Available online at www.elixirpublishers.com (Elixir International Journal)

Information Technology

Elixir Inform. Tech. 72 (2014) 25863-25866

ISSN: 2229-712X

Clustering and mining multi-versioned XML documents

Arti A Chaudhari*, Shweta A Gaikwad, Sujata V Patil and Surekha V Barke
Information Technology, University of Pune, India.

ARTICLE INFO

Article history:
Received: 18 March 2014;
Received in revised form:
15 July 2014;
Accepted: 26 July 2014;

Keywords
Banking services costs,
CAIS' components,
Competitive advantage.

ABSTRACT

Clustering is a process that partitions data in such a way that homogeneous data items are grouped into sets which is referred to as clusters. When the content or structure changes over time of the multi-version XML documents then Clustering is done. In real-world applications, the number of changes from one version of an XML document to another version of an XML document cannot be predicted. It is always possible that an initial clustering solution becomes obsolete after the modification take place in document. XML clustering algorithms is use to calculate pair-wise distances between documents. A time-efficient technique determined the pair-wise distances in a timely manner. In this paper we proposed a time-efficient technique to reassess pair- wise distances between clustered multi-version XML documents which change over a time, without performing redundant calculations. While performing redundant calculations we considering the previously known distances and the set of changes which might have affected the documents versions Mining is the process of searching the XML documents from the formed clusters and extracting the particular data from that searched XML document. For mining we have one Metric which has details that in which cluster a particular record should be. So when we want any reports we directly go to metric and see where we will find that records and directly access files inside that cluster.

Introduction

XML document is use for data storage and data exchange between applications. Types of XML documents are static XML documents and dynamic XML documents. Static XML documents do not change or modify their content and structure over time. For eg. an XML document containing details of papers presented at a conference or journals. Dynamic or multi-versioned XML documents can modify or change their structure And content over time. For eg. if the content of an online banking were represented in XML format, it would change daily depends upon the e-customer behavior. XML [Extensible Markup Language] has play important role in increasingly extension use of it as standard language for information representation and data exchange on the web. Most web applications deal with web data by translating them into XML document format, In order to organize these data efficiently grouping XML documents because of their structure, content and semantics hidden inside them is a possible solution.

In mining literatures one organizing process is referred as clustering which group similar XML data across heterogeneous once. Clustering is an intelligent technique for mining XML documents has been utilised as an excellent way of grouping the documents by their content or structure. A distance based XML clustering algorithms is use to calculate pair-wise distances between documents. Normally, a time-efficient technique requests the pair-wise distances to be determined within a time. In case of dynamic or multi-version XML documents, the amount of changes between various document versions cannot be predicted. Therefore, in case of dynamic XML documents, if hanges were minimum or if they affected only some of the clustered documents, recalculating pair-wise distances each time would be highly redundant.

We will propose a time-efficient technique to reassess pair-wise distances between clustered dynamic XML documents which changes in time, without performing redundant calculations. But it is consider only the previously known distances and the set of changes which affected the documents versions. In distance-based clustering techniques, every object from the given set is first assigned to a cluster. Then the distances between pairs of clusters are computed, and the closest clusters (the most similar) are grouped to form a new cluster. In other words, when two XML documents are more similar compared to other pairs of XML documents then the distance between them is smaller. therefore, they can become members of the same cluster.

RELATED WORK

In this we discuss the existing work in the area of clustering XML documents, stressing the fact that any of the existing system does not deal with the efficiently reassessing clusters of multiversion XML documents. Three main directions of research can be noticed, in this regard:-

1. Techniques for the clustering static (not changeable) XML documents.
2. Techniques for the clustering series (streams) of XML documents .
3. Clustering dynamic XML documents.

1. Clustering of static XML documents : In this technique the documents to be clustered are all known or familiar and available in advance, before running the clustering algorithm. The static XML documents do not change, hence the pair-wise distances are calculated only once, the resulting clustering solution is static i.e not changeable.

