Flexible Querying of Web data to Simulate Bacterial Growth in Food

Patrice Buche, Olivier Couvert, Juliette Dibie-Barthelemy, Gaëlle Hignette, Eric Mettler, Lydie Soler

Patrice Buche, Olivier Couvert, Juliette Dibie-Barthelemy, Gaëlle Hignette, Eric Mettler, et al.. Flexible Querying of Web data to Simulate Bacterial Growth in Food. Food Microbiology, 2011, 28 (4), pp.685-693. 10.1016/j.fm.2010.07.002 . lirmm-00538961

Flexible querying of Web data to simulate bacterial growth in food

Patrice Buche*[1],[1,7,8], Olivier Couvert[2,3,7], Juliette Dibie-Barthélemy[4,5], Gaëlle Hignette[4,5], Eric Mettler[6,7], Lydie Soler[4]

[1]INRA-UMR IATE, 2 place Viala, 34060 Montpellier Cedex 2, France

E-mail: buche@supagro.inra.fr

[2]ADRIA Développement, Creac'h Gwen, 29196 Quimper Cedex, France

[3]Université Européenne de Bretagne, Université de Brest, LUBEM EA 3882 - UMT
Physiopt 08.3, 6 rue de l'Université, F-29334 Quimper Cedex, France

[4]INRA Mét@risk 16, rue Claude Bernard 75231 Paris Cédex 5 France

[5]AgroParisTech, UFR Informatique, 16, rue Claude Bernard, 75 231 Paris Cedex 05, France

[6]Soredab (Groupe SOPARIND BONGRAIN), La Tremblaye,
78125 La Boissière-Ecole, France

[7]Groupement d'Intérêt Scientifique Sym'Previus, 147 rue de l'Université,
F-75007 Paris, France

[8]LIRMM, CNRS-UM2, F-34392 Montpellier, France

---

[1] corresponding author

Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible
querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI :
10.1016/j.fm.2010.07.002

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.

Keywords : Web data, flexible querying, ontology, predictive microbiology

Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI : 10.1016/j.fm.2010.07.002

Introduction

A preliminary step in microbial risk assessment in foods 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 (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

Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible
querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI :
10.1016/j.fm.2010.07.002

Web data tables were all together indexed by the ontology, it is therefore possible to use the terminology 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 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.

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.

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 Figure 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 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.

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 use in our instanciation method, then we present how we instanciate numeric types, symbolic types and relations.

Fuzzy sets:

We use the definition of fuzzy sets given by Zadeh, 1965 and Zadeh, 1978. The notion of fuzzy set is an extension of classical subsets. In the classical case, elements of a reference set X which have some properties belong to a subset A, and elements which do not have these

Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI : 10.1016/j.fm.2010.07.002

properties belong to the complementary subset of A in X. In a fuzzy set, elements can belong partially to the fuzzy set, with a membership degree included between 0 (element which is not part of the fuzzy set) and 1 (element which is completely part of the fuzzy set). The membership degree of an element x of the reference set X for the fuzzy set A is denoted $\mu_A(x)$. When X is defined on a continuous domain, we talk about a continuous fuzzy set; when X is defined on a discrete domain, we talk about a discrete fuzzy set. The support of a fuzzy set A defined on a reference set X is the set (in the classical definition) of elements x of X such that $\mu_A(x) > 0$. The kernel of a fuzzy set A defined on a reference set X is the set (in the classical definition) of elements x of X such that $\mu_A(x) = 1$.

A trapezoid fuzzy set TFS is a special continuous fuzzy set which is described only by its support sup = $[min_{sup}, max_{sup}]$ and its kernel ker = $[min_{ker}, max_{ker}]$. The membership degree of a numeric value x in the reference set is then defined as follows:

– if $x \leq min_{sup}$ or $x \geq max_{sup}$ then $\mu_{TFS}(x) = 0$;

– if $min_{ker} \leq x \leq max_{ker}$ then $\mu_{TFS}(x) = 1$;

– if $min_{sup} \leq x \leq min_{ker}$ then $\mu_{TFS}(x) = \dfrac{x - min_{sup}}{min_{ker} - min_{sup}}$;

– if $max_{ker} \leq x \leq max_{sup}$ then $\mu_{TFS}(x) = \dfrac{x - max_{ker}}{max_{sup} - max_{ker}}$;

There are several semantics for fuzzy sets, defined in Dubois and Prade, 1997:

- Preferences: the elements with the higher membership degrees are the preferred elements. This is used in the MIEL++ querying system by the user to define query preferences;
- Uncertainty or imprecision: there exists a “true” value, but we do not know it. The higher is the membership degree of a value x, the more possible is x to be the “true” value. This is used in our instanciation method to represent the imprecision of the original data in the tables (in the instanciation of numeric types);

- Similarity: a new value is represented by its similarity with known values. The higher is the membership degree of a known value x, the more it is similar to the new value. This is used in our instanciation method to represent the similarities between a term from the web and terms from the ontology (in the instanciation of symbolic types).

Instanciation of numeric types

Let us consider that a Web data table has been annotated by a semantic relation. There are three possibilities in order to instanciate (see definition in Table 1) a numeric type t of the signature of the relation in a given row of the Web data table:

1. There is one column in the table (thus one cell in the row to annotate) that was annotatetd by the numeric type t. In this case, the values in the cell are used to instanciate the type: it can be an isolated value, an enumeration of isolated values, an interval or a mean with a standard error. Intervals and mean with standard error are recognised using specific patterns; if those patterns are not recognised, then all numeric values in the cell are considered as isolated.

2. There are several columns in the table that were annotated by the numeric type t. In this case, we have to find the relations between the columns: it is done by looking for keywords in the columns titles. A column can represent a minimum value, a maximum value or an optimum value (included between the minimum and maximum); it can also represent a mean value or a standard error.

3. There is no column in the table that was annotated by the numeric type t. If the numeric type t has a defined unit, we search for occurences of a numeric value followed by this unit, in the table title or in the columns titles: those occurences are then considered as isolated values.

An instance of a numeric type is represented by a continuous fuzzy set of which the reference set is the numeric range of the numeric type defined in the ontology. This fuzzy set is built from trapezoid fuzzy sets, each being created as follows:

Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI : 10.1016/j.fm.2010.07.002

- when recognising an isolated value x in the table, we construct a trapezoid fuzzy set with sup = ker = [x, x];
- when recognising an interval [a, b] in the table, either in one cell or when a is the value in a cell recognised as minimum and b is the value in a cell recognised as maximum with no cell recognised as optimum, we construct a trapezoid fuzzy set with sup = ker = [a, b];
- when recognising a cell as minimum, its minimum numeric value being min, a cell as maximum, its maximum numeric value being max and a cell as optimum, its values being included in the minimum interval [a, b], we construct a trapezoid fuzzy set with sup = [min, max] and ker = [a, b];
- when recognising a mean m and a standard error e, we construct a trapezoid fuzzy set with sup = $[m - e, m + e]$ and ker = $[m,m]$.

Once all trapezoid fuzzy sets have been created, the instanciation of the numeric type is the union of all those sets (for example, there can be a union of several isolated values).

