www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 3 March 2013 Page No. 759-764

Information Retrieval From Web Document Using Clustering Techniques

Mr..Keole.Ranjit R * ***Dr..Karde.Pravin.P***
Assistant Prof. Department of Information Technology H. V. P. M. College of Engineering and Technology, Amravati.
ranjitkeole@gmail.com p_karde@rediffmail.com

Abstract -With more than two billion pages created by millions of Web page authors and organizations, the World Wide Web is a tremendously rich knowledge base. The knowledge comes not only from the content of the pages themselves, but also from the unique characteristics of the Web, such as its hyperlink structure and its diversity of content and languages. A considerably large portion of information present on the world wide web (www) today is in the form of unstructured or semi-structured text data bases. It becomes tedious for the user to manually extract real required information from this material. Large document collections, such as those delivered by Internet search engines, are difficult and time-consuming for users to read and analyze. The detection of common and distinctive topics within a document set, together with the generation of multi-document summaries, can greatly ease the burden of information management. Clustering is useful technique in the field of textual data mining. Cluster analysis divides objects into meaningful groups based on similarity between objects. This paper focus on this problem of mining the useful information from the collected web documents using clustering techniques of the text collected from the downloaded web documents.

Keywords: - web mining, cure, birch, rock, fuzzy clustering, text mining.

INTRODUCTION

With more than two billion pages created by millions of Web page authors and organizations, the World Wide Web is a tremendously rich knowledge base. The knowledge comes not only from the content of the pages themselves, but also from the unique characteristics of the Web, such as its hyperlink structure and its diversity of content and languages. A considerably large portion of information present on the World Wide Web (www) today is in the form of unstructured or semi-structured text data bases. The www instantaneously delivers huge number of these documents in response to a user query. However, due to lack of structure, the users are at a loss to manage the information contained in these documents efficiently. The www continues to grow at an amazing rate as an information gateway and as a medium for conducting business. Web mining is the extraction of interesting and useful knowledge and implicit information from artifacts or activity related to the www [9]. In this context, the importance of data/text mining and knowledge discovery is increasing in different areas like: telecommunication, credit card services, sales and marketing etc [3]. Text mining is used to gather meaningful information from text and includes tasks like Text Categorization, Text Clustering, Text Analysis and Document Summarization. Text Mining examines unstructured textual information in an attempt to discover structure and implicit meanings within the text. One main problem in this area of research is regarding organization of document data. This can be achieved by developing nomenclature or topics to identify different documents. However, assigning topics to documents in a large collection manually can prove to be an arduous task. Clustering is the proven technique for document grouping and categorization based on the similarity between these documents [13]. Documents within one cluster have high similarity with each another, but low similarity with documents in other clusters [8]. Over the last decade there is tremendous growth of information on World Wide Web (www).It has become a major source of information. Web mining is an application of data mining techniques to discover and extract

information from Web. Data mining has the role of searching large volumes of data and extracting Knowledge from data [12]. Clustering is one of the possible techniques to improve the efficiency in information finding process. It is a Data mining tool to use for grouping objects into clusters such that the objects from the same cluster are similar and objects from different cluster are dissimilar.

II. WEB MINING

Web mining is the use of Data mining techniques to automatically discover and extract information from World Wide Web. Web mining is a comprehensive technology, related to web, data mining, computer linguistics, information theory. It is the application of data-mining techniques to extract knowledge from web data, in which at least one of structure or usage (web log) data is used in the mining process [11]. Web mining tries to extract interesting, potentially useful and hidden information from the documents & help people abstract knowledge from WWW by way of data mining. Based on the different emphasis and different ways to obtain information, web mining can be divided into three major parts: Web Contents Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM). Web Contents Mining can be described as the automatic search and retrieval of information and resources available from millions of sites and on-line databases though search engines or web spiders. Web Structure mining pertains to mining the structure of hyperlinks within the web itself [11]. Web Usage Mining can be described as the discovery and analysis of user access patterns, through the mining of log files and associated data from a particular Web site.

