ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

Web Content Mining: An Implementation on Social Websites

Pooja Rohilla[1], Ochin Sharma[2]

Dept of Computer Science and Engineering, Manav Rachna International University, Faridabad, India[1,2]

Abstract: Web Content Mining is a component of Data Mining. Whenever we talk about data, we conclude that there is a vast range of data on World Wide Web. And to manage this vast range of data, we often need many tools that can retrieve the data as per our criteria. There are various tools available on the internet which mines the data according to their types like whether they are in a structured format or semi structured or unstructured data. In this paper, I would like to discuss my project work which has an ability to mine the data from web efficiently. In this paper, the previous Web Content Mining Techniques will also be discussed and the tree structure of a webpage document will also be discussed.

Keywords: Social Web Mining, Web Content Mining, Web Mining, and Application of Web content Mining.

I. INTRODUCTION

World Wide data is stepping up. The information stored on World Wide Web is rising day by day as the data is of 25 years old. Data growth is rapidly increasing by some modern techniques like OLTP (Online Transaction Processing), E-Commerce and also by data warehouse. Data is unmanageable because of its rapid growth day by day. There are different techniques, algorithms and tools available for mining the data from the web. Some of the techniques available are HTTP Programming, HTML Parsers, DOM Parsers, Web Scraping Software, web page analyzers.

II. WEB MINING

Web Mining is the process of retrieving the data from different web pages according to the given criteria. Web Mining is further divided in to three categories

A. Web Content Mining – The seeable data on the web pages or any type of information which includes text, audio, video, images, HTML, XML is known as the content. To extract these types of data from different web pages comes under Web Content Mining. Web Content Mining comprises of excavating structured data, semi structured data or non structured data.

B. Web Usage Mining – Web Usage Mining is the litigation of eliciting any type of information from server logs [7]. It is the process of analyzing the curiosity of the users on the internet i.e. in what type of data they are interested for. For instance some users are interested in text type data or some other users are interested in audio, video or images. With the help of Web Usage Mining, we can study the behavior of the user. Using Web Usage Mining, users can get the different type of suggestions for which they are looking for .E.g. Property Search, Online Shopping sites for a particular product etc.

C. Web Structure Mining- It is a tool practiced to discover the link between two or more webpage associated with information. The main intention of web structure mining is to take out the previously known relationships between the webpage.

It basically uses the graph theory with various nodes and the connection link to all the nodes.
In the field of business or E-Commerce, a group of users i.e. clusters can be made for searching similar type of data on the web which results in improvement in several businesses very efficiently and increase in the production of sale.

D. Web Mining Classification :- Web mining is divided in to three parts :-

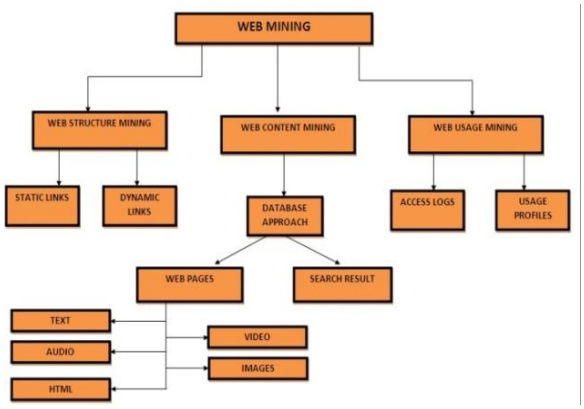

Fig 1 Web Mining Classifications

III. WEB CONTENT MINING TECHNIQUES

The techniques are briefly divided into two major categories-

A. Traditional Data Mining – In this, data is stored in the form of horizontal rows and vertical columns in a form of a table. The data in the tables are relational to each other and also it is structured in a well defined order.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

ID	PWD	FATHER'S NAME	MOTHER'S NAME
Reena.khosla	Abc123	lkjhgfds	eryuioo
Poonam.virmani	Bcd234	yuioptrr	gtnnbiikn
Jyoti.maggu	Hjgu456	ghjooytbn	vtfujknii
Seema.arora	Oiuy987	Gio0lkjgr	Fyujijk

TABLE 1 Structured Data

B. Modern Technique – In this, data is in semi structured format or it is totally unstructured data. To retrieve this type of data, we need to define proper patterns and clustering. In this, there are no tables in proper ordering. It can contain audio, videos, images etc in non structured format.

Fig 2 Unstructured Data

IV. MOTIVATION

A Person having multiple accounts on different social websites will find it difficult to manage all accounts with different account control settings under different code of conduct of social websites. So, I have created a common website known as Integrated Website i.e. INFUSION for all accounts that itself is enough to manage all other accounts of social websites [7].

