XML Data Mining: Different Techniques for Clustering

Monika, Roopal Mamtora
ITM University, Gurgaon, India

Abstract

Nowadays, XML has become a popular standard for data representation and data exchange over the web because of its varied applicability in number of applications. So XML mining is the important domain for research. Out of many XML mining processes, clustering is the most challenging process. This paper on XML data mining explains several concepts related to clustering XML documents and presents some commonly used similarity measures and techniques available for XML data mining.

Keywords: XML, Mining, Clustering, Similarity measure.

1. Introduction

XML documents have rich and flexible format for information representation and data exchange on the web. Because of its simplicity, self-describing and flexible nature, it is used in a variety of web application. These applications exchange their data in XML format over the Internet as well as on the Intranet. Database System provides tools to store, deliver, integrate and query them. Still XML-oriented database are not ubiquitous. But developers are moving forward by adding XML compatibility to their products. So, the increasing use of XML format raises new challenges for organizing and managing the XML data and retrieves these XML documents in large collections. Many organizing and managing processes are used in mining. One of the challenging process is Clustering. Clustering is the grouping of similar XML documents in the same subset without any prior knowledge about the dataset. So, it is also called “unsupervised learning”. It is called an unsupervised learning task as no class values denoting an a priori grouping of the data instances are given. XML documents can be static or dynamic. These documents can be clustered based on type of similarity – structural or content.

2. Similarity measures for Clustering XML documents

XML is widely used in many information retrieval applications. To measure the similarity between the XML documents are the crucial issue of XML Clustering. Similarity measures of XML documents can be computed based on the content, structure or both. Traditionally for document clustering methods, only content information is used to measure the documents similarities. The structural information contained in XML documents is totally ignored. But the clustering based on content is not so good.

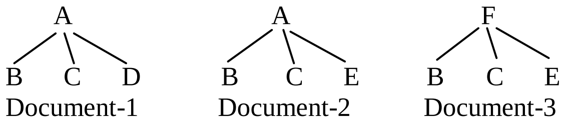
2.1. Similarity based on XML structure

The methods of computing similarities between structures of XML documents vary according to the representation of XML documents. If the representation of XML documents is based on tree then tree edit distance is used to measure the similarity between the structures of XML documents. If its representation is based on graph then similarity is computed on the basis of the set of Edges. If its representation is based on set of paths or edges or tags then similarity is computed on the basis of paths and so on. Some commonly used similarity measure are:

2.1.1. Tree–Edit Distance Approach. XML documents can be represented as labeled trees. To measure the similarity between two trees, compute the distance between trees which is known as tree–edit distance. The tree–edit distance generally computed using five different operations. The set of edit operations with the lowest total cost that transform one document into the other are Re-label, Insert, Delete, Insert tree and Delete tree.

Disadvantages of tree–edit distance :

i. Clustering quality produced by this method is poor.
ii. If the tree distance between documents that are structurally different will be same then it is not possible to distinguish these documents.

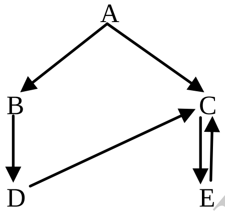


The tree transformation cost between Document-1 and Document-2 will be same as Document-2 and Document-3. If Document-1 and Document-2 cluster together, the DTD would be <! ELEMENT A(B,C,D,E)> which has four edges and if Document-2 and Document-3 cluster together, the DTD would be <! ELEMENT A(B,C,E)> and <! ELEMENT F(B,C,E)> which has six edges. It is better to cluster Document-1 and Document-2 but tree–edit distance may not be able to distinguish the structural difference.

2.1.2. Graph–based Approach. The problems in tree-edit distance are removed by new approach called "Graph–based Approach". In this approach structure–graphs (s–graphs) are used which are derived from XML documents, not from their DTD's.
Given a set of XML document D, the s–graph of D, sg(D) = (N,E) is direct graph where N is the set of all elements and attributes and E is the parent–child relationship.

```
<A>
  <B>
     <D>
        <C/>
     </D>
   </B>
   <C>
      <E>
         <C/>
      </E>
    </C>
</A>
```

XML Document



s-graph

The similarity between two given documents D_1 and D_2 is computed as follows:-

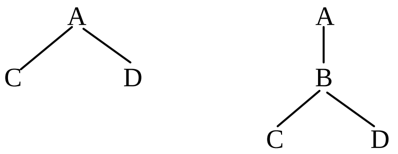
$$Sim(D_1,D_2) = \frac{|\, sg(D_1) \cap sg(D_2)|}{Max\{ \ |sg(D_1)| \, , |sg(D_2)| \ \}}$$

Where $|sg(D_i)|$ is the cardinality of edges in $sg(D_i)$, I = 1,2. Intersection(∩) gives the set of common edges in $sg(D_1)$ and $sg(D_2)$.

