VOL. 13, NO. 1, JANUARY 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

UNSUPERVISED LEARNING OF XML DOCUMENTS BY VISUALIZED CLUSTERING APPROACH (VCA)

K Rajendra Prasad and R Obulakonda Reddy
Department of Computer Science and Engineering Institute of Aeronautical Engineering Dundigal, Hyderabad, India
E-Mail: maulida@usu.ac.id

ABSTRACT

Clustering of XML documents is playing a vital role in web mining. The similarity between pairs of XML documents is measured by different distance measures such as Euclidean, Cosine etc. In the XML document clustering, we compute the similarity features by either comparing XML document structures, or XML semantics, or XML content. The XML structures are derived from their x paths. Thus, the process of extracting the XML structures and content is done before clustering for identifying of similar documents. The distinct XML document clusters information is assessed by visualizing clustering method. The aim of this paper is to solve the issue of clustering tendency by a visualized clustering method for the purpose of producing the efficient clustering results. The clustering results of XML documents are evaluated in respect to the performance measures during the experimental study.

Keywords: clustering, XML document, xpath, visualized clustering method.

1. INTRODUCTION

The process of clustering is completely an unsupervised approach. Extraction of web pages and finding the similar web documents is part of the web mining problem. Clustering is widely used in web mining for detecting the similar websites. Classical algorithm of k-means [2] has capable of detecting of similar web documents. But the problem of k-means is to initialize the k value of random choice. Due to this fact, there is less chance for producing effective clustering results by k-means. Other clustering algorithms such as hierarchical clustering algorithms [2], density based clustering algorithms [3], graph-based clustering algorithms [1] are also suffering from the same problem. For this reason, this paper concentrates initially on how to solve the problem of clustering tendency. The clustering tendency [7] has the knowledge about assessment of correct number of clusters for giving unlabeled data. It is useful for finding the k value in k-means. There is a large scope for the retrieving of better clustering results from this known k value in the clustering.

The term web mining has targeted the extract the required web pattern which has been generated by a user from web pages based on the similarity matching of web structure, and web content. However, we can also perform the web mining based on semantics in some applications. The main targeted extracted elements are content such as targeted text or targeted hypertext, multimedia data which include graphics, images, etc. In targeted structure, we extract either the XML tags also called as xpaths or HTML tags. Currently, we place the web page into XML format; because it is flexible and scalable. There are many tools available today to deal with the XML structure for extraction of appropriate data from XML file.

In this paper, we present literature in the methodology of web content and web structure mining by XML documents. For simplifying the process of structure mining, we use only; an elements names and their properties are used for the matching of structure. We represent a structure of XML documents as a tree-based structure and which has broken down as a collection of different valid paths. These paths are called as xpaths and these are used for measuring the structural similarity between documents. For given a possible set of XML documents {d1, d2 ...dn} denoted by D and set of xpaths are {p1, p2 ...pm} denoted by P are always extracted from D. Any XPath contains the element names from root node to a leaf node, and leaf node is must be a textual content. In the structure modelling of XML, we use a path vector for the document i is {p1i, p2i ...pMi}. In similarity matching, we use the path vectors, for example, we use path vector {p1x,p2x,.....pMx} for document x, path vector {p1y,p2y,.....pMy} for document y, and these two vectors are used to measure the similarity between two documents x ,and y based on the following Euclidean formula

$$\mathrm{structsim}(d_x, d_y) = \sqrt{\sum_{i=1}^{M} (p_{ix} - p_{iy})^2} \quad (1)$$

In content similarity matching, we collect the distinct terms from D as {t1, t2 ...tL}, and the frequency values of the term-document matrix is derived from the data of XML documents which is described in [12]. The structure and content similarity matrices are used for finding the effective clustering results of XML documents. The process diagram is shown in Figure-1. The classical algorithm of k-means may chance for producing poor results when there is no prior knowledge of clustering tendency (number of clusters) of XML documents. However, by using a visualized clustering approach, it is possible for finding the prior number of clusters; this prior knowledge could help the improving of XML documents clustering results. Section 2 presents the relevant work for the existing technique of visual access tendency (VAT).

