Clustering Algorithm for XML Documents Based on Similarity Content And Structure

1 Tayebeh Sahraneshin 2 Ali Haroun Abadi 3 Vahid Sattari Naeini

1. Department of Computer, Science and Research Branch, Islamic Azad University, Hormozgan, Iran, Email: nedaas@yahoo.com
2. Associate Professor: Dr. Ali Haroun Abadi, Email: a.harounabadi@gmail.com
3. Dr. Vahid Sattari Naeini, Email: Vnaeini@yahoo.com

Abstract

XML documents have taken into account due to their high flexibility and self-descriptive nature as one of the available methods for displaying and transferring data and many of these information resources use this method. Employing the methods that help to enhance the quality of search results is mandatory according to the increasing development of usage of the internet and search for information using search engines. One way of enhancing the quality of search engines is the usage of the techniques that conduct the search purposefully and present to the user only the related contents. In this paper, a method has been proposed for clustering XML documents that use of hierarchical clustering (based on the content and structure). Using XML documents clustering can help to enhance the quality of search results. In the proposed method, in the first stage, the documents have been classified based on the content similarity and then, documents of each class are classified based on structure similarity. K-means clustering algorithm was used for document content clustering and XCLS++ algorithm was used for structural clustering.

Keywords: XML document, Hierarchical clustering, Contextual clustering, Structural clustering, XCLS++ algorithm.

1-Introduction

Data mining is derivation or extraction of knowledge from a set of data, in other words, data mining is a process that extracts knowledge from a set of data using intelligent techniques.

Data mining algorithm requires a series of preprocessing on a training set and a series of postprocessing on the extracted patterns in order to be able to perform the knowledge extraction very well.

1-1-Data Mining stages

In general, stages of data mining can be categorized as follows:
Data cleansing: At this stage, the invalid data are removed from the training data set. Noisy data, incomplete data and etc are examples of data that cleanup should be done about them.

Data Integration: At this stage, the data multiple sources are combined together.

Data selection: Related data to the data mining process are separated from the other. This topic can be considered as part of the process of data reduction.

Data conversion: Data convert to a used format for data mining. The actions that take place at this stage are summarization or calculation of cumulative amounts.

Data mining: The main part of process in which takes place the extraction of knowledge patterns using specific methods and techniques.

Patterns evaluation: Desired correct recognition is done in this stage. The accuracy of patterns is evaluated based on a set of attractiveness criteria.

Knowledge representation: in this section, a set of visual tools are used to present the extracted knowledge to the user.

1-2-Data preparation

Since the related data were collected, it is very important to be aware that these data require to the cleanup and we should spend a lot of time for this. Because many error resources can be created during data collection from multiple databases in an analytical database and a good analyst is forced to perform many of data validation checks on the extracted data. It occurs rarely that there has not been any problem in the collected data.

1-3-Data reduction

Data reduction is to generate a smaller set of initial data that presents almost similar results with data mining results on the initial information under data mining operation.

This action can be performed through the elimination of non-related features with type of related data mining operation.

The elimination of related features that is made due to error in evaluating the amount of their relationship with data mining operation, can lead to inefficiencies in data mining process and extraction of incomplete rules and as a result worthless rules.

Removal failure of non-related features can considerably increase the time of data mining operation.

There are three methods for selecting related features of data mining:

- Progressive selection: At each stage, the feature that has the most relationship is chosen.
- Backward selection: At each stage, the feature that has the least relationship is chosen and is removed.
- Combined method: a combination of both progressive and backward method.
Different methods of data mining can be categorized as follows:

- Supervised learning algorithms (prediction method)
- Unsupervised learning algorithms (Description Methods)

The method will be introduced in the following is a method that uses hierarchical clustering (based on the content and structure) for clustering XML documents. Using XML documents clustering can help to enhance the quality of search results. In the proposed method, in the first stage, the documents are classified based on their content similarity and then, documents of each class are classified based on structural similarity.

