Food Microbiology 28 (2011) 685–693

Contents lists available at ScienceDirect

Food Microbiology

journal homepage: www.elsevier.com/locate/fm

Flexible querying of Web data to simulate bacterial growth in food

Patrice Buche [a,g,h,*], Olivier Couvert [b,c,g], Juliette Dibie-Barthélemy [d,e], Gaëlle Hignette [d,e], Eric Mettler [f,g], Lydie Soler [d]

[a] INRA-UMR IATE, 2 place Viala, 34060 Montpellier Cedex 2, France
[b] ADRIA Développement, Creac'h Gwen, 29196 Quimper Cedex, France
[c] Université Européenne de Bretagne, Université de Brest, LUBEM EA 3882 – UMT Physiopt 08.3, 6 rue de l'Université, F-29334 Quimper Cedex, France
[d] INRA Mét@risk 16, rue Claude Bernard, 75231 Paris Cédex 5, France
[e] AgroParisTech, UFR Informatique, 16, rue Claude Bernard, 75 231 Paris Cedex 05, France
[f] Soredab (Groupe SOPARIND BONGRAIN), La Tremblaye, 78125 La Boissière-Ecole, France
[g] Groupement d'Intérêt Scientifique Sym'Previus, 147 rue de l'Université, F-75007 Paris, France
[h] LIRMM, CNRS-UM2, F-34392 Montpellier, France

ARTICLE INFO

Article history:
Received 13 March 2010
Received in revised form
28 June 2010
Accepted 3 July 2010
Available online 21 July 2010

Keywords:
Web data
Flexible querying
Ontology
Predictive microbiology

ABSTRACT

A preliminary step in microbial risk assessment in foods is the gathering of experimental data. In the framework of the Sym'Previus project, we have designed a complete data integration system opened on the Web which allows a local database to be complemented by data extracted from the Web and annotated using a domain ontology. We focus on the Web data tables as they contain, in general, a synthesis of data published in the documents. We propose in this paper a flexible querying system using the domain ontology to scan simultaneously local and Web data, this in order to feed the predictive modeling tools available on the Sym'Previus platform. Special attention is paid on the way fuzzy annotations associated with Web data are taken into account in the querying process, which is an important and original contribution of the proposed system.

1. Introduction

A preliminary step in microbial food risk assessment is the gathering of experimental data (Tamplin et al., 2003; Baranyi and Tamplin, 2004; McMeekin et al., 2006).

In the framework of the Sym'Previus project (Couvert et al., 2007 and http://www.symprevius.org), we have designed a complete data integration system opened on the Web which allows a local database (Buche et al., 2005) to be complemented by data extracted from the Web (Buche et al., 2006; Hignette et al., 2008). The local data were classified by means of a predefined vocabulary organized in taxonomy, called ontology. This ontology is used to extract pertinent data from the Web. We focus on the Web data tables as they contain, in general, a synthesis of data published in the documents. Our aim is to integrate the data tables found on the Web with the local data by means of a flexible querying system which allows the end-user to retrieve the nearest local and Web data corresponding to his/her selection criteria. With our solution, the end-user may simultaneously and uniformly query local and Web data in order to feed the predictive modeling tools available on the Sym'Previus platform.

These developments have been introduced in the predictive modeling program Sym'Previus (www.symprevius.org). Actually, to take into account the food matrix effect, predictive models need raw data obtained from food product. Considering the large diversity of foods, a local database seems to be too limited (i) to gather information for all food products, and (ii) to have enough and adequate data to take into account the food variability. The simultaneous querying in local and Web data increases the accuracy and the pertinence of the simulation results.

We first remind the semi-automatic annotation method (implemented in the @WEB tool, see @Web demo) which allows data to be retrieved from data tables found in scientific documents on the Web and to be annotated thanks to the ontology. As the local data and the Web data tables were all together indexed by the ontology, it is therefore possible to use the vocabulary defined in the ontology in order to query simultaneously those two sources of information. Second, we present the original contribution of the paper, which consists in the design of the flexible querying system, called MIEL++. This system allows the end-user to query simultaneously and in a transparent way the local data and the semantic

* Corresponding author at: INRA-UMR IATE, 2 place Viala, 34060 Montpellier Cedex 2, France.
E-mail address: buche@supagro.inra.fr (P. Buche).

0740-0020/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.fm.2010.07.002

annotated Web data, thanks to the ontology. It is flexible because (i) it allows the end-user to express preferences in his/her selection criteria and (ii) it takes into account, in the answers content, the different kinds of fuzziness of the semantic annotated Web data. This second point is essential to deal with the uncertainty of the Web data and with the imperfection of their annotations. Third and finally, experimental results are presented and discussed.

2. Materials and methods

Our annotation method which allows the Web data tables to be indexed thanks to the vocabulary defined in the ontology has already been presented in details by Hignette et al. (2008). It is briefly recalled in the first paragraph of this section. The content of the Web data tables must be indexed according to the ontology in order to be queried. This indexation associates a set of annotation graphs with each row of a Web data table. This method is presented in the second paragraph. Then, we present in the third paragraph, the automatic querying method which uses the index associated with the Web data tables in order to perform the MIEL++ query. Finally, we present in the fourth paragraph, the way the experimental data, extracted thanks to the MIEL++ querying system, are used to estimate the parameters of the simulation model.

