International Journal of Global Engineering (IJGE) E-ISSN: 2456-3099

VOL 2 ISSUE 3 (2018) PAGES 27 - 35

Received: 20/02/2018. Published: 22/03/2018

Web Content Mining –Strategies, tools and its Comparison

[1]K.Mahalakshmi, P.G Scholar,Department of MCA, Ganadipathy Tulsi's Jain Engg. College, Vellore,Tamilnadu,India

[2]A.Chitra Devi , P.G Scholar,Department of MCA, Ganadipathy Tulsi's Jain Engg. College, Vellore,Tamilnadu,Inda.

[3]A.Appandairaj, Asst prof, Department of MCA, Ganadipathy Tulsi's Jain Engg. College, Vellore,Tamilnadu,Inda

Abstract

Today, there are many billions of markup language documents, footage & transmission files obtainable on the net. The net has big steady in current years and his content is dynamical each day thanks to heterogeneousness and unstructured nature of the info obtainable on the Web(WWW). There's a necessity of ways to assist us extract data from the content of websites. One answer to the present downside is exploitation the info mining techniques that's called website mining, that is outlined as "the method of extracting helpful data from the text, pictures and different styles of content that compose the pages". Website Mining may be an element of information Mining. The most uses of website mining area unit to collect, categorize, organize and supply the most effective potential data obtainable on the net to the user requesting the knowledge. The main target of this paper is to gift ways and numerous tools applied for internet mining/web content mining and its comparison.

Keywords- Internet, Data mining, Web content mining, structured data mining, unstructured data mining, semi-structured data mining.

1. INTRODUCTION

Internet may be a network of worldwide level, perpetually dynamic and non-structured [1]. Net may be a in style and interactive medium with intense quantity of knowledge freely on the Market for users to access. It's a group of documents, text files, audios, videos and different multimedia system information [2]. Totally different the various} sorts of information ought to be organized in such how that different users will with Efficiency access it. Data processing means that extraction of knowledge in terms of patterns or Rules from vast quantity of knowledge [3]. The term net mining was coined by Etzioni in 1996, to denote the utilization of knowledge mining techniques to mechanically discover net documents, extract info from net resources and uncover general patterns on the net. The analysis within the field of net is classed on 2 aspects: the retrieval and therefore the mining. The retrieval focuses on retrieving

International Journal of Global Engineering (IJGE) E-ISSN: 2456-3099

VOL 2 ISSUE 3 (2018) PAGES 27 - 35

Received: 20/02/2018. Published: 22/03/2018

relevant info from giant repository whereas mining analysis focuses on extracting new info already existing information [4].

In past, techniques like data extraction, data retrieval and machine learning were wont to discover new information from immense quantity of knowledge obtainable on internet. Data extraction focuses on extracting relevant facts whereas data retrieval focus selects relevant document. Now, internet mining may be a a part of each data extraction and data retrieval. Internet mining supports machine learning as a result of it improves the classification of text [4]. The most aim of internet mining is to extract data. Internet mining is integration {of data of data of knowledge} that's gathered by ancient data processing techniques with information gathered over World Wide internet. Internet mining is rotten into following subtasks [1]:

Resource Discovery: It helps in retrieving services and unacquainted documents on internet.

Information selection and preprocessing: It mechanically selects and preprocesses specific data from the online sources.

Generalization: It uncovers general pattern at individual websites similarly as across multiple sites.

Analysis: It validates and interprets the deep-mined pattern.

Visualization: It presents the end in visual and simple to grasp method.

2. WEB MINING CATEGORIES

Web mining is divided into three main classes counting on the sort of knowledge as Website Content mining, Web Structure mining and Web usage mining [3].

• Web Content Mining: looking the contents of the online pages, including: text, figures, tables and etc.

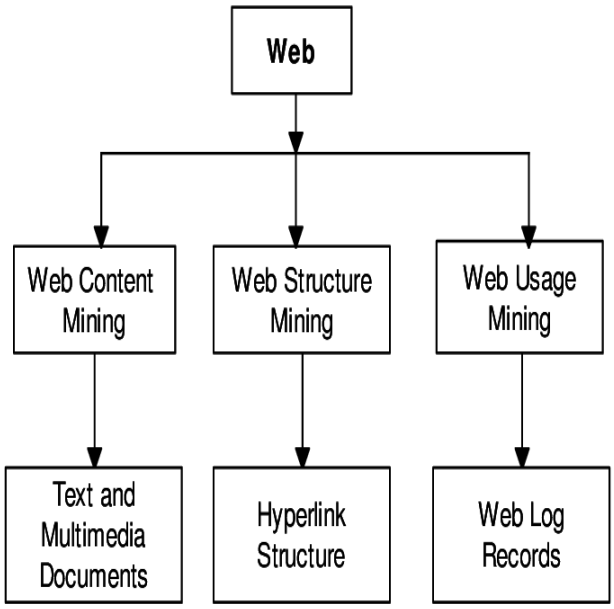
• Web Structure Mining: is looking a hyperlink, connections and links between sites.

• Web Usage Mining: is looking users' web usage pattern by analysis of Access Log Files.