Section 3 describes the proposed visualized clustering approach and XML document clustering. Section 4 presents the discussion of XML datasets and results analysis. Section 5 presents the conclusion along with future scope.

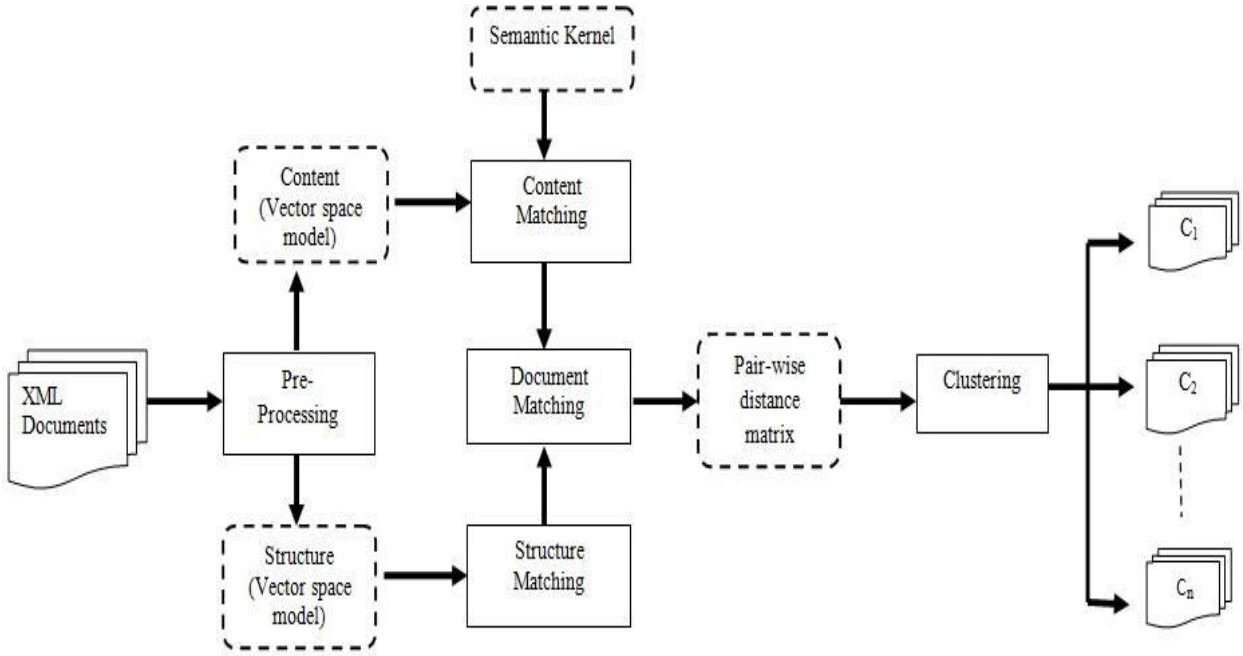

Figure-1. Overview of the proposed clustering approach.

2. RELEVANT WORK

Efficient clustering results from XML documents are an emerging research in web mining. The limitation of k-means procedure is to enable the clustering results without attempting the exact k value. Bezdek *et al*. [4], [5], [7] describes the methodology for the purpose of accessing the clustering tendency. This procedure is called as the visual access tendency (VAT), and this is extended as iVAT in [4], and as SpecVAT [4]. VAT procedure takes the dissimilarity matrix as prerequisites and it reorders the indices of dissimilarity matrix and produces the reordered dissimilarity image (RDI), this is called as VAT Image. In VAT Image, the square shaped dark blocks are appearing along the diagonal. Thus, total number of clusters is detected by counting the square shaped dark blocks. The iVAT is same as VAT, but it uses path-based distance measure for finding of dissimilarity matrix. This is most effective procedure, especially when the input data pattern is in the form path-based data. In complex data sets, SpecVAT produces better clarity of results than others; because SpecVAT uses a spectral approach. The procedure of VAT describes as follows.

