Mining XML data using K-means and Manhattan algorithms.

Wria Mohammed Salih Mohammed

Abstract— over the last two decades, XML has astonishing developed for describing semi-structured data and exchanging data over the web. Thus, applying data mining techniques to XML data has become necessary.
K-means clustering is one of the most popular algorithms in the clustering of data mining. Recently, there have been some researches undertaken on the mining XML data.
In this paper, applying k-means algorithm, which is one of the clustering algorithms, on XML data is proposed. K-means as an algorithm chooses centroids and then clustering the XML data into groups according to the centroids. The comparison distances between each element vary with every centroid and will make groups of elements together. The closest elements from each others will be in the same group. The distances are measured using the Manhattan algorithm. In this research a specific application has been build, the application allows the user to upload an XML file, choose the target field and select the number of clusters. As a result, the application shows the clusters and centroids used in all of the steps.

Index Terms— ASP.net, Centroids, Cluster, Data Mining, K-means, Manhattan, XML.

1 Introduction

With regardes to developing web technology, XML data has been developing over the past two decades. It has been mostly used with web technologies to transport and store data. As a result, XML has received a great deal of attention from the database community because XML is written by humans not computers. The properties of an XML file are:

- It is understandable and readable
- Users can create their own elements and attributes
- XML is also easier to code than html.
- Data can be easily exchanged between dissimilar applications
- XML is well formed because there is a rule to create an XML file, which makes XML well-structured [14][15].

However, the data mining community has focused on extracting data from XML files more than applying data mining of XML such as classification, clustering or association algorithms. Also, some papers have been written about clustering on XML files using the nearest neighbor, but there are a few papers about K-means algorithms. However, in this paper we are going to illustrate the mining of XML data using K-means algorithms. Another problem exists when any researchers want to research about XML mining. They need to transform XML files to traditional dataset using some tools, such as DOM or SAX builder in Java language programming or C#. But, in this research, we built an application to collect both steps in one step; it means that we can do converting and mining of XML to traditional dataset in our application without using any other tools.
In XML we have two main value types of data. One of them is element and another is attribute. There are some tools to retrieve elements without attributes and there is a tool in asp.net (XMLDataSource) to obtainattributes, but it cannot retrieve the elements. To solve this problem, we used an XML reader method to get both attributes and elements, and showing all of them in the traditional dataset. Finally, we can apply K-means algorithm with Manhattan on the traditional dataset.
It is necessary to use modern and valuable tools to achieve data mining in the case of XML, because of dissimilarities between XML data and relational databases in some aspects for example, an XML database is self-describing, but a relational database is not. XML needs to be inherited and ordered. However, this is not necessary for a database. Additionally, a relational database has a regular structure and it is flat. In contrast, an XML file contains many levels of nested documents [5]. XML, as a mark-up language, mixes mark-up and text together in a single file. Also, it has the following structures:

- XML declaration: e.g.
  <?XML version="1.0"?>
- Tags and elements.
  <Author>Wria Mohammed**</ Author >**
- Attributes
  … b is an attribute
- References:
  &amp = &, &gt = greater than.

• Wria Mohammed *has masters degree in Databases and web-based systems, he is also A. lecturer at the university of Sulaimani, Kurdistan-Iraq. E-mail: wria.salih@univsu.edu.iq*

- Text [16]

2 PRPBLEM STATEMENT

The problem of this research is to implement one of the unsupervised data mining techniques on XML data, implementing data mining on XML data requires two different techniques; one of them is converting an XML data to traditional dataset and another is to mine the traditional dataset.

This paper is concerned with the clustering technique of K-means algorithm on XML data. Applying K-means to XML data, we built a specific application to convert XML to traditional dataset and mine it.

Clustering is a process of partitioning a set of large group data into subsets, and each subset called a cluster. Also, all objects in a cluster are similar and dissimilar to objects in other clusters [1]. Another benefit of clustering is to reduce the size of the data that makes more understandable [2]. Furthermore, clustering is widely used in security, biology, web search, image patterns, and business intelligence, for examples, In business intelligence, clustering can be used to organize large numbers of customer into groups according to their similarities [1]. In biology, grouping of animals according to their features is an example of clustering [3]. In libraries, clustering is used for book ordering.

