Semantic annotation of Web data applied to risk in food

G. Hignette\textsuperscript{1,2}, P. Buche\textsuperscript{1}, O. Couvert\textsuperscript{4}, J. Dibie-Barthélemy\textsuperscript{1,2}, D. Doussot\textsuperscript{1,2}, O. Haemmerlé\textsuperscript{3}, E. Mettler\textsuperscript{5} and L. Soler\textsuperscript{1}

\textsuperscript{1}INRA MIA, Unité Mé@risk UR 1204, 16, rue Claude Bernard, 75 231 Paris Cedex 05, France (hignette@inapg.fr, buche@inapg.fr, dibie@inapg.fr, doussot@inapg.fr, lsoler@inapg.fr)
\textsuperscript{2}AgroParisTech, UFR Informatique, 16, rue Claude Bernard, 75 231 Paris Cedex 05, France
\textsuperscript{3}Département de Mathématiques-Informatique, UFR Sciences, Espaces et Sociétés, Université Toulouse le Mirail, 5, Allée Antonio Machado, F-31058 Toulouse Cedex 1, France (olivier.haemmerle@univ-tlse2.fr)
\textsuperscript{4}ADRIA Développement, Creac’h Gwen, 29196 Quimper Cedex, France (olivier.couvert@adria.tm.fr)
\textsuperscript{5}Soredab (Groupe SOPARIND BONGRAIN), La Tremblaye, 78125 La Boissière-Ecole, France (eric.mettler@soredab.org)

Abstract

A preliminary step to risk in food assessment is the gathering of experimental data. In the framework of the Sym’Previus project (http://www.symprevius.org), we have designed a complete data integration system, composed of data provided by industrial partners and data extracted from papers published in the main scientific journals of the domain. Those data have been classified by means of a predefined vocabulary, called ontology. Our aim is to complement the database by searching data on the Web. We have designed a semi-automatic acquisition tool, called AQWEB, which retrieves scientific documents from the Web. AQWEB only extracts data tables, which contain, in general, a synthesis of data published in the documents. In this paper, we explain how the columns of the data tables are automatically annotated with data types of the ontology. We also give the results of our experimentation to assess the quality of such an annotation.

Keywords

Information extraction, database, ontology

Introduction

A preliminary step to risk in food assessment is the gathering of experimental data: Tamplin \textit{et al.} (2003), Baranyi and Tamplin (2004), Le Marc \textit{et al.} (2005). In the framework of the Sym’Previus project (see Couvert \textit{et al.} (2007) and http://www.symprevius.org), Buche \textit{et al.} (2005) have designed a complete data integration system composed of data provided by industrial partners and data extracted from papers published in the main scientific journals of the domain. Those data have been classified by means of a predefined vocabulary, called ontology. This ontology is composed of data types meaningful in the field of risk in food. Data types are described in the ontology in two different ways depending on whether their associated values are symbolic (Food product, Microorganism …) or numeric (Temperature, Time …). Symbolic types are described by taxonomies of possible values (for example, a taxonomy of microorganisms); those possible values are called terms. Numeric types are described by the set of units in which they can be expressed (for example, °C or °F for Temperature, but no unit for pH or a\textsubscript{w}), and eventually their numerical range (for example, [0,14] for pH).

Our data integration system has been designed in order to take into account an important characteristic of the data, their incompleteness. Data are relatively rare due to confidentiality and acquisition cost. We propose two solutions to deal with that problem. The first solution relies on an extended querying system, called MIEL, which allows the user to retrieve the nearest data stored in the database corresponding to his/her selection criteria: the ontology is used in order to assess which data can be considered as “near” to the user’s selection criteria. The second solution, which is under construction in the framework of the WebContent project (http://www.webcontent.fr), is detailed in this paper. It consists in searching data on the Web to complement the database. We have designed a semi-automatic acquisition tool, called AQWEB, which retrieves scientific documents from the Web. AQWEB only extracts data
tables, which contain, in general, a synthesis of data published in the documents. In this paper, we explain how the columns of the data tables are automatically annotated with data types of the ontology. To find the type of a column, we use both the title of the column and its content. In a first step, we distinguish the columns with numeric data from those with symbolic data. To find the types of symbolic columns, the data of the column are annotated with terms from the taxonomies of symbolic types in the ontology. To find the types of numeric columns, we use the numeric values and units available in the column which are compared with the description of numeric types in the ontology. Once the tables are correctly annotated, they can be queried using the ontology in the same way as the existing database in MIEL presented in Buche et al. (2006).