The structure of this article is that at first in chapter 2 is presented the basic concepts such as clustering then, in chapter 3 the proposed method will be introduced in order to XML documents clustering. In chapter 4 the practical implementation results are presented and finally, chapter 5 includes conclusions and recommendations within the proposed method.

2-Basic Concepts

2-1 Clustering

Clustering is one of the most used techniques in data mining. Data clustering includes the methods for finding the small groups among large databases. Clustering is a tool for detecting data and finding a reliable and acceptable structure for data classification [1]. Clustering is often considered as the first step and one of the main important methods of data analysis. Clustering is a process in which objects can be classified into groups of similar objects. Each group or cluster contains objects that are similar to each other and are different from other group’s objects. Clustering is a form of data modeling that has roots in mathematics and statistics [1]. Unlike classification that is a supervised learning method, clustering is considered an unsupervised learning method, because data in classification have class label (it is a data that specifies a data object belongs to which group), but class label in clustering for data is unclear. Purpose in clustering is to minimize of inter-cluster data distance and maximization of data distance between the different clusters and hence it’s considered an optimization problem. Sometimes partitioning and segmentation expressions are considered instead of clustering. Clustering algorithms can be in forms of divisive (top to down) or agglomerative (down to top).

One of the proposed algorithms in the field of unsupervised clustering is the famous k-means algorithm [2].

2-2 Introduction of XML

In recent years, XML (eXtensible Markup Language) has become a standard for data representation and data exchange on the Internet. Many investigations have been done on storage and retrieval of XML documents. If XML document is mapped to a relational database, consequently, data can be retrieved using relational queries (conversion query into SQL language). But most of XML documents are stored in form of the XML database. The most important work in this database, is indexing on the documents. XML document indexing improves the speed of document retrieval during the query. Different optimization techniques have been developed for solving retrieval problems, updating and optimizing query.
Extensible Markup Language (XML) has become a suitable and efficient method for representing and exchanging data on the Web. Some of XML features which have created this interest are: self-defining nature of XML, its ability in representation of database scheme or object-oriented hierarchical and being distinguished of formatting features (tags) and real data.

Currently, a large amount of documents on the Web are provided in XML format. Content clustering techniques are employed for organizing improvement and query of these sets automatically. Movement towards the use of XML for data display indicates that XML documents set volume in the web has increasing growth each day. Therefore, it is required to develop techniques for clustering these documents so that information search is performed efficiently and appropriately.

Since XML documents are also content documents, the natural tendency is to utilize of standard methods of information retrieval in content documents for clustering these documents. Some other clustering techniques only consider the structure of XML document and are based on the theory that: existence of specific type of a structure in XML document is related to probability of its belonging to a particular cluster (category). This method completely ignores the content of fields. Therefore it is better the distinction is based on the scheme for clustering the documents and not based on the content. For this purpose, it is necessary that field names be meaningful.

2-3 Study of different segmentation for XML document clustering algorithms

XML documents (eXtensible Markup Language) have taken into account as one of the most important methods for data representation and transfer due to high flexibility and self-description nature [3] and many information sources or use it or seek to use it. One of the main problems of XML documents is how to display and to model them that can be performed into three different species of tree structure, graph and time series. Among these structures, tree structure is the best method to model, because XML documents have a user-defined hierarchical structure that be easily represented and implemented with tree structure. In this thesis, we will focus on the methods which use a tree structure [4].

XML data clustering is very different and more complex than conventional contextual data, because XML provide this possibility to add structural and semantic aspects to document content. A XML document contains labels and data that place among the labels. The label which describes the element names; contains concepts in form of the contextual data. Labels in addition to defining the document structure also show the relationship between the elements.

XML processing and management is an important research fields. The main important problem in the management of these documents is indexing them in order to optimal storage and retrieval. In the case of clustering these documents this is one of the important research issues, different algorithms have been proposed and discussed. In some clustering methods, XML documents are modeled as Rooty labeled trees. In this method, the amount of similarity between documents is calculated by edit distance between two trees [5].

