ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Vector Space Model in Clustering Web Documents

Ms. Anjali B. Raut*
Department of CSE
HVPM's COET
Amravati,India.
anjali_dahake@rediffmail.com

Dr. G. R. Bamnote
Professor & Head
Department of CSE, PRMITR ,Badnera
Amravati ,India.
grbamnote@rediffmail.com

Abstract: Web Mining is the use of Data Mining techniques to automatically discover and extract information from web. In Web Text Mining clustering is of the data mining tool used for grouping web documents into clusters such that the objects from the same cluster are similar and objects from different cluster are dissimilar .Vector space model is the most Widely used model used to represent document. In this paper we present the basic concept of web mining and its different types also we give different methods of implementing vector space model.

***Keywords*:** Web Mining, Clustering, Vector space model ,Information Retrieval ,Data Mining

I. INTRODUCTION

Over the last decade there is tremendous growth of information on World Wide Web (WWW).It has become a major source of information. Web creates the new challenges of information retrieval as the amount of information on the web and number of users using web growing rapidly. It is practically impossible to search through this extremely large database for the information needed by user . Hence the need for Search Engine arises. Search Engines uses crawlers to gather information and stores it in database maintained at search engine side. For a given user's query the search engine searches in the local database and very quickly displays the results.

The ability to form meaningful groups of objects is one of the most fundamental modes of intelligence. Human perform this task with remarkable ease. Cluster analysis is a tool for exploring the structure of data. The core of cluster analysis is clustering; the process of grouping objects into clusters such that the objects from the same cluster are similar and objects from different cluster are dissimilar. The need to structure and learn vigorously growing amount of data has been a driving force for making clustering a highly active research area.

Web Mining is the use of Data Mining techniques to automatically discover and extract information from web. 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.

The paper is structured as follows: section 2 describes some related work about Web Mining, Web mining taxonomy, its different processes and clustering algorithms. Section 3 gives different methods of vector space model Section 5 shows the results. In section 4, conclusion is given.

II. RELATED WORK

Data Mining has emerged as a new discipline in world of increasingly massive datasets. Data Mining is the process of extracting or mining knowledge from data. Data Mining is becoming an increasingly important tool to transform data into information. Knowledge Discovery from Data i.e. KDD is synonym for Data Mining.

A. Web Mining:

World Wide Web is a major source of information and it creates new challenges of information retrieval as the amount of information on the web increasing exponentially. Web Mining is use of Data Mining techniques to automatically discover and extract information from web documents and services [1].

Oren Etzioni was the person who coined the term Web Mining first time. Initially two different approaches were taken for defining Web Mining. First was a "process-centric view", which defined Web Mining as a sequence of different processes [1] whereas, second was a "data-centric view" , which defined Web Mining in terms of the type of data that was being used in the mining process [2]. The second definition has become more acceptable, as is evident from the approach adopted in most research papers[3][5]. Web Mining is also a cross point of database, information retrieval and artificial intelligence [4].

B. Web Mining Process:

Web mining may be decomposed into the following subtasks:

a. Resource Discovery: process of retrieving the web resources.
b. Information Pre-processing : is the transform process of the result of resource discovery
c. Information Extraction: automatically extracting specific information from newly discovered Web resources.
d. Generalization: uncovering general patterns at individual Web sites and across multiple sites[3].

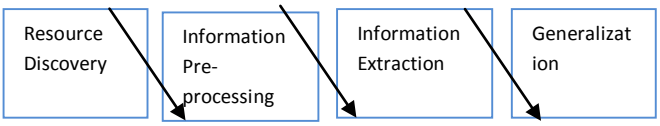

Figure 2.1: Web mining Process

C. Web Mining Taxonomy:

Web has different facets that yield different approaches for the mining process:

© 2010, IJARCS All Rights Reserved

CONFERENCE PAPER
National Conference on Information and Communication Technology for Development
Organized by PRMITR, Amravati (MS) India
http://mitra.ac.in/forthcoming.html

a) 1.Web pages consist of text.
b) 2.Web pages are linked via hyperlinks
c) 3.User activity can be monitored via Web server logs .

This three facets leads to the distinction into three categories i.e. Web content mining, Web structure mining and Web usage mining [4-7]. Following Fig 2.2 shows the Web Mining Taxonomy.