Clustering series of XML documents : In this technique the documents to be clustered are not familiar in advance. They

Tele:
E-mail addresses: shwetagaikwad03@gmail.com

become available one by one and the algorithm used to recalculates the distance between each incoming new document and the existing clusters. These techniques are not applied for the dynamic xml documents as in the multiversion XML document, content of the document is continuously changes[1].

3. Clustering dynamic XML documents : In this dynamic XML document technique if some or all of the clustered XML documents already change their structure or content in time, then in order to reflect the dynamic of the application. When the clustered documents are change the previous clustering composition might become no use longer if changes were in the document so significant that the modified documents were reallocated to different clusters or if new clusters were formed[1].

In mining XML documents are searched without clusterwise .The distance of a document to be search is a calculated first then this distance is match with the number of documents stored in XML document database, start matching the distance from the first document of database then second upto the N no. of documents until the document gets match but this process requires a lots of time for searching the document[3].

PROPOSED WORK

Distance measurement :

We are proposing an intelligent and time- efficient technique for determining the distances between clustered dynamic XML documents after they change, not by running full pair-wise comparisons but by calculating the effects of the changes on the previously known distances, that is on the distances

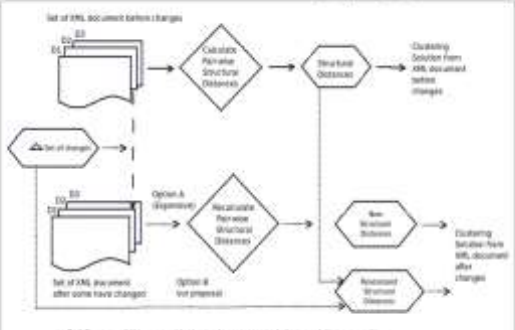
Set of XML document before changes
D1
D2
D3
Calculate Pairwise Structural Distances
Structural Distances
Clustering Solution from XML document before changes
Set of changes
D1
D2
D3
Option A (Expensive)
Recalculate Pairwise Structural Distances
New Structural Distances
Clustering Solution from XML document after changes
Set of XML document after some have changed
Option B (our proposal)
Reassessed Structural Distances

Fig1 Overview of the proposed technique to reassessed clustering solution composition

As shown in Fig. 1 an overview of the identified problem. As it can be noticed, one straight forward option (option A) would be to recalculate, after each set of changes and the distances between the XML documents by doing a full pairwise comparison of them. This option would be very expensive from the operational point of view, because there is no distinction made between documents affected more or less by the set of changes; hence, in case of: (i) new versions of documents carrying only a small amount of changes or (ii) documents not modified at all, some or all operations involved in the full comparison of each pair of documents would be unnecessarily repeated. The second option (option B – i.e our proposal) is to make use of the already known distances between pairs of XML documents in the clustering composition before the changes and the set of temporal changes, and use them together to determine the new modified distances. In short We are going to perform following modules[1].

Example of similar and dissimilar XML documents :

In Fig. 2 there are three XML documents in which, document DA and DB are highly similar. That is there is a similarity in a attributes of these XML documents, but the document DC that is the third document is not similar to either DA or DB. Documents DA and DB contains the information about two student John and Merry respectively, those includes attributes like year of study, subjects, exams and student names. Where the third document DC list the information about book, which includes the attributes like title, ISB number, and author names. In Fig. 2, any queries regarding students' details are applicable only to the relevant documents (that is, DA and DB) and not to any other documents which contain a different kind of information, such as DC. Documents DA and DB are grouped in a cluster, while DC forms a cluster by itself[3].