In this paper, I would like to discuss my work i.e. Web Content Mining applied on social websites. There are several techniques used to mine the content from different websites. There is a technique called Web Scraping which is used to extract the data from different websites.
Web Scraping is the process of extracting information in a structured way from a website. There are various algorithms which can be used to mine the data using Web Scraping. There is different type of HTML Parsers, DOM Parsers, and HTTP Programming methods.

I have used the Web Scraping using Tree Pattern Matching. HTML document consist of various elements. Some are allowed to b nested but some are not. HTML documents have a Tree like structure.

Any change in the HTML Document will reflect the same in its tree structure also, so we need to use Tree Edit Distance algorithm. The most desirable algorithm for Tree Edit Distance for HTML documents is Zhang and Sasha's algorithm.

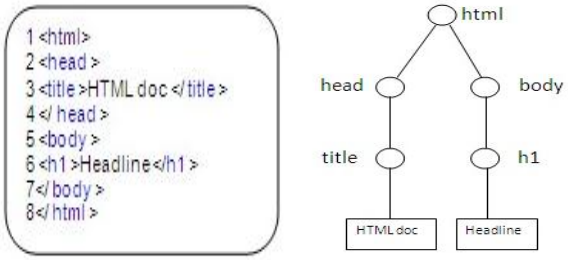

Fig. 3 HTML document and its tree structure

Zhang and Sasha's algorithm allows changes and also wildcards in the pattern of an HTML document.

As there are different social websites available on the web like Twitter, Facebook, Google Plus, Orkut etc , one cannot have enough time to check all the accounts from different websites. There are different social websites available to maintain our personal relations as well as our personnel relations. In this hectic schedule, no one has enough time to access all the accounts to see the update details of their personal and their personnel network.

V. IMPLEMENTATION

In this, I have designed three websites for applying Web Content Mining. Implementation of Web Content Mining consists of three websites named

A. Stay Connected- It is a website which works as a dummy social website in which user can perform various functions :

1) Register- Using this feature a user can add a new account to the website database. User must enter all the details like Name, Mobile Number, Email Id, Father's Name, Mother's Name, User ID and Password.
2) Login- Using this feature, a user can log in to his account by entering authorized User Id and Password.
3) View Friends- Using this feature, a user can see the friends that are added in his/her account with their display pictures.
4) View Timeline- Using this feature, a user can see the timeline update data done by his/her friends. For instance, Change their profile picture, like anyone's comment or picture, commented on anyone's status etc.

B. Social Net - It is a website which works as another dummy social website in which user can perform various functions:-

1) Register- Using this feature a user can add a new account to the website database.User must enter all the details like Name, Mobile Number, Email Id, Father's Name, Mother's Name, User ID and Password.
2) Login- Using this feature, a user can log in to his account by entering authorized User Id and Password.
3) View Friends- Using this feature, a user can see the friends that are added in his/her account with their display pictures.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

4) View Timeline- Using this feature, a user can see the timeline update data done by his/her friends. For instance, Change their profile picture, like anyone's comment or picture, commented on anyone's status etc.

C. Infusion- This is the integrated website which will fetch the data from the two above discussed website using User Id and Password. This website performs various functions:-
1) Login- Using this feature, a user can log in to his account by entering authorized User Id and Password.
2) Merging Friends- This feature has option of checkboxes through which friends of other two social websites simultaneously can be seen in this website.
3) Merging Timeline – This feature has option of showing the updated timeline view of all the friends belong to other social websites.
4) Sorting- This feature enables to view the timeline in an ascending order so that latest update will come first and so on. This sorting feature uses the date and time for sorting.
Therefore, INFUSION is a website application that uses the concept of web content mining in which the data is being fetched by other social websites.

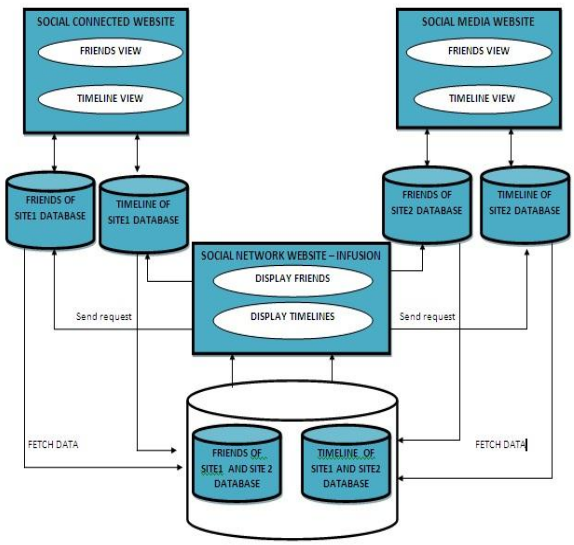

Fig. 4 Working of INFUSION.

