An Innovative Approach for Clustering of Web Pages Based on Transduction

[I]Muneer K., [II]Syed Farook K.
[1]M.Tech. Scholar, MES College of Engineering, Kuttippuram, Kerala, India.
[2]Assistant Professor, MES College of Engineering, Kuttippuram, Kerala, India.

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. Detection of near duplicate web pages also plays a vital role in today's world. This paper proposes an algorithm for forming clusters of near-duplicate web pages. Near duplicate pages differ only slightly in content in advertisements, timestamps etc. Even though they are not bitwise identical, they are remarkably similar. 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 determining the quality of 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.

Keywords

Web Mining, Clustering, Transduction, Feature Weighting, Near-duplicates

I. Introduction

The size of the WWW is growing rapidly. As the web keeps on growing, filtering out search results that are relevant to the user has become a challenging task to the search engines. Any one of the subsequent features: different character sets, formats, and inclusions of advertisement or current date may be the reason behind the dissimilarity among identical pages served from the same server [1]. Web crawling is employed by the search engines to populate a local indexed repository of web pages which is in turn utilized to answer user search queries. Business has become more proficient and fruitful owing to the ability to access contents of interest amidst huge heaps of data. 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 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 pre-processing 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. Web contains duplicate pages and mirrored web pages in abundance. Despite the fact that near duplicates are not bit wise identical, they are strikingly similar [5]. Near duplicates posses minute differences and so are not regarded as exact duplicates. Typographical errors, versioned, mirrored, or plagiarized documents, multiple representations of the same physical object, spam emails generated from the same template and the like are some of the chief causes for the prevalence of near duplicate pages. Such near duplicates contain similar content and vary only in minimal areas of the document like the advertisements, counters and timestamps. Web searches consider these differences as inappropriate. Various studies have identified a substantial portion of web pages as near duplicates [3].

Forming clusters of near-duplicate web pages is an important task in web mining. It can be helpful for search results visualization, plagiarism detection and for improving the quality and diversity of query results. This paper proposes an innovative algorithm to form clusters of near-duplicate web pages.

Rest of the paper is organised as follows. Section II discusses the related literature. Section III explains the proposed method in detail. Section IV discusses the experimental results and section V concludes the paper.

II. Literature Survey

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 the objects in the 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. *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[9] 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 xi to another document xk is affected by the relevance of other related documents to xk. Using the relevance matrix R, the clustering of the data can be determined[9]. Hung et.al.[9] proposed an algorithm for clustering search results using transduction. The snippets of web pages appearing in the search results are used as input to the algorithm. TF-IDF matrix is formed from the snippets and transductive clustering is performed on it. So, the processing time is less, while the accuracy is less than those algorithms which use the entire contents in the web pages.

B. Fuzzy Transduction based Clustering Algorithm (FTCA)

This is the fuzzy counterpart of TCA[9]. Here, a document can have varying degrees of membership to multiple clusters. The framework of both TCA and FTCA is shown in the Figure 2.1.

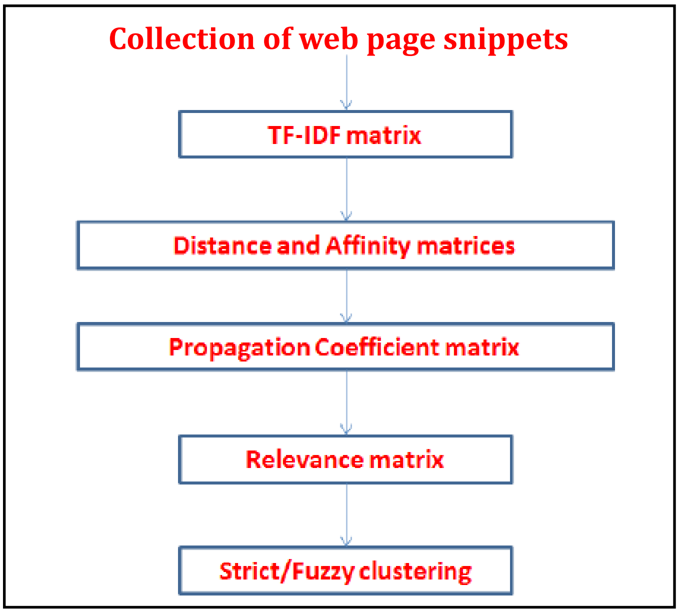

Collection of web page snippets
TF-IDF matrix
Distance and Affinity matrices
Propagation Coefficient matrix
Relevance matrix
Strict/Fuzzy clustering

Fig. 2.1: Framework of TCA and FTCA

In addition to the clustering algorithms, there are a number of algorithms for the detection of near-duplicate web pages in WWW[3][5][6][8]. An algorithm has been proposed by Muneer K. et.al[12] for detecting near duplicates using fuzzy logic based feature weighting and filtering approaches. Rasia Naseem et.al[11] proposed a method based on analytical feature weighting and Minimum Weight Overlapping.

