ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. 2, Special Issue I, January 2016 in association with
KARUR VELALAR COLLEGE OF ARTS AND SCIENCE FOR WOMEN, KARUR
Conference on Emerging Trends and Functional Applications of Computer Technology - 12th January 2016

ENHANCED TECHNIQUE OF WEB PAGE CLUSTERING AND SEGREGATION UNDER WEB MINING

Mrs.C.Menaka M.C.A.,M.Phil.,V.S.B Engineering College,Karur,Tamilnadu,India
Dr.N.Nagadeepa ,Principal,Karur Velalar College of Arts and Science for Women,,Karur,Tamilnadu,India

Abstract :

The development of internet plays a vital role in current industry. Web contents have been increased and updated in every second. Also, web includes huge amount of data so that number of users are growing at the maximum. Web mining is the application of data mining techniques to acquire knowledge from web data, including web documents and hyperlinks between documents, user log details of web sites. This paper focuses advanced technique of clustering process and categorizing the objects using agent based method in web content mining. Clustering analysis is a very most commonly used method in data mining. With the clustering analysis, a heterogeneous data group can be partitioned into significant subgroups. This paper contains various sections.

I. INTRODUCTION

Various data mining techniques have been proposed to implement in web mining to extract information from web. Web content mining is one among the three types of web mining[1][2]. Each type can serve its own specific task. Various tasks are: 1. Resource finding 2.Information selection & preprocessing.3.Generalization 4.Analysis.

Web Content mining implies the discovery of information from web content such as text, images videos etc. Two approaches used in web content mining are Agent based approach and database approach[1][3]. The three types of agents are intelligent search agents, Information filtering/Categorizing a gent, personalized web agents [13]. Intelligent Search agent s automatically searches for information according to a particular query using domain characteristics and user profiles. Information agents used number of techniques to filter data according to the predefine instructions. Personalized web agents learn user preferences and discovers documents related to those user profiles.

In Database approach, it consists of well formed database containing schemas and attributes with defined domains. Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data.

Fig.1 Taxonomy of Web mining

Web Content mining also contains generation of wrappers. Wrapper is nothing

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. 2, Special Issue I, January 2016 in association with
KARUR VELALAR COLLEGE OF ARTS AND SCIENCE FOR WOMEN, KARUR
Conference on Emerging Trends and Functional Applications of Computer Technology - 12th January 2016

but rules to be followed to extract from web pages[4][5]. Rules can be used for both manually and automatically.

Collection of data to be integrated may contain images, texts, audios or videos etc... this web content mining involves document tree extraction, data classification, and data clustering and with these all must be labeled the attributes for results[5][6]. Research activities are going on in information retrieval methods, natural language processing and computer vision [7].

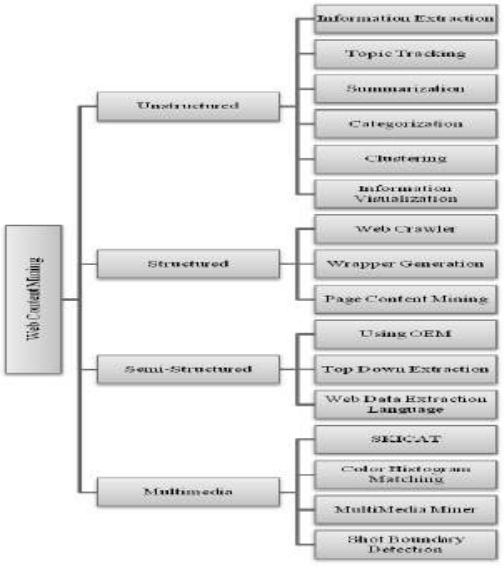


Fig.2 Web content mining techniques

Web Content mining technique can be used under structured and semi structured methods. In general, web page occupies maximum of text content information and the remaining are images, icons, video and audio files. Some of the techniques used in text mining are Information Extraction, Topic Tracking, Summarization, Categorization, Clustering and Information Visualization.

