Clustering Techniques for XML Data Using Data Mining

Chun-Sik Kim[*1)]

Abstract

Many studies have been conducted to classify documents, and to extract useful information from documents. However, most search engines have used a keyword based method. This method does not search and classify documents effectively. This paper identifies structures of XML document based on the fact that the XML document has a structural document using a set theory, which is suggested by Broder, and attempts a test for clustering XML document by applying a k-nearest neighbor algorithm. In addition, this study investigates the effectiveness of the clustering technique for large scaled data, compared to the existing *bitmap* method, by applying a test, which reveals a difference between the clause based documents instead of using a type of vector, in order to measure the similarity between the existing methods.

1. Introduction

Most Web pages have been constructed using HTML. HTML is a kind of language, which was developed to present the layout of a page. A Web page designer presents information using HTML, in which characters or positions of documents can be defined using this HTML.

HTML is a descendant of SGML. However, HTML is not used as a general method to define meta data differentiated from SGML. Although HTML uses meta data, HTML is only used to present the layout of a document. Thus, it has no function to structurally present a document. Therefore, a keyword becomes an important factor to search a document in the Web based on this reason. In other words, the importance of a document can be decided by the number of keywords, which are contained in a query, to search a document. The W3C was organized in July, 1996 to develop a specific standard. The objective of this organization is to build a markup language, which is a simplified version of SGML, in order to operate it on the Web. As a result, a XML 1.0 version was built by the recommendation of the W3C in Feb. 10, 1998.

In recent years, XML [1] based on a self-description property, which is a type of document description, was used as a standard of web documents. A user can search a document for exactly whatever he wants, if the user uses the structural property of XML.

Information and time are the most valuable resources in the 21st century. It's often said that we live in an age of the flood of information. Data mining is a type of technique implemented to select necessary and meaningful information among this flood of information. Data mining can be used to support a decision making under various situations.

It is expected that document classification can be effectively performed for the application of a document clustering algorithm of data mining to XML.

In addition, it seemed that measuring the

1) *Department of Digital Media Engineering, Anyang University

distance between documents using the structure of a document becomes a meaningful study.

XML uses a tag, which includes a certain meaning. Thus, a similar document uses a similar tag. This paper uses an algorithm proposed by Broder [2] in order to identify the similarity of a document. This algorithm presents an excellent result in a document classification test to measure the similarity between documents. This paper performed an experiment based on the fact that a classification of XML structure using the algorithm proposed by Broder is possible for effective clustering.

2. Related work

The study referenced in [3] presents various algorithms for clustering documents. A progressive hierarchical clustering algorithm is currently the most frequently used method. A linear time algorithm, such as K-means [4] algorithm, has the merit of online clustering. Based on the assumption that the order of words has a certain meaning, a clause can be usefully applied for clustering a lot of documents available on the Internet [5]. A bitmap index is usefully applied to optimize a query [6, 7, 8].

This paper applies a clustering method for the structure of XML. A clustering method, which generally used the content of document, is a kind of method used in the existing method because a HTML document mainly plays a role in the expression of the content of document, rather than that of the structure of a document. However, in the case of the XML document, a tag plays an important role in the classification of documents because it identifies the structure of a document. Thus, this paper proposes a clustering method, which compares the structure of a document using XML tags.

3. Preprocessing for document structure

A XML document clustering method is useful for many applications. This clustering method can applied to most applications that require hierarchical structure management, because XML is encoded with hierarchical data. In addition, most XML documents consist of data without any DTD. Thus, studies have been conducted to identify the automatic extraction of DTD, such as XTRACT [9], in order to compare document structures.

The extracted document structure becomes an important factor for an actual classification of documents. Although a document, which has the same document structure, has a different content, there is no doubt that a document, which has the same document structure, has the same interest on the document.

