Available at: www.researchpublications.org

Web Mining: An Approach towards Information Retrieval From Web with Cloud Mining

Mrs.Sonal Gore
Assistant Professor
Rahul Pitale
ME Student
Pimpri Chinchwad College Engineering ,Nigdi, Pune
sonalgore@gmail.com
rahulpitale3@gmail.com

Abstract: This Paper gives an introduction about Web mining which is use to extract useful information on the web. This paper also categories web mining including usage, content, structure of the web. web mining techniques use to extract information from web data. This paper also explain some basic algorithms used in web mining to retrieve most relevant pages on the top and less relevant at the bottom and comparison between this algorithms. we also described cloud mining concept which is the future of web mining. we also explain one of the cloud technique Sas(Software-as-a-Service) which is used to reduce the cost of web mining.

Keywords

Web Mining, webusage, web structure, PageRank, Weighted PageRank, HITS, CloudComputing, CloudMining, SaS (Software– as -a –Service).

I.Introduction:

Now a days World Wide Web(WWW) is growing tremendously. As on today WWW is the largest information repository for knowledge reference. In last 12 years , Web has grown tremendously and the usage of the web is unpredictable[1]. So it is important to understand and analyze the data structure of the Web for effective. Information Retrieval. Now a days User wants much more Information from WWW it is becoming moe difficult task to manage the information on WWW and satisfy the user needs. Therefore users are looking for better information retrieval tools, to find , extract useful information form web. Now a days most of the users use search engines to find information from the WWW. Many search engines available in web market but google,yahoo,Bing,etc these are the most popular search engine because of their crawling and ranking mechanisms. Search engines download, and store hundreds of millions of web pages and they answers millions of queries every day. So Web mining and ranking mechanism becomes very important for effective information retrieval. The Sample architecture[2] of search engine is shown in Fig .1

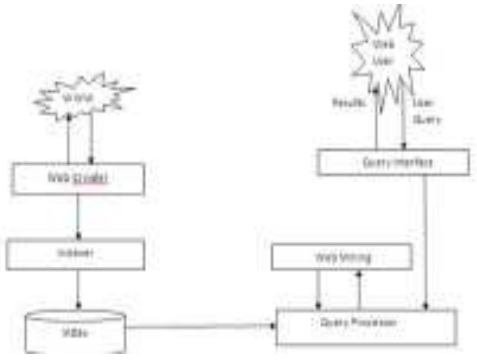


Fig.1 Sample Architecture of Search Engine

There are 3 important components in a search engine. They are Crwaler, Indexer, and Ranking mechanism. Crwaler is also called as spider that traverses the web and downloads the web pages. The downloaded pages are sent to an indexing module that parses the web pages and build the index based on keywords in those pages. When a user types a query using keywords on the search engine, the query processor component match the query keywords with the index and returns the URLs of the pages to the user. But before presenting the pages to the user, a ranking mechanism is done by search engines to present the most relevant pages at the top and less relevant ones at the bottom. It makes the search results easier for the user[1].

This paper is organized as follows-Web Mining is introduced in Section II. The areas Web Mining i.e.Web Content Mining, Web Structure Mining and Web Usage Mining. In III Section we are focusing n Web Structure Mining because most of the Page Rank algorithms are based on Web Structure Mining. various Page Rank Section IV gives future idea about Web Mining i.e Cloud Mining.

II.Web Mining

Web Mining is the data mining techniques to extract and discover usefull information from the World Wide Web (WWW). According to Kosala et al[3]. Web mining consists of following tasks:

- Resource finding: It is a task of retrieving intended Web documents.
- Information Extraction: It is a task which automatically selecting and pre-processing specific information from retrieved Web resources
- Generalization: It discovers genral patterns at individual Web sites as well as multiple sites
- Analysis:Validation and interpretation of the mined patterns

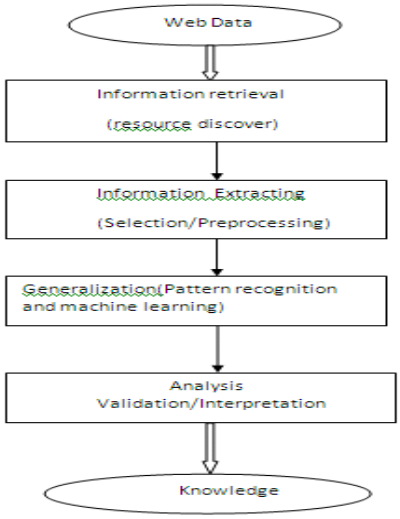


Fig 2.Web Mining Process

Web Mining Categories:

There are three areas of Web mining according to the usage of the Web data used as input in the data mining process, Web Content Mining (WCM), Web Strcuture Mining(WSM), Web Usage Mining (WUM). All of the three categories focus on the process of knowledge discovery and potentially useful information from the Web. Even though they are three areas of Web mining .the differences between them are narrowing because they are all interconnected.Fig.3 Shows General classification of web mining [4]