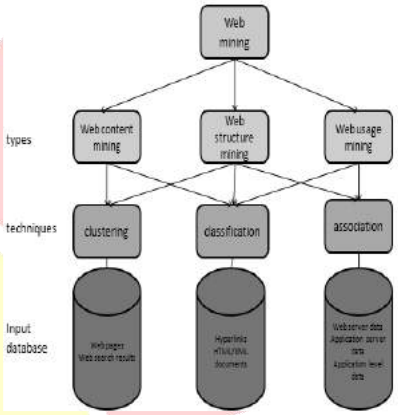


Fig.3 Categories of Web mining

II. Types of Clustering

The process of Clustering is one of the major data analysis methods and deals with the organization of a set of objects in a multidimensional space into cohesive groups, called clusters [8].

With cluster hypothesis, clustering can increase the efficiency and the effectiveness of the retrieval.

Moreover, clustering can be used as a very powerful mechanism for extracting collection of documents (e.g. scatter/group) or for exhibiting the results of the retrieval.

The various clustering techniques are as follows.

A. Text based Clustering:

According to the content each document can be characterized under this text-based document clustering approach. Suppose the text - based web document clustering approaches characterize each document according to its content, i.e. the words contained in it (or phrases or

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. 2, Special Issue I, January 2016 in association with
KARUR VELALAR COLLEGE OF ARTS AND SCIENCE FOR WOMEN, KARUR
Conference on Emerging Trends and Functional Applications of Computer Technology - 12th January 2016

snippets)[8][9]. The concept is that if two documents contain many similar words then it is very possible that the two documents are also very similar. According to the clustering method this issue can be further categorized with the following categories:

1. Partitional 2.hierarchical 3.graph based 4.neural network based and 5. Probabilistic algorithms

B. Partitional Clustering:

- Algorithm begins by taking k cluster centroids. Moreover the cosine distance between each document in the collection and the centroids is calculated and the document is assigned to the cluster with the nearby centroid.
- At last new cluster of centroids are recalculated and it performs iterative process until some criterion is reached.

C. Hierarchical Clustering

- Sequence of nested partitions. Each pair of documents is stored in a n x n similarity matrix.
- The algorithm either merges two clusters or splits a cluster in two.
- Resultant of the process can be displayed in a tree like structure termed as dendogram.
- Here, collection of many clusters at the bottom with one common cluster at the top document.

D. Graph Based Clustering:

- Clustered documents can be viewed as group of nodes and edges with its relationship.
- Edges – Weight, denotes the degree of relationship.
- A minimal spanning tree of a connected graph G = (V, E) is a connected sub graph with minimal weight that contains all nodes of G has no cycles.

E.Neural Network based Clustering:

- One of the most commonly used unsupervised neural network model.
- Input layer with n input nodes, which belongs to n documents.
- Output layer with k output nodes which correspond to k decision regions.
- Weight vector is assigned to k output units.

F. Fuzzy Clustering:

- Best clustering approach for handling more than one cluster.
- Best approach for optimizing a certain criterion function.
- Membership vector can be calculated fro each document in which i-th element indicates the degree of membership of the document in the i-th cluster.

G. Probabilistic Clustering:

- To deal with uncertainty of data these algorithms have been proposed.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. 2, Special Issue I, January 2016 in association with
KARUR VELALAR COLLEGE OF ARTS AND SCIENCE FOR WOMEN, KARUR
Conference on Emerging Trends and Functional Applications of Computer Technology - 12th January 2016

- Statistical models are used to calculate the similarity between the data.
- Probabilities for membership can be assigned for each document.
- As per the probability document can belong to more than one cluster.

III. Terminology and Proposed System

Many algorithm and techniques were used for clustering but in addition to that new tool named **Magnified Content Extractor** have been developed for the purpose of clustering and segregating the related and unrelated objects.

Added feature of this tool is it can be applicable to deal with structured and unstructured data as well. Also, it will generate the report on the basis of which the corresponding object is belonging to.

Actually all the algorithms have been implemented for various data retrieval techniques. In this tool distinct feature is going to be merging to work in efficient way and fast retrieval of information under clustering.

Name of the Tool	**Magnified Content Extractor**
Records the data	Yes
Extracts Structured & unstructured data	Yes
User Friendly	Yes
Efficiency	Yes
Performance	Contents can be clustered and segregated fast
Report	Fast report generation

By taking the URL into this tool particular web page has clustered and it generates report after completing the clustering process for the whole document. In the next step, the same tools have used to segregate the objects into category wise. In other words, if the document contains the various components like text, image, video and audio files. Every object is segregated with their type as per that if similar types of images are stored in one file. Distinct type of image file can be stored in another file.

