Indian Journal of Science and Technology, *Vol 7(9), 1355–1358, September 2014*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Context-based Classification of XML Documents in Feature Clustering

A. Mary Posonia[1*] and V. L. Jyothi[2]

[1]Department of Computer Science and Engineering, Sathyabama University, Chennai, India; soniadelicate@gmail.com
[2]Department of Computer Science and Engineering, Jeppiaar Engineering College,Chennai, India

Abstract

Text classification is the process of automatically sorting a set of documents into categories from a predefined set. Feature clustering is a powerful method to reduce the dimensionality of feature vectors for text classification. After pre-processing, the document can be clustered in the schema level based on the occurrence of the words relatively. Clustering process group the words based on the pattern. In proposing a feature clustering mechanism finds the pattern match with the number of relevant data present in the database.

Keywords: Feature Clustering, Feature Selection, Information Retrieval, Text Mining

1. Introduction

An automatic text classification can efficiently analyze the set of documents and organize the data based on certain categories[2]. Document clustering is defined as the automatic division of the data or grouping a set of objects in a way such that objects in the same group are more similar to each other than to those in another group[1]. Clustering is the main task of preliminary data mining, and a used technique for data analysis which is used in many fields such as machine learning, pattern recognition and also analyzed the data from the images.

In the emerging concept of XML document management, the clustering mechanism provides an efficient solution to organize the document content. Clustering involves organizing XML documents depends on their similarity without knowing the structural representation of XML documents[4].

Analyzing the data from a particular database or data repository will result to the set of data similar to the searched string is derived from the cluster. A Cluster is the form of data that which is collected at a particular point from many data sets or documents which can be retrieved from the process. The input data to a process is too large to be processed and it is suspected to be ignominiously redundant, then the input data will be transformed into a reduced representation set of features. Analyzing and making the data into the reduced form of size or content with features of original data is called feature extraction.

In text mining the sentence similarity has been measured based on the semantic relations expressed in external resources such as dictionaries or thesauri. Cluster evaluation may be either supervised or unsupervised for the goodness of the clustering. In the following, L ¼ fw1; w2; g is the set of clusters, C fc1; c2; g is the set of classes (for supervised evaluation), and N is the number of objects[6].

2. Related Work

Theodore Dalamagas[5,11] proposed a characteristic that distinguishes XML documents from semi structured data can be represented as DTDs. The structural representation of an XML document can be determined by its DTD. A valid XML document is one that has a DTD and conforms to it. A DTD, besides enabling exchange of

*Author for correspondence

documents through common vocabulary and standards, can generate relational schemas to store and query XML documents in relational database systems. Skabar[6] proposed a method to calculate similarity values s_{ij} for the affinity matrix we use a modified version of the measure proposed by Li et al. Proposed the approach document similarity, In this individual sentence has been compared with each document rather than using a common vector space of dimension n, where n is the number of distinct non-stop words appearing in the two sentences.

Support Vector Machines (SVM)[7,8] have been proved as one of the most successful classification methods for many applications including text classification. Even though the learning ability as well as computational complexity of training in support vector machines may be independent of the dimension of document feature space, minimizing complexity is an essential issue to efficiently handle a large number of terms in practical applications of text classification adopts novel dimension reduction methods to reduce the dimension of the document vectors dramatically.

Many approaches to retrieve XML data have been proposed in recent years, this approaches mainly focusing on XML document clustering based on structure. The motivation behind this is clustering of semi-structured data in web-based application[3]. The fuzzy relational clustering propose the page rank score of an individual objects within the cluster and the page rank treated as a important parameter to determine cluster membership function[6].

3. Proposed Approach

3.1 An Overview

The architecture shows the entire workflow of the proposed system. In the first stage the collection of XML documents should undergo pre-processing step and the collection of pre-processed data will be stored in the database. In the second phase the clustering process is based on feature extraction, the feature can be stated as highly repeated word or the different words with the occurrence of the pattern (Figure 1).

3.2 Preprocessing of XML Documents

The pre-processing converts the data in the tags or the schema to the form of Boolean retrieval. The processed data will be stored in the database. The Pre-processing of XML documents can be stated as retrieving the data from

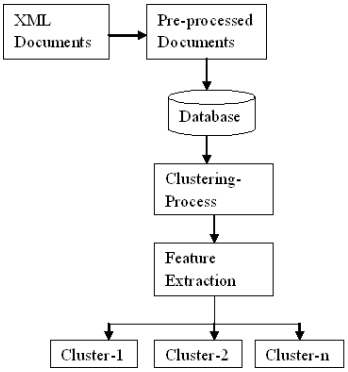

Figure 1. Proposed architecture.

the XML files. Generally the XML files consist of data in the form of tags and the elements. The data will be present in the tags, so the tags are to be retrieved as a column's name in the database. Each record will be stored in the database as the process of removing the tags.