Example: Table 2 is annotated with the semantic relation of the ontology *GrowthParameterAw*, with the access type *Microorganism* and the result type a_w. This result type is instanciated, for the first row of the data table, as a unique trapezoid fuzzy set with sup = [0.943, 0.97] and ker = [0.95, 0.96]. This fuzzy set is represented in Figure 2.

Instanciation of symbolic types

Let us consider that a Web data table has been annotated by a semantic relation. In order to instanciate (see definition in Table 1) a symbolic type t of the signature of the relation in a given row of the Web data table, we construct a discrete fuzzy set. The reference set of this fuzzy set is the taxonomy of possible values of the symbolic type t. The membership degree of a term x of the ontology in the fuzzy set is the term similarity between x and the term in the cell that was annotated by the symbolic type t. In the corpus of tables we used for experimentations, it did not happen that several columns in a table were annotated by the same symbolic type t, however, would that happen, we would construct a union of fuzzy sets (one fuzzy set for each column).

The term similarity between a term x of the ontology and the term in the cell has already been presented in Hignette *et al*. (2008) and is briefly recalled in the following. Both terms are transformed into weighted vectors: the coordinate system of the vectors is the set of all possible words (i.e. all words in the ontology plus the words of the terms to compare with the ontology), the coordinate values associated with a given vector represent the weight of those words in the term (1 if the word is present in the term, 0 otherwise). The similarity between both terms is computed as the cosine similarity measure between the two weighted vectors, which is one of the most popular similarity measures described by Lin, 1998.

Example: Table 2 is annotated with the semantic relation of the ontology *GrowthParameterAw*, with the access type *Microorganism* and the result type *aw*. The access type is instanciated, for the first row of the data table, as a unique fuzzy set, defined as follows: {0.5/*Clostridium perfringens*, 0.5/*Clostridium botulinum*} which means that the term, *Clostridium*, in the cell is similar to the terms *Clostridium perfringens* and *Clostridium botulinum* of the ontology with a similarity measure of 0.5. If other Clostridium are defined in the ontology, they will also appear in the fuzzy set.

Instanciation of semantic relations

Once all the types of the signature of a semantic relation have been instanciated for a given row of a Web data table, we can instanciate the semantic relation for this row: we create an instance of the relation which is composed of the instances of the numeric and symbolic types of its signature which were created for the row.

Example: Let us consider that Table 2 has been annotated by the semantic relation *GrowthParameterAw*. Figure 3 shows the RDF graph which represents the instanciation of this relation in the first data row of Table 2. In Figure 3, the RDF graph expresses that the row (having the identifier *uriRow1* in the RDF graph) is annotated by a discrete fuzzy set, called *DFSR1*. This fuzzy set has a semantic of similarity and indicates the list of the closest semantic relations of the ontology used to annotate the first row. Only the semantic relation *GrowthParameterAw* belongs to this fuzzy set with the pertinence score of 1.0, which expresses the degree of certainty associated with the semantic relation recognition by our annotation method. The access type of the relation, which is an instance of the symbolic type *Microorganism*, is instanciated by a discrete fuzzy set, called *DFS1*. This fuzzy set has a semantic of similarity and indicates the list of the closest terms of the ontology compared with the term *Clostridium*. Two terms (*Clostridium Perfringens* and *Clostridium Botulinum*) belong to this fuzzy set with a membership degree of 0.5. The result type of the relation, which is an instance of the numeric type *aw*, is instanciated by a continuous fuzzy set, called *CFS1*. This fuzzy set has a semantic of imprecision and indicates the possible growth limits ([0.943, 0.97]) and the possible optimal growth limits ([0.95, 0.96]).

In the following, we call annotations of a Web data table the instanciations of the semantic relations which have been recognised in the table.

Simultaneous flexible querying of the RDF graph database and the local database

The MIEL++ querying system relies on the domain ontology used to index the local data and the Web data tables. MIEL++ allows the end-user to retrieve the nearest local and Web data corresponding to his/her selection criteria expressed as fuzzy sets and representing his/her preferences. The ontology -more precisely the taxonomies of values associated with symbolic types- is used in order to assess which data can be considered as "near" to the user's selection criteria. A query is asked to the MIEL++ system through a single graphical user interface, which relies on the domain ontology. The query is translated into the query language of each data source: an SQL query in the relational local database (see Buche *et al.*. 2005 for more details) and a SPARQL query in the RDF graph base. SPARQL is the querying language recommended by the W3C to query annotations expressed in RDF graphs. The global answer to the query is then the union of the local answers in each data source, which are ordered according to their relevance to the query selection criteria. In this paper, we focus on three original aspects of the SPARQL querying: (1) the use of the taxonomies of values associated with the symbolic types to enlarge the querying, (2) the way comparisons between the user's selection criteria and fuzzy annotations of Web data tables are done, (3) the total order on the answers defined to retrieve the most pertinent data to the user.

Example: Let us consider a MIEL++ query *Q* expressed in the relation *GrowthParameterAw* and having for selection criteria *(aw=awPreference) and (Microorganism=MicroPreferences)*. The continuous fuzzy set *awPreferences*, which is equal to [0.9, 0.94, 0.97, 0.99], means that the end-user is first interested in a_w values in the interval [0.94, 0.97], but he/she accepts to enlarge the querying till the interval [0.9, 0.99]. The discrete fuzzy set *MicroPreferences*, which is equal to {1.0/*Gram+*, 0.5/*Gram-*}, means that the end-user is interested in microorganisms which are first *Gram+* and then *Gram-*. This fuzzy set defines implicitly user's preferences for microorganisms which are kinds of Gram+ and Gram-. Besides, the taxonomy of values associated with the symbolic type *Microorganism* contains the terms *Clostridium Botulinum*, *Clostridium Perfringens* and *Staphylococcus Spp*. which are kind of *Gram+* and *Salmonella spp*. which is a kind of *Gram-*. In order to take those implicit preferences into account in the querying, we propose to perform a closure of the fuzzy set *MicroPreferences* (see Thomopoulos *et al*. 2003 and Thomopoulos

Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI : 10.1016/j.fm.2010.07.002

et al. 2006 for more details). Intuitively, the closure propagates degrees of preferences to more specific values of the taxonomy. By example, the closure of the fuzzy set *MicroPreferences* is: {1.0/*Gram+*, 0.5/*Gram-*, 1.0/*Clostridium Botulinum*, 1.0/*Clostridium Perfringens*, 1.0/ *Staphylococcus Spp.*, 0.5/*Salmonella*}.

In order to build the answer to a query, selection criteria representing user's preferences expressed as fuzzy sets must be compared with fuzzy annotations, which are associated with the Web data tables to query. But, as we saw previously, the fuzzy sets used in the annotations have two different semantics: similarity for fuzzy sets associated with the instanciations of symbolic types and imprecision for those associated with the instanciations of numeric types. Consequently, we propose to realise those comparisons separately using two different measures: (i) a possibility degree of matching (noted Π) and a necessity degree of matching (noted N) which are classically used (see Dubois & Prade 1988) to compare a fuzzy set having a semantic of preference with a fuzzy set having a semantic of imprecision and (ii) an adequacy degree as proposed by (Baziz *et al.* 2006) to compare a fuzzy set having a semantic of preference with a fuzzy set having a semantic of similarity.