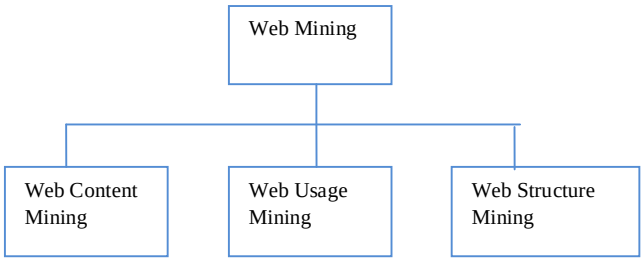

Figure 2.2: Web mining Taxonomy

a. Web Content Mining (WCM):

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed such as lists and tables. Application of text mining to web content has been the most widely researched.

b. Web Structure Mining (WSM):

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web.

c. Web Usage Mining(WUM):

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data. Web usage data includes data from web server logs, browser logs, user profiles, registration data, cookies etc.

WCM and WSM uses real or primary data on the web whereas WUM mines the secondary data derived from the interaction of the users while interacting with the web.

D. Clustering:

The Web is the largest information repository in the history of mankind. Finding the relevant information on www is not an easy task. The information user can encounter the following problems when interacting with the web[2].

a. low precision: Today's search tools have the low precision problem , which is due to the irrelevance of many search results. This results in a difficulty finding the relevant information.
b. Low recall: It is due to the inability to index all the information available on the web . This results in a difficulty finding the un indexed information that is relevant.

Clustering is one of the Data Mining techniques to improve the efficiency in information finding process. Many clustering algorithms have been developed and used in many fields. A. K. Jain, M. N. Murty and P. J. Flynn[8] provides an extensive survey of various data clustering techniques. Clustering algorithms can be broadly categorized into partitional and hierarchical techniques.

Agglomerative hierarchical clustering (AHC) algorithms are most commonly used .It use a bottom –up methodology to merge smaller cluster into larger ones , using techniques such as minimal spanning tree . These algorithms find to be slow when applied to large document collection. It has different variants such as single-link, group-average and complete-link. Single-link and group-average methods typically takes $O(n^2)$ time while complete-link method typically takes $O(n^3)$ time.

Partition algorithm such as K- means are linear time algorithms . It try to divide data into subgroups such that the partition optimizes certain criteria , like inter – cluster distance or intra- cluster distances. They typically take an iterative approach. The time complexity of this algorithm is O(nkt) , where k is the number of desired clusters and T is the number of iterations.

Most of the document clustering algorithm worked on BOW (Bag Of Words)model[5].Oren Zamir and Oren Etzioni[9] in their research listed the key requirements of web document clustering methods as relevance, browsable summaries, overlap, snippet tolerance, speed and accuracy. They have given STC (Suffix Tree Clustering) algorithm which creates clusters based on phrase shared between documents. Michael Steinbach, George Karypis and Vipin Kumar[10] presented the result of an experimental study of some common document clustering algorithms. They compare the two main approaches of document clustering i.e. agglomerative hierarchical clustering and K-means method. Nicholas O. Andrews and Edward A. Fox[11] presented the recent developments in document clustering .

A single object often contains multiple themes like a web document on topic Web Mining may contain different themes like Data Mining, clustering and information retrieval. Many traditional clustering algorithms assign each document to a single cluster, thus making it difficult for the user to retrieve information. Based on this concept clustering algorithm can be divided into hard & soft clustering algorithm. In traditional clustering algorithm each object belongs to exactly one cluster where as in soft clustering algorithm each object can belongs to multiple clusters [12].

The conventional clustering algorithms in Data Mining have difficulties in handling the challenges posed by the collection of natural data which is often vague and uncertain. The modelling of imprecise and qualitative knowledge, as well as handling of uncertainty at various stages is possible through the use of fuzzy sets. Therefore a fuzzy clustering method was offered to construct clusters with uncertain boundaries, so this method allows that one object belongs to multiple clusters with some membership degree. Pawan Lingras , Rui Yan and Chad West[13] applied fuzzy technique to discover usage pattern from web data. The fuzzy c-means clustering was applied to the web visitors of educational websites. The analysis shows the ability of the fuzzy c-means clustering to distinguish different user characteristics. Anupam Joshi and Raghu Krishnapuram[14] developed a prototype Web Mining system which analyzes web access logs from a server and tries to mine typical user access pattern. Maofu Liu, Yanxiang He and Huijun Hu[15] proposed a web fuzzy clustering model.

