Clustering of Web Pages through Feature Weighting: A Survey

Muneer K

***Abstract*— Web page clustering is the process of grouping up of related web pages. Web page clustering has applications in various fields like information extraction, taxonomy design, similarity search, search result visualization and it can assist to the evaluation of results by search engines. Besides the particular clustering algorithm, the different term weighting functions applied to the selected features to represent the web pages is a main aspect in clustering task. The purpose of this review is to explore different methods for dimensionality reduction and feature weighting functions used for web document clustering and to report their appropriateness for clustering large sets of web documents. The review also covers transduction based approaches for web document clustering.**

***Index Terms*— Clustering, Web mining, Dimensionality reduction, Feature weighting, Transduction.**

I. INTRODUCTION

The motivation behind clustering any set of data is to find inherent structure in the data, and expose this structure as a set of groups, where the data objects within each group exhibit a large degree of similarity. The size of the web is increasing day by day in a very large scale. The web page clustering algorithms are very useful to apply to tasks such as automatic grouping before and after the search, search by similarity, and search result visualization on a structured way[2]. Before clustering, all web page preprocessing steps such as noise removal, stop word elimination, stemming etc need to be performed. Two aspects are important in order to obtain good web page clustering results: the clustering algorithm, and the term weighting function applied to the selected features of the web pages.

II. LITERATURE SURVEY

Various approaches for dimensionality reduction, feature weighting and web document clustering are discussed here.

Manuscript received Mar,2016.
Muneer K, Transmission Executive, All India Radio, Thrissur

A. *Dimensionality reduction*

One of the main problems in representation and later clustering is the high number of features that have to be taken into account when documents are dealt with. The objective of dimensionality reduction is to find a subset of features that have all characteristics of the full feature set. Various methods for dimensionality reduction are discussed below.

a) *Min-Max Reduction*

Only the features that appear more than *FFmin* times in more than *DFmin* web pages, and less than FF_{max} times in less than *DFmax* web pages are selected[7]. This was proposed by Fresno et.al.

b) *Entropy based reduction*

In this approach proposed by Klose et.al [3], the entropy of all terms are computed. Words that occur in many documents will have low entropy. As index words, a number of words having a high entropy is selected.

c) *Mutual Information based reduction*

This technique proposed by Shine.N.Das et.al[9] aims at helping document classification based on the maximal relevancy at minimum feature set. Mutual information based feature selection is employed here.

d) *Filtering based reduction*

If clustering is to be performed based on a similarity threshold, various filtering techniques proposed by Xiao et.al can be used for minimizing the number of tokens to be indexed[8].

B. Feature weighting

The tokens/terms present in the web page after preprocessing is referred to as the features of the webpage. Different feature weighting functions has been proposed by researchers.

a) *TF-IDF*

In [1], Salton proposes combining Term Frequency(TF) and Inverse Document Frequency(IDF) to weight terms. This is one of the classical methods for feature weighting.

b) *Tag based Term weighting*

This idea proposed by Shine.N.Das et.al[10] is purely based on the field where the term is present. HTML tags are considered for determining the field.

c) *Analytical Combination Criteria*

This was proposed by Fresno et.al[6]. The ACC is a linear combination criteria. It considers four functions frequency function of a word on a web page, frequency function of a word in the title, emphasized function of a word and position function..

d) *Fuzzy Combination Criteria*

Fresno et.al proposed the FCC weight function is a fuzzy system for the assignation of feature weights and their combination[6][7]. Thus, the linguistic variables of the fuzzy system are Text frequency, Title frequency, Emphasis and Global position.

e) Fuzzy Term Weighting through Standard questions

TF-IDF has the disadvantage of not considering the degree of identification of the object if only the considered index term is used and the existence of tied keywords. To overcome this,Ropero et.al[11] proposed an FL based Term weighting. It is based on four questions[11] that must be answered to determine the term weight of the index term.

C. Clustering

Clustering involves dividing a set of n objects into a specified number of clusters k, so that objects are similar to other objects in the same cluster, and different from objects in other clusters. Clustering algorithms make unsupervised attempts to find documents that are similar to each other and group them together. This approach, with no predefined categories or training documents, is a good representation of the WWW where documents are fast changing and it is difficult to have a stable hierarchy of categories that can accurately cover and represent all documents on the web[5]. *Carrot* search engine is a good example for a clustering engine. Search engines such as *Bing* performs supervised classification.

a) Transduction based Clustering Algorithm(TCA)

Transduction based Clustering Algorithm(TCA) employs a Transduction based relevance Model(TRM) to consider local relationships between each web document. Rather than depending on a fixed distribution model, Transduction based Relevance Model(TRM) proposed by Matsumoto et.al generates relevance values using local relations. There are two key aspects to TRM, the generation of relevance and relevance transduction. Relevance is a function of distance, and is used in generating clusters and relevance ranking of results. The value of the relevance model is the transduction stage: the relevance of a document x_i to another document x_k is affected by the relevance of other related documents to x_k[12].Using the relevance matrix R, the clustering of the data can be determined[12].

b)Fuzzy Transduction based Clustering Algorithm(FTCA)

This is the fuzzy counterpart of TCA[12]. Here, a document can have varying degrees of membership to multiple clusters.

c) Optimised Transductive Clustering

A combination of Extended Fuzzy Combination Criteria (EFCC) and Inverse Document Frequency is used for feature weighting. Here, a weight will be assigned to the tokens present in the web page by considering in which parts of the web page they are present. Fuzzy logic is employed in assigning weights. A Term-Document Weight (TDW) matrix is created which stores the feature weights of the tokens in different web pages. Using this matrix, transduction based clustering is performed. The effect of neighbours
is also considered here while assigning a web page to a particular cluster. K.Muneer et.al[13] proposed this method.