Clustering is unsupervised learning because unlabeled data is input, and the output of cluster algorithms is segmented data. There are many algorithms for clustering, with the fundamental algorithms as follows:

- Partitioning method
- Hierarchical method
- Density-based method
- Grid-based method

The basic algorithm of clustering is the partitioning method that classifies data into some groups. However, one of the most famous clustering methods is the K-means method.

In the K-means algorithm, we should first decide how many clusters we would like to produce from the data. This is (K) value. Each K value is treated as a centroid. According to the nearest centroid, every object should be assigned. Thus, we have K clusters based on original K-centroids. After that, we recalculate the centroids for each cluster; then we will have new centroids for each cluster. This process repeats continuously until there is no change in the result of the previous step with the current step. In summary:

Choose the value of K randomly (how many centers we need)

Initialize the value for each K.

Assign each object to the nearest centroids.

Recalculate the centroids of each cluster.

Repeat the two previous steps until there is no change in the results [1], [4].

In the K-means algorithm:

Input:

K: the number of clusters

Dataset containing N objects.

Output:

A set of K clusters.

3 RELATED WORKS:

As demonstrated Recently, various clustering methods have been put forward. Saraee & Aljibouri[5] have proposed the XML data mining in clustering using the nearest algorithms. Also, as [6] mentioned, they use clustering algorithms for both homogenous and heterogeneous XML documents by using LevelEdge. The distance matrix between two XML documents was based on LevelEdge. Furthermore, they utilized partitioning clustering. Additionally, they [7] have studied the clustering, and they also focused on the feature selection. The paper [7] evaluates this approach with a collection of Inria activity1 for the year 2003. The author in [8] explained briefly XML data mining by using different techniques for clustering, including some clustering methods. For example, the partitioning method, hierarchical method, and density-based method, but paper [8] did not discuss K-means algorithms. However, paper [9] has mentioned K-means algorithm in a particular section and in [9] they proposed an approach for heterogeneous semantic clustering of an XML document and one of the methods which they used was the K-means. As [10] implemented the clustering algorithms on INDEX collection, they also used clustering tool implemented (CLUTO2) by George Karypis at Minnesota University. They also proposed K-means when $K \in \{5, 10, 15, 20, 25, 35\}$ for text, tags and text and tags together. However, in this project, we are going to focus on the implementation of K-means clustering algorithms only by using a particular tool that has been created by ourselves using asp.net with C# programming, since this tool can convert an XML data to traditional dataset and implement data mining on it. Also, It mentioned before in this paper, K-means clustering algorithm is used, for example:

Cluster :{ 2, 3, 6, 8, 15, 4, 7}, we suppose that k = 2, initial M1= 3 and M2=6 (k: number of clusters, M1: first centroid, M2: second centroid)

M1	M2	K1	K2

[1] Inria (the French National Institute for Research in computer science and control) publishes activity reports on the French parliament; it is also a paper document that is written by every Inria research team. It has the XML version that consists of 139 files (22900 lines)[7].

[2] -CLUTO, http://www-users.cs.umn.edu/~karypis/cluto/

3	7	{2,3,4}	{7,9,15,6}
3	9.25	{2,3,4,6}	{7,9,15}
3.75	10.33	{2,3,4,6,7}	{9,15}
4.4	12	{2,3,4,6,7}	{9,15}

Our answer is thus K1 = {2,3,4,6,7} and K2 = {9,15}.[11]

Centroids M1 = 4.4 and M2 = 12
XML is more complex than a traditional dataset because it is semi-structured data [12].

4 METHODOLOGY

K-means algorithms are applied by implementing ASP.net with C# codes.The application, which was created for this purpose, is made up of two main parts; first, converting XML dataset to traditional datasets, and secondly, applying K-means clustering algorithms to the traditional datasets. In this application, the user can work with elements, attributes, and the summary of how this application works is shows as followings

1- At the beginning of the web application, it allows the user to upload an XML file (only XML file) that they wish to be clustered.