III. WEB MINING AND INFORMATION RETRIEVAL

Some have claimed that resource or document discovery (IR) on the Web is an instance of Web (content) mining and others associate Web mining with intelligent IR. Actually IR is the automatic retrieval of all relevant documents while at the same time retrieving as few of the non-relevant as possible [11]. IR has the primary goals of indexing text and searching for useful documents in a collection and now a day's research in IR includes modeling, document classification and categorization, user interfaces, data visualization, filtering, etc. The task that can be considered to be an instance of Web mining is Web document classification or categorization, which could be used for indexing. Viewed in this respect, Web mining is part of the (Web) IR process. However, it is noted that not all of the indexing tasks use data mining techniques. IE has the goal of transforming a collection of documents, usually with the help of an IR system, into information that is more readily digested and analyzed. IE aims to extract relevant facts from the documents while IR aims to select relevant documents. While IE is interested in the structure or representation of a document, IR views the text in a document just as a bag of unordered words. Thus, in general IE works at a finer granularity level than IR does on the documents. However, the differences between the two become blurred if the interest of IR is in extraction and when used in the context of vague forms of information in which a full text IR system can provide some IE features. Building IE systems manually is not feasible and scalable for such a dynamic and diverse medium such as Web contents. Due to this nature of the Web, most IE systems focus on specific Web sites to extract. Others use machine learning or data mining techniques to learn the extraction patterns or rules for Web documents semi-automatically or automatically. Within this view, Web mining is part of the (Web) IE process. Other views regarding the relationship between (Web) IE and Web mining also exist. The results of the IE process could be in the form of a structured database or could be a compression or summary of the original text or documents. One could view for the former that IE is a kind of re-processing stage in the Web mining process, which is the step after the IR process and before the data mining techniques are being performed. In a similar view, IE can also be used to improve the indexing process, which is part of the IR process.

IV. TRADITIONAL INFORMATION RETRIEVAL TECHNIQUES

Traditional information retrieval methods represent plain-text documents using a series of numeric values for each document. Each value is associated with a specific term (word) that may appear on a document, and the set of possible terms is shared across all documents. The values may be binary, indicating the presence or absence of the corresponding term. The values may also be a non-negative integers, which represents the number of times a term appears on a document (i e. term

frequency). Non-negative real numbers can also be used, in this case indicating the importance or weight of each term. These values are derived through a method such as the popular inverse document frequency (tf-idf) model, which reduces the importance of terms that appear on many documents. Regardless of the method used, each series of values represents a document and corresponds to a point (i e. vector) in a Euclidean feature space; this is called the vector-space model of information retrieval. This model is often used when applying machine learning techniques to documents, as there is a strong mathematical foundation for performing distance measure and centroid calculations using vectors.

V. INFORMATION RETRIEVAL VIEW FOR UNSTRUCTURED DOCUMENTS

The unstructured documents are free texts. Most of the research uses bag of words to represent unstructured documents. The bag of words or vector representation takes single word found in the training corpus as features. This representation ignores the sequence in which the words occur and is based on the statistic about single words in isolation. The features could be Boolean (a word either occurs or does not occur in a document), or frequency based (frequency of the word in a document). Variations of the feature selection include removing the case, punctuation, infrequent words, and stop words. The features could be reduced further by applying some other feature selection techniques, such as information gain, mutual information, cross entropy, or odds ratio. Other preprocessing includes Latent Semantic Indexing (LSI) that seeks to transform the original document vectors to a lower dimensional space by analyzing the correlation structure of terms in the document collection such that similar documents that do not share terms are placed in the same topic, and stemming which reduces words to their morphological roots. For example the words "informing", "information", "informer", and "informed" would be stemmed to their common root "inform" and only the latter word is used as the feature instead of the former four. While those pre-processing variations are useful for reducing feature set size, the generality of their effectiveness over different domains for text categorization tasks are doubted. Other feature representations are also possible such as using information about word positions in the document, using n-grams representation(word sequences of length up to n) for example "the morphological roots" is a tri-gram, using phrases such as "the quick brown fox that run away", using document concept categories, using terms such as "annual interest rate "or "Wall Street", using hyponyms (linguistic term for the "is a" relationship - a dog is an animal, thus "animal" is a hyponym of "dog"), or using named entities such as people's names, dates, email addresses, locations, organizations, or URLs.

VI. THE WEB TEXT MINING PROCESS

The information coming from Web represents an important role in Knowledge Discovery Process. To examine web contents and to extract useful information to a purpose through techniques of Mining is scientifically called Web Text Mining the Web Text Mining Process is implemented through four steps. The first step foresees the recovery of the useful information from Web. This information consists in textual contents present in web pages. The references solution covers the first step of Web Text Mining process, dealing with automatic retrieval of all relevant documents. The recovery of the useful information is affected through crawling departing from one or more URL. The crawling is implemented by a Web Crawler. Web Crawler's architecture fits the guide lines of to Focused Crawler, as it is designed to only gather documents on to specific topic, thus reducing the amount of network traffic and downloads. It is composed from four components: Master, Slaves, Scribes, H-Information. Every component has a specific assignment. H-Information (Human Information) is the start of whole process of crawling. It allows the system administrator to interact with Web Crawler inserting a list of sites Web believed of particular importance for the service that the system must realize. Master is the kernel of Web Crawler that allows the start and the management of the whole process of recovery of the information. Master calls Slaves and passed them URLs of the pages in accord with the built list by the H-Information. Slaves accede to the Web for crawling the pages which are associated to URLs indicated by Master. Different Slaves can operate at the same time rendering more efficient, reducing the answer times and avoiding simultaneous accesses to the same resources. Slaves call Scribes. Scribes receive Web pages crawled from Slaves. For each Scribe there is a Slave. The second step is the pre-processing. It foresees the creation of a repository of the