VAT algorithm [7]
Input: Dissimilarity Matrix $R=[r_{ij}]$, for $1 \leq i,j \leq n$

Method:

Step 1: Set I = { };
J = {1,2,…,n};
P = (0, 0, …, 0);
Select $(i,j) \in \arg\max_{p \in J, q \in J} \{r_{pq}\}$;
Set P(1) = j;
ReplaceI = I $\cup$ {j} and J = J – {j}

Step 2: For t = 2,…, n:
Select $(i,j) \in \arg\min_{p \in I, q \in J} \{r_{pq}\}$;
Set P(t) = j;
ReplaceI = I $\cup$ {j} and J = J – {j}

Step 3: Form $R_{reordered} = [r_{ij}] = [r_{P(i)P(j)}]$ for $1 \leq i,j \leq n$

The positive aspect of VAT is that it can access an exact the number of clusters. Procedure for detecting the clustering results by this VAT is called visualized clustering method [8], [10], [11], which is described in the next section.

3. VISUALIZED CLUSTERING APPROACH (VCA)

By the indices of reordering nodes, we perform the clustering. First we sorted the nodes according to the label ordering. Next, we permute the clustering labels of VAT to match the given equivalent labels by the ground truth labels. We use a Khun-munkres algorithm for performing of clustering results as well as finding the clustering accuracy [14]. The outline of visualizing clustering approach is as follows.

Algorithm: Visualized Clustering Approach (VCA)

Input: Dissimilarity matrix-d[ ][ ],
Startindex[ ],Endindex[ ]
n-total number of objects
gnd-ground truth values
map- assign the clustering labels for reordered objects

Output:
Res-best mapping by khun-munkre
AC- Clustering Accuracy

1. [RV,C,I,RI]=VAT(d);
2. For i=1:n
   Compare the RI values with Startindex and Endindex of clusters and define values of map [11]
   Endfor
3. res = bestMap(gnd,map);
4. AC = length(find(gnd == res))/length(gnd);

In step1, we use a procedure of VAT for finding the reordered indices of objects. In step2, we assign the clustering labels for reordered objects. In step3, Khun-munkres [13] function is used to match the equivalent labels of map with ground truth labels in order to perform the clustering. These clustering results are generated by the concepts VAT and Khun-munkres. Hence, this is called as a visualized clustering approach.

XML documents clustering [12]

A Set of XML documents D= (d1, d2, d3,....dn) consists of two parts: XML structures and XML content. The Xpath of XML documents consists of the path from root to a leaf node. The leaf node contains the information about terms in textual form. The root and a set of intermediate nodes have the XML structural information. Therefore, we can effectively assess the subjective information about XML structure and its XML content; which is useful for computing the dissimilarity matrix of XML documents during clustering process.

Data matrices

Initially, we extract the paths of XML documents for generating of the structure based data matrix. In the same way, we can generate the content based data matrix for the same XML documents. The dissimilarity features of XML documents are computed by distance measures such as Euclide, Cosine etc.

The data matrices for structure-versus-documents, and terms-versus-documents are as follows:
Structure vs. documents data matrix (m structures x n documents)

$$\begin{bmatrix} td11 \ldots td1n \\ \ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots \\ tdk1 \ldots tdkn \end{bmatrix} \quad (2)$$

Terms vs. documents data matrix (k terms x n documents)

$$\begin{bmatrix} td11 \ldots td1n \\ \ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots \\ tdk1 \ldots tdkn \end{bmatrix} \quad (3)$$

Apply the distance measures for the matrices of equations 2 and 3 for obtaining of dissimilarity features of given XML documents in respect to the structure and contents. Based on these features, finally we retrieve the optimal clustering results by VCA. Respective experimental results and datasets are discussed in the next section.

