www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 7 Issue 8 August 2018, Page No. 24174-24180
Index Copernicus Value (2015): 58.10, 76.25 (2016) DOI: 10.18535/ijecs/v7i8.01

Analysis Web Mining Technology to Predict the User Behavior in Multi-Media

Navjot kumari[1], Vandana[2]

Department of computer science
Swami Vivekanand Institute of Engineering & Technology
Banur, Punjab, INDIA

Abstract

The fast growth of the Internet, we have entered a period of information explosion, there is a considerable measure of repetitive data in the Network. How to extract a useful part of this information from the massive information resources, dissecting the huge measure of data lastly get the potential information we need to extracted. Web mining innovation appeared and spared out the human from the data sea. This paper will break down the acknowledgment of Web content mining and Web structure mining, their essential algorithms standards, and their application areas.

Keywords- Web mining, structure mining, content mining; Multi-media mining, clustering.

I. Introduction

What is Web mining?

It is the development that discovers and extracts the valuable approach that users are involved from the immense Web information and behavior during the data mining tools [1]. Contrast to the well known Data mining, Web mining can be reached out to a more reflective and more extensive zones, the contrasts between them are likewise exceptionally self-evident: the protest of data mining is the data stored in database, in other words, the structured data; Web Mining goes for the substance or structure of Web verification, which has an element of wide-disseminated, dynamic and heterogeneous, and contains unstructured or semi-structured data. Based on the variety of information on the Web, Web mining is divided into the following types as shown in below

- Web structure mining
- Web content mining
- Web usage mining

These three mining methods are different in the feature of deals with the data, processing methods shown in Figure 1.

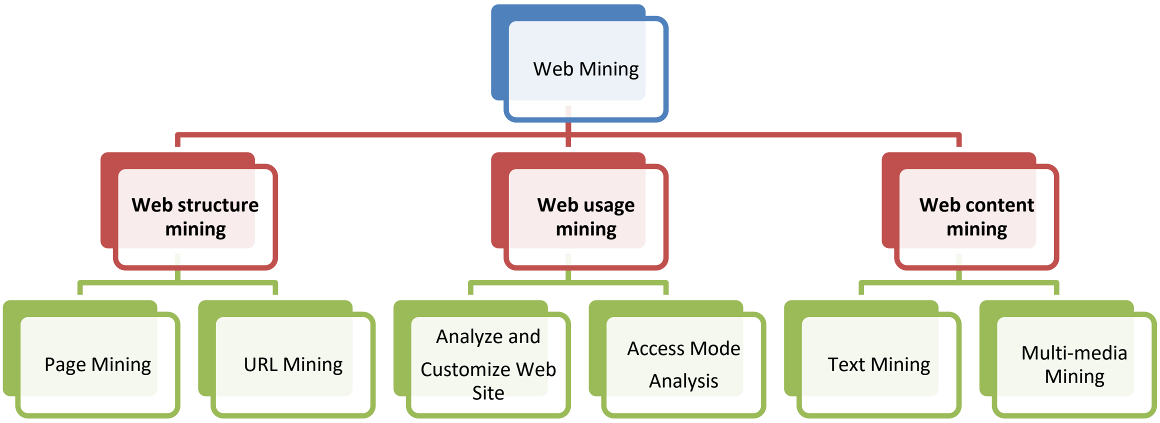

Figure 1 Web Mining Process

1) Web Structure Mining mainly contract with Web structure data, it can be divided into page structure mining and URL mining including with hyperlinks [3].

2) Web Content Mining mainly contract with unstructured data and semi-structured data, can be sophisticated into Web text mining and Web multimedia mining based on the content, in which multimedia mining is a popular research topic at current [4].

3) Web Usage Mining can be divided into commonly access mode examination and customize Web sites, it analyze Web sites logs to find some important information.

In this article we analyze the understanding of Web content mining and Web structure mining, their basic algorithm principles.