Let *(a=v)* be a selection criterion of the MIEL++ query *Q*, *v'* a fuzzy annotation of the attribute *a* (which is either a numeric type or a symbolic type of the ontology) stored in a RDF graph, $sem_{v'}$ the semantic of *v'*, μ_v and $\mu_{v'}$ their respective membership degrees defined on the domain of values *Dom* (see definition in Table 1) associated with the attribute *a* and *cl* the function which corresponds to the fuzzy set closure. The comparison result depends on the semantic of the fuzzy set *v'*:

- if $sem_{v'}$=imprecision, the comparison result is given by the **possibility degree of matching** between *v* and *v'* noted $\Pi(v,v')=\sup_{x\in Dom}(\min(\mu_v(x), \mu_{v'}(x)))$ and the **necessity degree of matching** between *v* and *v'* noted $N(v,v')=\inf_{x\in Dom}(\max(\mu_v(x), 1 - \mu_{v'}(x)))$ (see Figure 5 for a graphical representation);

- if $sem_{v'}$=similarity, the comparison result is given by the **adequacy degree** between *cl(v)* and *cl(v')* noted ad(cl(v), cl(v'))=$\sup_{x \in Dom}(\min(\mu_{cl(v)}(x), \mu_{cl(v')}(x)))$ (see Figure 6 for a graphical representation).

The comparison results of fuzzy sets having the same semantic (similarity or imprecision) and associated with different selection criteria are aggregated using the min operator. Therefore, an answer to a query is a set of tuples composed of (i) the pertinence score *ps* associated with the queried semantic relation, (ii) three comparison scores associated with the selection criteria of the query: a global adequation score ad_g and two global matching scores Π_g and N_g, and, (iii) the values associated with the answer attributes of the query. Based on those scores, we propose to define a total order on the answers which gives greater importance to the most pertinent answers compared with the ontology: answers are respectively sorted, in descendant order, according to *ps*, ad_g, N_g and Π_g.

Example: The answer to the MIEL++ query *Q* considered in the previous example and compared with the fuzzy annotations associated with the three rows of Table 2 is given below:

{*ps*=1, ad_g=0.5, N_g=1, Π_g =1, Microorg=(0.5/Clostridium Perfringens+0.5/Clostridium Botulinum), aw=[0.943, 0.95, 0.96, 0.97]},

{*ps*=1, ad_g=0.5, N_g =0.5, Π_g=0.68, Microorg=(0.5/Staphylococcus spp.+0.5/Staphylococcus aureus), aw=[0.88, 0.98, 0.98, 0.99]},

{*ps*=1, ad_g=0.5, N_g =0, Π_g=0.965, Microorg=(1.0/Salmonella), aw=[0.94, 0.99, 0.99, 0.991]}

Application in bacterial growth simulations

Microbial growth kinetics are usually described using primary models with four main parameters: x_0 is the initial bacterial concentration, x_{max} is the maximum bacterial concentration, lag is the lag time (h), and μ_{max} is the maximum specific growth rate (h^{-1}). The

Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI : 10.1016/j.fm.2010.07.002

two last parameters vary according to the physico-chemical food characteristics and the specific effect of the food matrix, whereas x_0 and x_{max} are considered as constant. The effects (pH, a_W, storage temperature and food matrix) on μ_{max} is described by a multiplicative function with interactions (Augustin *et al.*, 2005; Le Marc *et al.*, 2002) derived from the cardinal models of Rosso *et al.* (1995):

$$\mu_{\max}\left(T, pH, a_W\right) = \mu_{opt}.CM_2(T)CM_1(pH)CM_1(aw)\xi(T, pH, aw)$$

where

$$CM_n(X) = \begin{cases} 0 & , X \le X_{\min} \\ \dfrac{(X - X_{\max})(X - X_{\min})^n}{(X_{opt} - X_{\min})^{n-1}\left[(X_{opt} - X_{\min})(X - X_{opt}) - (X_{opt} - X_{\max})(X_{opt} + X_{\min} - nX)\right]} & , X_{\min} < X < X_{\max} \\ 0 & , X \ge X_{\max} \end{cases}$$

and $$\xi(T, pH, aw) = \begin{cases} 1 & , \phi \le 0.5 \\ 2 \cdot (1 - \phi) & , 0.5 < \phi < 1 \\ 0 & , \phi \ge 1 \end{cases}$$ with $$\phi = \sum_i \frac{\omega(X_i)}{2 \cdot \prod_{j \ne i} (1 - \omega(X_j))}$$ and

$$\omega(X) = \left(\frac{X_{opt} - X}{X_{opt} - X_{\min}}\right)^3,$$

where μ_{max} is the specific growth rate in the considered food and in the considered temperature, pH and a_W conditions, μ_{opt} is the optimal growth rate value when temperature, pH and a_W are set to their optimal values, and X_{min}, X_{opt} and X_{max} are minimal, optimal and maximal temperatures, pH, and water activities for growth. n is the shape parameter of the CM model.

The lag time (lag) is calculated according to the following equation:

$$lag = \frac{\mu_{opt}.lag_{\min}}{\mu_{\max}}$$

Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI : 10.1016/j.fm.2010.07.002

where lag_{min} is the lag time value when temperature, pH and aw are set to their optimal values. The optimal growth rate μ_{opt} and the minimal lag time lag_{min} depend on both strain and food matrix (Pinon *et al.*, 2004).

Cardinal values are characteristic parameters of the microorganisms and are independent from the food matrix. Consequently, $CM_n(X)$ functions are calculated using only bacterial species parameters and physico-chemical factors, whereas μ_{opt} and lag_{min} can be estimated from an experiental μ_{max} and lag related to microbial growth in the food to be assessed:

$$\mu_{opt} = \frac{\mu_{\max}(T, pH, a_W)}{CM_2(T).CM_1(pH).CM_1(a_W)}$$

$$lag_{\min} = \frac{\mu_{\max}.lag}{\mu_{opt}}$$

μ_{opt} and lag_{min} calculations need at least one μ_{max} and one lag value at certain temperature, pH and a_W conditions. In the Sym'Previus calculation tools, these data can be entered following two main ways. On the one hand, the user can enter manually values if he/she has this information (results of a challenge-test, bibliographic data, etc.). On the other hand, the user can select the food product in the ontology (figure 4), and submit a query simultaneously in local and web data. The results are returned in a synthetic table (figure 7) where experimental μ_{max} and lag values are proposed to be included in the μ_{opt} and lag_{min} calculation.

In the worst case (no food data available), simulations are carried out with known μ_{opt} and lag_{min} obtained in culture media in laboratory conditions.

Results and discussion

We use in this paper the same corpus as the one used in Hignette *et al.* (2008) which composed of 60 tables extracted from publications on food microbiology, in order to test our instanciation of semantic relations in Web data tables and the flexible querying of the Web data tables such annotated.

We have automatically instanciated the 119 relations which were correctly recognised to annotate the 60 tables in the experiment presented in Hignette *et al*. (2008).

Instanciation of numeric types

The instanciation of numeric types was analysed for the first data row of each table: we assume that the structure is enough homogenous inside a table, such that the instanciation of the first row can be considered as representative of what happens in the whole table. On the 119 relations, there were two errors on the extraction of numeric values (one was an error of type recognition; one was an error of numeric value recognition). For 5 tables (corresponding to 13 relations), the numeric type *Temperature* was not instanciated because its value was not present in the table but in the textual environment of the table in the original publication. There also were three errors in interval reconstruction (values were considered as isolated while it was an interval) and one error in the construction of a minimum/optimum/maximum trapezoid fuzzy set (values were considered as isolated). For all 100 remaining relations, all numeric values were correctly instanciated.