© 2010, IJARCS All Rights Reserved

CONFERENCE PAPER
National Conference on Information and Communication Technology for Development
Organized by PRMITR, Amravati (MS) India
http://mitra.ac.in/forthcoming.html

III. VECTOR SPACE MODEL

In Web Text Mining clustering is of the data mining tool used for grouping web documents into clusters such that the objects from the same cluster are similar and objects from different cluster are dissimilar .Vector space model is the most Widely used model used to represent document. In the statistically based vector-space model, a document is conceptually represented by a vector of keywords extracted from the document, with associated weights representing the importance of the keywords in the document and within the whole document collection; likewise, a query is modelled as a list of keywords with associated weights representing the importance of the keywords in the query. The weight of a term in a document vector can be determined in many ways. A common approach uses the so called *tf * idf* method, in which the weight of a term is determined by two factors: how often the term *j* occurs in the document *i* (the term frequency $tf_{i,j}$) and how often it occurs in the whole document collection (the document frequency df_j). Precisely, the weight of a term *j* in document *i* is

$$w_{i,j} = tf_{i,j} * idf_j = tf_{i,j} * \log N/df_j$$

Where *N* is the number of documents in the document collection and *idf* stands for the inverse document frequency. This method assigns high weights to terms that appear frequently in a small number of documents in the document set. Once the term weights are determined, we need to measure similarity between the document vectors. A common similarity measure, known as the *cosine measure*, determines the angle between the document vectors and when they are represented in a *V*-dimensional Euclidean space, where *V* is the vocabulary size. The similarity between a document D_i and D_q is defined as

$$sim(D_q, D_i) = \frac{\sum_{j=1}^{V} w_{q,j} * w_{i,j}}{\sqrt{\sum_{j=1}^{V} w_{q,j}^2 * \sum_{j=1}^{V} w_{i,j}^2}}$$

Where $w_{q,j}$ is the weight of term *j* in the query, and is defined in a similar way as $w_{i,j}$. The denominator in this equation, called the normalization factor, discards the effect of document lengths on document scores. Thus, a document containing {*x, y, z*} will have exactly the same score as another document containing {*x, x, y, y, z, z*} because these two document vectors have the same unit vector. Because the exact vector-space model is expensive to implement, here we have given some family of successively simpler approximations.

a. Exact Normalization Method (Method 1): The complexity of this method, the *full vector-space model*, depends on how we implement it. The document's vector representation is only conceptual. In practice, the full vector is rarely stored internally as is because it is long and sparse. The similarities between document vectors are calculated using following formula.

$$sim(D_q, D_i) = \frac{\sum_{j=1}^{V} w_{qj} * w_{ij}}{\sqrt{\sum_{j=1}^{V} w_{qj}^2 * \sum_{j=1}^{V} w_{ij}^2}}$$

b. Approximate Normalization (Method 2): For this second method to approximate the effect of normalization use instead the square root of the number of terms in a document as the normalization factor. While this still favours long documents, the effect of document size is not as significant as it is without any normalization. This normalization factor is much easier to compute than the original one; also, precomputation is possible. With the approximation, the formula becomes.

$$sim(Dq, Di) = \frac{\sum_{j=1}^{V} w_{qj} * w_{ij}}{\sqrt{Number\ of\ terms\ in\ Di}}$$

c. Inner Product Method (Method 3): This method lets us further simplify the computation by simply dropping the normalization factor:

$$sim(D_q, D_i) = \sum_{j=1}^{V} w_{qj} * w_{ij}$$

That is, the document score equals the inner product of the document vectors. Instead of computing the angle between the document vectors, this formula computes the length of the projection of the document vectors. It is quite clear that the document score is directly proportional to the length of the document vector.

d. If only Method (Method 4): This method only makes use of term frequencies in the calculation and ignores *idf*. It simplifies the computation as well as saving the file structure needed for storing the *df* values.

$$sim(D_q, D_i) = \sum_{j=1}^{V} w_{q,j} * tf_{i,j}$$

e. idf only Method (Method 5): This method ignores the term frequency (*tf*) information but retains the *idf* values in determining term weights :The *idf* values have the same effect as before. That is, they diminish the significance of words that appear in a large number of documents.

$$sim(D_q, D_i) = \sum_{j=1}^{V} w_{q,j} * w_{i,j}$$

$$\text{Where } w_{q,j} = \begin{cases} 1 & \text{if } j \in D_q \\ 0 & \text{otherwise} \end{cases}$$

$$w_{i,j} = \begin{cases} idf_j & \text{if } j \in D_i \\ 0 & \text{otherwise} \end{cases}$$

f. No tf and no idf Method (Method 6): This method is the simplest in the family. It ignores both *tf* and *idf* values and therefore measures the number of common terms in the documents.