III. Proposed Work

In order to form clusters of similar web pages, an innovative and effective algorithm is presented here. The process is mainly divided in to three phases – Feature Weighting, Relevance Matrix Generation and Cluster Formation.

The speciality of this algorithm is that the concept of transduction[9] is employed here in clustering. The effect of neighbours of a data item is also considered while asssigning a particular data item to a particular cluster. That is, in transduction based clustering, Relevance of a document x_i to a document x_k is affected by the relevance of other documents to $x_{k.}$

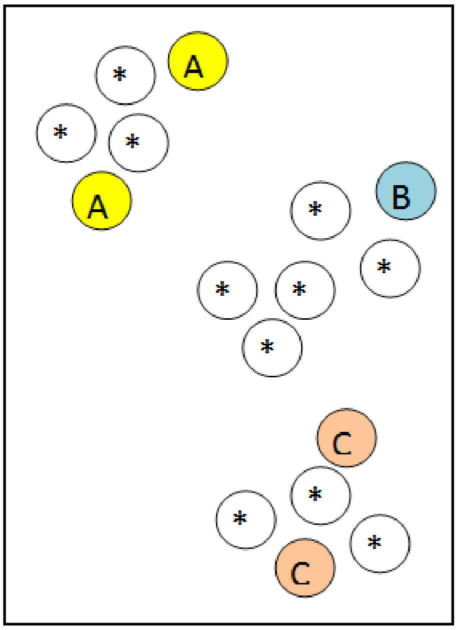


Fig. 3.1: Clustering by Transduction

For example, consider the above figure. Suppose the circles represent different data items. The labels are known only for some data items. The labels, i.e. cluster names are unknown for the items marked with "*". In this context, if inductive approaches such as distance based approaches are used for clustering, the two data items in the middle will be wrongly clustered and labelled as either 'A' or 'C', but it is evident that they belong to the cluster 'B'. While performing the clustering task, transduction based approaches consider all the points, instead of considering the labelled points only. Here, transductive approaches can label the non-clustered points according to the clusters to which they naturally belong. So, the two data items in the middle would be labelled as 'B' since they are packed very close to that cluster.

A. Feature Weighting

Fuzzy logic can be used for capturing human expert knowledge. Its knowledge base can be defined using a set of IF-THEN rules. A set of linguistic variables is defined with fuzzy sets, which describe the membership degree of an object to a particular class.

Web pages are very different from normal text files in many aspects. The importance of a token/term/feature in a web page depends on which parts of the web page that particular token is present. So, here feature weighting is done by employing an improved version of Extended Fuzzy Combination Criteria (EFCC) proposed by Alberto.P.Garcia-Plaza[10]. The EFCC is a fuzzy system for assigning feature weights with four linguistic variables.

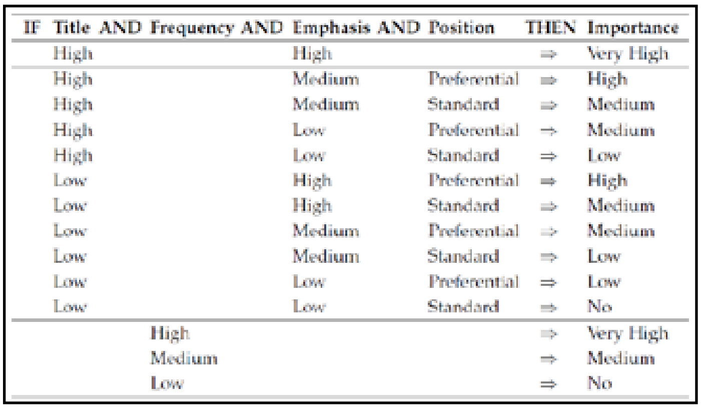
- **Text-Frequency** $Nf(d,w)$ is the number of occurrences of a token 'w' in document 'd'. It is normalized by dividing it with the greatest text-frequency value in document 'd'. Three labels "low", "medium" and "high" are assigned to this linguistic variable based on the normalized value.
- **Title-Frequency** $N_t(d,w)$ describes the number of occurrences of a token "w" in title, meta tags, headings etc in document "d". It is normalized by dividing this value with the greatest value of $N_t(d,w)$ in that document. Based on this value, two labels "low" and "high" are assigned to this linguistic variable.
- **Emphasis** HTML tags such as <b></b>, <i></i>, <u></u>, <code></code> etc emphasize parts of the text. $N_e(d,w)$ is the

number of occurrence of a token ‘w’ in emphasized parts in document ‘d’. It is normalized by dividing it with the greatest value of N_e(d,w) in that document. Then three labels “low”, “medium” and “high” are applied based on this value.