In other XML document clustering algorithms, clustering is performed based on document Structure and hierarchical relationships between document elements [6, 7, 8, 9]. In this class of algorithms, the similarity
between documents is considered based on common elements and structural relationships [10]. In addition to methods that have been mentioned, there are also many semantic clustering algorithms that their numbers are very small.

According to studies in the field of XML document clustering, the used algorithms in this field can be classified in different methods. Based on a classification, XML document clustering algorithms can be classified into the following two groups:

- Methods with pairwise operation
- Methods with increasing operation

The basic of algorithms based on pairwise methods (double) is as follows: first, a similarity matrix is created for each two documents of all documents which want to cluster. This matrix is filled between two documents by the scale for measuring the similarity. Finally, after completion of the matrix, one of the most common techniques of clustering can be used such as K-means algorithm and the documents can be clustered based on their similarities to each other. In increasing methods, the amount of document similarity is computed with existing clusters for entering any XML document, and if the similarity amount be greater that a user-defined threshold value then it places in desired cluster, otherwise, a new cluster is created and places in it.

2-4 Pairwise algorithms of XML document clustering

Pairwise algorithms use of tree edit distance criteria for computing the similarity of two documents that have been converted to tree display, the problem is to compute the minimum distance between T1 and T2 trees whereas it uses three edit operators for each of tree nodes such as insertion, removal and replacement edit operators. For each operator is considered costs. Therefore, for obtaining the similarity amount between two documents, it’s enough to compute the minimum cost for conversion tree T1 to tree T2 using the sequence of defined operations. For better words, some insertions and removal and replacement must be performed on the tree T1 to convert to tree T2 that the cost of this operation shows the amount of similarity between two trees.

The method were described, is the most basic method for achieving similarity amount of two XML documents structurally (tree display represents the XML document structure). However, the significant point is time complexity and high cost of this method for while that a large number of XML documents is supposed to cluster, because for every pair of document, the evaluation of distance amount must be performed between two documents.

Many algorithms have been proposed in order to reduce the time complexity that tried to reduce costs with making changes. In this category, it can be noted that two following cases have created a significant performance:

- In [11], a method was presented that based on it, obtained tree display of XML document is summarized, in this way, suitable performance achieved compared to previous methods by reducing iterations and Nested elements and definition a new criterion for determining the structural distance between two documents.
In [12] has been claimed that clustering procedure is performed with high performance with S-graph theory in order to display XML documents and proposal a distance criterion. Reason for this claim is to encode S-graph to a bit string that clustering procedure operates on this bit string simply.

2-5 Increasing algorithms of XML document clustering

As mentioned in previous section, pairwise algorithms have high cost and their time complexity is of order \(O(n^2)\) in the best case (n is the number of documents), therefore researchers are looking for new algorithms that have a linear time complexity by maintaining the performance of previous algorithms and obtained clustering have the appropriate accuracy. Three algorithms XCLS [4] and XEdge [13] and XCLS+ [14] have been recently proposed, which have linear time complexity and good performance. These three algorithms operate increasingly that are described in more details in the following.

The general procedure in increasing methods is this way that a XML document is entered at each stage and is determined that the entered document must be placed in which cluster and this operation is performed for all documents until clustering procedure ends.

2-6XCLS Algorithm

XCLS algorithm is one of the methods that performs clustering procedure increasingly, by considering the following definition, we will present XCLS method that is represented in [4].

We consider a set of document XML\(D=\{D_1, D_2, \ldots, D_n\}\), the result of clustering algorithm is set \(C=\{C_1, C_2, \ldots, C_q\}\) so that \([C_1\cup C_2\cup \ldots C_q = \{D_1, D_2, \ldots, D_n\}]\) and also for each \(1 \leq i, j \leq q\) we have \([C_i \cap C_j = \emptyset]\). Where \(n\) is number of XML documents and \(q\) is number of clusters and \(C_i\) represents each of the clusters in the used algorithm for clustering.

The overall scheme of XCLS method is illustrated in figure 1-2, at first each XML document or tree displaying equivalent with that is expressed as level structure and in continuing the clustering algorithm classifies the documents based on the level similarity criterion. An global criterion measures the similarity amount in clustering level considering hierarchical structure related to XML documents and for each new document, its similarity amount is compared to the existing clusters instead of calculating the similarity amount between each pair of documents.