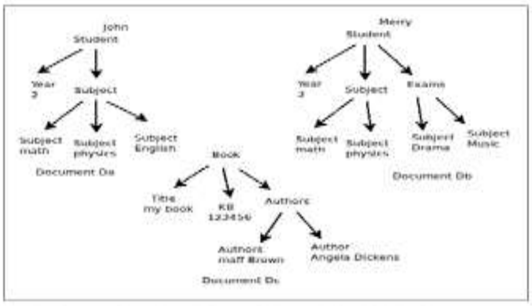

Fig 2: Example of similar and dissimilar XML documents

Fig 2: Example of similar and dissimilar XML documents

Clustering architecture

Clustering is very useful technique for grouping data objects such that objects within a same group or cluster have similar features, while objects in different groups are dissimilar. Architecture of an XML document clustering system can be illustrated as shown in Fig.3 [8]

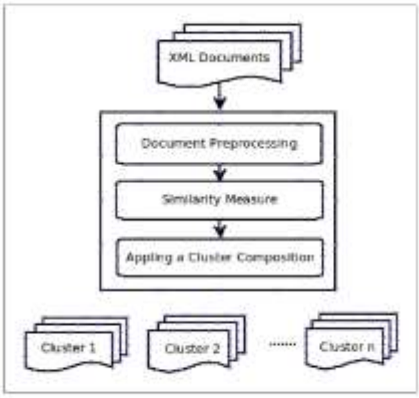

Figure 3 : XML document clustering architecture

(a) Document preprocessing : Documents are represented in a common data model then necessary preprocess is applied on structure and content of them to prepare them for retriving information for clustering. Different tasks are done based on the document representation..

(b) Similarity Measure : we should define an appropriate similarity measure due to the representation model in order to determine degree of similarity between pairs of objects.

(c) Clustering : The similar data objects are grouped together based on similarity measure using clustering algorithms[8]. We use Density Base algorithm to form a cluster. The key idea of the DBSCAN algorithm is nothing but for each point of a cluster, the neighbour of a given radius has to contain at least a minimum number of points which is the density in the neighbour has to exceed some already defined threshold. This algorithm needs three input parameters:

- k, the neighbour list size;
- E, the radius that delimitate the neighbourhood area of a point (E-neighbour);
- MinPts, the minimum number of points that must exist in the E-neighbour.

The process of XML document clustering is based on the classification of the points in the dataset as core points, border points and noise points, and on the use of density relations between points (directly density-reachable reachable, density-reachable, dens- ity- connected) to form the clusters.

Description of the Algorithm : In general the algorithm has two steps, choosing parameters E and cluster with varied densities. The procedure for this algorithm is as follows,

(i) It finds and stores k distance for each project and partition the k distance plots.

(ii) The number of densities is given primarily by k distance plot.

(iii) The parameter E is selected automatically for each density.

(iv) Scan the dataset and cluster various densities using corresponding E Display the valid cluster with respect to varied density.

Algorithm:

```
1.Partition k disancet plot.
2.Give thresholds of parameters E
   (i=1,2,.....n)
3.For each E
   (i=1,2,.....n)
   a)E = E
   b)Adopt DBSCAN algorithm for points that are no marked.
   c)Mark points as c
4.Display all the marked points as corresponding clusters[9].
```

Allocation Of XML Documents :

If we take input as first XML document then a cluster is form to store that document without considering any distance. To form a more clusters we require XML documents. So measure the distance of XML documents by measuring their attributes. After calculating the distance, give specific distance value to the cluster. That distance value is called threshold value. For e. g- If the distance of XML document is 10, so we consider the threshold value for the cluster more than 10, suppose it is 15. When next XML file is entered then we calculate the distance of that document. Compare that current distance with threshold value, if current value is less than threshold value then that XML file store in existing cluster otherwise new cluster form to store the XML file for this we use a cluster queue algorithm, which is explained below,

```
Distance(D) = CTv - cTv
if (D>CTv)
{
new cluster
}
else
{
14
Clustering and Mining Multi-version XML Documents
save in current cluster
}
```

where,

D=diffrence between Ctv-ctv

CTv=threshold value of current cluster.

cTv=threshold value of current document.