- **Global-position** To compute the position criteria, the web page is split into three parts according to the number of features in the <body> tag. The tokens/features belong to these three parts- Introduction, Body and Conclusion, with fuzzy membership values. Then the global position is obtained by means of an auxiliary fuzzy system[10].

Using these four linguistic variables, a rule base is defined to determine the relevance of a particular token ‘w’ in document ‘d’. The design details of EFCC rule base are explained in [10]. The output of the fuzzy module is a linguistic variable called “importance”, with associated labels “No”, “Low”, “Medium”, “High” and “Very High”. Term frequency based rules is separated from the rest of the rules in the rule base. At least one rule from both sets will be triggered for every token. Centre of Mass (COM) algorithm is used to determine the defuzzified value of the linguistic variable “importance”. Since the aim here is to find the near duplicate web pages with respect to an input web page and a threshold, the Inverse Document Frequency (IDF) is very important in feature weighting. The idea is that the tokens with greater values of document frequency will be commonly used words in all documents. So, in order to penalize words with higher values of document frequency, the “importance” obtained from EFCC is mapped to a numerical value and it is multiplied with the Inverse Document Frequency of that token k, thereby obtaining “global importance”, that is, term weight.

Now, the records corresponding to each web page in the repository needs to be canonicalized by sorting its tokens based on a global ordering. The term weight obtained from “global importance” is considered for sorting the tokens instead of considering the document frequency alone. A Term Document Weight (TDW) is to be created in such a way that it maps a token ‘x’ to a list of records that contain ‘x’ and stores the term weight. If a particular token is not present in a record, weight is considered as zero. Every record is represented as per the global ordering [10].



IF Title	AND Frequency	AND Emphasis	AND Position	THEN	Importance
High		High		⇒	Very High
High		Medium	Preferential	⇒	High
High		Medium	Standard	⇒	Medium
High		Low	Preferential	⇒	Medium
High		Low	Standard	⇒	Low
Low		High	Preferential	⇒	High
Low		High	Standard	⇒	Medium
Low		Medium	Preferential	⇒	Medium
Low		Medium	Standard	⇒	Low
Low		Low	Preferential	⇒	Low
Low		Low	Standard	⇒	No
	High			⇒	Very High
	Medium			⇒	Medium
	Low			⇒	No

Fig. 3.2. EFCC Rule Base

B. Relevance Matrix Generation

Clusters of near-duplicate web pages can be formed using the TDW matrix created earlier. The degree of similarity within the items in a cluster needs to be specified as the input. The output will be the sets of clusters with inter cluster similarity greater than or equal to the specified value.

The speciality of this algorithm is that the concept of transduction[9] is employed here in clustering. The effect of neighbours of a data item is also considered while assigning a particular data item to a particular cluster. That is, in transduction based clustering, Relevance of a document x_i to a document xk is affected by the relevance of other documents to x_k.

Transduction based Clustering Algorithm(TCA) employs a Transduction based relevance Model(TRM) to consider local relationships between each web document. TRM generates relevance values using local relations.

Relevance is a function of distance, and is used in generating clusters . Using the relevance matrix R, the clustering of the data can be determined.

First, the Cosine Similarity between each pair of documents is calculated by considering each row in the TDW matrix as a document vector.

$$\cos(x,y) = \frac{x.y}{||x||.||y||} \qquad (3.1)$$

Then, the distance matrix, D=d(i,j)nxn can be generated using (3.2).

$$d(i,j) = \frac{1}{\cos(i,j)+\alpha} \qquad (3.2)$$

α is a small value such as 0.001 which is used to avoid division by zero.

Affinity matrix is generated from the Distance matrix thus created. The Affinity matrix W=w(i,j)nxn with affinity between two documents xi and xj inversely proportional to the distance between xi and xj is defined as

$$w(i,j) = \frac{\exp\left(-(d(i,j))^2\right.}{\sigma^2} \qquad (3.3)$$

σ is the attenuation scale and is set to 1 here as in [9].