![Figure 1-2:Overall scheme of XCLS algorithm [12]](image-url)
XCLS algorithm starts by inferring of structural information of the document. Firstly, documents are converted to the ordered labeled tree display. Each label or element name which is used in document is specified with an integer number based on the appearance order in the document. The level structure indicates the element levels and labels in each level of tree structure. The level structure includes the information such as element name, their events and levels hierarchically.

3- Proposed method for clustering XML documents

The proposed method consists of two stages:

1) Document clustering based on meanings (content)
2) Structural classification of documents in each cluster.

1-3 Document content clustering

The goal of element content clustering is to survey the similarity of XML documents content. Two documents have more similarity if they have more similar contents. Figure 1-3 illustrates an example of content similarity.

As is shown in Figure 1-3, the contents of three documents $D_A$, $D_B$ and $D_C$ are:

$D_A$: {John, 2, Math, Physics, English}

$D_B$: {Mary, 3, Math, Physics, Drama, Music}

$D_C$: {My Book, 123456, Matt Brown, Angela Dickens}

By calculating two by two intersection of above sets, it can easily realize that $D_A$ and $D_B$ documents are similar to each other and they do not have any similarity to $D_C$ document (in terms of content).

In the proposed method, at first the contents of each document convert in form of a vector. Then, the created vectors are clustered with k-means clustering.
For converting content to vector, that’s enough the iteration number of each content is counted in a document. For example, assume that all the used contents in all documents contain $d_1$, $d_2$, $d_3$ and $d_4$. Then for a given assumed document $D$ that is shown in figure 2-3, its corresponding vector is $\langle 2, 1, 1, 1 \rangle$.

![XML codepiece](image)

Figure 2-3: An example of a XML codepiece.

As in the obtained vector is specified, the second vector element is equal to 2, because in the code piece of figure 2-3, $d_2$ is repeated twice.

### Document structural clustering

What is considered a similarity criterion in XML document structural clustering is document structure [16, 17, 18, 19]. Different methods are used for displaying the structure of a XML document. One of the most common methods is tree structure. An example of a tree structure is observed in figure 1-3. In this thesis also each document is mapped into a tree at first and evaluation of document similarity amount is performed based on their tree structure. Figure 3-3 shows the tree structure of figure 2-3.
The next stage of the proposed method includes the structural classification of clustered documents in the previous stage. Figure 4-3 shows general structure of the proposed method. Documents structural clustering is performed based on XCLS++ algorithm.

Figure 4-3, it shows the model of the proposed method.

4-Results

4-1 Evaluation Criteria

Four evaluation criteria have been used for evaluating of the proposed method performance including entropy, purity, precision and Remembering.
Entropy: Is total documents that have placed in cluster i and belong to class r. Therefore, the entropy value is calculated from equation 1-4.

\[
Entropy = \sum_{i=1}^{k} \frac{n_i}{N} \times E(C_i)
\]

\[
E(C_i) = \frac{1}{\log k} \sum_{i=1}^{k} \frac{n_i}{n_i} \log \frac{n_i}{n_i}
\]

In equation 1-4, variables \(C_i\), \(N\), \(k\), \(n\), and \(n_i^r\) represent cluster \(i\), total number of documents, number of clusters, number of documents of a cluster and number of document belonging to class \(r\) which have placed in cluster \(i\), respectively.

Whatever obtained entropy value be closer to zero, the better clustering has been done.

- **Purity Degree:** purity is sum of the maximum number of documents that has placed in cluster \(i\) and belongs to class \(r\). The purity degree is obtained from equation 2-4.

\[
Purity = \sum_{i=1}^{k} \frac{n_i}{\sqrt{N}} \times P(C_i)
\]

\[
P(C_i) = \frac{1}{n_i} \max(n_i^r)
\]

Whatever purity degree be closer to 1, the performance of clustering algorithms is better.