DOM and SAX (Simple API for XML) using methods can be used to extract a document structure by investigating XML [1].

```
<EMAIL>
<FROM>ADOIQKMS@nuwxmactz.com</FROM>
<TO>mipsan@msn.com</TO>
<SUBJECT>It is possible to detect bulk mail and enforce rules against it</SUBJECT>
<CONTENT>
It's true that it's easier to identify a single unsolicited message if you want to go after it, and that it takes more work to confirm a mailing was a bulk mailing. But it's far from impossible to do this, especially if the legal system is involved, as it would be in contract lawsuits or fraud or other anti-spam measures.
Read this essay on bulk detection for full details.
</CONTENT>
</EMAIL>
```

Fig 1. XML data (pmail.xml)

The SAX is supported by most XML processors, and is designed using a light structure of API, which doesn't generate an internal structure. If the SAX is used to extract a document structure, it will read the XML document as presented in Fig. 1, and extract the tag as presented in Fig. 2.

The extracted tag presented in Fig. 2 is a document structure. This tag has a meaning, which is different from HTML, and reveals

that a document, which contains a tag over a certain rate, has a similar document structure. Thus, it is possible to assume that the similar document structure presents the same content of the document. Thus, the clustering of XML using document structure presented in this paper is a requirement to classify or search documents. Therefore, this paper attempts to solve a problem, which exists in the large scaled clustering of a bitmap method, using the document comparison method proposed by Broder and k-nearest neighbor clustering method.

```
startElement : EMAIL
startElement : FROM
characters : ADOIQKMS@nuwxmactz.com
endElement : FROM
startElement : TO
characters : mipsan@msn.com
endElement : TO
startElement : SUBJECT
characters : It is possible to detect bulk mail and enforce rules against it
endElement : SUBJECT
startElement : CONTEN
characters : It's true that it's easier to identify a single unsolicited message if you want to go after it, and that it takes more work to confirm a mailing was a bulk mailing. But it's far from impossible to do this, especially if the legal system is involved, as it would be in contract lawsuits or fraud or other anti-spam measures.
Read this essay on bulk detection for full details.
endElement : CONTENT
endElement : EMAIL
```

Fig 2. Execution SAX parser

4. Document clustering

An automatic classification of a document means a process, which classifies documents using one of the two methods that were previously learned using a machine learning method. A statistical method and knowledge based method are used to perform these document classification methods.

A statistical method configures a classification category using the frequency of words, which is presented in a document. A knowledge based method performs the classification of documents according to the classification rule based on the content of documents, equal to what a human expert has done.

Although a statistical classification method presents a simple and fast means of classification, due to the fact that this method only uses the frequency of words without an analysis of the document's content, this method is limited in its accuracy because it doesn't analyze the content of a document.

In addition, although a knowledge based classification method presents a very high accuracy, due to the use of a specific classification rule, this method has the demerit that its lack of the classification rule increases the unclassification rate of documents.

A study related to this issue is a clustering method using various methods after mapping a mail using a vector. A k-means algorithm [10, 11] and k-nearest neighbor algorithm [12] were used to implement this method.

This paper attempts to calculate the distance between documents using the method proposed by Broder [2], and clustering XML documents using a k-nearest neighbor algorithm.

5. A design for XML document clustering

5.1 Measurement of the similarity of XML document

The similarity of a document compared to another document can be expressed as a mathematical concept.

The similarity $r(A, B)$ presents the degree of the similarity between the document A and the document B. This similarity is expressed as a range between 0 and 1, in which the closer value to 1 indicates an increased similarity [1].

$$d(A,B) = 1 - r(A,B) \cdots ①$$

As expressed in the equation, ① presents the distance between documents. When the document *A* existed, the distance between this document and other documents is to be calculated. As a result, the document *A* is to be contained in the nearest document group. It is necessary to introduce a concept, which measures the distance between documents, such as ①, in order to implement this process. The mathematical concept expressed as ① is required as an algorithm for clustering documents.

Based on the assumption that a document is *D*, a token exists in the document is expressed as *S(D,w)*. More than one neighbor characters included in *D* is called as shingle.

The character of w can be defined as the number of words included in *D*. The w-shingle means the number of w, which neighbors each other in a document [2].

The $S_w(A)$ is the entire w-shingle in the document *A*, and means a set. The similarity between the documents *A* and *B* can be defined as follows.

$$r_w(A,B) = \frac{|S(A,w) \cap S(B,w)|}{|(S(A,w) \cup S(B,w)|} \cdots ②$$

The number produced from ② existed between 0 and 1 recognizes that the two documents have a similar document structure.