2- When an XML file is uploaded successfully, it will be converted to traditional datasets by clicking on read XML file button. First the application read XML by using (ReadXML) method is entered into a dataset and then the datasets are shown with the Grid View tool.

3- The user can select the ID field, target field and the number of clusters with the drop down List. In the ID and Target Drop down List, all fields will appear and the user can select one of them.

- ID: the main field that only the result output shows.
- Target: the clustering will happen according to this field.

Also, the user can select how many clusters they want; in this application we just implemented two and three clusters.

4- The number of centroids depends on the number of clusters, if two clusters are chosen; the number of centroids are two as well. The first centroid will be the value of one-third (1/3) element, and second the value of three-quarters (3/4) elements. If the three clusters are selected, the centroids will have the value of one-quarter (1/4) element, middle element (1/2) and three-quarters element (3/4).

For example:
Cluster: {4,5,6,2,1,9,23,25,46}, when we will have two clusters. We need two centroids:
One-third (1/3) element is 6.
Three-quarters (3/4) element is 9.

5- The first output will be a dataset which includes all XML data (elements child entities and attribute values).

The second output will be the result of K-means algorithms mining that includes the centroids in each steps. They will be shown in the ListBox tools as well as each cluster element.

5 EXPERIMENTAL STEP

This web application was developed and tested using XML data. Below is a sample of the XML data:

```
<library>
 <book ISBN="978-0123814791">
  <author>M. K. J. P. Jiawei Han</author>
  <title>Data Mining: Concepts and
            Techniques</title>
  <Category>Computer</Category>
  <quantity>2</quantity>
  <publish_date>2011</publish_date>
  <description> this book is about data
            mining and it is one the most
      famous book in this area.
  </description>
 </book>
        .
        .
        .
</library>
```

It can be clearly seen that it includes both elements and an attribute (ISBN="978-0123814791" is an attribute and <author>is an element). In addition, this XML data file includes both numerical data and textual data e.g. <title> text and <quantity> is numerical data.
This web application used the Manhattan distance (/ X1 – X2/) to measure the distance between two elements. Also, every element is compared to all centroids; whichever centroid is closer, will be its group. For example, the centroids are (7, 11) and there is (5) as an element:
X1 is first centroid, X2 is second centroid, and Xi is any element.
/X1 – Xi/ = / 5-7/ = 2
/X2 - Xi / = / 5 – 11 / = 6
It means that 5 will be a group member of 7 not 11, so to summarize:
If /X1 - Xi / < /X2 - Xi / then Xi is a group member of X1

If /X2 - Xi / < /X1 - Xi / then Xi is a group member of X2

This one only applies to two clusters, but if the user chooses three clusters, it will be as follows:

If /X1 - Xi / < (/X2 - Xi / and /X3 - Xi /) then Xi is a group member of X1

If /X2 - Xi / < (/X1 - Xi / and /X3 - Xi /) then Xi is a group member of X2

If /X3 - Xi / < (/X1 - Xi / and /X2 - Xi /) then Xi is a group member of X3

After comparing all points against the centroids, we will have two clusters or three clusters, depending on which one was selected. For the next step, each cluster will have new centroids by finding the mean3 for each cluster, and then new centroids will again be compared to each element. As a result, we will have new clusters. This process is continually repeated until the results of new centroids are the same with the previous results.

6 Experimental Results

First of all, the built application for this research can only accept XML file, it means only XML file can be uploaded, after that it asks you to select target, and the number of clusters that the user wishes to be clustered.