Instanciation of symbolic types

The quality of the instanciation of symbolic types was evaluated on 185 instances of food products extracted from the corpus of 60 tables. For those food products, the “best match” in the ontology (i.e. the term in the ontology that was the nearest to the meaning of the term in the table) was manually defined. The evaluation is done by looking at the position of the “best match” in the automatic instanciation, by order of descendant membership degree. The position is evaluated at worse, i.e. if there are several terms in the ontology having the same membership degree in the fuzzy set used for the instanciation, the "best match" is always considered as being in last position. This evaluation at worse comes from the need to manually validate the instanciations: if we ask a user to choose among the 5 terms having the best membership degree, we want to be sure that the "best match" is among those 5. On the 185 terms from the web, 78% had a "best match" for which the term similarity with the term from the web was not null, 46% had their "best match" in first position in the computed instanciation, while 66% had their "best match" among the five best positions. These results validate the approach of keeping a fuzzy set for instanciating the symbolic types, instead of keeping only the term in the taxonomy having the best similarity with the term in the table.

Flexible querying of the RDF database including Web data tables

In preliminary tests performed on the RDF graph database composed of more than 22.000 RDF triples (312 graphs), we have evaluated 5 queries (see Table 3) covering at least 50% of the database entries. Querying quality is assessed using two measures: precision and recall. Precision is the ratio of correct answers over the total number of computed answers. Recall is the ratio of correct computed answers over the number of expected answers. We obtain better results in the queries where the selection criterion concerns microorganisms than in the ones concerning food products. This is due to the fact that microorganism names are more standardized in Web data tables than food product names. Nevertheless, we obtain a precision of 100% for the two last queries concerning food product if we put a threshold of 0.7 on the terms similarity degrees.

Microbial growth simulation in food

Growth simulation in food requires data related to the bacterial species (cardinal values), physico-chemical properties of food (pH, water activity, storage temperature) and food matrix effect. The implementation of the food matrix in models is achieved thanks to the μ_{opt} and lag_{min} parameters. These two parameters can be estimated only from experimental data collected in food. In the absence of food experimental data, simulations are usually processed using μ_{opt} and lag_{min} obtained in culture media at laboratory conditions, giving approximate results.

The following example presents *Listeria monocytogenes* growth simulation in cold smoked salmon (pH 6 +/- 0.2 and water activity 0.970 +/- 0.003) and stored at 6°C (+/- 1°C), performed with the Sym'Previus probabilistic software (Couvert *et al.*, submitted article). The initial contamination for simulation is expected to -2 log CFU/g (+/- 0.2). On the one hand, without any experimental data related to *L. monocytogenes* growth in cold smoked salmon, the simulation includes physico-chemical properties and optimum growth parameters obtained in culture media (figure 8-A). On the other hand, a query in local and web data allows one to retrieve experimental data (figure 7) which are used to estimate μ_{opt} and lag_{min} and, consequently, taking into account the food matrix in simulations. The figure 8-B reproduces the previous simulation in cold smoked salmon, taking into account the food product parameters. Median population size after 21 days of storage reaches 4.4 log CFU/g in laboratory media simulations, and -0.09 log CFU/g taking into account bibliographic data in cold smoked salmon. These results demonstrate the importance of the food matrix effect on bacterial growth, and the necessity to link bibliographic data stored in databases with simulation softwares.

Discussion

Recent propositions in the Semantic Web community propose to extract, filter, annotate and query Web data tables (Ying *et al.* 2007, Cafarella *et al.,* 2008), but they have not been designed with the same objectives. TableSeer (Ying *et al.* 2007) for instance permits to extract a set of predefined metadata (caption, cell content, geographical position of the table in the HTML page, ...) from Web tables, but it does not compare the schema of the Web tables to preexisting schemas defined in an ontology. We can also cite WebTables (Cafarella *et al.,* 2008) which proposes a system to identify data tables in the huge amount of tables included in HTML documents and index them, in order to query and rank them. Nevertheless, the WebTables querying language is only composed of a set of key-words which are compared with the titles of the columns present in the Web tables. The row content of the Web tables is not used in the querying process which is only based on global co-occurrences frequencies statistics of the terms appearing in the titles of the columns. Therefore, it is not possible to compare the results we obtain with our annotation and instanciation methods with other Web table mining methods because they don't have the same aim.

Conclusion and perspectives

Probabilistic simulations of Sym'Previus software needs a lot of data in food products to take the food matrix into account and to assess food variability in bacterial growth simulations. A prototype of the @WEB and the MIEL++ tools will be soon integrated with the predictive modeling tools of the Sym'Previus project. These automatic links between web data and simulation tools are a major contribution to enhance risk assessment. In the near future, we will study ontology evolution in order to take into account the evolutions of the predictive modeling tool. For example, (i) the impact of the packaging and gaz transfer on the behaviour of the microorganism in the food matrix should be studied or (ii) the possible use of the growing number of ontologies which are available on the Web in close domains (OBOE, OUM, …) and which may enhance the quality of our annotation and instanciation methods if they are merged with our own ontology. Therefore, in order to take into account this new information, the domain ontology of our system should evolve, and, new methodologies and

Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI : 10.1016/j.fm.2010.07.002

tools should also be developed according to this ontology evolution. Another perspective is to extend this work in order to also be able to extract pertinent information represented as graphics or directly from the text.

Acknowlegements

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

References

Augustin, J. C., Zuliani, V., Cornu, M., Guillier, L., 2005. Growth rate and growth probability of *Listeria monocytogenes* in dairy, meat and seafood products in suboptimal conditions. Journal of Applied Microbiology. 99, 1019-1042.

Baranyi, J., Tamplin, M., 2004. ComBase: A Common Database on Microbial Responses to Food Environments. Journal of Food Protection. 67, 1834-1840.

Baziz, M., Boughanem, M., Prade, H., Pasi, G., 2006. A fuzzy logic approach to information retrieval using a ontology-based representation of documents. In Sanchez, E. (Ed), Fuzzy logic and the Semantic Web, Elsevier, pp. 363–377.

Buche, P., Dervin, C., Haemmerlé, O., Thomopoulos, R., 2005. Fuzzy querying of incomplete, imprecise, and heterogeneously structured data in the relational model using ontologies and rules. IEEE Transactions on Fuzzy Systems. 13, 373-383.

Buche, P., Dibie-Barthélemy, J., Haemmerlé, O., Hignette, G., 2006. Fuzzy semantic tagging and flexible querying of XML documents extracted from the Web. Journal of Intelligent Information System. 26, 25-40.

Cafarella, M. J., Halevy, A. Y., Zhe Wang D., Wu, E., Zhang, Y., 2008. WebTables: exploring the power of tables on the web. Proceedings of the VLDB Endowment Vol 1, Issue 1 538-549.

Couvert, O., Augustin, J. C., Buche, P., Carlin, F., Coroller, L., Denis, C., Jamet, E., Mettler, E., Pinon, A., Stahl, V., Zuliani, V., Thuault, D, 2007. Optimising food process and formulation through Sym'Previus, managing of the food safety. Proceedings of 5th International Conference Predictive Modeling in Foods.