Mining of XML Document :

For searching the XML document, we have to search the documents with cluster and without cluster. In the mining with cluster, we have some threshold value to cluster. Then for searching the documents with cluster, we compare the distance of documents that is threshold value Clustering and Mining Multi-version XML Documents of documents with the threshold value of cluster. If the threshold value of document is less than the threshold value of cluster then search that document in a respective cluster. Otherwise search that document in the another cluster. For this overall process we have to use the searching algorithm. As shown in fig.5 below we do the data mining on XML documents and this process is called as XML mining and from that we will get the knowledge discovery of respective XML documents.

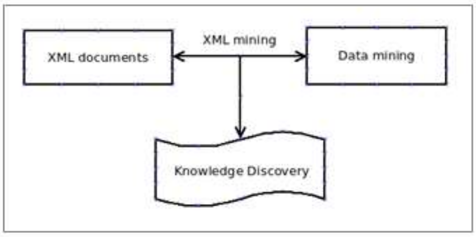

Figure 4 : Role of XML Mining

For mining we have one Metric which has details, in which cluster a particular record should be.

For example :

Consider whether report as a xml document.

In weather report we have,

_RMX : Max Rain Only

_RMX_SMX : Max Rain and Max Snow

So when we want max rain reports we directly go to metric and see where we will find the Max rain records and directly access files inside that cluster. Metric will form at runtime because we don't know how many clusters are present at that particular moment. This mining method is called as Information Based Metric Mining. Every time anybody asks for data(mining) Metric will form every time as per available clusters And for non clustered data we go record by record and see is record match with our requirement of mining.

CONCLUSION

In Our proposed technique allows the user to reassess the pair-wise Distances between XML document. Instead of fully comparing each new pair of versions in the clustering solution,we will determining the effect of the temporal changes on the previously known distances between them. This approach used for both time and I/O effective, as the number of operations involved in distance reassessing is greatly reduced. For mining we have one Metric which has details, in which cluster a particular record should be. So when we want any reports we directly go to metric and see where we will find the that records and directly access files inside that cluster.

ACKNOWLEDGEMENT

We express our sincere and humble thanks to Sandip Institute of Technology and Research Centre for making resources available. We especially thanks to our guide prof. Vijay Sonawane sir for mentoring, guiding and igniting our minds throughout the project. We also thanks to all are friends who supports and encourage us during the compilation of the pro

REFRENCES

[1] Rusu L.I., Rahayu W. and Taniar D., Intelligent Dynamic XML Documents Clustering, In Proceed of The 22nd International Conference on Advanced Information Networking and Applications.(IEEE-2008).

[2] Rusu L.I., Rahayu W. and Taniar D., Extracting Variable Knowledge from Multi-versioned XML Documents, In Proceed of The 6th International Conference on Data mining.(IEEE-2006)

[3] Laura Irina Rusu, XML data mining, Part 3: Clustering XML documents for improved data mining, May 2012.

[4] Mining XML Documents with Association Rule Algorithm - Gorkem Gurel.

[5] Costa, G., Manco, G., Ortale, R. and Tagarelli, A., A tree-based Approach to Clustering XML documents by Structure, PAKDD 2004, LNAI 3202, 137-148, Springer 2004

[6] Dalamagas, T., Cheng, T., Winkel, K.J. and Sellis, T., 2004, Clustering XML documents by Structure, SETN 2004, LNAI 3025, 112-121, Springer 2004

[7] Rusu L.I., Rahayu W. and Taniar D., Mining Changes from Versions of Dynamic XML Documents, (2011)- Springer-Verlag

[8] Elaheh Asghari, Mohammad Reza, Keyvan Pour, XML document clustering: techniques and challenges, Springer 2013

[9]A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases

Web-sites :

• www.cs.washington.edu/datasets - SIGMOD XML dataset