II. Web Structural Mining

A. Introduction of Structural Mining

In the era of information retrieval web sites compose the intact Internet network, and all pages in these Web sites wholes are includes with hyperlinks, that pages are contains some information. The intention of Web structure mining is to mine out the unseen knowledge, so that it can be entirely useful. Examinations the hyperlink structure between Web pages. Web Structure Mining is worried with finding the model key the association structures of the web. It is used to consider the topology of the hyperlinks. This model can be used to sort site pages and is useful to make information, for instance, the resemblance and connection between different locales. While Web Content Mining attempts to explore the structure within a record (intra-report structure), Web Structure Mining ponders the structures of chronicles within the web itself (between record structure). In this article expose briefly introduce two structure mining algorithm, PageRank and HITS.

B. PageRank Algorithm

PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. The PageRank algorithm outputs a probability (0,1) distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. The algorithm is not related to the user's query, that is to say, the algorithm is unrelated to the idea, it full uses PageRank value as a site evaluation criteria. This is the basic idea of the PageRank algorithm [5].

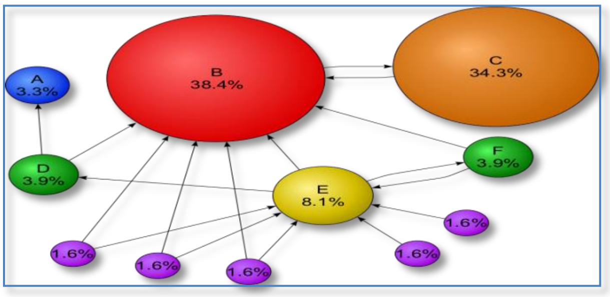

Figure 2 Process of PageRank algorithm

PageRank can be calculated for collections of documents of any size. The page rank of a given page is calculated as

$$PageRank\ of\ site = \sum \frac{PageRank\ of\ inbound\ link}{Number\ of\ links\ on\ that\ page}$$

If the only links in the system were from pages B, C, and D to A, each link would transfer 0.25 PageRank to A upon the next iteration, for a total of 0.75.

$$PR(A) = PR(B) + PR(C) + PR(D)$$

$$PR(A) = (1 - d) + d\,\frac{PR(T1)}{C(T1)} + \cdots +$$

$$PR(A).\,PageRank\ of\ a\ Website, d\ .\,Damping\ Factor\ T1, \ldots..Tn\ .\,Links$$

C. HIT (Hyperlink-Induced Topic Search) based algorithm

A HIT involved with two essential concepts: the content weight ,the number of a certain Web page being referenced by other web pages; the link weight the number of a certain page positions to other webpage [6]. In the HITS algorithm, if a page's content weight is high, then the page is pointed by many other pages, which indicates that its content is of high quality; and if a page's Hub is high, then this page points to many other high-quality pages. When the search results are sorted at the end of the search, the pages are sorted from high to low according to the score of the Authority, and several pages with the highest weight are taken out as the search result of the response users' query. This is the basic idea of the HITS algorithm.

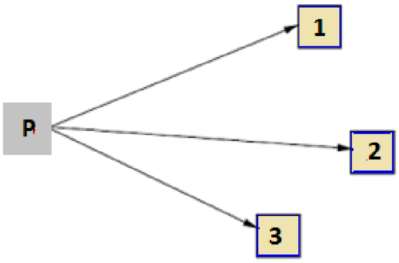

Figure 3 Hit algorithm

E. HillTop Algorithm

The HillTop algorithm at around 2000 and later authorized Google and facilitates. It is complete an important procedural inform in web search ranking. HillTop approves the essential principle of PageRank, that is, conclude the sorting weight of explore results through the number and quality of link in the page, the difference is that HillTop use term input by user in the query to determine the ability of a page [8].

Figure 4 is the flow chart of Hilltop algorithm, including two main processes which are specific page search and main object page sorting.

