A Semantic Enrichment of Data Tables Applied to Food Risk Assessment

Hélène Gagliardi, Ollivier Haemmerlé, Nathalie Pernelle, and Fatiha Saïs

LRI (UMR CNRS 8623 - Université Paris-Sud) / INRIA (Futurs),
Bâtiment 490, F-91405 Orsay Cedex, France
{gag, ollivier, pernelle, sais}@lri.fr

1 Introduction

Our work deals with the automatic construction of domain specific data warehouses. Our application domain concerns microbiological risks in food products. The MIEL++ system [2], implemented during the Sym'Previus project, is a tool based on a database containing experimental and industrial results about the behavior of pathogenic germs in food products. This database is incomplete by nature since the number of possible experiments is potentially infinite. Our work, developed within the e.dot project[1], presents a way of palliating that incompleteness by complementing the database with data automatically extracted from the Web. We propose to query these data through a mediated architecture based on a domain ontology. So, we need to make them compatible with the ontology. In the e.dot project [5], we exclusively focus on documents in Html or Pdf format which contain data tables. Data tables are very common presentation scheme to describe synthetic data in scientific articles. These tables are semantically enriched and we want this enrichment to be as automatic and flexible as possible. Thus, we have defined a Document Type Definition named SML (Semantic Markup Language) which can deal with additional or incomplete information in a semantic relation, ambiguities or possible interpretation errors. In this paper, we present this semantic enrichment step.

2 An Automatic Approach to Enrich Tables Semantically

The data tables which are extracted from the Web are first represented in an XML format using purely syntactic tags : rows and cells. Besides, when it is possible, titles are extracted. We have then to express these data using the vocabulary stored in the ontology. The Sym'Previus ontology contains a taxonomy of 428 terms and a relational schema which describes 25 semantic relations of the domain. In a SML document rows are not represented by cells anymore but by a set of semantic relations between columns.

The semantic enrichment of tables is done in two steps: the first step consists in identifying the semantic relations appearing in the data table. The second step consists in instantiating semantic relations discovered in the table.

[1] Cooperation between INRIA, Paris South University, INRA and Xyleme.

A. Hoffmann, H. Motoda, and T. Scheffer (Eds.): DS 2005, LNAI 3735, pp. 374–376, 2005.

In order to extract semantic the relations of the table, we first identify the *A-terms*[2] which represent each table column. We look for an A-term which subsumes most of the values. [4] and [6] showed that such techniques give good results when one searches for schema mappings for relational data bases or XML. If the values do not help, we exploit the title of the column. If no A-Term has been found, we associate a generic A-term named *attribute* with the column. Thus we obtain a schema for the table. The schema *tabSch* of the table *Table. 1* is: {*(1,food) (2,attribute) (3,lipid),(4,calorie)*}.

```
<table> <table-title>Nutritional Composition of some food products </table-title >
<column-title> Product </column-title> ... <content> <rowRel additionalAttr=”yes”>
<foodLipid relType=”completeRel”><food attrType=”Normal”>
<ontoVal indMap=”intersection”> whiting Provencale</ontoVal>
<ontoVal indMap=”intersection”> green lemon </ontoVal>
<ontoVal indMap=”intersection”> whiting fillets </ontoVal>
<originalVal> whiting with lemon </originalVal></food>
<lipid attrType=”Normal”> <ontoVal indMap=”notFound”/>
<originalVal> 7.8 g</originalVal> </lipid>
<attribute indMap=”notFound” attrType=”generic”> <ontoVal/>
<originalVal> 100 g</originalVal></attribute> </foodLipid>

<foodAmountLipid relType=”partialNull”> ... <amount attrType=”null”>...
</amount></foodAmountLipid> </rowRel> ... </content> </table>
```

Fig. 1. *SML Representation of the nutritional composition of food products*

Table 1. *Nutritional Composition of some food products*

Products	Qty	Lipids	Calories
whiting with lemon	100 g	7.8 g	92 kcal
ground crab	150 g	11.25 g	192 kcal
chicken	250 g	18.75 g	312 kcal

Then we propose an automatic identification of the semantic relations as flexible as possible. A relation is ***completely represented (CR)*** if each attribute of its signature subsumes or is equal to a distinct A-term of the table schema. A relation is ***partially represented (PR)*** if it is not completely represented and if at least two attributes of its signature subsume or are equal to a distinct A-term of the table schema. In such cases one of the missing attributes may correspond to a constant value which appears in the title of the table. The missing attributes are represented in the SML document by means of an empty tag or by a constant. For example the relation *foodAmountLipid* shown in figure 1 is a **partially represented relation**, where the attribute *amount* is represented by an empty tag. When no relation has been found, a generic relation is generated in order to keep semantic links between values. Fig.1 shows a part of the SML document which is automatically generated from the table shown in Table. 1.

[2] An A-term is a term of the taxonomy that appears at least once as an attribute of a relation signature in the relational schema of the ontology.

Once the relations are extracted, we instantiate them by the values contained in the table. Besides, terms of the ontology are associated with each value when it is possible. The SML formalism allows us to associate several terms that can be found by different mappings procedures. The first one uses simple syntactic criteria. The second one is the unsupervised approach PANKOW [3].

The SML representation of a relation is composed of the set of attributes that appear in the signature of the relation described in the ontology (e.g. *foodLipid*(*food*, *lipid*)). A set of terms represented inside the XML tag *ontoVal* is associated with each value. Thus, three different terms are proposed for *whiting with lemon* : *whiting Provençale*, *green lemon* and *whiting fillets*. The original value is kept inside the XML tag *originalVal* and this value can be shown to the user.

In order to evaluate our approach, we have collected 50 tables from the Web and we have compared the recall, the precision and the F-measure for the different kinds of semantic relations. This result shows that the recall significatively increases when partially identified relations are kept (recall(CR)=0.37 and recall(CR&PR)=0.60) and that the precision do not fall much (0.61 to 0.56).

3 Conclusion

Our method allows one to enrich semantically tables extracted from heterogenous documents found on the Web. The semantic enrichment is completely automatic and is guided by an ontology of the domain. Thus, that processing cannot lead to a perfect and complete enrichment. The SML representation we propose keeps all the possible interpretations, incompletely identified relations and original elements of the context. Contrarily to previous approaches like [1], we cannot base the search for information on a common structure discovered among a set of homogeneous documents.

References

[1] Arvind Arasu and Hector Garcia-Molina, *Extracting structured data from web pages*, Proceedings of the 2003 ACM SIGMOD international conference on Management of data, ACM Press, 2003, pp. 337–348.

[2] Patrice Buche, Juliette Dibie-Barthélemy, Ollivier Haemmerlé, and Mounir Houhou, *Towards flexible querying of xml imprecise data in a dataware house opened on the web*, Flexible Query Answering Systems (FQAS), Springer Verlag, june 2004.

[3] Philipp Cimiano, Siegfried Handschuh, and Steffen Staab, *Towards the self-annotating web*, WWW '04: Proceedings of the 13th international conference on World Wide Web, ACM Press, 2004, pp. 462–471.

[4] AnHai Doan, Ying Lu, Yoonkyong Lee, and Jiawei Han, *Profile-based object matching for information integration*, Intelligent Systems, IEEE **18** (2003), no. 5, 54– 59.

[5] e.dot, *Progress report of the e.dot project*, http://www-rocq.inria.fr/gemo/edot, 2004.

[6] Erhard Rahm and Philip A. Bernstein, *A survey of approaches to automatic schema matching*, The VLDB Journal **10** (2001), no. 4, 334–350.