Semi Automatic Ontology Generation about XML Documents

Mi Sug Gu, Jeong Hee Hwang, Keun Ho Ryu, Doo Yeong Jung, Keum Woo Lee

Database Laboratory, Chungbuk National University, Korea
Cheongju Chungbuk, 361-763, Korea
{gumisug,jhhwang,khryu}@dblab.chungbuk.ac.kr, fiorgeo@trut.chungbuk.ac.kr, gmoo@etri.re.kr

Abstract: Recently XML (eXtensible Markup Language) is becoming the standard for exchanging the documents on the web. And as the amount of information is increasing because of the development of the technique in the Internet, semantic web is becoming to appear for more exact result of information retrieval than the existing one on the web.

Ontology which is the basis of the semantic web provides the basic knowledge system to express a particular knowledge. So it can show the exact result of the information retrieval. Ontology defines the particular concepts and the relationships between the concepts about specific domain and it has the hierarchy similar to the taxonomy.

In this paper, we propose the generation of semi-automatic ontology based on XML documents that are interesting to many researchers as the means of knowledge expression. To construct the ontology in a particular domain, we suggest the algorithm to determine the domain. So we determined that the domain of ontology is to extract the information of movie on the web. And we used the generalized association rules, one of data mining methods, to generate the ontology, using the tag and contents of XML documents. And XTM (XML Topic Maps), ISO Standard, is used to construct the ontology as an ontology language.

The advantage of this method is that because we construct the ontology based on the terms frequently used documents related in the domain, it is useful to query and retrieve the related domain.

Keywords: XML (eXtensible Markup Language), Semantic Web, Ontology, XTM (XML Topic Maps)

1. Introduction

In these days, the amount of the information is increasing because of the development of Internet technology. Semantic web is appearing for the exact result of information retrieval compared with the existing retrieval system on the web. Ontology which is the basis of the semantic web provides the basic knowledge system to express a particular knowledge. So it can show the exact result of the information retrieval to the user.

Ontology defines the particular information and the relationships between the information and it has the hierarchy similar to the taxonomy.

In this paper, we propose the generation of semi-automatic ontology based on XML documents that are interesting to many researchers as the means of knowledge expression. To construct the ontology, we determined that the domain of ontology is to extract the information of movie. So we transformed the information of movie on the web to XML documents and used them to generate ontology. XML documents are composed of the information, using tags such as director, actor, writer, country, time, genre, and so on.

There is a problem how we will determine the domain of ontology when we construct ontology using XML documents. Ontologies have shown their usefulness in application areas such as intelligent information integration, information brokering and natural-language processing. However, their wide-spread usage is still hindered by ontology engineering being rather time-consuming and expensive.

The existing ontology was usually constructed by the domain expert, but in this paper we propose the method which determines the domain of ontology automatically using the generalized association rules. It is one of data mining methods and constructs the ontology using the tag and contents of XML documents. Through generalized association rule, we can learn the relations and then use them to construct the ontology. The relations learned in this way are applied to XTM, the standard of ISO, to generate the ontology automatically.

For example, when the user poses the query; "Retrieve comedy among the movies that Brad Pitt has starred", ontology with the hierarchy is generated, finding the correlations of "Actor", "Genre", "Director", and so on.

The result of information retrieval system of the existing web using keyword has some drawbacks which just display the web page simply. But the ontology based on XML documents is the basis of Semantic web. And it is useful to query and retrieve the related domain, because the ontology is constructed based on the terms frequently used documents related in the domain.

2. Related Works

[1] explained the Generalized Association Rule which is extended Association rule. And in [3, 4] the algorithm is utilized to construct the ontology. This algorithm finds associations that occur between items, e.g. supermarket products, in a set of transactions. The basic association algorithm computes association rules $X_k \Rightarrow Y_k$ that meas-

This work was supported by ETRI in Korea.

ures for support and confidence that exceed user-defined thresholds, given a set of transactions $T:=\{t_i|i=1...n\}$. Support of a rule $X_k \Rightarrow Y_k$ is the percentage of transactions that contain $X_k \cup Y_k$ as a subset, and confidence for $X_k \Rightarrow Y_k$ is defined as the percentage of transactions that Y_k is seen when X_k appears in a transaction.

