IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. IV (Jan – Feb. 2015), PP 55-59
www.iosrjournals.org

Challenging Issues and Similarity Measures for Web Document Clustering

S. Mahalakshmi
Research Scholar, Bharathiar University, Coimbatore, India.

Abstract: *Web itself contains a large amount of documents available in electronic form. The available documents are in various forms and the information in them is not in organized form. The lack of organization of materials in the WWW motivates people to automatically manage the huge amount of information. Text-mining refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining framework contains Information Retrieval, Information Extraction, Information Mining and Interpretation. During Information Retrieval, so many web documents are retrieved. In that how we can find out similar documents among retrieved? This paper deals with the challenging issues and similarity measures for web document clustering .*
Key words: *Text Mining, Information Retrieval,Framework,Information Extraction,Similarity,Clustering*

I. Introduction

The growth of information in the web is too large, so search engine come to play a more critical role to find relation between input keywords. Similarity Measure is widely used in Information Retrieval (IR) and also it is important component in various tasks on the web such as relation extraction, community mining, document clustering, and automatic metadata extraction. The main goal of clustering is to partition documents into homogeneous groups according to their characteristics and abilities. Text Clustering is to find out the groups information from the text documents and cluster these documents into the most relevant groups. Text clustering groups the document in an unsupervised way and there is not label or class information. Clustering methods have to discover the connections between the document and then based on these connections the documents are clustered.

Given huge volumes of documents, a good document clustering method may organize those huge numbers of documents into meaningful groups, which enable further browsing and navigation of this corpus be much easier. A basic idea of text clustering is to find out

which documents have many words in common and place these documents with the most words in common into same group Each cluster is a collection of data objects that are similar to one another are placed within the same cluster but are dissimilar to objects in other clusters. This paper discuss about the role of similarity in clustering and how to find similarity between the retrieved web documents and the problems faced during similarity measures.

- Section 2 presents Review of Related Work.
- Section 3 introduces Challenging Issues in Text Mining.
- Section 4 Describes Challenging Issues In Web Clustering Based On Similarity
- Section 5 describes Conclusion.

II. Review Of Related Literature

Information on the Web is present in the form of text documents (formatted in HTML), and that is the reason many Web document processing systems are rooted in text data mining techniques. [14]Due to the growth of information in web leads to drastic increase in field of information retrieval. Efficient information retrieval and navigation is provided by document clustering. Document clustering is the process of automatically grouping the related documents into clusters. Instead of searching entire documents for relevant information, these clusters will improve the efficiency and avoid overlapping of contents. Relevant document can be efficiently retrieved and accessed by means of document clustering.

String based similarity contains similarity measures like character based and term based similarity [1]. Wael H. Gomaa , Aly A. Fahmy explained different types of similarity approaches and finally they conclude that hybrid similarity will give better results[2].Finding similarity between documents will be useful in clustering and clustering is used to find intrinsic structures in data and organize them into subgroups

[3]. Mark.Dixon[4] describes about the framework for text mining and it is closely related with IE and IR.Different similarity measures were explained by Anna Huang [5]. As claimed by [6], the ambiguity is still the major "world problem" in text mining applications. As a result, most approaches for clustering non-segmented documents consist of two phases: a text mining process to extract the keywords, and a document clustering process to compute the similarity between the input documents[8]. The vector space representation of text is an incredibly powerful tool. Any text can be treated as a vector in a V-dimensional vector-space (Jaime Arguello) [11]. Documents are matched with a query based on their similarity. If a document is similar to the query, it is likely to be relevant. Non-binary weights for index terms in queries and documents are used in the calculation of degree of similarity. Decreasing order of this degree of similarity for the retrieved documents gives the ranked documents with partial match (Manwar et al.) [12]. In research paper[13], Wei Ning proved that in some document corpus K-Means might even achieve a better performance with the help of SVD although including SVD into the clustering processing might result in more time consumption. Thereby we suggest K-Means a good candidate on text mining and organization of large document corpus. Further research might be made on the feasibility of combining Frequent Itemset and K-Means.