Dubois, D., Prade, H., 1988. In: Possibility theory- An approach to computerized processing of uncertainty. Plenum Press, New York.

Dubois, D., Prade, H., 1997. The three semantics of fuzzy sets. Fuzzy Sets and Systems. 90(2), 141-150.

Hignette, G., Buche, P., Couvert, O., Dibie-Barthélemy J., Doussot, D., Haemmerlé, O., Mettler, E., Soler, L., 2008. Semantic annotation of Web data applied to risk in food. International Journal of Food Microbiology. 128, 174-180.

Le Marc, Y., Huchet, V., Bourgeois, C. M., Guyonnet, J. P., Mafart, P., Thuault, D., 2002. Modeling the growth kinetics of *Listeria* as a function of temperature, pH and organic acid concentration. International Journal of Food Microbiology. 73, 219-237.

Lin, D., 1998. An information-theoretic definition of similarity. In: ICML '98 : Proceedings of the Fifteenth International Conference on Machine Learning, 296-304. Morgan Kaufmann Publishers Inc., San Francisco.

McMeekin, T. A., Baranyi, J., Bowman, J., Dalgaard, P., Kirk, M., Ross, T., Schmid, S., Pinon, A., Zwietering, M. H., Perrier, L., Membré, J. M., Leporq, B., Mettler, E., Thuault, D., Coroller, L., Stahl, V., Vialette, M., 2004. Development and Validation of Experimental Protocols for Use of Cardinal Models for Prediction of Microrganism Growth in Food Products. Applied and Environmental Microbiology. 70, 1081-1087.

Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI : 10.1016/j.fm.2010.07.002

Rosso, L., Lobry, J. R., Bajard, S., Flandrois, J. P., 1995. Convenient Model To Describe the Combined Effects of Temperature and pH on Microbial Growth. Applied and Environmental Microbiology. 61, 610-616.

Tamplin, M., Baranyi, J., Paoli, G., 2003. Software programs to increase the utility of predictive microbiology information. In: McKellar, R.C., Lu, X. (Eds), Modeling Microbial responses in Foods. CRC, Boca Raton.

Thomopoulos R., Buche P., Haemmerlé O., 2003. Different kinds of comparisons between fuzzy conceptual graphs. Lecture Notes In Artificial Intelligence. 2746, 54-68.

Thomopoulos R., Buche P., Haemmerlé O., 2006. Fuzzy sets defined on a hierarchical domain. IEEE Transactions on Knowledge and Data Engineering. 18(10), 1397-1410.

Ying, L., Kun, B., Prasenjit, M., Lee Giles, C., 2007. TableSeer: automatic table metadata extraction and searching in digital libraries. Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, 91-100 (ISBN:978-1-59593-644-8).

Zadeh, L., 1965. Fuzzy sets. Information and control. 8, 338-353.

Zadeh, L., 1978. Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems. 1, 3-28

Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI : 10.1016/j.fm.2010.07.002