International Journal of Global Engineering (IJGE) E-ISSN: 2456-3099

VOL 2 ISSUE 3 (2018) PAGES 27 - 35

Received: 20/02/2018. Published: 22/03/2018



A. Web Content Mining

Web Content Mining is that the method of mining helpful info from the contents of websites and internet documents, that area unit principally text, pictures and audio/video files. It includes extraction of structured data/information from websites, identification, similarity and integration of data's with similar which means, read extraction from on-line sources, and idea hierarchy, data incorporation [5].

Web content mining analyzes the content of internet resources. Content knowledge corresponds to assortment of facts an online page was designed to convey to the users. Most of the knowledge the information out there on the online is unstructured data. Two totally different points of read of web page mining are: the data retrieval read and also the information read.

B. Web Structure Mining

The structure of a typical internet graph consists of websites as nodes, and hyperlinks as edges connecting connected pages. Internet structure mining is that the method of discovering structure data from the online. This is often additional divided into 2 sorts that's supported the type of structural data used [7].

Hyperlinks: Hyperlinks facilitate in connecting websites to totally different completely different} location either in same website or on different website.

A link is split into 2 classes i.e. intra-document link and inter-document link. Intra-document link connects totally different completely different} a part of identical page whereas inter-document link connects 2 different pages.

International Journal of Global Engineering (IJGE) E-ISSN: 2456-3099

VOL 2 ISSUE 3 (2018) PAGES 27 - 35

Received: 20/02/2018. Published: 22/03/2018

Document Structure: The content at intervals the online page may be organized in tree structure that's supported numerous markup language and XML tags.

C. Web Usage Mining

Web usage mining is that the application knowledge of information mining techniques to find fascinating usage patterns from internet usage data. It tries to find helpful data secondary information derived from the interaction of users whereas water sport internet [6].

There square measure 3 phases of internet usage mining. The 3 phases square measure [8]:

•Preprocessing: It helps in retrieving the data from internet resources and so processes the info.

•Pattern Discovery: once preprocessing the info, the info is employed for locating patterns

• Pattern Analysis: once discovering the pattern, the pattern is analyzed and so the pattern is checked. If the pattern is correct then it's enforced on internet to extract the knowledge from internet.

3. WEB MINING APPLICATIONS

The various fields where web mining is applied are:

- E-Commerce
- Information filtering
- Fraud detection
- Education and research

E-Commerce: In e-commerce, web mining helps in generating user profiles by customizing the choice of users. For example, web mining enables a user to search for an advertisement and information regarding a product of his interest. Internet advertising is one of the major fields in e-commerce, where web mining is widely used. Advertising in a specific domain of an e-commerce web site or a general web site and is considered as one of the major application area of web mining.

Information filtering: Information filtering is the method to identify the most important results from a list of discovered frequent set of data items for which you can make use of web mining.

Fraud detection: Fraud detection can be performed using web mining by maintaining a list of signatures of all the users. Web mining is also applied for plagiarism detection and research works.

4. WEB CONTENT MINING STRATEGIES

Web Content Mining Approaches: Two approaches used in web content mining namely, Agent based approach and database approach [4], [5].

A. *Agent based approach:* The three types of agents are

International Journal of Global Engineering (IJGE) E-ISSN: 2456-3099

VOL 2 ISSUE 3 (2018) PAGES 27 - 35

Received: 20/02/2018. Published: 22/03/2018

- *Intelligent search agents* :Intelligent Search agents automatically searches for information according to a particular query using domain characteristics and user profiles

- *Information filtering/Categorizing agent:* Information agents used number of techniques to filter data according to the predefine information.

- *Personalized web agents:* Adapted web agents learn user preferences and discovers documents related to those user profiles [4], [5].

B. *Database Approach:* The database approach for Web mining tries to develop techniques for organizing semi- structured data stored in the Web into more structured collections of information resources.

Web content mining has the following approaches to mine data

1 Unstructured text mining,
2 Structured mining,
3 Semi-structured text mining, and
4 Multimedia mining. [8]

Unstructured Text Data Mining: Most of the web content information is of unstructured text information. Content mining needs application of information mining and text mining techniques [7]. The analysis around applying data processing techniques to unstructured text is termed information Discovery in Texts (KDT), or text data processing, or text mining. A number of the techniques utilized in text mining are

- Information Extraction,
- Topic Tracking,
- Summarization, Categorization,
- Clustering and
- Information Visualization.

Structured Data Mining: The Structured data on the Web represents their host pages. Structured data is easier to extract when compared to unstructured texts. The techniques used for mining structured data are

- Web Crawler,
- Wrapper Generation,
- Page content Mining.

Semi-Structured Data Mining: Semi-structured data evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real world objects without sending the application writer into contortions. HTML is a special case of such intra-document structure.