6.1 K-means clustered results

When an external XML file uploaded successfully, all values of element children and attributes are shows as traditional dataset, as shows in the following graph-1.

author	title	Category	quantity	publish_date	description
M. K. J. P. Jiawei Han	Data Mining: Concepts and Techniques	Computer	2	2011	advances in data mining technology have made extensive data collection much easier.
M. Bramer	Principles of Data Mining	Computer	4	2013	A former architect battles corporate zombies, an evil sorceress, and her own childhood to become queen of the world.
R. M. Monika	XML Data Mining: Different Techniques for Clustering	International Journal	10	2013	After the collapse of a nanotechnology society in England, the young survivors lay the foundation for a new society.
G. Q. T. a. D. U. Pasquale De Meo1	An Approach for Clustering Semantically Heterogeneous XML Schemas	Data Mining paper	12	12 9 2005	G. Q. T. a. D. U. Pasquale De Meo1, "An Approach for Clustering Semantically Heterogeneous XML Schemas," pp. 32 - 346, 12 9 2005.
A. Doucet and H. Ahonen¬Myka	Naive clustering of a large XML document collection	Data Mining paper	3	2001-09-10	The two daughters of Maeve, half-sisters, battle one another for control of England. Sequel to Oberon's Legacy.
Steven Feuerstein , Bill Pribyl	Oracle PL/SQL Programming	Programming	20	1 Oct 2009	This book is the definitive reference on PL/SQL, considered throughout the database community to be the best Oracle programming book available.
Todd Burpo , Sonja Burpo , Colton Burpo	Heaven Is for Real: A Little Boy's Astounding Story of His Trip to Heaven and Back	true story	30	2 Nov 2010	A Little Boy's Astounding Story of His Trip to Heaven and Back A young boy emerges from life-saving surgery with remarkable stories of his visit to heaven.
Jamie Reid	Doped: The Real Life Story of the 1960s Racehorse Doping Gang	Real Story	11	27 Sep 2013	Doped is the gripping true-story racing thriller set in Britain in the 1950s and early 1960s.
Joe Fawcett, Danny Ayers, Liam R. E. Quin	Beginning XML	Computer Science	25	6 July 2012	Dive into the key aspects of XML to deliver data on the web From simple data transfers to providing multi-channeled conten

Graph-1 converting XML to traditional dataset.

After selecting the target field and number of clusters, you just need to click on K-means, and clustering will happen and all centroids will be shown.

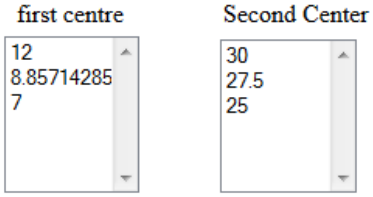
In this example, we will choose (quantity) field as a target and (two) clusters as the particular number of the clusters.

Quantity = 2 , 4 , 10 , 12 , 3 , 20, 30 , 11 , 25

The first centroids are 12 and 30 as the application chose.

The first centroids: 12 , 8 , 7

The second centroids: 30 , 27 , 25



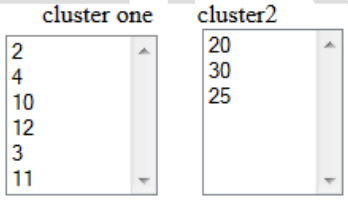
Graph-2 centroids.

New clusters and new centroids are continually repeated until we will have the repeated results, then the loop will stop. As a result, we need only (7) as a final result of the first centroid and 25 as the second.

Finally the cluster will be clustered into two groups according to the final centroids;

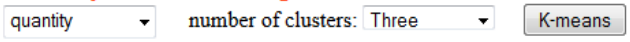
First cluster = 2 , 4 , 10 , 12 , 3 , 11

Second cluster: 20 , 30 , 25.



Graph-3 clusters.

If we select three clusters:



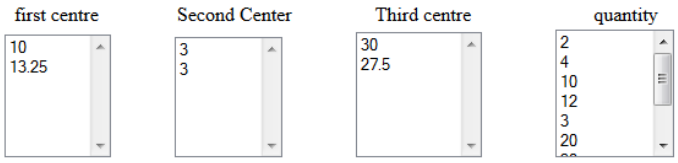
Then we will the following results:

First centroids = 3, 3

Second centroids= 10, 13.25

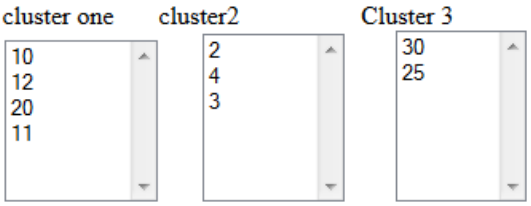
Third centroids= 30, 27.5

[3] - Add up all the numbers and then divide them by how many numbers there are [13]



Graph-4: centroids in three clusters

Graph 4 shows all centroids in each step, with the quantity column is selected as a target field, in this example, in two steps the centroids are found and the loop will stop.



