Towards a Fuzzy Approach for Mining XML Mediator Schemas

Anne Laurent*, Pascal Poncelet**, Maguelonne Teisseire*

*LIRMM, Université Montpellier 2
161, rue Ada – 34392 Montpellier Cedex 5 - France

**LGI2P, EMA
Site EERIE – Parc scientifique G. Besse – 30035 Nîmes Cedex 1 – France

{laurent, teisseire}@lirmm.fr, Pascal.Poncelet@ema.fr

Simple Abstract

As highlighted by the World Wide Web Consortium, XML has been proposed to deal with huge volumes of electronic documents and is playing an increasing important role in the exchange of a wide variety of data on the Web. However, when dealing with such large and heterogeneous data sources, it is necessary to have an idea on the way these data sources are structured. This information is indeed essential in order to build mediator schemas. These mediator schemas are required to query data in a uniform way. Moreover, this information is interesting since it provides users with a semantic structure of the data they can query. Recently Schema Mining approaches have been proposed to extract in an efficient way the commonly occurring schemas that appear in a collection. Nevertheless, according to the semantic point of view, such approaches suffer from different drawbacks. In this work, we propose thus a fuzzy approach, showing why and how fuzziness is useful in order to extract frequent approximate schemas.

Key Words: *fuzzy logic, semi-structured data, mediator schemas, semantic structures, XML, frequent patterns.*

Introduction

XML has become crucial for the representation of the information on the Web and for data exchanges. Large amounts of data are available in XML format, and even if large volumes of "legacy" data are still marked up in HTML, efficient approaches have been proposed to transform HTML documents into XML documents. These XML documents, stored in many sources distributed over the Web, contain useful information. When accessing such a database, users have to be provided with a mediator schema which is a shared structure through which queries can be defined. However, no automatic tool is available to extract semantic knowledge from these large amounts of distributed and heterogeneous data. In this work, we consider the problem of extracting XML mediator schemas from a database perspective. Database concepts can indeed help mining mediator schemas. We show that Schema Mining approaches have some drawbacks and that a fuzzy approach is very useful to mine approximate schemas. Moreover, mined mediator schemas provide an interesting source of information about the data available on the web since these schemas can be seen as semantic structures describing the information.
Our approach is based on the definition of the tree inclusion. We define several ways to introduce fuzzy logic in this problem. The main idea is to propose a definition of soft inclusion, meaning that a tree is no more *included or not* in another one, but *gradually included* within it. A degree of inclusion is defined, depending on the way the fuzzy inclusion is considered.
Finally, we introduce fuzzy frequent patterns which aim at representing the *strength* of the links from a frequent tree.

Schema Mining Principles

As presented previously, mining frequent subtrees from a database of trees is of great interest in order to mine mediator schemas. A subtree is said to be *frequent* if it occurs more than a user-defined number of times in the database. Existing approaches to automatically mine frequent subtrees are mainly based on *levelwise* algorithms.
In the framework of semantic web in general, and XML mediator schema mining in particular, several possibilities are considered, depending on the way trees are considered and mined. Trees can indeed be considered as being ordered (*i.e.* the children of every node are ordered) or unordered.
Moreover, existing approaches can be divided depending on the way they consider ancestor relationships. Some approaches deal indeed only with parent relationships while some other ones consider ancestor-descendant relationships.
However, all the existing approaches consider crisp inclusion when mining frequent subtrees. We propose thus to deal with fuzzy inclusion in order to evaluate to which extent a subtree is embedded within a tree.

Why and How Considering Fuzzy Approaches

In order to avoid the crisp inclusion problem, we propose to fuzzify this definition in order to better describe the data available on the Web. It is indeed very interesting to mine approximate schemas in order to have a better idea of the semantic structures we are provided with.
We propose thus (i) to define the notion of fuzzy tree inclusion and (ii) to mine fuzzy frequent patterns.

Fuzzy Tree Inclusion

We define four ways to consider fuzzy inclusion of a tree within another one. These definitions aim at describing the extent to which a tree is embedded within another one. While crisp approaches deal with crisp inclusion, meaning that a tree *is* or *is not* embedded within another one, we propose to use a degree, defined between 0 and 1. The four ways this degree can be obtained are described below.