Sample XML Document

```
<? Xml version = "1.0">
<table ID = "Customer">
<T>
    <C_CUSTKEY> 1 </ C_CUSTKEY>
    < C_NAME> Customer #0000001 < / C_
    NAME>
    < C_ADDRESS> IvhZiApRb 0t,C,E < / C_
    ADDRESS>
    < C_NATIONKEY> 15 </ C_NATIONKEY>
    < C_PHONE> 25-989-741-2988 < /C_PHOME>
    < C_ACCBAL> 711.56 < /C_ACCTBAL>
    < C_MKTSEGMENT> BUILDING < /C_
    MKTSEGMENT>
    < C_COMMENT> Special Packages,slyly reg </
    C_COMMENT>
</ T>
```

Algorithm : For Preprocessing

1 **Input** : XML File
2 **Output**: Preprocessed XML Document
4 For each value in the XML File Do
5 for i = 1 to nodes.getLength() do
6 if (document element ≠ 0) {
7 Copy tag name and split the values from XML file
8 Node node = nodes.item(i);
9 }endif }endwhile
10 the XML file is already pre-processed, update the records or skip the process.
11 stop the process

The above algorithm uses the sample XML file of customer details. The data in this format will be processed by the pre-processing technique and the output of the process will be stored in the database. Some example XML tags are listed as C_CUSTKEY which consists of the key number, C_NAME which contains the name of the customer, C_ADDRESS which contains the address of the customer and C_MKTSEGMENT which contains the market segment details of the customer. After the pre-processing the processed data have been stored in the database.

3.3 Clustering of XML Documents

Clustering analysis organizes data by abstracting underlying structure either as individual's or as a group. The clusters can then be analyzed to see if the data group, according to preconceived ideas or to suggest new methods. Clustering mechanism can analyze the structure of the data that does not require the assumptions common to most statistical methods. It is called "unsupervised learning" in the literature of the pattern recognition and artificial intelligence[5].

Clustering of the documents is based on the values present in the table by which the values align with the particular value. Clustering of XML document is not possible with the tags present in it. So, taking the pre-processed data which contains the tags which are removed. The clusters are formed based on the occurrence of the data in the document. This experiment mainly based on the context present in the XML document.

Algorithm: For Clustering of XML Documents

1 **Input** : Preprocessed XML File

2 **Output**: Clusters

4 Select the distinct data from the preprocessed data.

5 For each record in the database compares the distinct data from the database do

5.1 Compare the string with the next record value if it matches increment the flag.

5.2 Else store the new word in the array and add in cluster list.

5.3 count the no of clusters.

6 Represent the data in the form of clusters.

4. Experiments

The information retrieved from the pre-processing has been grouped into the clusters based on the data on the most relevant pattern that which has occurred more in the document. In Table 1 the data's are grouped by the data which contains the market segment details of the customer (C_MKTSEGMENT) that which contains the following set of information.

Table 1. Preprocessed data

Category	Name	Attributes	Value
Customer	T	C_CUSTKEY	1500
		C_NAME	#0001500
		C_ADDRESS	4Z0W43
		C_NATIONKEY	5
		C_PHONE	15-260-872-4790
		C_ACCBAL	6910.79
		C_MKSGMENT	MACHINERY
		C_COMMENT	Quickly even packages

Table 2. Clustered data

Clustering - Category	Total no. Of Documents
Building	10
Furniture	7
Automobile	4
Household	5
Machinery	2

4.1 Clustered Data

The data's are grouped by the data which contain the market segment details of the customer (C_MKTSEGMENT) that which contain the following information. The set of tests data taken from the XML bench mark dataset and based on the clustering mechanism it can be grouped by the following category (Table 2).

5. Conclusion

In the self-constructing feature clustering algorithm introduced an incremental clustering approach to reduce the dimensionality of the features in text classification and features that are similar to each other are grouped into the same cluster. If a word which is not similar to any existing cluster, a new cluster has been created for this word. Once entire words have been analyzed, a desired number of clusters are formed automatically. This study shows the efficient clustering mechanism and preliminary evaluation done with the synthetic data and bench-mark data set.

6. References

1. Jardine N, Rijsbergen VC. The use of hierarchic clustering in information retrieval. Information Storage and Retrieval. 1971; 7(5):217–40.
2. Denoyer L, Gallinari P. Bayesian network model for semi-structured document classification. Inform Process Manag. 2004; 40(8):807–27.
3. Kollios G, Terzi E. Clustering large probablistic graphs. IEEE Trans Knowl Data Eng. 2013; 25(2):325–36.
4. Tran T, Nayak R. Evaluating the performance of XML Document Clustering by Structure only, Comparative Evaluation of XML information Retrieval Systems. Lecture Notes in Computer Science. 2007; 4518:473–84.
5. Dalamagas T, Cheng T, Winkel K-J. A Methodology for Clustering XML Documents by Structure. J Inform Syst. 2006 May 1; 31(3):187–228.
6. Andrew Skabar, Khaled Abdalgader. Clustering sentence-level text using a novel fuzzy relational clustering algorithm. IEEE Trans Knowl Data Eng. 2013; 25(1):62–75.
7. Joachims T. Text Categorization with Support Vector Machine Learning with Many Relevant Features. University of Dortmund; 1998. Technical Report:LS-8-23
8. Cortes C, Vapnik V. Support-Vector network. Mach Learn. 1995; 20(3):273–97.