Graph-5 all clusters.

7 DISCUSSION

This research is to find the clustering XML data using one of the most well-known algorithm which is K-means. First of all, after importing XML data into the application, then the application converts the XML data into traditional dataset including all atrributes and element children as shows in graph-1. Next, it can be clearn seen from graph-3 that if user choose two clusters, then the cetroids will be two lists of centers and two groups of data, the last centers of both lists is the centers of both clusters, it means all dataset items will be grouped according to the last centers from the list of centers. However, we can select three clusters, then the results will have three different centers, and the dataset will be divided into three clusters as shows in graph-4 and graph-5.

8 CONCLUSION

This application is a tool to mine XML data into groups by using K-means clustering algorithm according to centroids. We have used the Manhattan distance to measure the distance between two elements. This application has a Graphical User Interface to make the users understandable.in the first step. The centroids will be shown and then the data will be clustered according to centroids. In this paper, we tried to use our own tools to cluster XML. The future works of this research is that we are going to try to use fuzzy sets instead of CRISP K-means algorithm. Another future works of this research is using another algorithm to measure distance instead of Manhatan.

REFERENCES

[1] M. K. J. P. Jiawei Han, Data Mining: Concepts and Techniques, Waltham: Morgan Kaufmann, 2011 .

[2] R. Chattamvelli, Data Mining Methods, Oxford: Alpha Science International, 2009.

[3] "www.home.deib.polimi.it," [Online]. Available: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/.[Accessed 20 December 2013].

[4] M. Bramer, Principles of Data Mining, Portsmouth: Springer-Verlag London, 2013.

[5] M. Saraee and M. A. Joanna , "Mining XML data : a Clustering approach," in Internation conference on data mining, Las Vagas, USA, 2005.

[6] P. Antonellis, C. Makris and N. Tsirakis, "clustering homogeneous and heterogeneous XML documents using edge summaries," ACM, pp. 1081-1088 , 2008.

[7] T. Despeyroux, Y. Lechevallier, B. Trousse and A.-M. Vercoustre, "Experiments in Clustering Homogeneous XML

[8] R. M. Monika, "XML Data Mining: Different Techniques for Clustering," International Journal of Engineering Research & Technology (IJERT), vol. 2, no. 11, pp. 1546 - 1549, November - 2013.

[9] G. Q. T. a. D. U. Pasquale De Meo1, "An Approach for Clustering Semantically Heterogeneous XML Schemas," A. Doucet and H. Ahonen-Myka, "Naïve clustering of a large XML document collection,

[10] Doucet and H. AhonenMyka, "Naïve clustering of a large XML document collection, " http://citeseerx.ist.psu.edu, pp. 90 - 95, 2002.

[11] "http://axon.cs.byu.edu/Dan/478/Reading/ Clustering.pdf," [Online].

[12] F. Ya-qin and F. Wen-yong, "XML in Web Data Mining Application," in WASE International Conference on Information Engineering, Changchun, 2010.

[13] "www.mathsisfun.com," Rod Pierce DipCE BEng, [Online]. Available: http://www.mathsisfun.com/mean.html. [Accessed 2 January 2014].

[14] http://publib.boulder.ibm.com/, IBM, [Online]. Available: http://publib.boulder.ibm.com/infocenter/iseries/v5r4/index.jsp?topic=%2Frzamj%2Frzamjintroadvantages.htm. [Accessed 2 January 2014].

[15] L. Dykes and E. Tittel, XML For Dummies, John Wiley & Sons, 9 May 2011.

[16] "http://www.XMLnews.org," XML and news production, 1999. [Online]. Available: http://www.XMLnews.org/docs/XML-basics.html. [Accessed 3 January 2014].