IJCSN International Journal of Computer Science and Network, Volume 3, Issue 2, April 2014
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 0.274

Evaluation of Web Mining Approach

[1] Sharnjeet Kaur, [2] Nancy Khanna

[1]Assistant Professor, Asian Educational Institute, Patiala, Punjab, India

[2]Assistant Professor, Asian Educational Institute, Patiala, Punjab, India

Abstract - With abundance of data on system it is complicated to accessing facts oriented information from the WWW and provided a reasonable outcome to the users. It has become mandatory to utilize some approach so that valuable information can be extracted. Web mining is most promising solution in such circumstances. The emerging of web mining focuses on bringing knowledge representation capabilities to the web. However knowledge oriented information can be extract by using some sophisticated technique such as semantic web and intelligent agent. Web Mining is one of the emerging research areas in semantic web and highly potential for research.

Keywords - *Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining.*

1. Introduction

Web mining utilizes the data mining approach to automatically find and retrieve information from Web documents [1]. It is an appropriate technique to use for attains the knowledge about user's requirements while searching information to the web. Web Mining is contributing towards the development of extract knowledge oriented information from the web and optimized search results. However some major issue arises in front of web mining approach like huge volume of the information to be mined require for development of adaptive websites and issues related to privacy of users are still prevailing in this area.

Nowadays, web has appropriate popular medium to search information. Whenever some information is preferred from the web, we come crossways a huge quantity of information, out of which categorization the relevant information is left for the user. This condition is termed as information overload, which leads to complexity in finding relevant information [9]. Web Mining is used as a tool to eliminate the problem of information overload while searching information over the web. This paper has been broadly divided into five sections. Section 2 elaborates the classification of web mining system. Third section defines the characteristics of web mining system and section four concludes by presenting open research challenges. The upcoming segment justifies the amalgamation of ontology and web mining in the existing model of wisdom web.

2. Classification of Web Mining System Web Content Mining

Web Content Mining utilizes data mining technique to automatic discovery useful information from the web contents and resources, these web resources are documents, usually as HTML (semi structured) and plain-text (unstructured). It uses especially representative to the attributes of text when it occurs in Web resources [11]. Therefore, the central aim of web mining system is to discovery of patterns in large document collection, and in frequently the collections of document changing. Further the techniques of content mining will be employed for ontology learning, matching and integration ontology.

Research activities in web mining field have drawn a lot of methods developed in other disciplines such as Knowledge Retrieval (KR) and NLP. While there survives a considerable part of work in retrieving information from images in the fields of image processing and computer visualization, the application of these techniques to Web content mining has been limited [16]. The web content mining technique develops in two forms: agent based approach and database approach.

Agent-Based Approach

Normally, agent-based Web mining systems can be developed into the following three parts:

- **Intelligent Search Agents:**
  Agents are software entities that can work on behalf of others. Characterized with autonomy, learning ability, goal directedness, mobility, reactivity and pro activity and many other appealing attributes, they have provide competent result for web mining performance [22]. They are extensively used in e-commerce based applications, providing special assist to users and many web based applications. Various intelligent Web agents have been developed, which find for appropriate knowledge using domain characteristics and user profiles to systematize and construe the discovered knowledge.
- **Information Classification:**
  There is several number of Web agents employ in information extracting methods and

distinctiveness of open hypertext Web documents to automatically extract, filter, and mine them.

- **Personalized Web Agents:**
  These types of agents study user preferences and find knowledge resources based on these preferences.

Database approach

Database approaches concern with systematizing the semi-structured data on the Web into more prepared compilations of sources.

- **Multilevel Databases**
  The central aim of multilevel database is that the lowest level of the database holds semi-structured information saved in assorted Web repositories, such as hypertext documents.
- **Web Query Systems**
  It use ordinary database query languages such as SQL, structural information about Web documents, and even natural language processing for the queries, which are utilizes in World Wide Web searches.

3. Web Structure Mining

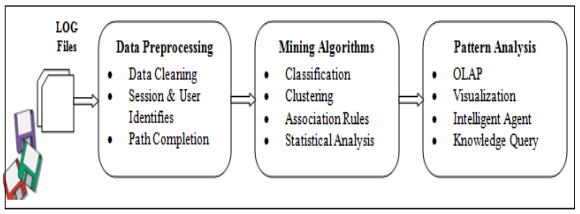
Web Structure Mining works with hyperlink structure of web. It is a method for study of the link structure of the Web in order to classify relevant documents [21]. The graph structure can give knowledge about a pages ranking and improve search outcome through filtering. The underlying proposal is to observe a hyperlink as an appearance of approval and therefore, obtain benefit of the combined conclusion of a page in the assessment of its quality [31]. The Google search engine utilizes web structure mining technique in the process of analysing the relevance of a page and uses page rank algorithm. The pages mapping a set of keywords are ranked according to evaluate of the human interest and attention devoted to each page. It is a research field, focused on two aspects.