[2] is the specification of XTM (XML Topic Maps), the standard of ISO. XTM specifies the concepts of the subjects and defines the relationships between concepts. At first it was devised as the data model for electronic index. But at present, it is used as a knowledge map of the knowledge management system, contents map of the contents management system, and data model of Semantic Web ontology. Topic Map model is composed of three elements, Topic, Occurrence, and Association. Topic means subjects and association is the relationships between subjects. And occurrence plays a role to point to the information of the subject resources.

[5] introduces association rules for XML data based on XPath and XQuery. The operator proposed in this paper allows us to express complex mining tasks, compactly and intuitively.

[6] addresses the problem of deriving general association rules from XML data and proposes an approach to perform the task.

[7] shows that extracting association rules from XML documents without any preprocessing or post-processing using XQuery is possible and analyzes the XQuery implementation of the well-known Apriori algorithm.

3. Process of ontology generation

3.1 XML document

XML is becoming the standard of exchanging documents in the present web. Because HTML documents were easy for the users to use, they were used by many users. But they didn't contain the meta-data. So XML documents which can express the related meta-data systematically is used the means of the standard for web documents, public documentations, and so on. In this paper we propose the ontology generation method using XML documents which are expected to be used in many areas. Through the web sites providing the information of movie, we got the information about movie and then transformed it to XML documents.

We just applied data mining algorithm to XML documents using the tags and contents in them. We learned the relationships of them by joining the documents. And then we can construct the ontology proper to the users' query and show the result of the information retrieval to the user. See the XML documents about meta-data of the movie in the following Fig.1.

To generate the ontology, we need preprocessing for the filtering of the similar word. Usually there are many kinds of expressions that indicate the same meaning. That is, semantic-based filtering model using WordNet [http://www.cogsci.princeton.edu/~wn/wn2.0] filters the similar word into one concept of the tags and contents in XML documents.

```
<?xml version "1.0" encoding "UTF-16"?>
<!-- file name: movie.xml -->
<Movie>
        <Title>Legends of the Fall</Title>
        <Year>1994</Year>
        <Directed by>
                <Director>Edward Zwick</Director>
        </Directed by>
        <Genres>
                <Genre>Drama</Genre>
                <Genre>(more)</Genre>
        </Genres>
        <Country>America</Country>
        <Cast>
                <Actor>
                        <Male Actor>
                                <Name>BradPitt</Name>
                        </Male Actor>
                </Actor>
                <Actor>
                        <Male Actor>
                                <Name>Anthony Hopkins</Name>
                        </Male Actor>
                </Actor>
                <Actor>
                        <Female Actor>
                                <Name>Julia Ormond</Name>
                        </Female Actor>
                </Actor>
                </Cast>
</Movie>
```

Fig. 1. XML document of movie

In this paper we suggest the ontology generation method using XML documents which is used as a tool to express the knowledge in many areas, compared to the existing ontology based on the text.

In the next section, we will explain the algorithm used for data mining.

3.2 Learning Algorithm for ontology generation

In this paper we used the generalized association rule [3 4] to generate the ontology about XML documents semi-automatically. Generalized association rule algorithm does not only detect relations between concepts, but also determines the appropriate level of abstraction at which de define relations. It is crucial how many and what type of conceptual relationships should be modeled in a particular ontology, when we intend to construct ontology about particular subjects. Because it can reduce more time consumption and effort by the expert, we used the automatic ontology generation algorithm.

This algorithm determines associations at the right level of a taxonomy, by a taxonomic relation $H \subset C \times C$. It extends each transaction to include each ancestor of a particular item. And then it computes confidence and support for all possible association rules $X_k \Rightarrow Y_k$[3,4]. It prunes the result of the association rule after determining whether it is proper.

For the discovery of conceptual relations of XML documents, we build the schema in the following four steps.

1. Determine the transaction T; $T := \{t_i|i=1...n\}$.
2. Determine support for all association rules $X_k \Rightarrow Y_k$, where $|X_k|=|Y_k|=|1|$.
3. Determine confidence for all association rules $X_k \Rightarrow Y_k$ that exceed user-defined support in step 2.