III. ANALYSIS

The accuracy of methods using snippets for clustering will be low as compared to methods using entire contents in the web page. The weighting scheme such as TF-IDF are not well suitable to reflect the notion of importance of a term in a particular web page. Better term weighting functions can be used here. Also for scalability, the complexity of the clustering algorithm should be low without affecting the quality of clustering.
The problem can be formally defined as follows. Given a set of web pages ψ, return clusters of web pages $\{\varphi\}$, $i = 1,2, \ldots n$ where n is the number of clusters. Apply a weighting function to the terms so that the notion of importance of the terms in the web pages is reflected in the representation of the web pages in the vector Space Model. The system should also be able to find clusters of near duplicate web pages when the degree of similarity, t, $0 \leq t \leq 1$ within the items in the clusters is specified.

Quality of clustering depends on the quality of the data representation. When snippets are used for clustering web search results, the accuracy will be low as compared to clustering using the entire contents of the web page. But when the entire contents are considered, the time complexity will be higher. So, we need to reach at a trade off between accuracy and time. So, instead of considering the entire terms in a web page, select most representative terms from the web page for dimensionality reduction. When the web page contents are considered instead of snippets for clustering, better term weighting functions other than TF-IDF can be applied thereby increasing the quality of the data representation; and which in turn will lead to better clustering. Identification of near duplicate web pages can also be performed efficiently when the web page contents are considered instead of snippets. So, here clusters of near duplicate web pages can also be formed. Even though transduction based clustering offers better accuracy in clustering, it is more time consuming and complex. So there is a scope for a simpler but effective clustering algorithm.

IV. CONCLUSION

Clustering is a very important task in web mining. There are various algorithms for dimensionality reduction of the tokens to be indexed. TF-IDF is a classical method used for term weighting. Fuzzy logic based methods are found to outperform other term weighting functions. Min-max method and mutual information based methods give better results. For search results clustering, transduction based clustering and its fuzzy counterpart are found to give more accurate results. Better representation of data leads to better clustering. So the clustering of web pages will be more efficient if better term weighting functions are used. Even though considering the entire contents of the web pages increases time complexity as compared to the method using snippets, this increases quality and it makes possible formation of clusters of near duplicate web pages.

REFERENCES

[1] Klose. A. Interactive text retrieval based on document similarities. In *Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy*, 2000.

[2] R. C. Dubes A. K. Jain. Algorithms for clustering data. In *John Wiley Sons*, 1988.

[3] A.Leuski. Improving interactive retrieval by combining ranked lists and clustering. In *Proceedings of RIAO*, 2000.

[4] Raquel Martinez. Alberto P.Garcia-Plaza, VictoF fresno. Web page clustering using fuzzy logic based representation and self-organizing maps. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2008.

[5] Arul Prakash Asirvatham. Web page classification based on document structure. In *IEEE National Convention*, 2001.

[6] Xuemin Lin Chuan Xiao, Wei Wang. Efficient similarity joins for near duplicate detection. In *17th international conference on World Wide Web*, 2008
.

[7] M. A Hearst. Clustering versus faceted categories for information exploration. In *Communications of the ACM*, 2006.

[8] Pramod K.Vijayaraghavan Midhun Mathew, Shine N Das. A novel approach for nearduplicate detection of webpages using tdw matrix. In *International Journal of Computer Applications 19(7):16-21*, 2011.

[9] Ropero.J. A fuzzy logic intelligent agent for information extraction: Introducing a new fuzzy logic based term weighting scheme. In *Expert Systems with Applications*, 2011.

[10] Pramod Vijayaraghavan Shine N Das, Midhun Mathew. An approach for optimal feature subset selection using a new term weighting scheme and mutual information. In *International conference on advanced science, engineering and information technology*, 2011.

[11] Edward Hung Takazumi Matsumoto. A transduction-based approach to fuzzy clustering, relevance ranking and cluster label generation on web search results. In *J Intell Inf Syst (2012) 38:419448*, 2012.

[12] Soto Montalvo Victor fresno, Raquel Martinez. Improving web page clustering through selecting appropriate term weighting functions. In *1st International Conference on Digital Information Management*, 2006.

[13] Muneer.K, Syed Farook K. An Innovative Approach for Clustering of Web Pages Based on Transduction. *International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014). Vol. 2, Issue 3 (July - Sept. 2014)*