1. Considering fuzzy indirect links within trees. Let us have a closer look on embedded and induced subtree. More formally, given the trees *S* and *T*, if *S* and *T* are isomorphic then *S* is called an *induced* subtree of *T*, *i.e.* the parent relationship is preserved. S is an *embedded* subtree of *T* if the ancestor-descendant relationship is preserved. According to the fuzzy view point, it is clear that an approximate relationship allows us to consider these two relationships in a similar way. For instance, example 1 from Figure 1 shows a tree S which is embedded within a tree T. We consider fuzzy membership functions describing the ancestor-descendant relationship depending on the number of nodes separating the two nodes being considered.

2. Considering fuzzy level inclusion. When considering ordered trees in the crisp approaches, a subtree S is embedded within a tree T only if all nodes of S can be mapped to nodes of T in the same order. In our approach, we propose to soften this definition by considering the proportion of nodes being included and well-ordered. For instance, example 2 from Figure 1 shows an ordered tree S which is not embedded within an ordered tree T in the classical approaches since one of the node is misordered. We consider S as being embedded within T with a certain degree since the other nodes satisfy the inclusion.

3. Considering partial node inclusion. In a general way, all the nodes of a subtree S must be present in a tree T if S is embedded within T. In our approach, we propose to define partial inclusion by considering the proportion of nodes of S being present in T. For instance, example 3 from Figure 1 shows a tree S having 75% of its nodes embedded within T.

4. Considering fuzzy similarities. Let us now consider that we are provided with knowledge on the data. One of drawbacks of schema mining approaches is that the inclusion detection is only performed on nodes having same labels. According to the Semantic Web point of view, this restriction is usefulness since two different labels could describe similar concepts. By using fuzzy approaches we can overcome this drawback. For instance, example 3 from Figure 1 shows a tree S that should be matched with T if we know that the concept D is close to B.

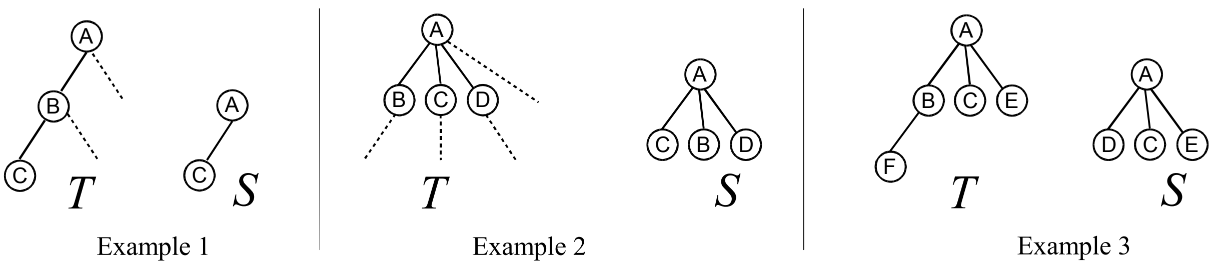

FIGURE 1. Fuzzy Inclusion

Fuzzy Frequent Subtrees

Classical approaches mine frequent trees that do not provide much information about the occurrences whithin the database except the fact that these subtrees are embedded in a sufficient number of trees from the database (depending on the user-defined minimum support value). We propose thus to extend the knowledge on the mined semantic structures by providing fuzzy links within the frequent trees. These fuzzy links help knowing whether they are *very shared, middle shared* or *a little shared*, as illustrated on Figure 2.

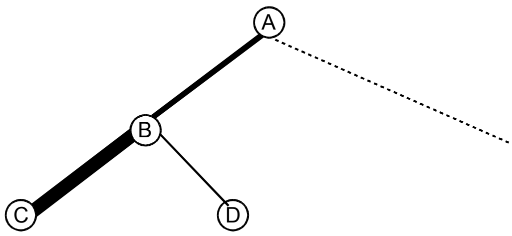

FIGURE 2. Fuzzy Frequent Subtree

Conclusion

Data Mining is of great interest for Semantic Web. Nevertheless, existing approaches do not handle the problem of approximate pattern mining. Fuzzy Data Mining is thus a promising approach. In this framework, mining fuzzy XML mediator schemas is crucial (i) in order to be automatically provided with a mediator schema to query data and (ii) in order to get semantic structures from the huge amounts of distributed and heterogeneous data.

The approach we propose here can be used in order to mine frequent web structures from web sites. Moreover, mediator schemas, fuzzy or not, can be used in the framework of fuzzy queries over web data.