- **Precision:** The rate of documents which were clustered correctly than documents that have placed in a cluster correctly or incorrectly.
- **Remembering:** The rate of documents which were clustered correctly than documents which have placed in a cluster correctly or have not placed in a cluster incorrectly.

2-4 Obtained results of evaluation

Evaluated XML documents have been converted to the following files after the separation:


After the contextual clustering, bib_3.xml, bib_7.xml, bib_11.xml, bib_14.xml and AL_opuses2.xml files have been placed in cluster 1 and the others in cluster 2 using k-means clustering.

Table 1-4 shows the evaluation results of the proposed method for the used documents.
Table 1-4 Evaluated results of the proposed method

<table>
<thead>
<tr>
<th></th>
<th>entropy</th>
<th>purity</th>
<th>Precision</th>
<th>Remembering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.02</td>
<td>0.972</td>
<td>1</td>
<td>0.972</td>
</tr>
</tbody>
</table>

As shown in Table 1-4, Structural clustering has been also done with very high precision. In this evaluation cluster 1 has been clustered based on the clustering structure completely correct and in cluster 2 only one of the incorrect documents has been clustered (based on structure).

3-4 Comparison the proposed method with the similar methods

Table 2-4 shows the results of the comparison of the proposed method with similar methods.

Table 2-4: The results of comparison the proposed method with the similar methods.

<table>
<thead>
<tr>
<th></th>
<th>entropy</th>
<th>purity</th>
<th>Precision</th>
<th>Remembering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Method Cluster 1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Proposed Method Cluster 2</td>
<td>0.02</td>
<td>0.972</td>
<td>1</td>
<td>0.972</td>
</tr>
<tr>
<td>Structural Clustering using XCLS++ algorithm</td>
<td>0.18</td>
<td>0.91</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Structural Clustering using XCLS+ algorithm</td>
<td>0.28</td>
<td>0.83</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>nearest neighbor method (structural and contextual clustering) [26]</td>
<td>-</td>
<td>-</td>
<td>0.925</td>
<td>0.9289</td>
</tr>
</tbody>
</table>

As can be seen in Table 2-4, the performance of the proposed method is very good and is better than other methods.
5 – Conclusion

Considering XML documents use of the simple and obvious structure, nowadays they widely use as a database. XML documents define with a simple structure and can be stored in a contextual file with low volume. Using these documents as databases has led to introduce new methods for extracting information from them.

One of the important information that can be extracted from a XML documents database is dependency and similarities between these documents. Considering that nowadays this virtual world plays a crucial role in generation and transfer of knowledge, employing the methods are required that somehow make the use of available information easier and more efficient. Clustering of XML documents is one of the activities that can be very effective in the field of improving search on the Internet (by search engine). By clustering these documents, the search results of a phrase or a document will be more meaningful, because instead of providing a large number of web pages that may have no connection with each other, a set of pages (documents) are presented to the user that have a close connection with search topic.

XML document clustering can be based on content, structure or both of them [20, 21]. In content-based clustering, the criterion for evaluating the similarity of two documents is variable values (labels). Document structure is evaluated in structure-based clustering. This structure can express in based on tree, graph, time series and etc. Both criteria are evaluated in content-based and structure-based clustering.

In this thesis, a method has been presented in order to clustering of XML documents. This method uses of hierarchical clustering. Clustering is performed in two stages: in the first stage, documents are clustered based on the content and documents that have the similar content, place in the same cluster and in the second stage, documents of each cluster are clustered in terms of structure. The results of evaluating the proposed method indicate the precision of the method in documents clustering.

1-5 Suggestions

As mentioned in previous sections, the proposed method in this thesis, was created of two phases:

1- Content-Based Clustering
2- document clustering in the obtained clusters of the previous stage based on the structure,

it is obvious that because at first the contextual clustering is performed, in this method, the priority of contextual clustering is more than structural clustering, because the documents are compared structurally that have the same content.

One of the tasks that can be done to improve the proposed method is to determine the impact factor for content and structure. Therefore, according to the documents which are supposed to cluster, it can be determined which type of clustering is performed sooner (content or structure).
References