In essence, document clustering is to group documents based on how relative they are. To cluster documents correctly, it is very important to measure how much a document is relative to another. And the extent of relativeness should be some real numbers and can be compared[13].Also they explained about quality measures for clustering . Basically there are two measures to evaluate how good a clustering algorithm is. One is Precision rate and the other is Recall rate.

III. Challenging Issues In Web Clustering Based On Similarity

The major challenging issue in text mining arise from the complexity of the natural language itself.The natural language is not free from the ambiguity problem.Ambiguity means the capability of being understood in two or more possible senses or ways.One word may have different meanings.One phrase or sentence may have multiple meanings.One phrase or sentence can be interpreted in various ways,thus various meanings can be obtained.Although a number of Researches have been conducted in Resolving the ambiguity problem, the work is still immature.

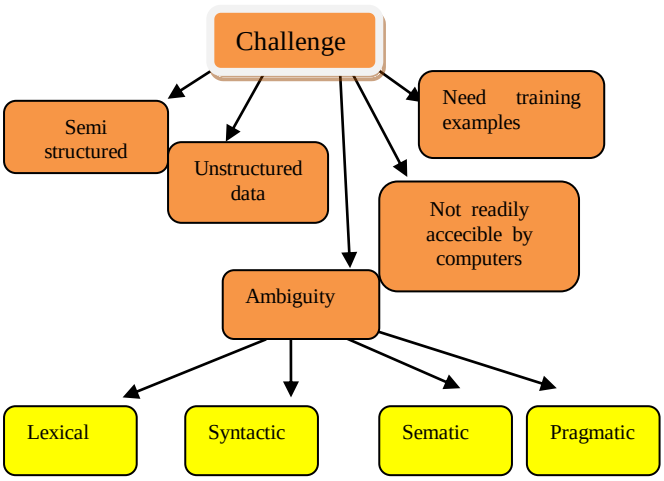

Fig 1: Challenging issues in Web document clustering

Major problems and the impact of similarity measures in web clustering are discussed below:

i) The World Wide Web is huge, widely distributed; global information service centre in this retrieving accurate information for users in Search Engine faces a lot of problems. This is due to accurately measuring the semantic similarity between words is an important problem, and also efficient estimation of semantic similarity between words is critical for various natural language processing tasks such as word sense disambiguation, textual entailment, and automatic text summarization. In information retrieval, one of the main problems is to retrieve a set of documents that is semantically related to a given user query. Retrieving accurate information to users to such kind of similar words is challenging. Some existing system proposed an architecture and method to measure semantic similarity between words, which consists of snippets, page-counts and two class support vector machine. Context-Aware Semantic Association Ranking can be applied to enhance the results with ranking [7].

i) Cluster Validity: Recent developments in data stream clustering have heightened the need for determining suitable criteria to validate results. Most outcomes of methods are depended to specific application. However, employing suitable criteria in results evaluation is one of the most important challenges in this area.

ii) Authors LIN, Zhenjiang [9] explained about challenges in similarity measures using link based methods. They explained that unimportant neighbors are pruned using MatchSim and PageSim Algorithms. They first proposed two novel neighbor-based similarity measures called MatchSim and PageSim, respectively. MatchSim takes the similarity between neighbors into account by defining recursively the similarity between objects by the average similarity of their maximum matched similar neighbors. PageSim measures the influences of indirect neighbors by adopting feature propagation strategy.They also proposed the Extended Neighborhood Structure (ENS) model which defines a bi-directional and multi-hop model, to help neighbor-based methods achieve higher accuracy.

So first there is a challenge for the researches to improve the efficiency of MatchSim algorithm in order to make it practical. Second, in MatchSim and PageSim, we prune unimportant neighbors according to the PageRank scores. There are other possible ranking methods, such as IDF-like weighting scheme, which may help to produce better results. Third, many kinds of properties of objects can be exploited to measure similarity, so how to integrate the link-based methods or similarity results with others is always a practical issue for us. So there is a challenge for researchers to integrate IDF-like weighting scheme with link based methods to produce better results.