4. Output association rules that exceed user defined confidence and are not pruned

Based on the common user query, we analyzed the information using the generalized association rule to find the related tags and contents pairs.

For example, if there is a user query; "Find out the comedy that Brad Pitt was the main character." The actor, Brad Pitt, is related to comedy, drama, action and so on. Using the tag in XML documents, the correlation of the concepts such as Brad Pitt and comedy, drama, or action is derived.

The algorithm determines the support and confidence to evaluate the correlation of three pairs and then finds out the abstract relation of the upper level such as genre and Brad Pitt. While the improper relations between tags and contents are pruned by the algorithm, the proper abstraction level which represents the relation of concepts is determined.

4. Example

The hierarchy of abstraction level is built by using the tags and contents in XML documents. However, when we intend to construct the ontology using this hierarchy, it's not the appropriate structure of ontology. So it is needed to learn the relations between tags and contents in XML documents fit to the user query by using data mining algorithm. And the relations without any correlations are deleted to make the appropriate hierarchy for ontology generation.

The following queries are common in web sites about movie.

1. Retrieve SF one among Stephen Spielberg's movies.
2. Which director has Brad Pitt usually made movie with?
3. Which male actor has made movie with Meg Ryan?
4. What's the male actor's name in the movie, Indiana Jones?

After learning the four common query sentences, we can find out the pairs of related terms and attributes in XML documents.

For example, given a set of 100 texts each describing a particular movie in detail. Each movie may have an elaborate description of the different types of genre, actor, director, and so on, resulting in 1.000 concept pairs returned from linguistic processing.

SF is one of genres of movie and Speilberg is one of the directors in number 1. And in the other example queries, we can find the relations of terms and attributes.

As a result, we can find out the correlation of the concepts below in table 1. Based on the correlation of the concepts, the hierarchy of the abstraction level is derived. After that, the algorithm finds out the appropriate relations of the concepts for generating ontology. And there is an abstraction level hierarchy scenario in the following figure 3.

Table 1. Related Pairs of Concepts

term	attribute	term	attribute
SF	Genre	Spielberg	Director
Director	Director	Brad Pitt	Actor
Male Actor	Actor	Meg Ryan	Female Actor
Male Actor	Actor	Indiana Jones	Title

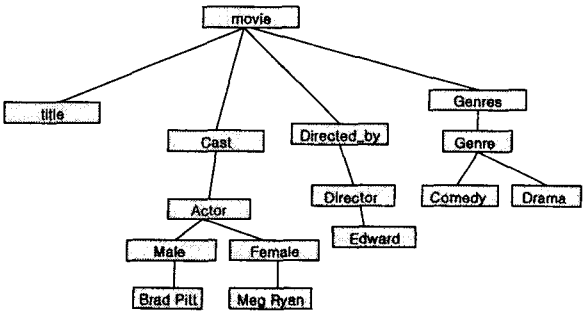

Fig. 2. Abstraction Scenario

Through the confidence and support, the pairs of concept relations with low percentage are pruned because they are not appropriate for ontology generation. In the following table 2, (Male Actor, Drama) and (Edward, Meg Ryan) are pruned, because (Actor, Genre) and (Director, Actor) are the ancestral relations of the above pairs respectively. And (Actor, Genre) and (Director, Actor) pairs have their confidence and support higher. The following table represents the deleted pairs of concept relations drawing the line on them.

Table 2. Discovered relation

Discovered relation	Confidence	Support
(Actor, Genre)	0.39	0.04
~~(Male actor, Drama)~~	~~0.1~~	~~0.03~~
(Director, Actor)	0.35	0.03
~~(Edward, Meg Ryan)~~	~~0.14~~	~~0.02~~
~~(Brad Pitt, Jolling)~~	~~0.20~~	~~0.03~~
(Actor, Writer)	0.32	0.02
(Writer, Director)	0.36	0.05

In this way, ontology is generated using the related concept pairs regarded as frequent.

And we used XTM to construct ontology. XTM specifies the concepts of the subjects and defines the relationships between concepts. Topic Map model is composed of three elements, Topic, Occurrence, and Association [2, 8].

