Web Mining

Sunkara Nagasivaanjaneya Reddy[1], R. Nagarjuna Yadav[2], Alka Choksi[3]

[1]Student, Parul Institute of Engineering & Technology, Parul University, Gujarat, India
210511207026[at]paruluniversity.ac.in

[2]Student, Parul Institute of Engineering & Technology, Parul University, Gujarat, India
210511211045[at]paruluniversity.ac.in

[3]Assistant Professor, Parul Institute of Engineering & Technology, Parul University, Gujarat, India
alka.choksi1484[at]paruluniversity.ac.in

Abstract: ***Web- the name which is on the tip of everyone's tongue now a days and we all know it is the largest database in the world. The first recollection that we get when we think of it, is a vast Data Source where, finding the necessary data is a bit difficult. To overcome this problem and with the goal of providing user satisfaction the concept of Web Mining has been introduced. Simply, Web Mining is the Data Mining Techniques that are incorporated to manage WWW. The Web Mining research is at the crossroads of research from several communities such as database, sub areas of machine learning and natural language processing. However, there is a lot of confusion over the use of it when compared with various research views. But the Web Mining due to its vast size has found great scope of development. Keeping all these facts in mind, in this paper we elaborated on the types of Web Mining, its usage, popularity along with its applications and limitations. Finally, we will conclude with some future directions to improve it for effective and efficient Data retrieval.***

1. Introduction

The World Wide Web (WWW) is a prevalent and a promising interactive medium for the dissemination of information to multiple users across worldwide [1. The massive growth in the Internet applications paved a way to obtain the information about a particular topic from the WWW within few seconds. This leads to the increase in the complexity of web search due to the generation of enormous data and information overload. Hence, extraction of data becomes a tedious task. The web mining techniques are used for solving the overloading issues either in a direct or indirect way. Also, the Information Retrieval (IR) approach, Information Extraction (IE) techniques, Machine Learning (ML) algorithms, Natural Language Processing (NLP), and Web document community are used for extracting the knowledge from the huge amount of data in web . shows the process of web mining. The remaining sections of this review are organized as follows: Section II presents an overview of the web mining. Section III describes the taxonomy of web mining including web content mining, web structure mining and web usage mining. Section IV illustrates the web content mining tools and section V shows the web usage mining tools. Section VI involves the literature survey of various web mining techniques. Section VII concludes the review.

2. Purpose of Study

Web mining is using data mining technique to discover and extract information automatically from documents and Web services. Web mining aims to discover and retrieve useful and interesting patterns from large data sets, as well as in the classic data mining [3]. Big data act as data sets on web mining.

Web Mining

Web mining defines the process of using data mining techniques to extract beneficial patterns trends and data generally with the help of the web by dealing with it from web-based records and services, server logs, and hyperlinks. Web mining aims to discover the designs in web information by grouping and data to receive important insights.

Web mining can widely be viewed as the application of adapted data mining methods to the web, whereas data mining is represented as the application of the algorithm to find patterns on mostly structured data fixed into a knowledge discovery process**.**

There are various applications of web mining which are as follows:

1) Web mining is used to discover how users navigate a website and the results can help in improving the site design and making it more visible on the web.
2) In Customer Relationship Management (CRM), Web mining is the unification of data gathered by traditional data mining approaches and techniques with data gathered over the World Wide Web. Web mining can learn user be compute the effectiveness of a specific Web site, and provide quantify the success of a marketing campaign.
3) The popularity of digital images is quickly increasing because of enhancing digital imaging technologies and convenient availability supported by the web. However, how to find customer- intended images from the web is nontrivial. The main reason is that the web images are generally not annotated utilizing semantic descriptors. It is used to fetch web images from the internet, web mining is utilized.
4) Web mining is used for key phrase extraction. Key phrases are beneficial for several purposes, such as

summarizing, indexing, labelling, categorizing, clustering, featuring, scanning, and searching. The task of automatic key phrase extraction is to select key phrases from within the text of a given document. Automatic key phrase extraction creates it feasible to make key phrases for the large number of files that do not have manually assigned key phrases.

5) Web mining is used for social network analysis. A social network is the study of social entities (person in an organization, known as actors), and their connections and relationships

3. Definitions

3.1 Related Technologies

1. R:

R is a language or a free environment for statistical computing and graphics. It has been made accessible from scripting languages like Python, Ruby, Perl, etc.