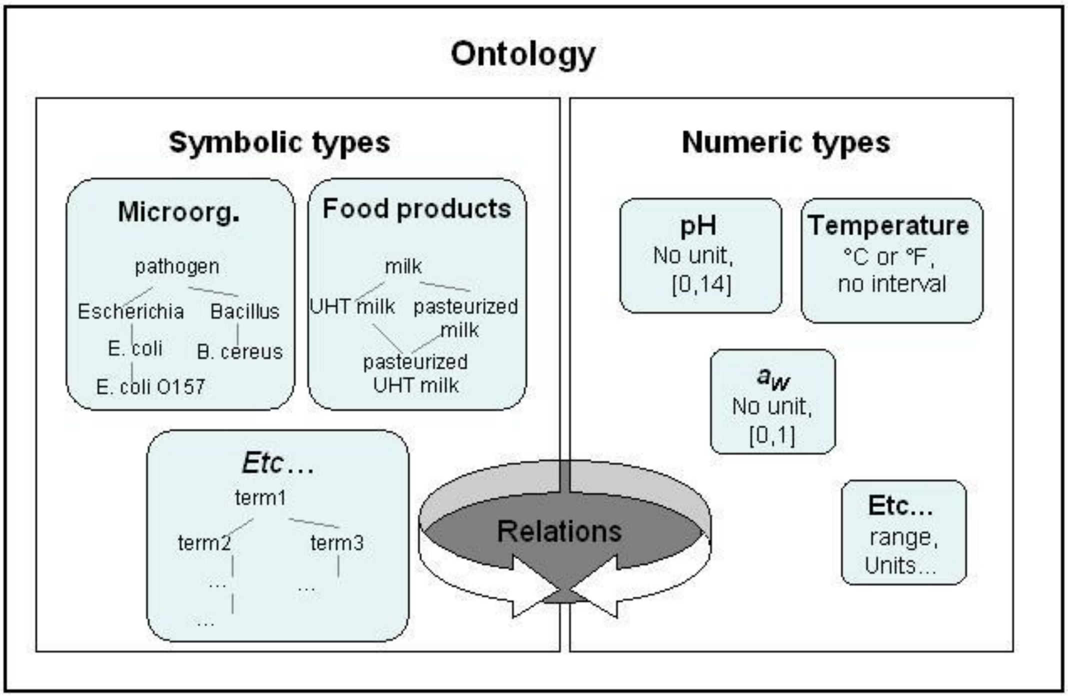


Figure 1

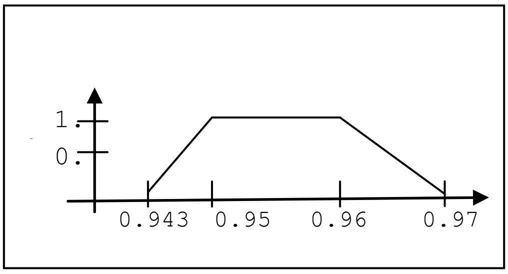


Figure 2

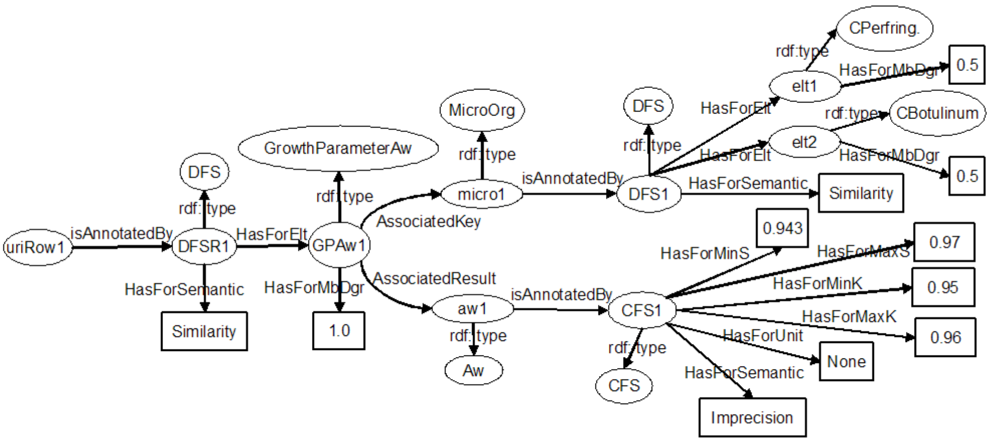


Figure 3

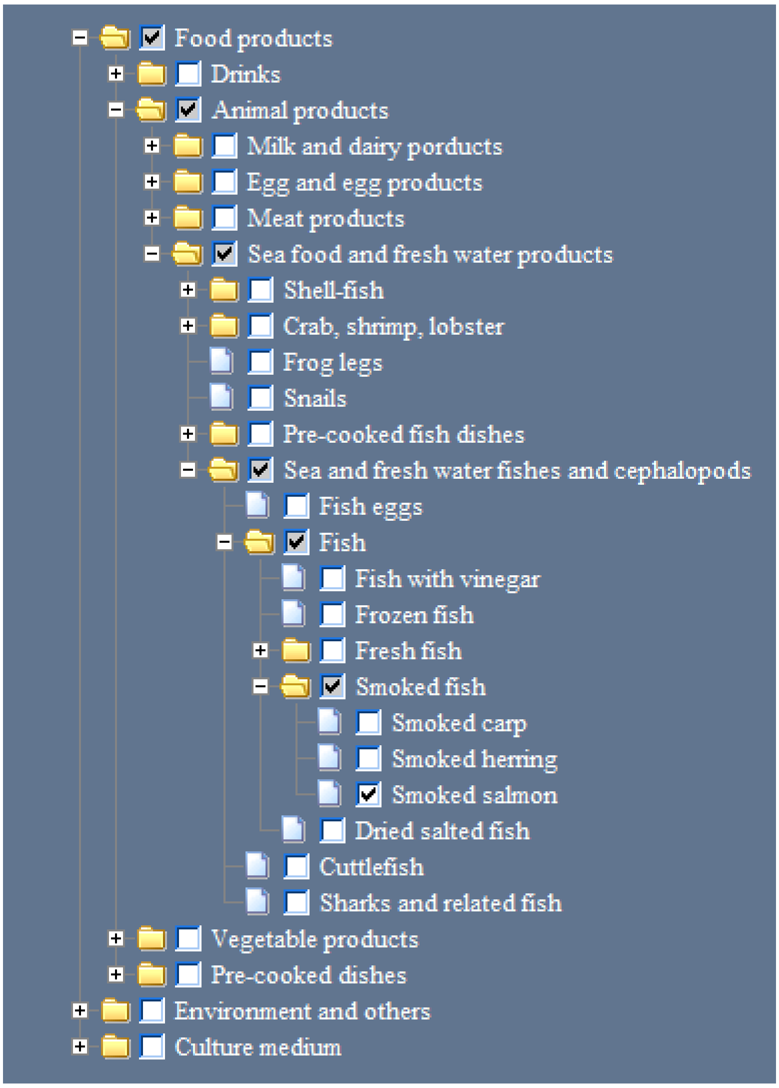
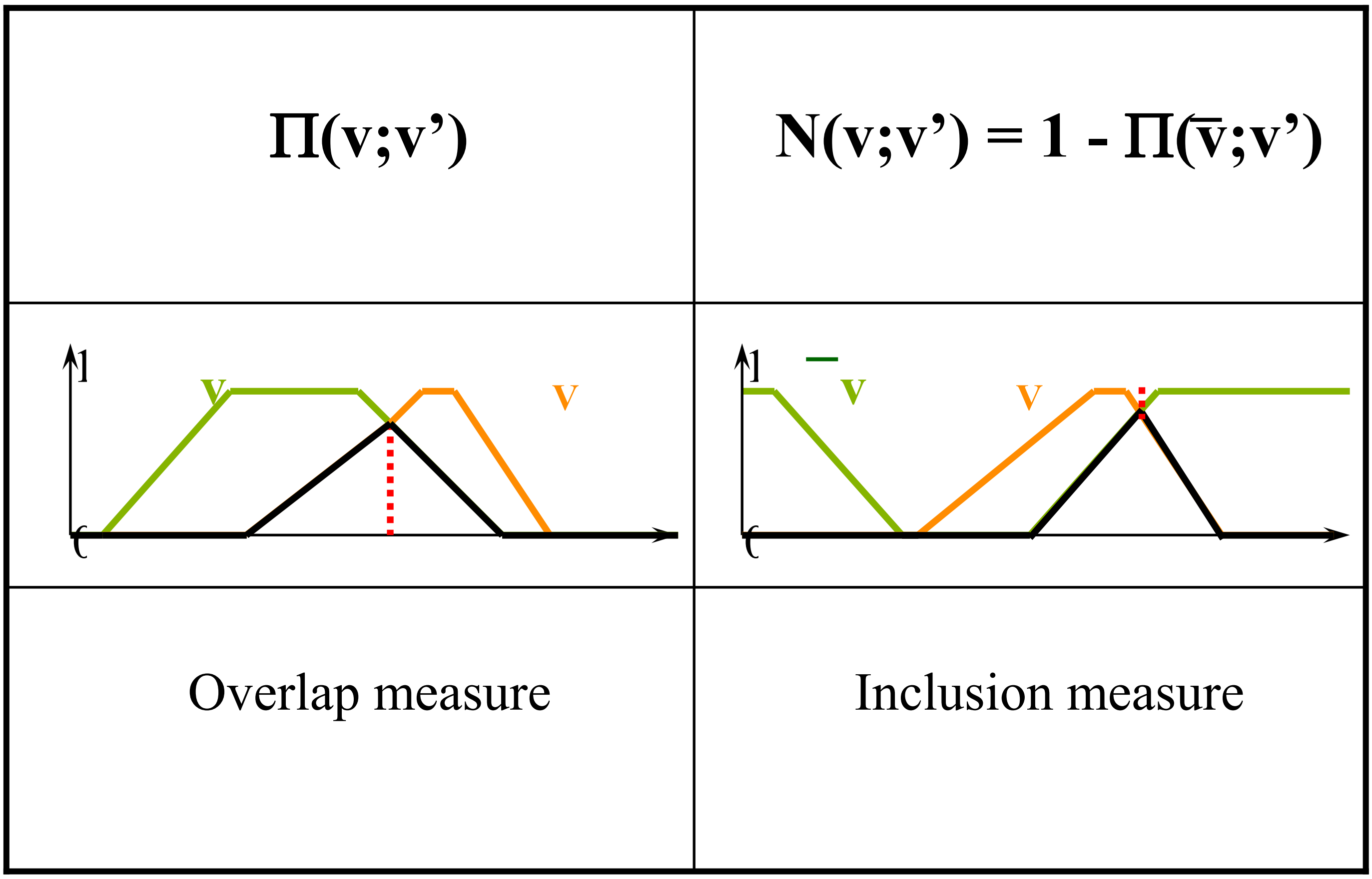

- ☑ Food products
  - ☐ Drinks
  - ☑ Animal products
    - ☐ Milk and dairy porducts
    - ☐ Egg and egg products
    - ☐ Meat products
    - ☑ Sea food and fresh water products
      - ☐ Shell-fish
      - ☐ Crab, shrimp, lobster
      - ☐ Frog legs
      - ☐ Snails
      - ☐ Pre-cooked fish dishes
      - ☑ Sea and fresh water fishes and cephalopods
        - ☐ Fish eggs
        - ☑ Fish
          - ☐ Fish with vinegar
          - ☐ Frozen fish
          - ☐ Fresh fish
          - ☑ Smoked fish
            - ☐ Smoked carp
            - ☐ Smoked herring
            - ☑ Smoked salmon
          - ☐ Dried salted fish
        - ☐ Cuttlefish
        - ☐ Sharks and related fish
  - ☐ Vegetable products
  - ☐ Pre-cooked dishes
- ☐ Environment and others
- ☐ Culture medium