(1) Extracting patterns from hyperlinks in the web
(2) Mining the document structure

4. Web Usage Mining

Web usage mining is vital techniques to analysis the user web suffering behavior, when user interacts with the Web. It is a procedure of retrieving useful and meaningful knowledge from several server logs, client logs and proxy logs and determines how user interacts with the Internet and which types of contents, where users are more interested. It is useful techniques both to the Web Administrators and to the individual user. When considerate the users preferences, illustrated by the surfing log of users, the Web administrator can enhance the site creation according to the requirement of customers as well as their business objectives. These goals used to personalize Web pages, introduce new links, to increase the average time a user spends in the site, page chasing mechanism on server or on proxy server or to introduce new pages in places which make them highly visible [32].

More complicated systems and methods for detection and examination of patterns are now emerging. These tools can be placed into various forms, as discussed below.



Preprocessing Tasks

The first phase of data preprocessing is data cleaning. Data cleaning is a process of eliminate the irrelevant and inconsistent data on server logs. Removal of inconsistent data can be logically proficient by examination the suffix of the URL name.

Discovery Techniques on Web Transactions

This is the major element of the Web mining, which meet the algorithms and methods from various research areas like data mining, machine learning, statistics, and pattern recognition. These all techniques are elaborate in given below section.

- **Statistical Analysis**
  Statistical methods are the much prevailing techniques in retrieving information about visitors to a Web site. The summarization can complete dissimilar kinds of expressive statistical analyses based on dissimilar content when summarizing the session data. By summarizing the statistical data hold in the interrupted Web system report, the retrieved report can be helpful for enhancing the system presentation, increasing the security of the system, facilitation the site adaptation job, and giving help for marketing judgments [5].
- **Association Rules**
  This technique can be utilized to find unordered relationship among data items found in a database processing. The term Web usage mining, the association rules related to sets of pages, which are entranced together with help significance beyond some particular threshold.
- **Clustering**
  Clustering is a method to cluster together data pages with the same properties. Clustering of user data can assist the expansion and implementation of future marketing approaches. It helps to find the group of users,

who have same web suffering behavior. It's very helpful for assuming user navigational to execute market segmentation in E-commerce applications or give personalized Web content to the individual users.

- **Classification**
  Classification is the system to draw a data item into one of numerous predefined classes. In the Web domain, Web master will have to utilize this method, if they covet to begin a profile of users belonging to a particular group. The categorization can be completed by using supervised inductive learning algorithms such as decision tree classifiers and naïve Bayesian classifiers. The table 1 illustrates the comparative analysis of web mining categories.

Table 1: Web Mining Categories

	View of Data	Method	Algorithm	Application
Web Content Mining	Structured, Unstructured, Semi-structured, DB	Machine Learning, Statistical (including NLP), Proprietary algorithms, Association rules	IA	Categorization, Finding extract rules, Finding patterns in text
Web Structure Mining	Link Structure	Proprietary algorithms	Page Rank, Weighted Page Rank	Categorization, Clustering
Web Usage Mining	Interactivity	Machine Learning, Statistical, Association rules	Statistical Algorithm	Site Construction, adaptation and Management, Marketing, User Modeling

Characteristic of Web Mining System

Web mining provides the wide range of facilities in different areas of research. Some characteristic are given below:

- Web mining system is able to summarize huge amounts of click stream information from offline sources and apply sophisticated evaluation for web personalization and other interactive marketing programs.
- Personalization for a user can be attained by maintaining path of earlier retrieved pages. These pages can be employed to verify the characteristic browsing behavior of a user and subsequently to calculate preferred pages.
- By analyzing recurrent access behavior for users, desirable links can be recognized to enhance the overall act of potential accesses.
- Information relating to regularly accessed pages can be employed for caching.
- Web mining can be utilized to assemble business aptitude to enhance Customer desirability, Customer preservation, sales, marketing and advertisement.
- It can help in the learning of how browsers are work and the user's interaction with a web browser interface.
- Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web.Web mining supports in civilizing the magnetism of a Web site, in terms of content and structure.
- Web mining and data mining is also helpful for perceiving interruption, deception, and effort break-ins to the system.

4. Conclusion

Alteration of Web to Web mining needs amendment in the approach of web information attaining system and thus requires new technique/architecture for search engines. Contrasting the liberated information access from the search engines, there should be account oriented access for searches so that the user exploring the information may be recorded and provided with updates and new information on their concern domains. This amelioration in search engine architecture may guide to appropriate web based knowledge oriented information, can help in giving knowledge, based commendation to an individual and thus can contribute towards web mining.

References

[1] R. Cooley, B. Mobasher, and J. Srivastava,, "Web Mining: Information and Pattern Discovery on the World Wide Web", pp 1-10.

[2] C. M. Brown, B. B. Danzig, D. Hardy, U. Manber, and M. F. Schwartz. The harvest information discovery and access system. In Proc. 2nd International World Wide Web Conference, 1994.