Supported Operating Systems: UNIX platforms, Windows, MacOS
Area of Web Mining: Web Usage Mining

2. Octopars:

Octoparse is a simple but powerful web data mining tool that automates web data extraction. It allows you to create highly accurate extraction rules. (You know I will definitely mention our tool.) Crawlers run in Octoparse are determined by the configured rule. The extraction rule would tell Octoparse: which website is to go to; where the data is you plan to crawl; what kind of data you want, etc.

Supported Operating Systems: Windows XP/7/8/10
Area of Web Mining: Web Content Mining.

3. Oracle Data Mining (ODM):
Oracle Data Mining is data mining software by Oracle. Oracle Data Mining is implemented in the Oracle Database kernel, and mining models are first-class database objects. Oracle Data Mining processes use built-in features of Oracle Database to maximize scalability and make efficient use of system resources.

Supported Operating Systems: Microsoft Windows

Area of Web Mining: Web Usage Mining

1) Data streaming technologies

Data streaming technologies, such as Apache Kafka and Apache Flink, are used to process and analyze real-time data streams. They can be used in conjunction with big data and cloud computing to handle real-time data processing and analysis.

2) Containerization technologies

Containerization technologies, such as Docker and Kubernetes, are used to deploy and manage applications and services in the cloud. They can be used to deploy big data applications in a scalable and efficient manner. Overall, these related technologies play a critical role in enabling organizations to manage and process large amounts of data efficiently and effectively using cloud computing. By leveraging these technologies, organizations can gain valuable insights from their data and make better decisions based on that data.

3) Autonomic Computing

Autonomic computing is an approach to computing that seeks to create self-managing systems, self healing, self-optimizing, and self-configuring. The goal of autonomic computing is to create systems that can operate with little or no human intervention, thus reducing the need for manual maintenance and improving system efficiency.

Autonomic computing is based on the principles of feedback control theory, which is used to model and control complex systems. In an autonomic computing system, feedback mechanisms are used to continuously monitor the system's performance and make adjustments as needed to ensure that the system is running at optimal efficiency.

Autonomic computing is particularly important in cloud computing environments, where large-scale systems must be managed and maintained efficiently to meet user demands. By automating many of the management tasks involved in cloud computing, autonomic computing can help reduce costs, improve system performance, and ensure high levels of reliability.

Some of the key technologies and techniques used in autonomic computing include machine learning, artificial intelligence, and advanced analytics. These technologies are used to analyze system performance data and make predictions about future performance, allowing the system to adapt and adjust as needed to maintain optimal performance.

Overall, autonomic computing is a critical approach to computing in the era of big data and cloud computing. By creating self-managing systems, autonomic computing can help organizations improve system efficiency and reduce the need for human intervention in managing and maintaining complex systems.

Types of web mining

Web mining is the application of machine learning (data mining) approaches to web-based data for the goals of learning or deriving knowledge. Web mining methodologies can be defined into one of three distinct elements which are as follows −

Web Usage Mining − Web usage mining is a kind of web mining that enables the set of Web access data for Web pages. This usage data supports the direction leading to accessed Web pages. This data is gathered automatically into connection logs via the Web server. CGI scripts provide useful data including referrer logs, user subscription data, and survey logs. This category is essential to the complete use of data mining for organization and their internet/ intranet-based applications and data access. Usage mining enables companies to make productive data about the future of their business serviceability. Various data can be derived from the collective data of lifetime user value, product cross marketing approaches, and promotional campaign effectiveness. The usage data that is gathered provides the organization with the ability to make results more efficient for their businesses and enhancing of sales. Usage records can also be beneficial for creating marketing skills that will out-sell the competitors and enhance the company's services or product on a larger level.

Web Structure Mining − Web structure mining is a tool that can recognize the relationship among Web pages linked by data or direct link connection. This structure information is discoverable by the arrangement of web structure schema through database approaches for Web pages .This connection enables a search engine to pull records relating to a search query directly to the connecting Web page from the website the content rests upon. This completion takes place through the need of spiders browsing the websites, fetching the home page, then, and connecting the information through reference links to bring forth the definite page including the desired data. The goal of structure mining is to derive previously unknown relationships among Web pages. This structure of data mining supports to use of a business to link the data of its website to allow navigation and cluster data into site maps. This enables its users the ability to access the desired data through keyword relations and content mining.