4. XML DATASETS AND RESULT ANALYSIS

The Wikipedia and IEEE XML datasets are available from the INEX 2006 Document Mining Challenge [9]; these are used for evaluating the clustering approach. The final clusters are labelled according to the textual or content theme which makes the content based similarity measure and it is more important than the structure based similarity measure. The IEEE XML dataset is also derived from the same structural definition, hence all documents consists of the same set of element names.

Table-1 shows the detail of our datasets. A subset of Wikipedia dataset is used in our experiments. Wikipedia and IEEE XML datasets are the homogeneous dataset; the documents in the dataset are being conformed to only structural definition.

Table-1. Datasets.

Datasets	Total documents taken	Number of true categories taken
Wikipedia	2500	50
IEEE	4050	16

For the evaluation, we have taken the methods are micro-average and macro-average, the respective formulas are available in [12]. The evaluation result of these two measures by our method is presented in the Table-2.

Table-2. Experimental results on Wikipedia datasets.

Similarity/Evaluation measure	Micro-average	Macro-average
Content-based similarity	0.267	0.235
Structure-based similarity	0.11	0.01

5. CONCLUSION AND FUTURE SCOPE

In this paper, we attempt the problem of clustering tendency of XML documents. This issue is solved by the procedure of VAT before clustering process.

The classical k-means algorithm doesn't know the clustering tendency initially. This limitation is solved by VAT and the relevant idea of the visualized clustering method is proposed and it uses a Khun-munkres concept for obtaining of efficient clustering results. We have done the clustering of XML documents from their structures and contents information. The semantic based clustering is the next future scope of this paper.

REFERENCES

[1] Xiaochun Wang; D. Mitchell Wilkes. 2009. A Divide-and-Conquer Approach for Minimum Spanning tree-Based Clustering. Member, IEEE Transactions on knowledge and data engineering. 21(7).

[2] J. Han and M. Kamber. 2002. Data Mining: concepts and techniques. Elsever.

[3] M. Ester; P. Kriegel; J. Sander; X. xu. 1996. A density based algorithm for discovering clusters in large databases with noise. Int. Proc 3rd International conference on Knowledge discovery and data mining. pp. 226-23.

[4] L. Wang, T. Nguyen, J. Bezdek, C. Leckie and K. Rammohanarao. 2010. iVAT and aVAT: Enhanced visual analysis for clustering tendency assessment. In: Proc PAKDD, India.

[5] Timothy C. Havens, James C. Bezdek. 2011. An efficient formulation of the improved visual assessment of cluster tendency. IEEE Trans on Knowledge and Data Engineering.

[6] Jain and R. Dubes. 1988. Algorithms for clustering data. Prentice-Hall.

[7] J. Bezdek and R. Hathaway. 2002. VAT: A tool for visual assessment (cluster) tendency. In: Proc. IJCNN, Honolulu, Hi. pp. 2225-30

[8] B. Eswara Reddy, K. Rajendra Prasad. 2012. Reducing runtime values in minimum spanning tree based clustering by visual access tendency, International Journal of Data Mining & Knowledge Management Process. 2(3): 11-22.

[9] Doucet A. & Lehtonen M. 2006. Unsupervised classification of text-centric xml document collections. In: 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX'. pp. 497-509.

[10] K. Rajendra Prasad and Dr. P. Govinda Rajulu. 2010. A Survey on Clustering Technique for large databases using Efficient Graph Structures. is accepted for Publication in the Journal International Journal of Engineering Science and Technology(IJEST). 2(7) , ISSN 0975-5462

[11] K. Rajendra Prasad and Dr. B. Eswara Reddy. 2013. Assessment of clustering tendency through progressive random sampling and graph-based clustering results. 3rd IEEE Int'l Advanced Computing Conference, Ghaziabad, India.

[12] Tien *et al*. 2008. Combining structure and content similarities for XML document clustering.

[13] L. Lovasz and M. Plummer. Matching theory. Budapest, North Holland, 1986

[14] B. Rama Subbaiah, M. Suleman Basha, B. Raviteja. 2014. A Review of Biometric Idenfication in Signal Processing. International Journal of Engineering Research. 3(2): 22-25