iii)Clustering is a widely used technique to partition a set of heterogeneous data to homogeneous and well separated groups. Its main characteristic is that it does not require a-priori knowledge about the nature and the hidden structure of the data domain. In this thesis [10], they investigate clustering techniques and their applications to Web text and video information retrieval. In particular they focus on: web snippets clustering, video summarization and similarity searching. For web snippets, clustering is used to organize the results returned by one or more search engines in response to a user query on the fly. The main difficulties concern: the poor informative strength of snippets, the strict time constraints and the cluster labeling.

iv)In thesis [10], Filippo Geraci focused on the problem of clustering in the web scenario. To reduce the processing time in adding points to clusters they used Medoid Furthest Point First heuristic algorithm. In similarity searching of semi structured text documents one should want to allow the user to assign different weights to each field at query time. This requirement raises the problem of vector score aggregation which complicates preprocessing because at that time weights are unknown and thus one should build a data structure able to handle all the possible weight assignments. We conclude this thesis discussing some preliminary ideas about on-going research directions. We observed that clustering algorithms spend most of the processing time in adding points to the clusters, searching for the closest center. This time can be reduced using a keen similarity searching scheme. We plan to develop a general scheme for approximate FPF that takes advantage from the approximate similarity searching to reduce the algorithm running time without any data dependence.

IV. Similarity Measures for Text Document Clustering

Clustering is a useful technique that organize a large quantity of unordered text documents into a small number of meaningful and coherent clusters. A wide variety of document distance functions and similarity measures have been used for clustering such as Euclidean Distance and relative entropy.

Accurate clustering requires a precise definition of the closeness between a pair of objects ,in terms of either the pair wise similarity or distance. A wide variety of similarity or distance measure have been proposed and widely applied such as cosine similarity And the Jaccard Correlation coefficient.

Meanwhile similarity is often conceived in terms of dissimilarity or distance as well. Measures such as Euclidean Distance and Relative entropy have been applied in clustering to calculate the pair wise distances.

4.1 String Based Similarity

4.1.1 Character Based Similarity

- Longest Common Substring (LCS). This algorithm considers the similarity between two strings is based on the length of contiguous chain of characters that exist in both strings.

- Damerau-Levenshtein Jaro

The Levenshtein distance between two strings is defined as the minimum number of edits needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character.

- Jaro–Winkler

In computer science and statistics, the Jaro–Winkler distance is a measure of similarity between two strings. It is a variant of the Jaro distance metric and mainly used in the area of record linkage. The higher the Jaro–Winkler distance for two strings is, the more similar the strings are. The Jaro–Winkler distance metric is designed and best suited for short strings such as person names. The score is normalized such that 0 equates to no similarity and 1 is an exact match.

- Smith-Waterman

The Smith–Waterman algorithm performs local sequence alignment; that is, for determining similar regions between two strings or nucleotide or protein sequences. Instead of looking at the total sequence, the Smith–Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure

- N-gram

In the fields of computational linguistics and probability, an *n*-gram is a contiguous sequence of *n* items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The *n*-grams typically are collected from a text or speech corpus.

4.1.2 Term based Similarity

- Block Distance is also known as Manhattan distance, boxcar distance, absolute value distance, L1 distance, city block distance and Manhattan distance. It computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Block distance between two items is the sum of the differences of their corresponding components
- Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them.
- Dice's coefficient is defined as twice the number of common terms in the compared strings divided by the total number of terms in both strings .
- Euclidean distance or L2 distance is the square root of the sum of squared differences between corresponding elements of the two vectors.
- Jaccard similarity is computed as the number of shared terms over the number of all unique terms in both strings .

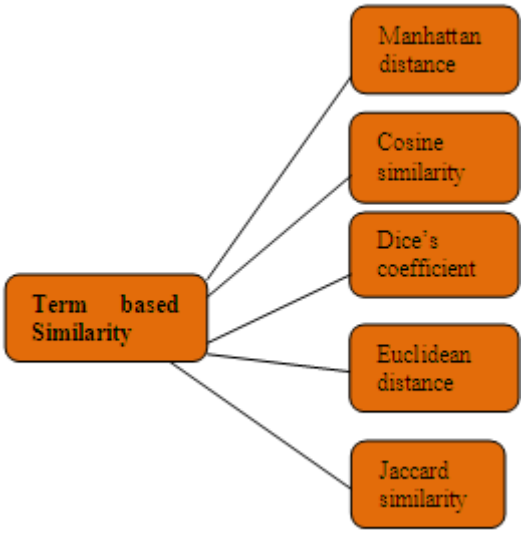

Fig 2: Term based Similarity

4.2 Corpus Based Similarity

Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora. A Corpus is a large collection of written or spoken texts that is used for language research.