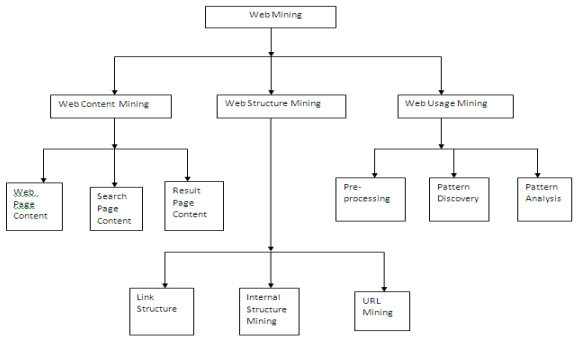


Fig 3.Classification of web mining

Web Content Mining(WCM):

Web Content Mining is the process of extracting usefull information from the contents of web documents. The web documents may consist of text, images, audio, video, or structured records like tables and lists. Web Content Mining is related to Data Mining because many Data Mining techniques can be applied in Web Content Mining. It is also related with text mining because much of the web contents are text. Mining can be applied on web documents as well as results pages produced from search engine. Web Content Mining could be differentiated from two points of view: the agent based approach or database approach. The First approach aims on imporoving the information finding and filtering and colud be placed into the following three categories:

1. Intelligent Search Agents: These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information
2. Information Filtering/Categorization: These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve,filter,and categorize them.
3. Personalized Web Agents: These agents learn user preferences and discover Web information based on these preferences, preferences of other users with similar interest.

The Second approach aims on modeling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it.

Web Usage Mining(WUM):

It is discovery of meaningful pattern from data generated by client server transaction on one or more web localities. A web is collection of interelated files on one or more web servers. It extracts data stored in access logs,referrer loges, agent logs, client-side cookies, user profile and meta data. Web Usage Mining is categories in three phases:

- Preprocessing
- Pattern Discovery
- Pattern Analysis

Preprocessing: According to client, server and proxy server it is first approach to retrieves the raw data from web resources and processed the data .it is automatically transformed the original raw data

Pattern Discovery: According the data preprocessing discovered the knowledge and implements the techniques to discover the knowledge like as machine learning and data

mining prodeures are carried out at this stage

Pattern Analysis: Pattern analysis is the process after pattern discovery. Its check the pattern is correct on the web and how to implement on the web to extract the information on web search/extract knowledge from web.

III.Web Strcuture Mining(WSM):

Web Strcuture Mining focuses on the hyperlink structure of the web. It is use to generate structural summary about the web site and web page[11]. Web Structure Mining will categorize the web pages and generate the information like similarity and relationship between different web sites.There are many algorithms use to foucs on the link structure of the web to find the importance of the web pages

- PageRank Algorithm
- Weighted Page Rank Algorithm
- The HITS algorithm

Page Rank Algorithm

Brin and Page developed[5] Page rank algorithm to calculates the importance of the web pages using link structure of the web. Page rank algorithm is based on citation analysis.Page Rank algorithm is used by the most famous search engine Google, They applied citation analysis in web search by treating the incoming links as citations to the web pages Page rank algorithms provides more advance way to compute the importance of web pages than simply counting the number of pages that are linking to it.(calld as" backlinks") If a backlink comes from an important pages then that backlink is given a higher weighting than those backlins comes from non-important pages. The Algorithm of Page Rank is as follows:

- Page rank takes the backlinks into account and propagates the ranking through links. A page has a higher rank,if the sum of the ranks of its backlinks is high.Fig.4 shows an example of backlinks where page A is backlink of page B and page C and page B and page C are backlinks of page D.[1-10]

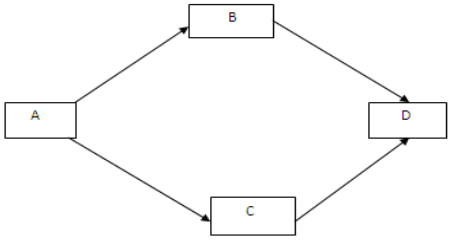


Fig 4 Example of Backlinks

The original Page Rank algorithm is given in following equation

PR(P)=(1-d)+d(PR(T1)/C(T1)+……..PR(Tn)/C(Tn))

Where,PR(P)=PageRank Of page P, PR(Ti)=Page Rank of page Ti which links to page,
C(Ti)=Number of outbound links on page T
D=Damping factor which can be set between 0 and 1(normally it sets 0.85)

Weighted Page Rank Algorithms

Wenpu Xing and Ali Ghorbani[6] proposed a Weighted PageRank Algorithm which is an extension of the PageRank algorithm.