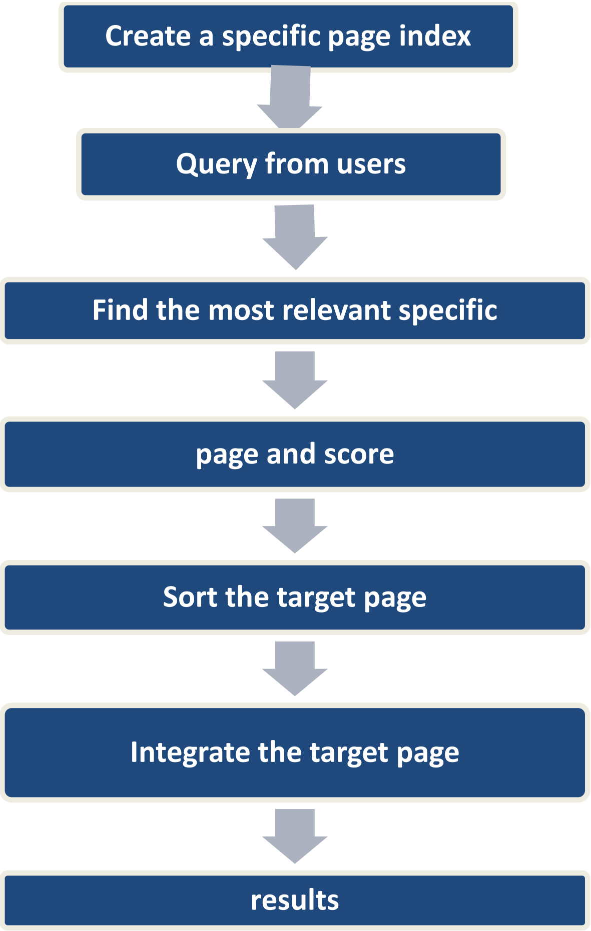
Create a specific page index
Query from users
Find the most relevant specific
page and score
Sort the target page
Integrate the target page
results

Figure 4 Hiltop algorithm flow Chart

The number of advantages of HillTop is judging high-quality webpages with precise and aims techniques, which increase the search efficiency. This algorithm also has many weaknesses, such as operational efficiency and scalability are low comparatively, and cannot circumvent some websites intentionally cheating in order to increase the ranking of websites. At the present the HillTop algorithm and the PageRank algorithm have been combined to provides Google for better results [10].

III. Web Content Mining

A. Web Content Mining Introduction

Web Content Mining is a method of discovering important knowledge and extract useful possible information from immense Web data information. It can be separated into two types firstly is content, including text, hypertext data and multimedia documents like such as video, audio, images, graphics and other multimedia data. The second type is content mining is also separated into two ways which are text mining and multimedia mining, text mining has always extract meaningful information from the given content, but in current topic that study of multimedia mining has progressively been played more attention from research.

B. Text Mining

Text Mining is the development of discovering and extracting implied knowledge from Web documents and finally managing the information that can be use by the user. The procedures of data are semi-structured and unstructured data, in which mainly include free text, HTML tags. The development of text knowledge discovery can be abbreviation as shown in figure 5: text preprocessing, text mining, pattern assessment and representation.

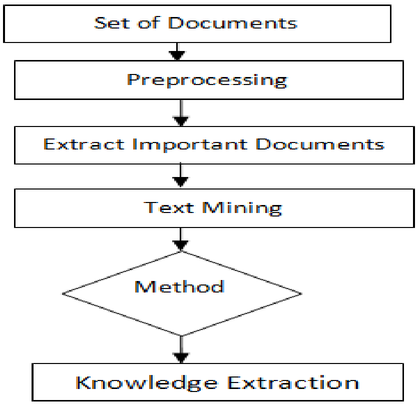

Figure 5 Process of Knowledge discovery

The major techniques of text mining for the large number of Web documents include text summarization, text classification, text clustering, and text association rule analysis and so on. The classification and clustering of text are the most important and basic in Web text mining. Text mining is divided into Information Retrieve method and Database method.