The similarity between documents means the distance between documents. In addition, there is a useful algorithm, which identifies the characteristics of these documents, and classifies these documents into a proper document group for a given new document. This algorithm is *"Nearest Neighbor Clustering"*.

This paper applies a clustering algorithm, which is used for clustering with similar documents using XML documents.

For example, if a given document is expressed as D_A={A, B, C}, D_B={A, B, C, D, E}, the degree of similarity between D_A and D_B is 60% by following the equation ②. Thus, it is revealed that the two documents D_A and D_B are related documents.

Input :
° A threshold t on the nearest neighbor distance
° A set of data points $\{x_1, x_2, \ldots x_n\}$
Algorithm :
° Initialize assign set i=1, k=1 x_i to C_k
° Set i=i+1 Find nearest neighbor of xi among points already assigned to clusters
° Let the nearest neighbor be in cluster m
° If distance to the nearest neighbor is < t
- Assign x_i to m
- Else increment k and assign x_i to C_k
- If all points are assigned then stop

Fig 3. Nearest Neighbor Clustering

Fig. 3 presents an algorithm for clustering a XML document. The *K-NN* algorithm can be expressed as follows.

This algorithm finds the number of learning document *k* that is the most similar document for the number of clustering documents *n*. The closest document group is to be decided using the information of a learning document.

(1) Finds the number of upper neighbored document *k* , which has a high similarity, in the learning group using the equation ② for the document, which is to be clustered.

(2) Identifies the closest document group for the document, which is to be clustered, and assigns the document, which is to be clustered, to the closest document group.

(3) If the value of score, which is produced from the calculation, presents a lower value than the specific level, a new document group will be generated, even though the number of neighbor documents *k* is found in the learning group.

(4) The process is repeated until the document, which is to be clustered, fails to exist.

5.2 System design

Fig. 4 presents the overall system

configuration for clustering XML documents.

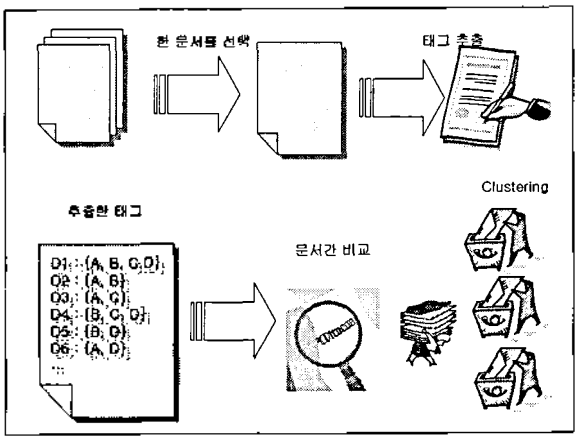

Fig 4. System structure

The configuration presented in Fig. 4 can be described step-by-step as follows.

First, a document is selected among XML documents.

Second, a tag using the SAX parser for the selected document is extracted.

Third, the extracted tag is recorded, and removed from the overlapped tag.

Fourth, the distance between documents is calculated using the equation ②.

Fifth, documents are classified using the *KNN* algorithm.

The data for the test applied in this paper generated using a program. The number of generated XML data was 1,000, and was used to perform a clustering test.

As shown in Fig. 5, the XML data structure generated various and different mixed XML data, and classified these data according to the data, whether it had the same structure or not. Table 1 presents the results of classification.

Because the precision was over 90%, useful data can be generated, if a mining method, which is used to search for the content of document, is applied to the result of the group proposed in this paper.

6. Conclusions and future research

We live in an age of the flood of information. This means that we live with an abundant amount data in our daily life. New more advanced technologies are required to survive such a flood of information.