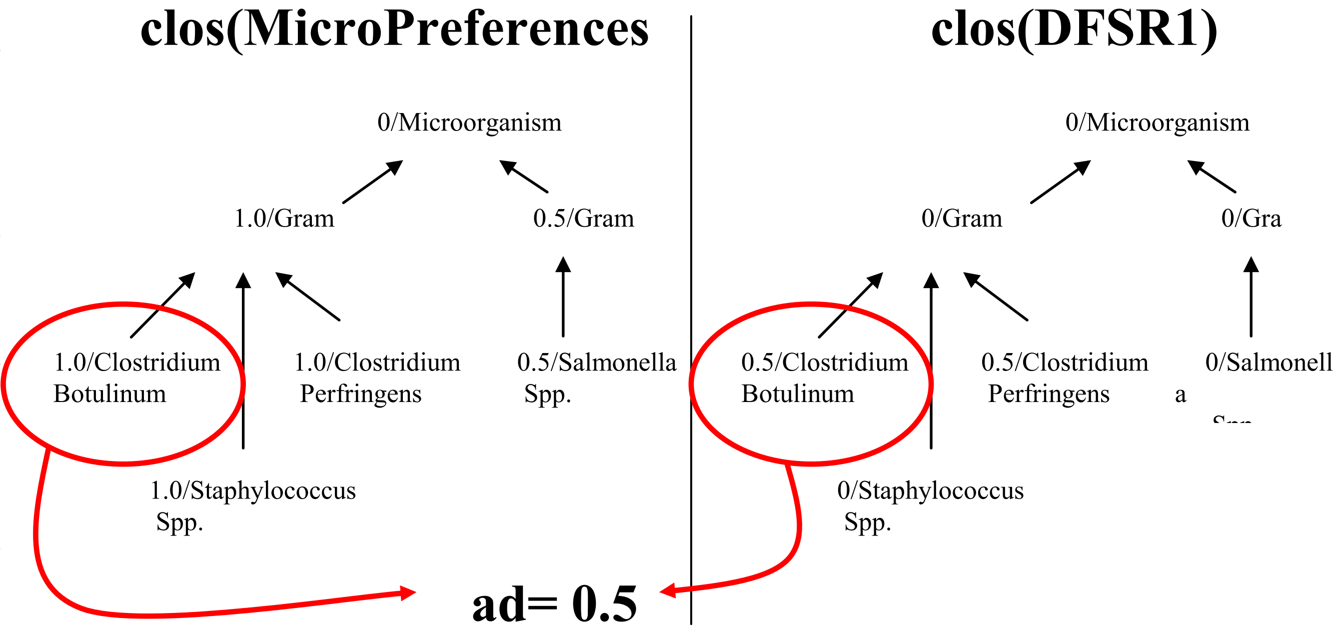
Figure 4

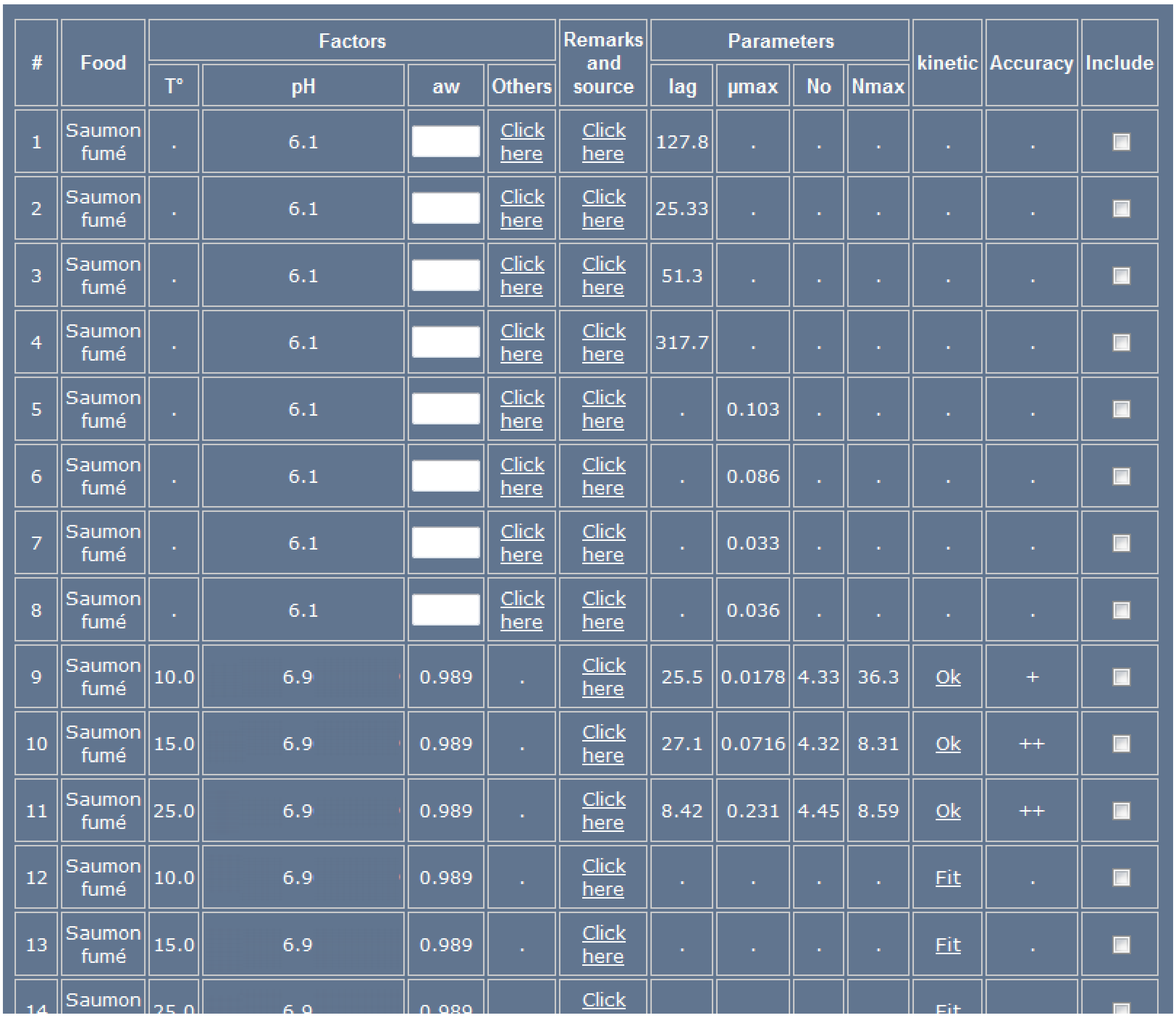
Figure 5



Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI : 10.1016/j.fm.2010.07.002

Figure 6


clos(MicroPreferences
0/Microorganism
1.0/Gram
0.5/Gram
1.0/Clostridium Botulinum
1.0/Clostridium Perfringens
0.5/Salmonella Spp.
1.0/Staphylococcus Spp.
ad= 0.5
clos(DFSR1)
0/Microorganism
0/Gram
0/Gra
0.5/Clostridium Botulinum
0.5/Clostridium Perfringens
0/Salmonell a
0/Staphylococcus Spp.