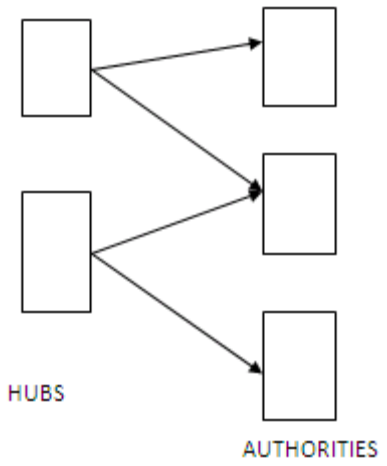
This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a pge evenly among its outgoing linked pages. each outgoing link geta a value proportional to its importantance. The importance is assigned in terms of weight values to the incoming and outgoing links and it is denoted as Win(m,n)and Wout(m,n)

The formula as proposed by Wenpu et al for the WPR is as shown below which is the modified PageRank formula

WPR(n)=(1-d)+d∑WPR(m)Win(m,n)Wout(m,n)

HITS Algorithm

Kleinberg[7] proposed HITS algorithm. according to him he said that there are two different forms of Web pages called hubs and authorities. Authorities are pages having important contents. Hubs are pages that act as resource lists, guiding users to authorities. Hubs and authorities are shown in Fig.5



IV. Cloud Mining Future of Web Mining

Cloud Computing is one of the most seductive technology now a days. The term 'cloud' is a symbol for the internet. Cloud Computing overlaps some of the concepts of distributed, grid,and utility computing.Cloud computing really is accessing resources and services needed to perform functions with dynamically changing needs. Basically Cloud Minning is new approach to faced search interface for your data.SaS(Software –as-a Service) is used for reducing the cost of web mining and try to provide security that become with cloud mining technique[8-9].

Reduce Web Mining cost by SaS-Cloud mining is born:

SaS distribution model helps to reduce costs by providing flexible license option and outsourcing the hardware effort.At SaS[5] , the software is not applied in the company , it lies at software service providers servers. That means the provider deals with the hardware, looks after software updates and maintains technically everything. In Cloud mining the servers that provide the software are the Cloud.

KEY CHARACTERISTICS OF SaaS

- Centralized feature updates : This obviates the need for downloadable patches and upgrades typical of an on-premise software installation.
- Single-instance, multi-tenant architecture: A one-to-many model implies a single physical instance with customers hosted in separate logical spaces. There can be multiple

variations of how a single instance really gets implemented and how multi-tenancy really gets achieved.

- Managed centrally and accessed over the Internet: There is no software component installed at the customer site. All applications can be accessed remotely over the web.
- Generally priced on a per-user basis: The minimum number of users that companies can sign up for varies from one SaaS vendor to another and also depends on what stage the SaaS vendor is in their evolution path as a company. Some do charge additional fees for extra bandwidth and storage.
- Mostly subscription-based, no upfront license costs: This implies that functional leaders (from sales, marketing, HR and manufacturing) do not have to go through their IT department to get them approved.

In terms of "mining" clouds the Hadoop and MapReduce communities who have developed a powerfull framework for doing predictive analytics against complex distributed information sources[8]

Applications and Future work

Web mining is used in various applications such as Health care, Student management, mathematics, Science, in various website. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users.

Here we explore the how the data mining tools like SAS are used in cloud computing to extract the information. Many researchers have looked for way of represent the web mining and future of web mining,some of these are said that cloud mining is the future of web mining.

Conclusion

In this paper we provide a survey about the area of Web mining and their categories like web content mining, web structure mining, web usage mining use for Information Retrieval. and also discussed future of web mining i,e cloud mining which is used to reduced the costs by applying Sas technique in cloud computing.

References:

[1] Ashutosh Kumar Singh, Ravi Kumar P,"A Comparative Study of Page Ranking Algorithms for Information Retrieval," International Journal ofElectrical and Computer Engineering ,2009.

[2] N.Duhan, A. K.Sharma and K.K.Bhatia," Page Ranking Algorithm: A Survey, Proceedings of theIEEE International Confernece on Advance Computing, 2009.

[3] R.Kosala, H.Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, Newsletter of the ACM Special InterestGroupon Knowledge Discovery and Data Mining Vol 2, No.1 pp 1-15,2000.

[4] R.Cooley, B.Mobasher and J.Srivastava," Web Mining: Information and pattern Discovery on the world wide web". Proceedings

of the 9th IEEE International Conference on Tools with Artificial Intelligence, pp.(ICTAI'97),1997

[5] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Computer Networks and ISDN Systems,Vol.30,Issue 1-7, pp. 107-117,1998

[6] W.Xing and Ali Ghorbani "Weighted PageRank Algorithm", Proc. of the second Annual Conference on Communication Networks and Services Research(CNSR,'04) 2004

[7] J.Kleinberg,"Authoritative Sources in Hyper-Linked Environment Journal of the ACM 46(5),pp 604-632,1999

[8] Kavita Sharma,Gulshan Shrivastava, Vikas Kumar, "Web Mining:Today and Tomorrow" 3rd International Conference on Electronics Computer Technology(ICECT 2011).

[9] Bhagyashree Ambulkar, Vaishali Borkar, "Data Mining in Cloud Computing" Proceedings published by International Journal of Computer Applications (IJCA)ISSN: 0975 – 8887 2012.

[10] Tamanna Bhatia, "Link Analysis Algorithms For Web Mining "IJCST Vol 2, Issue2, 2011.

[11] Miguel Gomes da Costa Junior, Zhiguo Gong, "Web Structure Mining: An Introduction", Proceeding of the 2005 IEEE International Conference on Information Acquisition 2005.