Disadvantages of Graph–based Approach:

i. Due to the presence of cyclic relationship between nodes, graph clustering is complex in nature.

ii. It relies on the loose grained similarity. Two documents having same s-graphs and still have significant structural differences.



According to definition similarity between two above given s-graphs is zero. Thus the measure fails to consider similar documents that do not share common edges even if they have many elements with the same label.

2.1.3. Path–based Approach. XML documents can be represented as a collection of paths. In this approach similarity measure between XML documents can be computed by finding the common paths. Various techniques are used for identifying the common paths. Some methods are XSD cluster, PCXSS, XClut, VSM model etc. Bit vector is used for constructing paths of the tree corresponding to an XML document. In this approach both node's name and node's position is consider in the path to measure the similarity between XML documents. The similarity between two given XML documents x and y is computed as follows:-

$$Sim(x,y) = \sum_{i=1}^{n} \sum_{y=1}^{m} Name(x_i,y_i) * min(Lev_{xi},Lev_{yi})$$

Where $Name(x_i,y_i)$ is the name weight of two nodes x_i and y_i and Lev_{xi} and Lev_{yi} are the Level weights.

Disadvantages of path – based Approach:

i. This approach fails to capture the sibling – relationship between the nodes in a tree which results in information loss.

ii. Partial path match, that is, the level information is not taken into account when nodes to be compared appear in different hierarchical level.

2.1.4. Sequence–based Approach. This approach is used to overcome the problem encountered in path– based approach. It stores the ancestor–descendent and sibling relation. XML trees are encoded based on sequence which establishes a one to one mapping between XML tree and sequence. Only common nodes are extracted based on sequence code instead of extracting all paths. This Approach is more effective.

2.1.5. Edge–based Approach. : It clusters both heterogeneous and homogeneous XML documents using edge summaries. Depending on the type of XML document, its proposed algorithm modifies its distance metric in order to properly adapt the special structure characteristics of homogeneous and heterogeneous XML documents. The main advantage of Edge – based approach is the preservation of structural relationships between nodes of consecutive levels of the XML documents form of edges.
The Similarity measure between two level Edge representation of homogeneous XML document is the proportion of total weight of the common edges

in same levels of the two level edges to the total weight of all distinct edges in both level edge.
The Similarity measure between two level Edge representation of heterogeneous XML document is the proportion of total weight of the common edges in levels of the two level edge to the total weight of all distinct edges in both level edge.

2.2. A methodology for the Choice of Similarity Measure

The presented approaches represent the current efforts of the research community in the evaluation of similarity between XML documents for clustering together similar XML documents.

i. Variants of tree edit distance are a good choice for structured XML collections when the structure is particularly relevant.
ii. If the structures of documents seem too complex, some kind of structural simplification such Structural summarization techniques can improve the results.
iii. If document collection is heterogeneous and XML documents do not share the same node tags, i.e., they belong to different semantic categories, a tag based similarity is suitable.
iv. If document collection is heterogeneous but XML documents share the same node tags, path based or edge based approach may be better than tag based approach.
v. If DTD information of document collection is available, bit vector based approach is suitable.

3. Clustering Methods

There are many well-known clustering algorithms. The main reason for having many clustering methods is the fact that the notion of “cluster” is not precisely defined. Farley and Raftery divides the clustering methods into two main groups: hierarchical and partitioning methods. Han and Kamber categorise the methods into additional three main categories: density-based methods, model-based clustering and grid-based methods. Some of clustering methods are :

3.1. Partitioning Method. Partitional clustering directly decompose the data set into a set of disjoint clusters. Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. In this method the number of clusters will be pre-set by the user. The criterion function that the clustering algorithm tries to minimize the local structure of the data by assigning clusters to the global structure. Typically, the global criteria involve minimizing some measure of dissimilarity in the samples within each cluster, while maximizing the dissimilarity of different clusters.