3. Automatic annotation method of a Web data table

Web data tables are semi-automatically annotated by means of a predefined vocabulary, called ontology (see definition in Table 1). This ontology is composed of data types meaningful in the domain of risk in food and semantic relations linking those data types. The structure of the ontology is presented in Fig. 1. Data types are described in two different ways depending on whether their associated values are symbolic (*Food product*, *Microorganism* …) or numeric (*Temperature*, *pH* …). Symbolic types are described by taxonomies of possible values (for example, a taxonomy of microorganisms). The taxonomy of possible values associated with a symbolic type defines its domain of values. Numeric types are described by their possible set of units (for example, °C or °F for *Temperature*, but no unit for *pH* or a_w), and their possible numeric range (for example, [0, 14] for *pH*). The numeric range associated with a numeric type defines its domain of values. Semantic

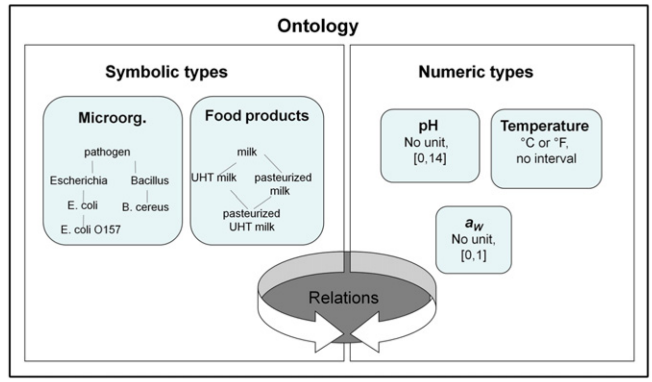

Fig. 1. The structure of the ontology shown on an excerpt of it.

relations (see definition in Table 1) are defined by their signature which is composed of a result data type and a set of access data types. For example, the relation *GrowthParameterAw*, representing the growth limits of a microorganism for any food product, has for access type the symbolic type *Microorganism* and for result type the numeric type a_w. Our annotation method first annotates the symbolic columns and the numeric columns and then uses these annotations to recognise the semantic relations present in the Web data tables (see Hignette et al., 2008 for more details).

Example. We consider a table having for legend "Reported prevalence of Campylobacter" and which is composed of two columns having respectively for title: "Product" and "Positive for Campylobacter (%)". The first row of this table is composed of the term "Chicken products" in the cell corresponding to the "Product" column and 0.07 in the cell corresponding to the "Positive for Campylobacter (%)" column. When annotating this table, the method finds that the first column is symbolic and the second one is numeric. Concerning the first column, the method annotates it by the symbolic type *Food product*. The second column is annotated by the numeric type *Samples Positive*. Finally the whole table is automatically annotated by the *Prevalence* semantic relation.

In the following, we explain how the semantic relations used to annotate a Web data table are instanciated for each row of the Web data table in order to index it, this indexation being a preliminary step to the flexible querying process.

Table 1
Glossary of technical terms.

Term	Explanation
Ontology	It is, for a given scientific domain, a set of concepts and semantic relations which link those concepts. By example, *Microorganism* and *Clostridium perfringens*, pH, a_w are concepts of the ontology. *Microorganism* is a concept classified as symbolic data type. pH is a concept classified as numeric data type. *Microorganism* and *Clostridium perfringens* are linked by the *a kind of* semantic relation.
Semantic relation	It is a relation which links concepts of the ontology. Semantic relations are defined by their signature which is composed of a result data type and a set of access data types. For example, the relation *GrowthParameterAw*, representing the growth limits of a microorganism for any food product, has for access type the symbolic type *Microorganism* and for result type the numeric type a_w.
Instanciation	The instanciation of a concept or a semantic relation is an occurrence of a concept (a numeric or a symbolic type) or a semantic relation used to annotate a given row of a given Web data table.
Domain of values (Dom)	A domain of values is defined for a symbolic type and a numeric type of the ontology. The domain of values of a symbolic type is its taxonomy of possible values in the ontology. The domain of values of a numeric type is its numeric range in the ontology.

4. Instanciation of a semantic relation in a Web data table into a RDF graph

Once a Web data table has been annotated by one or several semantic relations, it is indexed by instances of these relations which are associated with each row of the Web data table. The instanciation (see definition in Table 1) of a semantic relation in a Web data table is represented, for each row of the table, as a Resource Description Framework (RDF) graph. RDF is the language recommended by the W3C (World Wide Web consortium) to represent semantic annotations associated with Web resources. An instance of a semantic relation associated with a row of a Web data table is composed of the instances of the result data type and the access data types of its signature which are associated with the data present in the cells of the row. The generated instanciations are fuzzy: they allow one to take into account the imprecision of the initial data in the table (for example an interval for a numeric type), the similarity comparison between the vocabulary used in the table with the vocabulary of the ontology, and the uncertainty of the annotation of the table by semantic relations. We first present briefly the theory of fuzzy sets that we