International Journal of Global Engineering (IJGE) E-ISSN: 2456-3099

VOL 2 ISSUE 3 (2018) PAGES 27 - 35

Received: 20/02/2018. Published: 22/03/2018

The techniques used for semi structured data mining are

- Object Exchange Model (OEM),
- Top Down Extraction, and
- Web Data Extraction language.[8]

Multimedia Data Mining: The techniques of Multimedia data mining are

- SKICAT,
- Color Histogram Matching, Multimedia Miner and Shot Boundary Detection.
- Multimedia miner
- Shot Boundary Detection

5. WEB CONTENT MINING TOOLS

Web content mining tools helps to download the essential information. Some of them are Screen-scraper, Automation Anywhere, Web Info Extractor, Mozenda and Web Content Extractor, Rapid Miner.

Rapid Miner: Rapid Miner is open source software and it is a tool for extracting information from web, Contains inbuilt algorithm. It can generate algorithm by itself.

Features:

- Easy to use.
- Reduce time.
- Open source software.

Screen-scaper: Screen-scraping may be a tool for extracting/mining info from websites [10]. It is used for looking an information, SQL server or SQL information that interfaces with the software system, to attain the content mining necessities. The programming languages like Java, .NET; PHP, Visual Basic and Active Server Pages (ASP) can even be accustomed access screen hand tool.

Features: Screen-scraper gift a graphical interface permitting the user to designate URL's, knowledge parts to be extracted and scripting logic to traverse pages and work with strip-mined knowledge. Once this stuff are created, from external languages like .NET, Java, PHP, and Active Server Pages, is invoked.

This conjointly facilitates scraping of knowledge at periodic intervals. One among the foremost regular usages of this software system and services is to mine knowledge on merchandise and transfer them to a computer programmer. A classier example would be a meta-search engine wherever in an exceedingly

search question entered by a user is at the same time run on multiple websites in period once that the results area unit displayed in an exceedingly single interface

Automation Anywhere: It is a Web data extraction tool used for retrieving web data, screen scrape from Web pages or use it for Web mining [10].

Features:

- Unique SMART Automation Technology for fast automation of complex tasks.
- Record keyboard and mouse or use point and click wizards to create automated tasks quickly. Web record and Web data extraction.

Web Info Extractor: This is a tool for data mining, extracting Web content, and Web content analysis. It can extract structured or unstructured data from Web page, reform into local file or save to database, place into Web server.

Features:

- No need to learn boring and complex template rules and it is easy to define extract tool.
- Extract tabular as well as unstructured data to file or database.
- Monitor Web pages and extract new content when update.
- Can deal with text, image and other link file
- Can deal with Web page in all language
- Running multi-task at the same time
- Support recursive task definition.

6. COMPARATIVE STUDY OF WEB CONTENT MINING TOOLS

Table 1 shows the web content mining tools and the tasks these tools perform [12].

International Journal of Global Engineering (IJGE) E-ISSN: 2456-3099

VOL 2 ISSUE 3 (2018) PAGES 27 - 35

Received: 20/02/2018. Published: 22/03/2018

Name of Tool	Tasks			
	Records the data	Extract Structured data	Extract Unstructured data	User friendly
Automation Anywhere	Yes	Yes	Yes	Yes
Web Info Extractor	No	Yes	Yes	Yes
Web Content Extractor	No	Yes	Yes	Not for Unstructured data
Screen Scraper	No	Yes	Yes	No

CONCLUSION

This paper discusses the strategies and tools of web content mining. Web content mining has been proved very useful in the business world. The mining of web data still be present as a challenging research problem in the future. Web content mining solves this problem and helps the users to fulfill their needs At the end discusses about different tools that can be used in web content mining. Our hope is that this overview provides a starting point for fruitful discussion.

References

[1] Herrouz, A., Khentout, C., Djoudi, M. Overview of Visualization Tools for Web Browser History Data, IJCSI International Journal of Computer Science Issues, Vol.9, Issue 6, No3, November 2012, pp. 92-98, (2012).

[2] Singh, Brijendra, and Hemant Kumar Singh. “Webdata mining research: A survey." Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on. IEEE, 2010.

[3] R. Malarvizhi and K. Saraswathi.“Web Content Mining Techniques Tools & Algorithms-A Comprehensive Study." International Journal of Computer Trends and Technology (IJCTT), Volume 4, 2013.

[4] Deepti Sharda and Sonal Chawla. "Web Content Mining Techniques: A Study." International Journal of Innovative Research in Technology & Science.

[5] Han, J., Kamber, M. Kamber. “Data mining: concepts and techniques”. Morgan Kaufmann Publishers, 2000.

[6] Manoj Pandia, Subhendu Kumar Pani and Sanjay Kumar Padhi. "A Review of Trends in Research on Web Mining." International Journal of Instrumentation, Control and Automation, Volume 1, 2011.

[7] Srivastava, Prasanna Desikan and Vipin Kumar. "Web mining–concepts, applications and research directions." Foundations and Advances in Data Mining, Springer Berlin Heidelberg, 2005.

[8] Sharma, Kavita, Gulshan Shrivastava and Vikas Kumar. "Web mining: Today and tomorrow." Electronics Computer Technology (ICECT), 2011 3rd International Conference Volume 1, IEEE, 2011.

[9] Automation Anywhere Manual. AA, http://www.automationanywhere.comViewed 06 February 2013.

[10] Mozenda,http://www.mozenda.com/web- mining-software Viewed 18 February 2013.