Now in order to implement the concept of transduction, the influence I’s neighbours on the influence of xi to xj needs to be considered, and so the Relevance Matrix is generated.

$$r(i,j) = \frac{w(i,j)}{\sum_{k=1}^{n} w(i,k)} \qquad (3.4)$$

C. Cluster Formation

Now, different clusters of similar or near duplicate clusters can be formed using this Relevance matrix. Here, the number of clusters will be equal to the number of web pages in the repository. For that, a relevance value needs to be fixed as the threshold value. This relevance value specifies the inter cluster similarity. Then, from each row in the matrix, all the web pages which have relevance value above the threshold will be grouped together to form clusters. In this way, clusters of near-duplicate pages can be formed very effectively.

IV. Experimental Results

A large collection of web pages were fed into the data set. An online tool was created in PHP for retrieving the web pages and to perform the required pre-processing operations like noise removal, stemming etc. Then, the extracted tokens were stored into the TDW matrix along with their weight.

The clustering algorithm was implemented in MATLAB. Here, the number of clusters will be equal to the number of web pages in the repository. For that, a relevance value needs to be fixed as the threshold value. This relevance value specifies the inter cluster similarity. Then, from each row in the matrix, all the web pages which have relevance value above the threshold were grouped together to form clusters. The clusters thus formed will be sets of near-duplicate web pages. The experiments were conducted using 5 different data sets. The results are shown in Table 4.1.

Table 4.1: Experimental Results

Data Set	Precision	Recall
D1	92.1	91.9
D2	91.5	91.2
D3	90.0	92.0
D4	89.1	90.2
D5	91.1	91.0
Average	90.76	91.26

V. Conclusion

Extended Fuzzy Combination Criteria(EFCC) coupled with Inverse Document Frequency is used here for feature weighting. Transduction based clustering is performed using the TDW matrix where the effect of neighbours of a document on the relevance of a document to a particular cluster is also considered.

The algorithm works efficiently in determining and clustering near duplicate Web pages on the Web. The EFCC-IDF feature weighting approach not only takes up syntax of the page, but also considers in which parts of the page content the author intends to give emphasis. Clustering using transduction improves the accuracy of clustering since the effect of neighbours also is taken into account. The analysis of the results shows improved precision and recall values.

References

[1] *Ilyinsky, S., Kuzmin, M., Melkov, A., Segalovich, I., An efficient method to detect duplicates of Web documents with the use of inverted index, Proceedings of the Eleventh International World Wide Web Conference, 2002.*

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

[3] *Gurmeet Singh Manku, Arvind Jain and Anish Das Sarma, Detecting near duplicates for web crawling, Proceedings of the 16th International Conference on World Wide Web, pp. 14-150, Ban ,Alberta, Canada, ,2007.*

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

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

[6] *V.A. Narayana, P. Premchand and A. Govardhan, Effective Detection of Near Du-plicate Web Documents in Web Crawling, International Journal of Computational Intelligence Research, Volume 5, Number 1,pp. 8396 -2009*

[7] *Ropero.J., A fuzzy logic intelligent agent for information extraction: Introducing a new fuzzy logic based term weighting scheme, Expert Systems with Applications ,2001*

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

[9] *Edward Hung, Takazumi Matsumoto, A transduction based approach to fuzzy clustering, relevance ranking and cluster label generation on web search results, J Intell Inf Syst ,2012*

[10] *Alberto Prez Garca-Plaza, An Improved Fuzzy System For Representing Web Pages In Clustering Tasks. Phd Thesis. Departamento de Lenguajes y Sistemas Informticos. Universidad Nacional de Educacin a Distancia. ,2012*

[11] *Rasia Naseem, Sheena Anees, Muneer.K, Syed Farook.K, Near Duplicate Web Page Detection With Analytic Feature Weighting, Third International Conference on Ad-vances in Computing and Communications, pp:324-327, Kochi, India ,2013.*

[12] *Muneer.K., Syed Farook.K, Rasia Naseem and Balachandran K.P. Near Duplicate Web Page Detection With Fuzzy Feature Weighting. International Conference on Communication and Computing, June 2014. Bangalore, India.*

Muneer K. Received B.Tech. Degree in Computer Science and Engineering in 2011 from Cochin University of Science and Technology. Currently, he is pursuing M.Tech. Degree in Computer Science and Engineering from MES College of Engineering, Kuttippuram under Calicut University. He has presented various papers in three international conferences and has won Best Paper Award in ICRTCSE'12. His research interests include Data Mining, Web Mining, Information Processing and Big Data.

Syed Farook K. Pursued B.E. Degree in Computer Science and Engineering from Anna University, Chennai and M.E. Degree from KAHS, Coimbatore. He has a teaching experience of 8 years. Currently, he is working as an Assistant Professor in M.E.S. College of Engineering, Kuttipppuram, India. His research interests include Data Mining, Web Caching and Information Retrieval.