information from the web during the execution of the first step. During this phase it had to provide to affect a control and a cleaning (Text Cleaning) of the pages reached during the first step. The definition of Template, which contains keywords of the universe, allows a choice on the single content of the examined web page. If it contains information that correspond to one of the Template defined for the system in matter, it is accepted and on it a cleaning of the content is effected to eliminate all what to be held "garbage", that is to say tag, banner, information not tightly connected to the content of the same page, Otherwise the page is discarded. The third step is the fundamental step on which the whole Web Mining process founds him. In this step a first phase of Information Extraction and the following phase of Text Mining are affected. It founds on the use of Text Mining tools that receiving in input the data which are contained in the repository of the second step and they are able to extract "intrinsic information" contained in the document or in all documents. Rules are the most important part of Text Mining and their definition from the programmer will allow to get or not a good result of the whole Web Text Mining process. Before applying the rules, Text Mining software handles the phase of Information Extraction effecting a tokenization and lemmatization of the text that it will allow an accurate analysis of the document during the Text Mining. Tokenization is a process in which a text is split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. This tokenized representation is then used for further processing. The fourth is last step of the process; it foresees the phase of presentation of the obtained results. In our architecture, the information drawn out during the execution of the third step is stored in a second repository. This repository created during this phase will follow the rules of the Web Warehousing according to the Whoweda schemas. The repository is composed by URLs and nodes. URLs are the reference URLs of the document, that is to say, “referenced from" and "it references who" whereas the nodes are the skills found in the third step.

VII. DOCUMENT CLUSTERING

Clustering documents is a form of data mining that is concerned mainly with text mining. According to a survey by Kosala and Blockeel [11] on web mining , currently the term text mining has been used to describe different applications such as text categorization , text clustering ,empirical computational linguistic tasks ,exploratory data analysis , finding patterns in text databases ,finding sequential patterns in text ,and association discovery. Document clustering can be viewed from different perspectives, according to the methods used for document representation, document processing, methods used, and applications. From the viewpoint of the information retrieval (IR) community (and to some extent Machine Learning community), traditional methods for document representation are used, with a heavy predisposition toward the vector space model.

VIII. CLUSTERING TECHNIQUES

Various techniques for accurate clustering have been proposed, K-MEAN [16], CURE [6], BIRCH [17], ROCK [14].

K-MEAN clustering algorithm is used to partition objects into clusters while minimizing sum of distance between objects and their nearest center. In statistics and machine learning, *k*-means clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.

CURE (Clustering Using Representation) represents clusters by using multiple well scattered points called representatives. A constant number ‘c’ of well scattered points can be chosen from ‘2c’ scattered points for merging two clusters. CURE can detect clusters with non spherical shapes and works well with outliers. CURE achieves this by representing each cluster by a certain fixed number of points that are generated by selecting well scattered points from the cluster and then shrinking them toward the center of the cluster by a specified fraction. Having more than one representative point per cluster allows CURE to adjust well to the geometry of non-spherical shapes and the shrinking helps to dampen the effects of outliers. To handle large databases, CURE employs a combination of random sampling and partitioning. A random sample drawn from the data set is first partitioned and each partition is partially clustered. The partial clusters are then clustered in a second pass to yield the desired clusters.

BIRCH (Balance and Iterative Reducing and Clustering Hierarchies) is useful algorithm for data represented in vector space. It also works well with outliers like CURE [6]. BIRCH incrementally and dynamically clusters incoming multi-dimensional metric data points to try to produce the best quality

clustering with the available resources. BIRCH can typically find a good clustering with a single scan of the data, and improve the quality further with a few additional scans. BIRCH is also the first clustering algorithm proposed in the database area to handle "noise)' (data points that are not part of the underlying pattern) effectively.

ROCK (Robust Clustering Algorithm for Categorical Attributes) gives better quality clusters involving categorical data as compared with other traditional algorithms.

In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster.

HARD CLUSTERING, Hard c-means is better known as k-means and in general this is not a fuzzy algorithm. However, its overall structure is the basis for all the others methods. Therefore we call it hard c-means in order to emphasize that it serves as a starting point for the fuzzy extensions hard c-means is a crisp algorithm such that each object belongs to exactly one cluster. So, in hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster.