The following figure 3 is the part of the source of XTM ontology. Generating the ontology about movie is started

from the root, movie as a tag in XML document. Topics in Topic Maps are "Legends of the Fall" as the title of movie, "Director", "Actor", "Time", "Country", and so on. And association is "movie" as a root, and "Legends of the Fall" is the sub relation to "movie". And "Actor", "Director", "Time", "Country" are the sub relation of "Legends of the Fall". Using these structures of XTM Topic Maps, ontology which represents the information of movie with XML documents is generated.

```
<!-- movie -->
    <topic id="movie">
        <baseName>
            <baseNameString>Movie</baseNameString>
        </baseName>
    </topic>
<!-- movie_Legends of the Fall -->
    <topic id="Legends of the Fall">
        <instanceOf>
            <topicRef xlink:href="#movie"/>
        </instanceOf>
        <baseName>
            <baseNameString>Legends of the Fall</baseNameString>
        </baseName>
    </topic>
    <topic id="Director">
        <instanceOf>
            <topicRef xlink:href="#Legends of the Fall"/>
        </instanceOf>
        <baseName>
            <baseNameString>Director</baseNameString>
        </baseName>
        <occurrence>
            <resourceRef xlink:href="http://megalo.wo.to"/>
        </occurrence>
    </topic>
```

Fig. 3. XTM ontology

Using the ontology generated in this way, we show the exact result of the information retrieval to the previous user's query; "Retrieve comedy among the movies that Brad Pitt has starred ".

When we compare the result of information retrieval using ontology of XML document with that of the existing method on the web, there are several advantages. As the information retrieval of the existing web usually uses the key word information retrieval system, the result is just the display of the related web pages. So sometimes the user can be confused because of so many results of information that he wants go get.

But our system used the hierarchy of ontology, so we can get the exact result. Ontology describes the knowledge system using the hierarchy of taxonomy. In our proposed method, we can get the information about movie exactly that the user wants to get. And because we used XML documents which are the standard of exchanging of information and express the meta-data of the related information, we can retrieve the meta-data that explain the related data specifically compared with HTML documents.

In this way the semantic web can be implemented to help that the user can get the exact information.

Because of the ontology based on the terms frequently used documents related in the domain, it is useful to query and retrieve the related domain.

5. Conclusions

In this paper we proposed semi-automatic ontology generation method using data mining algorithm. To construct the ontology, we transformed the information of movie into XML documents. The existing web expressed in HTML has some disadvantages such that it can't find out the meta-data because of the simplicity. But at present XML documents are used to represent the knowledge in many areas. So in this paper we learned the correlations to construct the ontology using XML documents and generalized association rule. And based on this, we show the exact result of information retrieval to the user query after constructing the ontology of movie using XTM Topic Maps.

The advantage of this method is that because we construct the ontology based on the terms frequently used documents related in the domain, it is useful to query and retrieve the related domain.

In the future, we plan to extend this method to research the way of efficient retrieval using association rule. And we are going to apply many kinds of other data mining rule to construct the ontology.

References

[1] R.Srikant,R.Agrawal, 1995, "Mining Generalized Association Rules", VLDB

[2] S.Pepper,B.Moore, "XML Topic Maps(XTM) 1.0", TopicMaps.Org

[3] A.Maedche,S.Staab, 2001, "Discovering Conceptual Relations from Text", Institute AIFB, Karlsruhe University, Germany

[4] A.Maedche,S.Staab, 2000, "Semi-Automatic Engineering of Ontologies from Text", Institute AIFB, Karlsruhe University, Germany

[5] D.Braga,A.Campi,S.Ceri,M.Klemettinen,P.Lanzi, 2003, "Discovering interesting information in XML data with association rules", SAC

[6] Q.Ding,K.Ricords,J.Lumpkin, 2003, "Deriving General Association Rules from XML Data", SNPD

[7] Jacky W.W.Wan,G.Dobbie, "Mining Association Rules from XML Data using XQuery", 2004, ACM International Conference Proceeding

[8] Steve Pepper, 2000, "The TAO of Topic Maps", XML 2000 Conference & Exposition