Materials and methods
Our annotation algorithm is divided in three parts. First, we distinguish between columns containing numeric data and columns containing symbolic data. Then we annotate the symbolic columns, and then the numeric columns.

Distinction between numeric and symbolic columns
The distinction between numeric and symbolic columns is not as simple as it seems: symbolic columns may contain numbers (for example, the strain of a microorganism) and numeric columns often contain character strings such as units, etc. We thus propose a method that uses the units defined in the ontology to classify the columns.

Let \(\text{col}\) be a column of the table we want to annotate. We search \(\text{col}\) for all occurrences of numbers (in decimal or scientific format) and of units of numeric types described in the ontology. We also search \(\text{col}\) for all words, which are defined as alphabetic character sequences that are neither units nor “no result indicators”.

Let \(c\) be a cell of the column \(\text{col}\). We apply the following classification rules:

- if \(c\) contains a number immediately followed by a unit, or a number in scientific format, then \(c\) is numeric;
- else, if \(c\) contains more numbers and units than words, then \(c\) is numeric;
- else, if \(c\) contains more words than numbers and units, then \(c\) is symbolic;
- else (equal amount of words and numbers+units) the status of \(c\) is considered as unknown.

Once all cells of the column \(\text{col}\) have been classified using the above rules, \(\text{col}\) is classified as symbolic if there are more cells classified as symbolic than numeric. Else, the column is classified as numeric.

Symbolic column annotation
Once a column has been recognised as symbolic, we annotate each cell in the column with the terms from the taxonomies of each symbolic type in the ontology. For that, we use a similarity measure between a term from the web (found in the cell of a symbolic column), and a term from the ontology. All terms are transformed into weighted vectors: the coordinate axis of the vectors represent all possible words, the coordinate values represent the weight of those words in the term. Table 1 presents an example of such a vector representation of terms. For terms from the ontology, each word is manually weighted according to its importance in the meaning of the term. A weight of 1 means that the word is essential to the meaning of the term, a weight of 0.2 means the word is secondary to the meaning of the term. For terms from the web, each word has a weight of 1, as the meaning of the term is not known a priori.

Table 1: Terms represented as weighted vectors.

<table>
<thead>
<tr>
<th>Term from the web</th>
<th>Meaning of the vector axis</th>
<th>ground</th>
<th>meat</th>
<th>fresh</th>
<th>beef</th>
</tr>
</thead>
<tbody>
<tr>
<td>ground meat</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Term of the ontology</td>
<td>fresh meat</td>
<td>0</td>
<td>1</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>Term of the ontology</td>
<td>ground beef</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The similarity between a term from the web and a term from the ontology is computed as the cosine similarity measure between the two weighted vectors. Let \(w\) be a term from the web,
represented as the weighted vector \( w = (w_1, \ldots, w_n) \) and \( o \) a term from the ontology, represented as the weighted vector \( o = (o_1, \ldots, o_n) \). The similarity between \( w \) and \( o \) is computed as:

\[
sim(w, o) = \frac{\sum_{k=1}^{n} w_k \times o_k}{\sqrt{\sum_{k=1}^{n} w_k^2 \times \sum_{k=1}^{n} o_k^2}}
\]

For example, using the terms given in Table 1, we compute the following similarities:

\[
sim(\text{ground meat}, \text{fresh meat}) = \frac{\frac{1}{\sqrt{1^2 + 1^2 + 1^2}}}{\sqrt{1^2 + 1^2}} \approx 0.57
\]

\[
sim(\text{ground meat}, \text{ground beef}) = \frac{\frac{0.2}{\sqrt{1^2 + 1^2 + 0.2^2}}}{\sqrt{1^2 + 1^2}} \approx 0.11
\]

For each cell in the symbolic column, we compute the similarity measure with each term from the taxonomies of symbolic types of the ontology. We compute the sum of such similarities for each symbolic type, for each cell. A cell is considered as having the type which has the best sum of similarities. If several types have the same highest sum of similarities for a cell, this cell is considered as of unknown type.

When each cell of the column is assigned a type, we compute the score of a symbolic type \( \text{type} \) for the column \( \text{col} \) according to the column contents, noted \( \text{score}_{\text{content}}(\text{type}, \text{col}) \), as the proportion of cells in that column that were considered as having this type.