In FUZZY CLUSTERING, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

HYPER-SPHERICAL FUZZY C-MEANS (H-FCM), recently the Fuzzy c-Means (FCM) algorithm is modified for clustering text documents based on the cosine similarity coefficient rather than on the Euclidean distance. The modified algorithm works with normalized *k*-dimensional data vectors that lie in hyper-sphere of unit radius and hence has been named Hyper-spherical Fuzzy c-Means (H-FCM). The H-FCM algorithm for document clustering has shown that it outperforms the original FCM algorithm as well as the hard k-Means algorithm.

IX. CONCLUSION & FUTURE SCOPE

This paper presents an overview of web mining for information retrieval from web documents using clustering algorithms that could be potentially suitable for document clustering; we have surveyed HARD K –MEANS (HCM), Fuzzy C –MEANS (FCM) and HYPER-SPHERICAL FUZZY C-MEANS (H-FCM) clustering algorithms. The HCM algorithm has a tendency to get stuck in a local minimum, which makes it necessary to conduct several runs of the algorithm with different initializations. In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. The H-FCM generates clusters with a higher level of granularity and that the resulting clusters hierarchy successfully links clusters of the same topic. There are many areas in text mining, where one may carry the work to enhance various areas. Out of these, the labeling of the clusters is a very daunting challenge of this time. No remarkable effort has been made in this regard to get good result. That is why automatic labeling of the clusters is not so much accurate. A keen and concerted work has been done to remove this hurdle. It will certainly serve as a lime length for future researchers.

REFERENCES

[1] Bezdek J. C, "*Pattern Recognition with Fuzzy Objective Function Algorithms*", Plenum Press, New York, 1988.

[2] Hillol Kargupta, Anupam Joshi, Krishnamoorthy Sivakumar and Yelena Yesha, "Data Mining: Next Generation Challenges and Future Directions", MIT Press,USA , 2004 .

[3] Hsinchun Chen and Michael Chau, "Web Mining: Machine learning for Web Applications", *Annual Review of Information Science and Technology 2003.*

[4] Jingwen Tian, Meijuan Gao, and Yang Sun, "Study on Web Classification Mining Method Based on Fuzzy Neural Network", Shenyang, China August 2009.

[5] King-Ip Lin and Ravikumar Kondadadi, "A Similarity Based Soft Clustering Algorithm for Documents", *in Proceeding of the 7th International Conference on Database Systems for Advanced Applications (DASFAA-2001)*, April 2001.

[6] Linas Baltruns and Juozas Gordevicius, "Implementation of CURE Clustering Algorithm", *SIGMOD Seattle*, WA, USA ACM February 1, 2005.

[7] Mr. Rizwan Ahmad and Dr. Aasia Khanum, "Document Topic Generation in Text Mining by Using Cluster Analysis with EROCK", *International Journal of Computer Science & Security (IJCSS)*, Volume (4): Issue(2) Aug 2008.

[8] Nicholas O. Andrews and Edward A. Fox, "Recent Development in Document Clustering

Techniques", Dept of Computer Science, Virgina Tech 2007.
[9] Oren Etzioni, "The World Wide Web:quagmire or gold mine?" Communications of ACM", Nov 96.
[10] R. Cooley,B. Mobasher and J. Srivastava ,"Web Mining: Information and Pattern Discovery on the World Wide Web", *In the Proceeding of ninth IEEE International Conference on Tool s with Artificial Intelligence (ICTAI'97),*1997.
[11] R. Kosala and H.Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations , July 2000.
[12] Ramakrishna, M.T. Gowdar, L.K. Havanur, M.S. Swamy (2010), "Web Mining: Key Accomplishments, Applications and Future Directions", *International Conference on Data Storage and Data Engineering (DSDE), pp.187 – 191.*
[13] Schenker, M. Last and A. Kandel (2001)," A term-based algorithm for hierarchical clustering of Web documents" *in Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, vol.5, pp. 3076-3081, Vancouver, Canada, July 2001.
[14] Shady Shehata, Member, IEEE, Fakhri Karray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow, IEEE, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", *IEEE Trans. On Knowledge and Data Engineering*, Vol. 22, No. 10, Oct 2010.
[15] Shaoxu Song and Chunping Li, "Improved ROCK for Text Clustering Using Asymmetric Proximity", *SOFSEM* 2006, LNCS 3831, pp. 501–510, 2006.
[16] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, "An Efficient k-Means Clustering Algorithm: Analysis and implementation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 7, July 2002.
[17] Tian Zhang, Raghu Ramakrishna, Miron Livny, "BIRCH: An Efficent Data Clustering Method for Very Large Databases" *SIGMOD* '96 6/96 Montreal, Canada
[18] Valente de Oliveira and J. Pedrycz W," *Advances in Fuzzy Clustering and its Applications*", John Wiley & Sons, pp 3-30, 2007.
[19] WangBin and LiuZhijing , "Web Mining Research" , *in Proceeding of the 5th International Conference on Computational Intelligence and Multimedia Applications(ICCIMA'03)* 2003.