#	Food	Factors				Remarks and source	Parameters				kinetic	Accuracy	Include
		T°	pH	aw	Others		lag	µmax	No	Nmax			
1	Saumon fumé	.	6.1		Click here	Click here	127.8	.	.	.	.	.	☐
2	Saumon fumé	.	6.1		Click here	Click here	25.33	.	.	.	.	.	☐
3	Saumon fumé	.	6.1		Click here	Click here	51.3	.	.	.	.	.	☐
4	Saumon fumé	.	6.1		Click here	Click here	317.7	.	.	.	.	.	☐
5	Saumon fumé	.	6.1		Click here	Click here	.	0.103	.	.	.	.	☐
6	Saumon fumé	.	6.1		Click here	Click here	.	0.086	.	.	.	.	☐
7	Saumon fumé	.	6.1		Click here	Click here	.	0.033	.	.	.	.	☐
8	Saumon fumé	.	6.1		Click here	Click here	.	0.036	.	.	.	.	☐
9	Saumon fumé	10.0	6.9	0.989	.	Click here	25.5	0.0178	4.33	36.3	Ok	+	☐
10	Saumon fumé	15.0	6.9	0.989	.	Click here	27.1	0.0716	4.32	8.31	Ok	++	☐
11	Saumon fumé	25.0	6.9	0.989	.	Click here	8.42	0.231	4.45	8.59	Ok	++	☐
12	Saumon fumé	10.0	6.9	0.989	.	Click here	.	.	.	.	Fit	.	☐
13	Saumon fumé	15.0	6.9	0.989	.	Click here	.	.	.	.	Fit	.	☐
14	Saumon	25.0	6.9	0.989		Click					Fit		☐

Figure 7

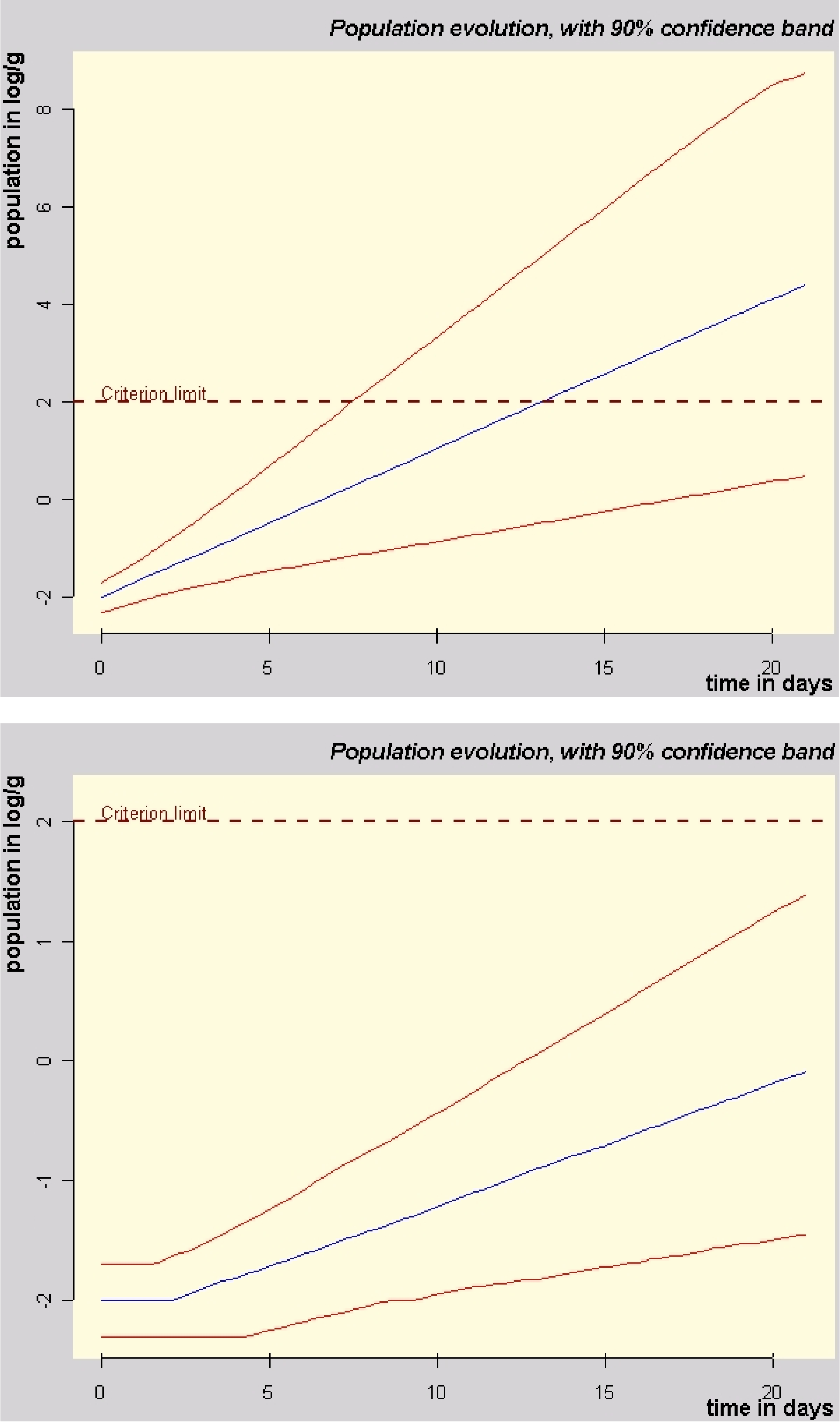

Population evolution, with 90% confidence band
population in log/g
8
6
4
2
0
-2
Criterion limit
0
5
10
15
20
time in days
Population evolution, with 90% confidence band
population in log/g
2
1
0
-1
-2
Criterion limit
0
5
10
15
20
time in days

Figure 8

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

Table 2: Cardinal values (growth boundaries).

Organism	**a_W minimum**	**a_W optimum**	**a_W maximum**
Clostridium	0.943	0.95-0.96	0.97
Staphylococcus	0.88	0.98	0.99
Salmonella	0.94	0.99	0.991

Table 3: Evaluation of query results

Queried relation	**Selection criteria**	**Precision-recall**	**Nb of answer graphs**
Lag Time	Microorganism=*L. monocytogenes*	100%-100%	47 graphs
Lag Time	Microorganism=*P. fluorescens*	100%-100%	29 graphs
Growth kinetics	Microorganism=*E. coli*	100%-100%	39 graphs
Lag Time	FoodProduct= Egg salad	50%-100%	24 graphs
Growth kinetics	FoodProduct= Salad	54%-100%	26 graphs

Comment citer ce document :
Buche, P., Couvert, O., Dibie, J., Hignette, G., Mettler, E., Soler, L. (2011). Flexible querying of web data to simulate bacterial growth in food. Food Microbiology, 28 (4), 685-693. DOI : 10.1016/j.fm.2010.07.002