[3] K. Hammond, R. Burke, C. Martin, and S. Lytinen, (1995),"Faq Finder: A case-based approach to knowledge navigation", In Working Notes of the AAAI Spring Symposium: Information Gathering from Heterogeneous, Distributed Environments. AAAI.

[4] T. Kirk, A. Y. Levy, Y. Sagiv, and D. Srivastava, "The information manifold. In Working Notes of the AAAI Spring Symposium: Information Gathering from Heterogeneous, Distributed Environments. AAAI Press.

[5] C. Kwok and D. Weld, Planning to gather information. In Proc. 14th National Conference on AI.

[6] E. Spertus. Parasite: mining structural information on the web. In Proc. of 6th International World Wide Web Conference, 1997.

[7] M. Balabanovic, Yoav Shoham, and Y. Yun, "An adaptive agent for automated web browsing. Journal of Visual Communication and Image Representation".

[8] D. K. Gi_ord. Hypursuit: a hierarchical network search engine that exploits content-link hpertexxt clustering. In The Seventh ACM Conference on Hypertext.

[9] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell. Webwatcher: A learning apprentice for the world wide web. In Proc. AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments

[10] P. Resnik, N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative _ltering of netnews. In Proc. of the Computer Supported Cooperative Work Conference, ACM.

[11] U. Shardanand and P. Maes. Social information filtering: Algorithms for automating "word of mouth". In Proc. of Conference on Human Factors in Computing Systems (CHI-95), pp 210-217.

[12] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining - Concepts, Applications & Research Directions", pp 1-21.

[13] Prasanna Desikan, Jaideep Srivastava, Vipin Kumar, and Pang-Ning Tan, "Hyperlink Analysis: Techniques and Applications", pp 1-42.

[14] Neelam Tyagi, Simple Sharm, (2012), " Comparative study of various Page Ranking Algorithms in Web Structure Mining (WSM)", (IJITEE), pp 14-19.

[15] Wenpu Xing and Ali Ghorbani, (2004), "Weighted PageRank Algorithm", IEEE, pp 1-10.

[16] P. Ravi Kumar and Ashutosh Kumar Singh, (2010), "Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval", American Journal of Applied Sciences, pp 840-845.

[17] Ke Wang, Liu, "Discovery of Typical Structures of Documents: A Road Map Approach", pp 1-9.

[18] Dong D, (2009), "Exploration of Web Usage Mining and its Applications", IEEE, pp 1-5.

[19] Singh A, Mishra R, (2012), "EXPLORING WEB USAGE MINING WITH SCOPE OF AGENT TECHNOLOGY", IJEST, pp 42834289.

[20] V.Chitraa, (2011), "A Novel Technique for Session Identification in Web Usage Mining Preprocessing", IJCA, pp 23-27.

[21] Jaideep Srivastava, Robert Cooley, (2000), "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD, pp 1-12.

[22] Vijay Rana and Singh G, "Evaluation of an Intelligent Approach for Semantic Web", IJCT, pp 478-482.

[23] Kosala R and Blockeel H, (2000), "Web Mining Research: A Survey", ACM SIGKDD, pp 1-15.

[24] Pierrakos D, Paliouras G, Papatheodorou C and Spyropoulos, "KOINOTITES: A Web Usage Mining Tool for Personalization", pp 1-6.

[25] COOLEY, R., TAN, P-N., AND SRIVASTAVA, J. 1999b. WebSIFT: The web site information filter system. In Proceedings of the Web Usage Analysis and User Profiling Workshop (WEBKDD'99), Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Boston, August).

[26] WU, K-L., YU, P. S., AND BALLMAN, A. 1998. SpeedTracer: A web usage mining and analysis tool. IBM Syst. J. 37, 1

[27] LIEBERMAN, H. 1995. Letizia: An agent that assists web browsing. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (Montreal).

[28] SPILIOPOULOU, M. AND FAULSTICH, L. C. 1998. WUM: A web utilization miner. In Proceedings of the International Workshop on the Web and Databases (Valencia, March)

[29] Yadav S, Ahmad K and Shekar J, (2011), "Analysis of Web Mining Applications and Beneficial Areas", IIUM, pp 185-195.

[30] MOBASHER, B., COOLEY, R., AND SRIVASTAVA, J. 2000a. Automatic personalization based on web usage mining. Commun. ACM, 142–151.

[31] Vijay Rana, Singh G, "Analysis of Ontology Matching System and their Countermeasures", International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT -2014, IEEE, pp 86-92.

[32] Vijay Rana, Singh G, "An Analysis of Semantic Heterogeneity Issues and their Countermeasures Prevailing in Semantic Web", International Conference on Reliability, Optimization and Information Technology -ICROIT 2014, IEEE.

[33] Manoj Manuja & Deepak Garg, (2011) Semantic Web Mining of Un-Structured Data: Challenges And Opportunities, International Journal of Engineering (IJE), Volume (5) : Issue (3), Pp 269-276.