Example: Image file format are of the type TIF, JPG, PNG, GIF. First, if the web document contains all these formats grouped into one file that means all the image files are at one place. Then it will scattered as per the file format with the help of this content extractor.

Web Data:

The information which is available in web is termed as web data. Web data contains Text, images, audio, video, eBook, 3D image, vector image, and Page layout, Spreadsheet, Database, and Game.

IV.USES OF WEB CONTENT MINING

Some uses of web content mining

1. To determine the relevance of the content to the search query.
2. To gather, categorize, organize and provide the best possible information available on the WWW to the user requesting the information.
3. Improve the navigation of information on the web provides productive marketing.
4. Produce a higher quality of information to the user.
5. Understand customer behavior, evaluate effectiveness of a particular

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. 2, Special Issue I, January 2016 in association with
KARUR VELALAR COLLEGE OF ARTS AND SCIENCE FOR WOMEN, KARUR
Conference on Emerging Trends and Functional Applications of Computer Technology - 12th January 2016

web site, and help quantify the success of a marketing campaign.

V.WEB CONTENT MINING TOOLS

Huge information and data are available on the web, it is essential to deal and make use of the essential information which is required for the user; various web content extractor tools are used. Tools are as follows, Screen -scraper, Automation Anywhere 6.1, Web Info Extractor, Mozenda, and Web Content Extractor [8][9].

Screen-Scraper:

Tool for extracting information which is specifically focuses on database server which interacts with the system [10][11].

It helps mainly for Graphical interface allowing the user to facilitate URL data elements to be extracted.

Automation Anywhere:

Tool which is used for complex tasks in a fast manner is AA [11][12].

Records keyboard and mouse or point click wizards to automate tasks quickly.

Web Info Extractor:

With this tool, new content can be extracted while updating documents. Also it has the capability of handling text, image and some other link file which is directed to other page. It supports all type of languages which is in web page [11][12].

Mozenda:

Tool which is used for extracting and managing web data. [12][13] User can set up agents that normally extract, store and also publish data to multiple destinations.

Also, it runs only on windows since it is a platform independent.

Web Content extractor:

Wizard driven interface which helps through the process of building a data interaction pattern. It supports to retrieve information from various sources like online stores, Trading, Real estate, and economic websites, commercial and job sites [13].

VI.WEB MINING APPLICATIONS

Web mining extends by combining other information with Web traffic data.

Practical applications of Web mining technology are abundant, and are by no means the limit to this technology. Web mining tools can be extended and programmed to answer almost any question. It can be applied in the following areas:

1. Web mining can provide companies managerial insight into visitor profiles, which help top management take strategic actions accordingly.
2. The company can obtain some subjective measurements through Web Mining on the Effectiveness of their marketing campaign or marketing research, which will help the business to improve and align their marketing strategies[15].
3. In the business world, structure mining can be quite useful in determining the connection between two or more business Web sites.
4. This allows accounting, customer profile, inventory, and demographic information to be correlated with Web browsing [14][15].
5. The company can identify the strength and weakness of its web marketing campaign through Web Mining, and then make strategic adjustments, obtain the feedback from Web Mining again to see the improvement.
6. Search engine Google provides advanced and efficient searching capabilities

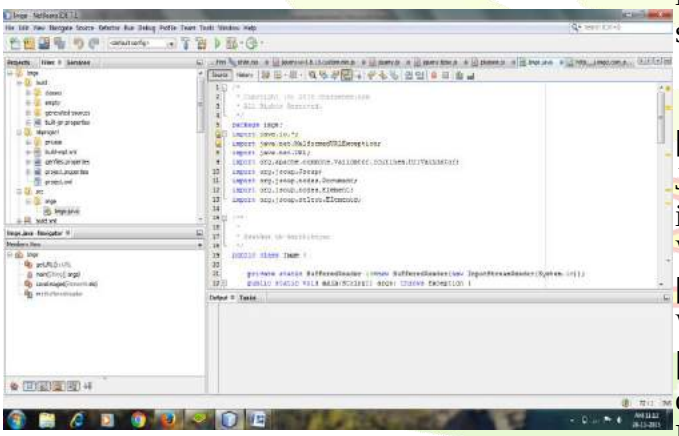
ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. 2, Special Issue I, January 2016 in association with
KARUR VELALAR COLLEGE OF ARTS AND SCIENCE FOR WOMEN, KARUR
Conference on Emerging Trends and Functional Applications of Computer Technology - 12th January 2016