- Latent Semantic Analysis

LSA is a technique of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. In the context of its application to information retrieval, it is called LSI.

- Explicit Semantic Analysis (ESA)

In natural language processing and information retrieval, explicit semantic analysis (ESA) is a vectorial representation of text (individual words or entire documents) that uses a document corpus as a knowledge base. Specifically, in ESA, a word is represented as a column vector in the tf–idf matrix of the text corpus and a document (string of words) is represented as the centroid of the vectors representing its words.

- Pointwise Mutual Information

Pointwise mutual information (PMI), or point mutual information, is a measure of association used in information theory and statistics.

V. Conclusion

This paper explains detailed description about text mining and its framework, Challenging issues in web clustering based on similarity, Different similarity measures such as string based, corpusbased, knowledge based and hybrid based similarity. Finally we come to know that, Accurate clustering requires a precise definition of the closeness between a pair of objects ,in terms of either the pair wise similarity or distance.

References

[1]. Wael H. Gomaa , Aly A. Fahmy " Short Answer Grading Using String Similarity And Corpus-Based Similarity",International Journal of Advanced Computer Science and Applications, Vol. 3, No. 11, 2012

[2]. Wael H. Gomaa , Aly A. Fahmy , " A Survey of Text Similarity Approaches ", International Journal of Computer Applications (0975 – 8887) Volume 68– No.13, April 2013

[3]. Kalaivendhan.K, Sumathi.P , "An Efficient Clustering Method To Find Similarity Between The Documents " ,International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Special Issue 1, March 2014

[4]. Mark.Dixon(1997), "An overview of Document Mining Technology",http://www.geocities.com/Research Triangle /Thinktank1997/mark/writing/dix 97-dm.ps

[5]. Anna Huang, "Similarity measures for Text document" ,Proceedings of the New Zealand CS Research Student Conference , April 2008, New Zealand.

[6]. Shaidah Jusoh and Hejab M. Alfawareh," Techniques, Applications and Challenging Issues in Text Mining",IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012

[7]. P.Ilakiya, "Discovering Semantic Similarity between Words Using Web Document and Context Aware Semantic Association Ranking" ,International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) ,Volume 2, Issue 6, June 2013,

[8]. Todsanai Chumwatana, Kok Wai Wong, Hong Xie, "A SOM-Based Document Clustering Using Frequent Max Substrings for Non-Segmented Texts ""J. Intelligent Learning Systems & Applications, 2010, 2, 117-125

[9]. LIN, Zhenjiang ,Phd Thesis on"Link-based Similarity Measurement Techniques and Applications" , Computer Science and Engineering ,The Chinese University of Hong Kong .September 2011

[10]. Filippo Geraci,Phd Thesis: "Fast Clustering For Web Information Retrieval", Anno Accademico 2007-2008

[11]. A. B. Manwar, Hemant S. Mahalle , K. D. Chinchkhede and Vinay Chavan , "A vector space model for information retrieval: matlab approach", Indian Journal of Computer Science and Engineering, Vol. 3, No. 2, pp. 222-229, 2012.

[12]. S. K. Jayanthi and S. Prema, " Facilitating Efficient Integrated Semantic Web Search with Visualization and Data Mining Techniques", Proceedings of International Conference on Information and Communication Technologies, pp. 437 – 442, 2010.

[13]. Wei Ning, Phd Thesis :"Textmining and Organization in Large Corpus", Kongens Lyngby 2005

[14]. G.Thilagavathi, J.Anitha, K.Nethra,"Sentence-Similarity Based Document Clustering Using Fuzzy Algorithm", International Journal of Advance Foundation and Research in Computer (IJAFRC) Volume 1, Issue 3, March 2014. ISSN 2348 - 4853