3.2. Hierarchical Method. These methods construct the clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. These methods can be subdivided as following:

i. **Agglomerative Hierarchical Method** — In this hierarchical clustering initially each object represents a cluster of its own. Then these clusters are successively merged to get the new cluster. This process is repeated until the desired cluster structure is obtained.

ii. **Divisive Hierarchical Method** — In this hierarchical clustering initially all objects belong to one cluster. Then the cluster is divided into sub-clusters, which are successively divided into their own sub-clusters. This process is repeated until the desired cluster structure is obtained.

The result of the hierarchical methods is a dendrogram. It is representing the nested grouping of objects and similarity levels at which groupings change. A clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level. The merging or division of clusters is performed according to some similarity measure. The hierarchical clustering methods could be further divided according to the manner that the similarity measure is calculated:

Single-link clustering: In this clustering the distance between two clusters to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, the similarity between a pair of clusters is considered to be equal to the greatest similarity from any member of one cluster to any member of the other cluster.
Complete-link clustering : In this clustering the distance between two clusters to be equal to the longest distance from any member of one cluster to any member of the other cluster.
Average-link clustering: In this clustering the distance between two clusters to be equal to the average distance from any member of one cluster to any member of the other cluster. Such clustering algorithms may be found in.
Disadvantages of the hierarchical methods:
i. Inability to scale well—The time complexity of hierarchical algorithms is at least $O(m2)$ (where m is the total number of instances), which is non-linear with the number of objects. Clustering a large number of a objects using a hierarchical algorithm is also characterized by huge I/O costs.
ii. Hierarchical methods can never undo what was done previously. Namely there is no back-tracking capability.

3.3. Density-based Method. Density-based methods assume that the points that belong to each cluster are drawn from a specific probability distribution. The overall distribution of the data is assumed to be a mixture of several distributions. The aim of these methods is to identify the clusters and their distribution parameters.

3.4. Model-based Methods. These methods attempt to optimize the fit between the given data and some mathematical models. Unlike conventional clustering, which identifies groups of objects, model-based clustering methods find characteristic descriptions for each group, where each of the group represents a concept or a class. The most frequently used methods are decision trees and neural networks.

In decision trees, the data is represented by a hierarchical tree, where each leaf refers to a concept. Each leaf contains a probabilistic Clustering Methods description of that concept.

In Neural Networks algorithm, each cluster is represented by a neuron or "prototype". The input data is also represented by neurons, which are connected to the prototype neurons. Each such connection has some weight, which is learned adaptively during learning.

3.5. Grid-based Methods. These methods divide the space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. Grid–based methods have fast processing time.

3.6. Soft computing Clustering. Traditional clustering approaches generate partitions; in a partition, each instance belongs to one and only one cluster. Hence, the clusters formed by a hard clustering are disjointed. Fuzzy clustering extends this notion and suggests a *soft clustering* schema. In this case, each pattern is associated with every cluster using some sort of membership function, namely, each cluster is a fuzzy set of all the patterns. For the assignment of the pattern to the cluster, larger membership values (μ) indicate higher confidence. A hard clustering can be obtained from a fuzzy partition by using a threshold of the membership value (μ).

Conclusion

XML Mining is used to retrieve the useful information from very large amount of web data. The increasing use of XML documents for data representation and data exchange has attracted a great deal of researchers for efficient data management and retrieval. In this article, we tried to describe some commonly used similarity measures and clustering methods.

4. References

[1] W3C (1998). Extensible Markup Language (XML).
[2] On Some Clustering Techniques - Bonner, R.

[3] Evaluating Structural Similarity in XML Documents - Nierman, Jagadish
[4] A Tree-Based Approach to Clustering XML Documents by Structure - Costa, Manco, et al. – 2004
[5] The Tree-to-Tree Editing Problem - Selkov - 1977.
[6] An Efficient and Scalable Algorithm for Clustering XML Documents by Structure - Lian, Cheung, et al. - 2004
[7] A Methodology for Clustering XML Documents by Structure - Dalamagas, Cheng, et al. - 2006