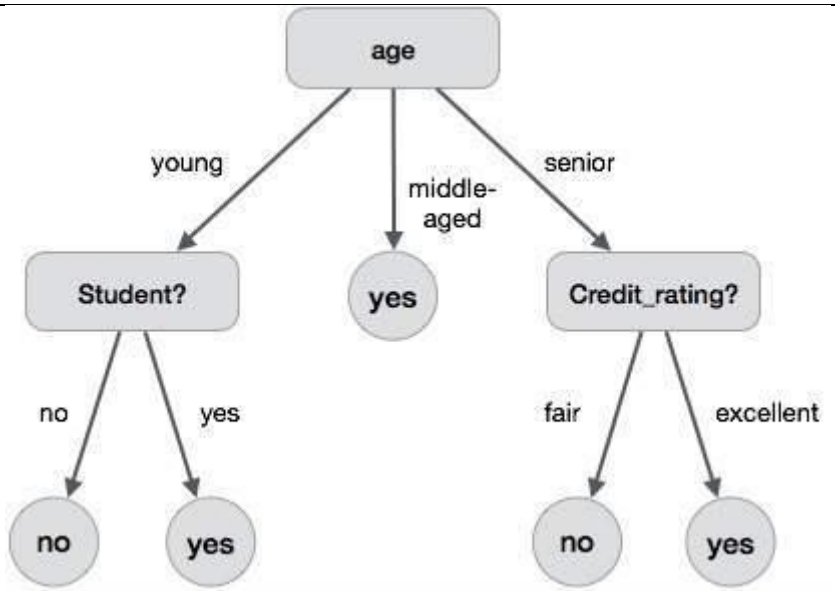

Fig 2 example for decision tree algorithm

b) Naïve Bayes Algorithm

Naive Bayes classifiers are a group of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them segment a common standard, i.e. every pair of features being classified is sovereign of each other. The fundamental features make on independent and equal contribution to outcome.

c) Support Vector Machine

The objective of the support vector machine algorithm is to discover a hyperplane in an N-dimensional space (N — the number of features) that definitely categorizes the data points. Its aim is to discovery a plane that has the maximum margin, i.e the maximum distance between data points of both classes

d) Neural Network

Artificial neural networks (ANN) or connectionist systems are calculating systems imprecisely inspired by the biological neural networks that create animal brains. Such systems "learn" to achieve tasks by considering examples, generally without being planned with task-specific rules.

IV. Web Mining Tools

Web mining tools are automatically extract the information from the world wide web

a) Data miner

The data miner is automatically extract the knowledge from the world wide web and provide the data into csv or excel format.

b) Google Analytic

It tracks and report web traffic. It is used to improve the performance of the websites.

c) Scrapy

This tool also called web content tool. This tool used to extract the content from the world wide web.

d) Similar Web

The similar web is web usage tool. The extract the user behavior from the web log server. It's find user interesting pattern from the web.

V. Issues on Web Data Mining

- Large Volume of Data

  Web Data set contains large volumes of data. To store this data, we need huge of amount of memory

- High Velocity of Data

  The web data is dynamic. So the updating the data is very difficult.

- Heterogeneous Data

  The different types of data exist in the web.

- Data Cleaning

  The automatic data cleaning is required.

- Limitations of Search Engine

  The search engines are limit the user searching query.

V. Conclusion

In this paper some basic thoughts of web data mining have been presented. This paper also deliberated the classifications of web data mining and some of the current research issues challenged by web data mining researchers.

References

[1] Muhammd Jawad Hamid Mughal "Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview" International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, 2018.

[2] Anurag Kumar and Ravi Kumar Singh, "Web Mining Overview, Techniques, Tools and Applications: A Survey," International Research Journal of Engineering and Technology (IRJET), vol. 03, no. 12, pp. 1543-1547, December 2016.

[3] Simranjeet Kaur and Kiranbir Kaur, "Web Mining and Data Mining: A Comparative Approach," *International Journal of Novel Research in Computer Science and Software Engineering*, vol. 2, no. 1, pp. 36-42, January - April 2015.

[4] Ahmad Tasnim Siddiqui and Sultan Aljahdali, "Web Mining Techniques in E-Commerce Applications," International Journal of Computer Applications, vol. 69– No.8, pp. 39-43, May 2013.

[5] Kshitija Pol, Nita Patil, Shreya Patankar, and Chhaya Das, "A Survey on Web Content Mining and extraction of Structured and Semistructured data," Emerging Trends in Engineering and Technology, pp. 543-546, July 2008.

[6] R. Malarvizhi and K Saraswathi, "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study," International Journal of Computer Trends and Technology (IJCTT), vol. 4, no. 8, pp. 29402945, Augest 2013.

[7] Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey," SIGKDD Explorations, vol. 2, no. 1, pp. 1-15, July 2000.

[8] Faustina Johnson and Kumar Santosh Gupta, "Web Content Mining Techniques: A Survey," International Journal of Computer Applications (0975 – 888), vol. Volume 47– No.11, pp. 44-50, June 2012.

[9] Abdelhakim Herrouz, Chabane Khentout, and Mahieddine Djoudi, "Overview of Web Content Mining Tools," The International Journal of Engineering And Science (IJES), vol. 2, no. 6, June 2013.

[10] Anurag kumar and Kumar Ravi Singh, "A Study on Web Content Mining," International Journal Of Engineering And Computer Science, vol. 6, no. 1, pp. 20003-20006, January 2017.

[11] Kavitha, Priyanka Mahani Dr. Neelam Ruhil,"Web Data Mining A Perspective of Research issues and challenges," Internatinal Conference on computing for Sustainable Global Development(INDIACom) 2016.

[12] Brijendra Singh, Hemanth Kumar Singh," Web Data Mining research: ASurvey"IEEE2010.

[13] Radha Krishnan, B., Vijayan, V., Parameshwaran Pillai, T. and Sathish, T., 2019. Influence of surface roughness in turning process—an analysis using artificial neural network. Transactions of the Canadian Society for Mechanical Engineering, 43(4), pp.509-514.

[14] Krishnan, B.R., Ramesh, M., Giridharan, R., Sanjeevi, R. and Srinivasan, D., Design and Analysis of Modified Idler in Drag Chain Conveyor. International Journal of Mechanical Engineering and Technology, 9(1), pp.378-387.

[15] Krishnan, B.R., Vijayan, V. and Senthilkumar, G., 2018. Performance analysis of surface roughness modelling using soft computing approaches. Applied Mathematics & Information Sciences, 12(6), pp.1209-1217.

[16] KRISHNAN, B.R. and PRASATH, K.A., 2013. Six Sigma concept and DMAIC implementation. International Journal of Business, Management & Research (IJBMR), 3(2), pp.111-114.