1. Information Retrieve method and Database method

The information retrieve method [5]chiefly utilizes the information question innovation to assess and enhance the nature of the query output information, and can likewise manage unstructured information and HTML structure of semistructured information, for the most part utilized as a part of text order, grouping and example revelation; database methods and information stockroom method utilize information extraction and transformation method to change over or outline unstructured Web information into organized information, and after that mine the information with information mining innovation.

2. Text summary

Text summary [4] alludes to showing the center information from the text as brief depiction with the goal that clients never again need to tap on the full text. The archive summary can be utilized to the inquiry result area of the web crawler.

3. Text categorization

Text categorization [3] is to recognize the discriminate work through preparing on arrangement rules in view of existing information, utilize this capacity to distinguish and characterize an assortment of obscure. Web information with various traits so it is more advantageous and successful for clients to inquiry and read Web reports. The present word division method can be finished up into the mechanical division and understanding division. Presently, the most develop word division ought to be Chinese lexical investigation framework ICTCLAS, which is created by the Chinese Academy of Sciences. There are numerous calculations for text arrangement, the most vital ones are Naive Bayesian grouping calculation, K-closest neighbor calculation[6], choice tree calculation, neural system calculation and bolster vector machine calculation. This innovation can consequently sort a substantial number of Web archives along these lines spare a lot of time. Text arrangement can be utilized for information recovery, text separating, record programmed order, advanced library and different fields, one of the vital applications is to distinguish and channel spam messages and messages. From the present outcomes, this innovation has brought awesome comfort for our work and life.

4. Text clustering

Text clustering[7] is an unsupervised enlistment machine learning issue. The current text clustering calculation can be generally ordered into two kinds: Hierarchical Clusters spoke to by G-HAC and Plane division calculations spoke to by K-Means. In PC programming, the means of the K-implies calculation are as per the following:

(1) Take K components from the first information as each focal point of the K bunches.

(2) Calculate the base separation between different components and the bunch focus, allow these components to the closest group.

(3) According to the clustering comes about, recalculate the separate focuses of the K groups. The method is to compute the math mean of the individual measurements of the considerable number of components in the group.

(4) Re-group every one of the components as per the new bunch focus.

(5) Repeat the past advance until the point that the clustering result never again changes, yield the outcome. The importance of text clustering is that it can isolate an expansive number of texts into a few classifications as per its trait or substance; this incredibly encourages the client's hunt and spares client's perusing time. As the text clustering does not require preparing process, so its adaptability and accommodation have made it a noteworthy method to manage Web reports. Text clustering can be connected to give a summary of huge scale records' substance, recognize comparability between shrouded reports, and facilitate the perusing of related information, and so on.

V. Multi-Media Mining

Multi-media mining allude to separating significant learning from the mass multimedia information in the Web. Through an extensive investigation of varying media highlights and semantics, find verifiable, powerful, significant, justifiable model, and get the pattern and pertinence of the business, therefore giving clients a critical thinking level of choice help limit [10]. Multimedia mining, for the most part, includes two research regions which are information mining and multimedia information preparing. From the point of view of prepared information, it by and large incorporates illustrations and picture information mining, video information mining and voice information mining. The procedure of multimedia mining can be generally separated into three stages:

(1) Extract the required metadata from an extensive number of multimedia information and arrange it into a meta a database;

(2) Use proper algorithmic systems for mining multimedia content;

(3) Present and clarify the mining comes about plainly.