We also compute the score of a symbolic type \( \text{type} \) for the column \( \text{col} \) according to the column title, noted \( \text{score}_{\text{title}}(\text{type}, \text{col}) \), as the cosine similarity measure between the column title and the type name.

Then the final score of a symbolic type \( \text{type} \) for the column \( \text{col} \) is computed as follows:

\[
\text{score}_{\text{final}}(\text{type}, \text{col}) = 1 - (1 - \text{score}_{\text{content}}(\text{type}, \text{col})(1 - \text{score}_{\text{title}}(\text{type}, \text{col}))
\]

The type of the column is then the type that has the best final score for this column. If there are several types that have the same highest score, then the column is considered as of unknown type.

**Numeric column annotation**

When a column has been recognised as numeric, we look at all the units that are presented in this column. Let \( \text{num} \) be a function that associates to a unit \( u \) the number \( \text{num}(u) \) of numeric types in the ontology that can be expressed with this unit. Let \( \text{units} \) be a function that associates to a numeric type \( \text{type} \) and a column \( \text{col} \) the set \( \text{units}(\text{type}, \text{col}) \) of all units that are present in the column \( \text{col} \) and that can be used to represent data of the type \( \text{type} \). Then the score of the numeric type \( \text{type} \) for the column \( \text{col} \) according to the contents of the column is:

\[
\text{score}_{\text{content}}(\text{type}, \text{col}) = \max_{u \in \text{units}(\text{type}, \text{col})} \frac{1}{\text{num}(u)}
\]

We also compute the score of a numeric type \( \text{type} \) for the column \( \text{col} \) according to the column title, noted \( \text{score}_{\text{title}}(\text{type}, \text{col}) \), as the cosine similarity measure between the column title and the type name.

Then the final score of a numeric type \( \text{type} \) for the column \( \text{col} \) is computed as follows:

\[
\begin{align*}
\text{score}_{\text{final}}(\text{type}, \text{col}) &= 0, \text{ if the numeric contents of the column are not compatible with the range defined in the ontology for the numeric type} \\
\text{else } \text{score}_{\text{final}}(\text{type}, \text{col}) &= 1 - (1 - \text{score}_{\text{content}}(\text{type}, \text{col})(1 - \text{score}_{\text{title}}(\text{type}, \text{col}))
\end{align*}
\]

The type of the column is then the type that has the best final score for this column. If there are several types that have the same highest score, then the column is considered as of unknown type.

**Experimental approach**

Our annotation algorithm was tested on 60 tables extracted from publications on food microbiology. The tables were manually annotated to give a type to each of the 349 columns: they were first separated between numeric and symbolic, then the symbolic columns were annotated with the types Microorganism, Food Product, Response or “other” if the column contained other precisions that did not match any of the symbolic types of our ontology. The numeric columns were annotated with one of the 18 numeric types of our ontology (all
numeric data could be manually classified as belonging to one of the types in the ontology. Then we ran our annotation algorithm, and compared the computed column types with the ones that had been manually chosen.

**Results and discussion**

Annotation quality is assessed using two measures: precision and recall. Precision is the ratio of correct computed annotations over the total number of computed annotations. Recall is the ratio of correct computed annotations over the number of manual annotations. Our annotation algorithm achieved 98% precision and recall for the distinction between numeric and symbolic columns. Then the annotation of symbolic columns gave a 93% precision and a 66% recall: this is due to the fact that the column is considered as unknown whenever there is a doubt on its type. The annotation of numeric columns gives better results, with 96% precision and 93% recall, which is mainly due to a lesser extent of variations in column titles (for example, Temperature is always called Temperature) and the use of some very indicative units (for example, cfu will only denote a microorganism concentration). Such annotation results can be considered as very good as they are obtained via a fully-automatic method.

**Conclusion**

We have presented an annotation algorithm that allows finding the types of the columns of data tables according to an ontology. Our aim is to find the semantic relations that are represented in a table (for example, a study of the growth rate of one specific microorganism in one specific food product at different temperatures). Those relations have to be annotated with our ontology in order to give responses to a user, when he queries the MIEL system, not only coming from the relational database that was manually fed, but also coming from tables from the web that were automatically annotated, thus reducing the problem of data rarity.

**Acknowledgements**

Financial support from the French National Research Agency (ANR) in the framework of the National Network for Software Technology (RNTL) is gratefully acknowledged.

**References**