Table 1. Test results for the precision and recall rate

XML data produced by the program	
precision	recall
91%	80%

```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <rss version="2.0">
- <channel>
<title>Yahoo! News: Science</title>
<copyright>Copyright (c) 2005 Yahoo! Inc. All rights reserved.</copyright>
<link>http://news.yahoo.com/news?tmpl=index&cid=753
</link>
<description>Science</description>
<language>en-us</language>
<lastBuildDate>Sun, 06 Feb 2005 03:59:56 GMT</lastBuildDate>
<ttl>5</ttl>
- <image>
<title>Yahoo! News</title>
<width>142</width>
<height>18</height>
<link>http://news.yahoo.com/</link>
<url>http://us.i1.yimg.com/us.yimg.com/i/us/nws/th/main_142b.gif</url>
</image>
- <item>
<title>Astronomers Find 'Hot' Vortex on Saturn (AP)</title>
<link>http://us.rd.yahoo.com/dailynews/rss/science/*http://story.news.yahoo.com/news?tmpl=story2&u=/ap/20050206/ap_on_sc/saturn_hot_spot
</link>
<guid isPermaLink="false">ap/20050206/
saturn_hot_spot</guid>
<pubDate>Sun, 06 Feb 2005 03:58:42 GMT</pubDate>
<description>AP - Astronomers using a giant telescope atop a volcano have discovered a hot spot at the tip of Saturn's south pole. The infrared images captured by the Keck I telescope at the W.M. Keck Observatory atop Mauna Kea on the Big Island suggest a warm polar vortex &amp;#151; a large-scale weather pattern likened to a jet stream on Earth that occurs in the upper atmosphere. It's the first such hot vortex ever discovered in the solar system.</description>
</item>
</channel>
</rss>
```

Fig 5. news for yahoo xml

One of these technologies is data mining. Data mining parallels work in the process of mining gold or diamonds, in which the mining requires a lot of work digging soils and rocks to find the real one.

A clustering method is one type of data mining. This clustering method is a process to cluster physical and abstract objects into a similar object group. In this process, a set of objects, which is similar, is called a

clustering. In the case of large scaled and complex data, the profile of an entire set of data can be configured by investigating a number of groups, which is a part of the entire data.

In recent years, XML data, which is a standard document of the Internet, has been largely produced. It is not easy to find an interest document among these documents. Thus, this paper proposed an algorithm to search these documents under the assumption that the required document has a specific classification rule, and tested this proposed algorithm.

The test revealed that the classification method using the structure of document becomes a useful classification standard for XML documents.

Future study must be continued to facilitate the development of the proposed algorithm and user interface for use with a commercial purpose.

References

[1] W3C, Extensible Markup Language(XML) 1.1, http://www.w3.org/
W3C Working Draft, April, 2002. (xml)

[2] A.Z. Broder. On the resemblance and containment of documents. In Compression and Complexity of Sequences (SEQUENCES'97), pages 21-29. IEEE Computer Society, 1997.

[3]P. Willet, Recent Trends in Hierarchical Document Clustering : a Critical Review, Information Processing and Management, 24:577-97, 1988.

[4] D. Hill, A Vector Clustering Technique, Mechanised Information Storage, Retrieval and Dissemination, North Holland, Amsterdam, 1968.

[5] J. Pei, J. Han, B. M. Asi, H. Pinto, "PrefixSpan : Mining Sequential Pattern Efficiently by Prefix-Projected Pattern Growth," Int, Conf, Data Engineering (ICDE), 2001.

[6] C. Chan and Y. Ioannidis, Bitmap Index Design and Evaluation, Proc. of Int'1 ACM SIGMOD Conference, 1998, 355-366

[7] P. O'Neil and S. Quass, Improved Query Performance with Variant Indexes, Proc. of Int'1 ACM SIGMOD Conference, 1997, 38-49.

[8] M. Wu, Query Optimization for Selections using Bitmaps, Proc. Int'1 ACM SIGMOD Conference, 1999, 227-238.

[9] Minos N. Garofalakis, Aristides Gionis, Rajeev Rastogi, S. Seshadri, and Kyuseok Shim. XTRACT: A System for Extracting Document Type Descriptors from XML Documents. In Proc. ACM SIGMOD, Dallas, Texas, USA, pages 165-176. ACM, 2000.

[10] G. manco, Em. Masciari, M. Ruffolo, and A. Tagarelli. Towards an adaptive mail classifier, 2002

[11] B.R. Segal and J.O. Kephart. Mailcat : an intelligent assistant for organizing email. In Proc. of the 3rd International Conference on Autonomous Agents, pages 276-282, 1999.

[12] T. Payne and P. Edwards. Interface agents that learn : an investigation of learning issues in a mail agent interface. Applied Artificial Intelligence, 11:1-32, 1997.