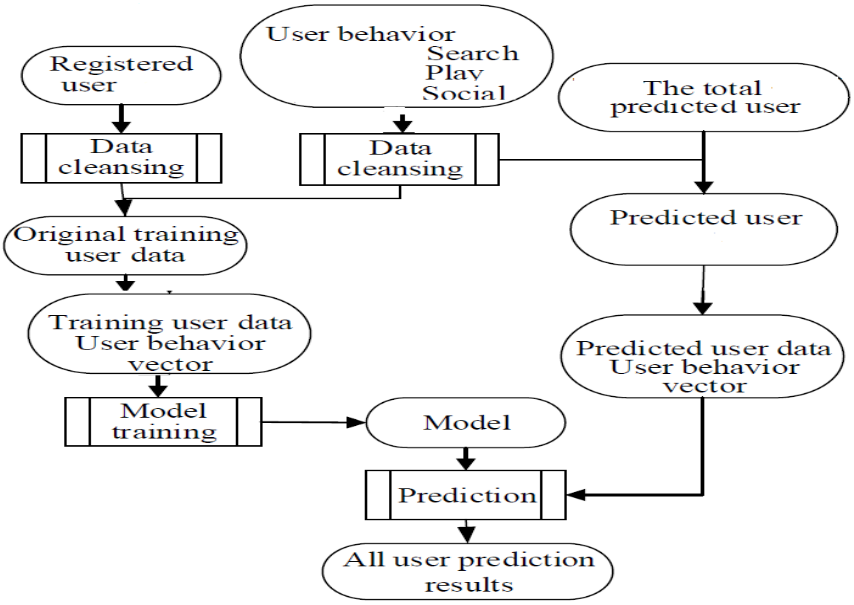

Figure 6 Multimedia data mining process

As appeared in figure 6, taking the video content as case, a progression of information, for example, client enlistment information, playback record, look conduct, social conduct are broke down, and the client's characteristics and conduct propensities are judged by cleaning, examination, and gauging to finish the client's picture, in this manner to accomplish the reason for exact promoting and suggesting site content and different business purposes. Multimedia mining includes the substance of numerous regions, however hard to look into; it is an extremely encouraging zone. These days, multimedia content involves an expansive extent of Web substance and has incredible mining esteem. There must be boundless prospects if the multimedia mining innovation can be connected in the business field.

V. Conclusion

With the rise and improvement of Web mining, this innovation isn't just utilized as a part of the field of web indexes, yet in addition includes in the internet business, web-based shopping, e-learning, e-government and different parts of social life. What's more, through Web mining innovation, individuals have another comprehension of man-made brainpower and have likewise made new leaps forward system security. As a sort of innovation to extricate and find information from enormous information, Web mining innovation has turned into the premise of a substantial number of rising Internet advancements and has made imponderable esteem. Web mining innovation is additionally encountering continually advance and refresh. With joint unremitting endeavors from trades, the fate of Web mining innovation will be more created and qualified to assist individuals with solving more issues.

References

[1] Kosala R, Blockeel H. Web mining research: a survey[J]. Acm Sigkdd Explorations Newsletter, 2015, 2(1): 1-15.

[2] Cooley B R W. Web usage mining: Discovery and application of interest in patterns from web data[J]. Campus-Wide Information Systems, 2010, 24(5):308-330.

[3] Zewen Li. Web-based Data Mining Technology[J]. Modern computer, 2011,3(15):51-58

[4] Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual web search engine [J]. Computer networks, 2012, 56(18): 3825-3833.

[5] Rong Zhang. Research on Technology of Web Mining[J]. Computer Engineering: 2006.08 vol.32(15)

[6] R.Lempel and S.Moran, SALSA: stochastic approach for link structure analysis and the TKC e ect[J]. ACM Trans, Information Systems: 2001,19:131-160

[7] Monika R.Henzinger and Krishna Bharat. Improved algorithms for topic distillation in a hyperlinked environment. Proeeedings of the 21’st Internationa1 ACM SIGIR Conference on Research and Development in IR, 1998.08

[8] Bishui Zhou, Yanhong Zhang. HillTop Algorithm analysis[J]. Computer age: 2005 vol.4

[9] Zhao Y. Association Rule Mining with R[J]. 2015

[10] Mahajan, S., & Rana, V. Spam Detection on Social Network through Sentiment Analysis. Advances in Computational Sciences and Technology, 10(8), 2225-2231 (2017).

[11] Osmar R, Zaiane, Jiawei Han, Ze-Nian Li, Jean Hou. Mining Multimedia Data. CASCON' 98: Meeting of Minds, Toronto, Canada, November,1998.83-96.