VI. ADVANTAGES OF WEB MINING

A. Increase in Profits- Whenever Web Mining used in businesses , it creates more publicity and will result in gaining profits from all over the world by selling its goods and services as all the web users are easily retrieving the results they want.

B. Relevant Information- When applying data mining techniques over the web content data, will result in finding the relevant information easily detecting the patterns of the content and hence makes it easier to get the work done.

C. Web Personalization- It refers to the technique that is used to track the number of visitors and the type of data they are looking for. It also relates with enhancing marketing strategies to provide the users for what they are looking for.

VII. APPLICATIONS OF WEB MINING

A. Electronic Commerce- Web mining helps in creating access logs and usage profiles by using several clustering techniques where the behavior of the user can be determined resulting in providing them the link they want in an easier way.
B. Hiring Profiles- Web mining helps in creating the profiles of the users who submit the curriculum vitae for the jobs and hence make easier for the recruiter to find the suitable candidate for the job and can easily call them without accessing their curriculum vitae in deep.

C. Credit Risk Analysis- Through Web Mining, the previous data of the users can easily be fetched who took bank loans earlier, which resulting in showing the behavior of the customer and his re-paying strategies results in creating risk models precisely.

VIII. CONCLUSION

Web Mining techniques are used for extracting relevant information from various type of websites like online shopping sites. And hence this can also be applied to elicit information from social websites. Any type of confidential data can easily be elicited from the social websites which results in more efficient working with social websites in lesser time and it is totally depend upon the user's requirement. Finally we can say that Web Mining has a huge advantage to extract the data as per the user's criteria.

REFERENCES

[1] S. Balan , P. Ponmuthuramalingam , " A study of various techniques of Web Content Mining Research Issues and Tools , International Journal of Innovative Research and Studies (IJRIS) , Volume 2 , Issue 5 , May 2013 , ISSN: 2319-9725.
[2] Abdelhakim Herrouz , Chabane Khentout , Mahieddine Djoudi , "Overview of Web Content Mining Tools , The International Journal of Engineering and Science (IJES) ,Volume 2 , Issue 6 ,June 2013 , ISSN: 2319 – 1813.
[3] Sanjeev Pippal, Lakshay Batra , Akhila Krishna ,Hina Gupta ,Kunal Arora , "Data mining in social networking sites : A social media mining approach to generate effective business strategies" , International Journal of Innovations & Advancement in Computer Science (IJIACS) Volume 3, Issue 2 April 2014 ISSN 2347 – 8616
[4] Ananthi.J , " A Survey Web Content Mining Methods and Applications for Information Extraction from online shopping sites" , International Journal of Computer Science and Information Technologies (IJCSIT), Voume 5 , Issue 3 , ISSN :0975- 9646.
[5] R.Malarvizhi, K.Saraswathi , "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study" , International Journal of Computer Trends and Technology (IJCTT) , volume 4 ,Issue 8, August 2013, ISSN: 2231-2803.
[6] Deepti Sharda, Sonal Chawla "WEB CONTENT MINING TECHNIQUES: A STUDY", International Journal of Innovative Research in Technology & Science (IJIRTS), ISSN: 2321-1156.
[7] Monika Yadav , Pradeep Mittal , "Web Mining: An Introduction ",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013 ISSN: 2277 128X.
[8] E.Raju , K.Sravanthi , "Analysis of Social Networks Using the Techniques of Web Mining" , International Journal of Advanced Research in Computer Science and Software Engineering , Volume 2, Issue 10, October 2012 ISSN: 2277 128X.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

BIOGRAPHIES

Pooja Rohilla was born in Faridabad (Haryana), India. She received her B. Tech degree in Computer Science and Engineering in 2011 From Manav Rachna College of Engineering, Faridabad, India. Presently she is an M. Tech Student in Computer Science and Engineering in Manav Rachna International University, Faridabad, Haryana, India.

Mr. Ochin Sharma has done M.Tech (CSE) from Kurukshetra University campus in 2007. He has worked as a Sr. Software Testing Engineer in QAInfotech Pvt Ltd., Noida, India. He Joined academics and taught in Banasthali University, Rajasthan, India. Presently he is working as an Assistant Professor in Manav Rachna International University, Faridabad, India. Along with, he is pursuing Ph.D in CSE from Banasthali University, Rajasthan. He has published several research papers in International Conferences and Journals. His areas of interest are System Testing, Data Mining and Big data.

DOI 10.17148/IJARCCE.2015.4724