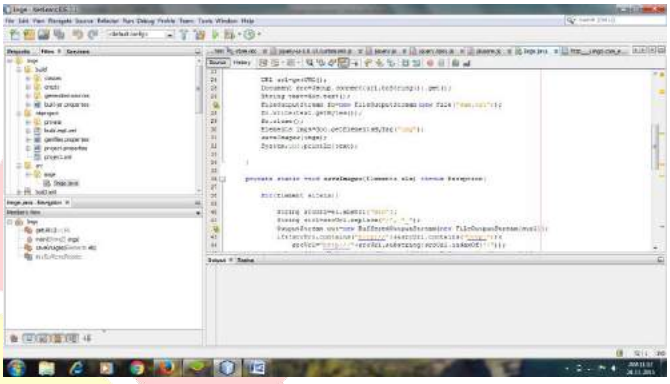
VII.CONCLUSION

This paper explains the basic purpose of various algorithms related to clustering and different content mining tools which have been already used for retrieval of web information. With that we can have the basic idea about the approach. Also it helps to add distinct feature by developing the new tool named Magnified Content Extractor to make the web access in efficient and in fast manner for both unstructured and unstructured document. The study of this paper focuses automation tool for better understanding and time consuming. This tool may help to segregate the objects as well with clustering process.

VIII. SAMPLE SCREEN SHOTS

In the below figure the url is taken as input and the executed code will produce the output like separating the images and grouped in one folder.





XI.FUTURE ENHANCEMENT

The Tool named MCE is the best to use with structured and unstructured data. In future the same tool can be applicable for extracting information along the same can be enhanced to only focus on images. For Example, images have the following file formats TIF, JPEG, JPG, PNG, GIF. If we are opening one web page in that we can acquire only images among the retrieved images particular type of file can be segregated towards to its file type.

References:

[1] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava, "Web Mining: information and Pattern Discovery on the WWW"

[2] Mary Garvin, "Data Mining and the Web: What They Can Do Together"

[3] Han J Kamber M, "Data Mining: concepts and Techniques", Second Edition Morgan Kaufmann publishers .2006

[4]Lieu, B., Web Data Mining Exploring Hyperlinks, Contents, and Usage Data (Springer-Verlag, Berlin, Heidelberg 2007).

[5] M.Zdravko, T.L. Daniel,, Data mining the Web : Uncovering patterns in Web content, structure & usage (WileyInterscience Publication, 2007).

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. 2, Special Issue I, January 2016 in association with
KARUR VELALAR COLLEGE OF ARTS AND SCIENCE FOR WOMEN, KARUR
Conference on Emerging Trends and Functional Applications of Computer Technology - 12th January 2016

[6] J. Srivastava, P. Desikan, V. Kumar, "Web Mining – Concepts, Applications and Research Directions", Studies in Fuzziness and Soft Computing, Volume 180, pp. 275–307, (2005).
[7] B.Masand, M.Spiliopoulou, J. Srivastava, O.Zaiane, ed. Proceedings of "WebKDD2002 –Web Mining for Usage Patterns and User Profiles", Edmonton, CA, 2002.
[8] M. Spiliopoulou, "Data Mining for the Web", Proceedings of the Symposium on Principles of Knowledge Discovery in Databases (PKDD), 1999.
[9] Screen-scraper, http://www.screen-scraper.com Viewed 19 February 2013.
[10] Automation Anywhere Manual. AA, http://www.automationanywhere.com Viewed 06 February 2013.
[11]Mozenda,http://www.mozenda.com/web-mining-software Viewed 18 February 2013.
[12] Web Content Extractor help. WCE, http://www.newprosoft.com/web-content-extractor.htm Viewed 18 February 2013.
[13] Raymond Kosala, Hendrik Blockee, "Web Mining Research: A Survey", ACM Sigkdd Explorations Newsletter, June 2000, Volume 2.
[14] Magdalini Eirinaki "Web Mining: A Roadmap "Http://WWW.engr.sjsu.edu/ meirinaki/papers/NEIS.pdf
[15] Qingyu Zhang & Richard S. Segall, "Web Mining: A Survey of Current Research", Information Technology and Decision Making, 7(4), 683- 720, 2008.