$$sim(D_q, D_i) = \sum_{j=1}^{V} w_{q,j} * w_{i,j}$$

$$\text{Where } w_{q,j} = \begin{cases} 1 & \text{if } j \in D_q \\ 0 & \text{otherwise} \end{cases}$$

$$w_{i,j} = \begin{cases} 1 & \text{if } j \in D_i \\ 0 & \text{otherwise} \end{cases}$$

IV. CONCLUSION

In this paper we presented family of six implementations of vector space model which we can use for clustering web documents. Following observations are regarding the different simplifications of the space-vector model[16].

a. The approximate normalization factor works just as well as the exact normalization factor. In fact, it is better than the exact normalization factor in case of when vector length is small. Only in one case, the vector length is large approximate normalization slightly worse than the exact normalization factor. This result indicates that vector lengths should not be completely discarded in calculating document scores.

b. The inner product (no normalization) is not as good as the first two methods, but provides a reasonable

© 2010, IJARCS All Rights Reserved

CONFERENCE PAPER
National Conference on Information and Communication Technology for Development
Organized by PRMITR, Amravati (MS) India
http://mitra.ac.in/forthcoming.html

compromise given that it does not require explicit computation and storage of the normalization factor.

c. The last three methods ignore some or all of the frequency information. In general, this group is not as good as the first three methods for natural-language queries. In particular method 4, which ignores *idf* without ignoring *tf*, is consistently worse than any of the first three methods.

d. Methods 5 and 6, which try to measure the overlap between the document and query terms, perform *extremely* well for the concept query, although they perform relatively poorly for natural-language queries

e. For concept queries, method 5, which takes *idf* into account, is better than method 6. This confirms that a concept that appears in many documents should be given a small weight. However, method 5 performs *extremely* poorly for natural-language queries. This may be attributed to the inconsistent use of frequency information. In other words, *idf* and *tf* each represent only half of the equation in computing term weights; therefore neither should be used alone.

V. REFERENCES

[1]. Oren Etzioni, "The World Wide Web: quagmire or gold mine?", Communications of ACM", Nov 96.

[2]. 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 Tools with Artificial Intelligence (ICTAI'97), 1997.

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

[4]. WangBin and LiuZhijing , "Web Mining Research" , In Proceeding of the 5th International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03) 2003.

[5]. R. Kosala and H.Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations ACM SIGKDD, July 2000.

[6]. Sankar K. Pal,Varun Talwar and Pabitra Mitra , "Web Mining in Soft Computing Framework : Relevance, State of the Art and Future Directions ", IEEE Transactions on Neural Network , Vol 13,No 5,Sept 2002 .

[7]. Andreas Hotho and Gerd Stumme, "Mining the World Wide Web- Methods, Application and Perceptivities", in Künstliche Intelligenz, July 2007. (Available at http://kobra.bibliothek.uni-kassel.de/)

[8]. A. K. Jain,M. N. Murty and P. J. Flynn, "Data clustering: A review," ACM computing surveys,31(3):264-323,Sept 1999.

[9]. O. Zamir and O. Etzioni, "Web document clustering: A feasibility demonstration", in Proceeding of 19th International ACM SIGIR Conference on Research and Development in Informational Retrieval , June1998.

[10]. Michael Steinbach, George Karypis and Vipin Kumar, "A Comparison of Document Clustering Techniques", KDD Worksop on Textmining, 2000.

[11]. Nicholas O. Andrews and Edward A. Fox, "Recent Development in Document Clustering Techniques", Dept of Computer Science, Virgina Tech 2007.

[12]. 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.

[13]. Pawan Lingras, Rui Yan and Chad West, " Fuzzy C-Means Clustering of Web Users for Educational Sites", Springer Publication ,2003.

[14]. Anupam Joshi and Raghu Krishnapuram , " Robust Fuzzy Clustering Methods to Support Web Mining", Proceedings of the Workshop on Data Mining and Knowledge Discovery , SOGMOD ,1998 .

[15]. Maofu Liu, Yanxiang He and Huijun Hu ," Web Fuzzy Clustering and Its Applications In Web Usage Mining", Proceedings of 8th International Symposium on Future Software Technology (ISFST-2004).

[16]. Lee,Chuang &Kent,"Document ranking and Vector space model",in the proceeding of IEEE Software April 1997

© 2010, IJARCS All Rights Reserved

CONFERENCE PAPER
National Conference on Information and Communication Technology for Development
Organized by PRMITR, Amravati (MS) India
http://mitra.ac.in/forthcoming.html