Manufacturing

Predictive manufacturing provides near-zero downtime and transparency. It requires an enormous amount of data and advanced

Various companies in the media and entertainment industry are facing new business models, for the way they – create, market and distribute their content.

- Predicting what the audience wants
- Scheduling optimization
- Increasing acquisition and retention
- Ad targeting
- Content monetization and new product development

Internet of Things (IoT)

Data extracted from ***IoT*** devices provides a mapping of device inter-connectivity. Such mappings have been used by various companies and governments to increase efficiency. IoT is also increasingly adopted as a means of gathering sensory data, and this sensory data is used in medical and manufacturing contexts.

5. Conclusion

The application of data mining techniques for the extraction of useful and hidden information from the WWW is referred as web mining. In this paper we presented an overview and taxonomy of web mining and a literature survey of recent web mining techniques used in various applications. The efficient improvement in the web mining approaches will provide fast and efficient services for the business, e-learning, digital libraries, e-government and e-commerce applications. Web has been adopted as a critical communication and information medium by a majority of the population.

- Web data is growing at a significant rate.
- A number of new Computer Science concepts and techniques have been developed.
- Many successful applications exist.
- Fertile area of research Web Content Mining Algorithms: Multiple techniques are used by web mining to extract information from huge amount of data bases. There are different types of algorithms that are used to fetch knowledge information, below are some classification algorithms are described

Bayes' theorem:

Support Vector Machine is a well-known and simple machine learning classification and algorithm. SVM is a method that can be used for linear and non-linear data sets . Optimal separating hyper plane (decision boundary) is just a line that is used to draw to separate the two classes depends on the different classification features. Neural network is another web content mining approach which use back propagation algorithm. The algorithm consist of multiple layers i.e. input layer, some hidden layers and then output layer, each feeds the next layer till last layer (output). Neuron is the basic unit of neural network. Inputs are fed simultaneously to units. From input layer, inputs are simultaneously feeding to hidden layers. Usually there is one.

Information

1) Extraction
2) Summarazation
3) Information visualization

4) Topic Tracking
5) Categorazation
6) Clustering

Structured is a technique that mines structured data on the web. Structure data mining is an important technique because it represents the host page on the web. Compare to unstructured, in structured data mining it is always easy to extract data.

Following are some techniques used for structured data mining;

1) Web crawler
2) Page Content mining
3) Wrapper Generation Web Content Mining Algorithms: Multiple techniques are used by web mining to extract information from huge amount of data bases. There are different types of algorithms that are used to fetch knowledge information, below are some classification algorithms are described:

Decision tress is a classification and structured based approach which consist of root node, branches and leaf nodes.

It is hierarchical process in which root node is split into sub branches and leaf node contains class label. Decision tress is a powerful technique. Naïve Bayes is an easy, simple, powerful algorithm for classification and also known as Native Bayes classifier.

Based on Bayes' Theorem. From predefined dataset Val probabilities are calculated.

References

[1] Kosala and H. Blockeel, "Web mining research: A survey," ACM Sigkdd Explorations Newsletter, vol. 2, pp. 1-15, 2000.

[2] P. Maes, "Agents that reduce work and information overload," Communications of the ACM, vol. 37, pp. 30-40, 1994.

[3] O. Etzioni, "The World-Wide Web: quagmire or gold mine?," Communications of the ACM, vol. 39, pp. 65-68, 1996.

[4] M. A. Hearst, "Untangling text data mining," in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp. 3-10.

[5] A.-H. Tan, "Text mining: The state of the art and the challenges," in Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, 1999, pp. 65-70

[6] A.-H. Tan, "Text mining: The state of the art and the challenges," in Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, 1999, pp. 65-70

[7] O. R. Zaiane, J. Han, Z.-N. Li, S. H. Chee, and J. Y. Chiang, "MultiMediaMiner: a system prototype for multimedia data mining," in ACM SIGMOD Record, 1998, pp. 581-583O. Etzioni, "The World Wide Web: quagmire or gold mine?," Communications of the ACM, vol. 39, pp. 65-68, 1996.

[8] S. Chakrabati, B. Dom, D. Gibson, J. Kleinberg, S. Kumar, P. Raghavan, et al., "Mining the link structure of the World Wide Web," IEEE Computer, vol. 32, pp. 60-67, 1999.

[9] M. A. Hearst, "Untangling text data mining," in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp. 3-M. A. Hearst, "Untangling text data minin10.