IUFoST 2006
DOI: 10.1051/IUFoST:20060659

Fuzzy Semantic Approach for Data Integration Applied to Risk in Food: an Example about the Cold Chain

Hignette, G[1], **Buche, P.**[2], **Dervin, C.**[2],
Dibie-Barthélemy, J.[1], **Haemmerlé, O.**[3] **and Soler, L.**[2]

[1]INA P-G, UER Informatique, 16, rue Claude Bernard, 75 231 Paris Cedex 05, France
hignette@inapg.fr, dibie@inapg.fr
[2]INRA MIA, Unité Mét@risk UR 1204, 16, rue Claude Bernard, 75 231 Paris Cedex 05, France
buche@inapg.fr, dervin@inapg.fr, lydie.soler@inapg.fr
[3]GRIMM-ISYCOM, Université Toulouse le Mirail, Département de Mathématiques-Informatique, 5, Allées Antonio Machado, F-31058 Toulouse Cedex, France
ollivier.haemmerle@univ-tlse2.fr

Abstract: A preliminary step to risk in food assessment is to gather experimental data. During the Sym'Previus project, we have designed a system for the integration of experimental data on food microbiology. Data provided by industrial partners and data extracted from experimental research results published in the main scientific journals of the domain are stored into a database. For that, data are indexed by means of a predefined vocabulary, called ontology.

An important characteristic of the data in the database is their incompleteness. To enlarge the number of answers given to a user query, we have built an extended querying system called MIEL, which allows the user to retrieve the nearest data to his selection criteria. In this paper we detail another solution to deal with data scarcity, which consists in searching data on the Web to complement the database. Data tables that contain, in general, a synthesis of experimental data published in the documents, are automatically extracted from documents found on the web. The data tables are then automatically annotated with terms of the ontology and stored into the database. Each term belonging to a given data table is associated with a fuzzy set defined by the set of nearest terms belonging to the ontology weighted by a similarity score. Two complementary scores, which take into account "semantic" information provided by the ontology, are presented in this paper. This process, called fuzzy semantic annotation, permits the MIEL user to query the database, including data tables retrieved from the Web, using the same vocabulary (the ontology).

We present an example of a query which can be useful in a cold chain study concerning E. coli O157 on ground beef.

Keywords: data integration, semantic annotation

Introduction

To assess the risk of micro-organisms developing in foods, we first need to gather experimental data about the growth of micro-organisms in foods as influenced by factors such as temperature, water activity or pH. We have designed a full data integration system within the framework of the French national project Sym'Previus[1] which was dedicated to building tools in predictive microbiology. Data provided by industrial partners and data extracted from experimental results published in the main scientific journals of the domain are stored in a database.

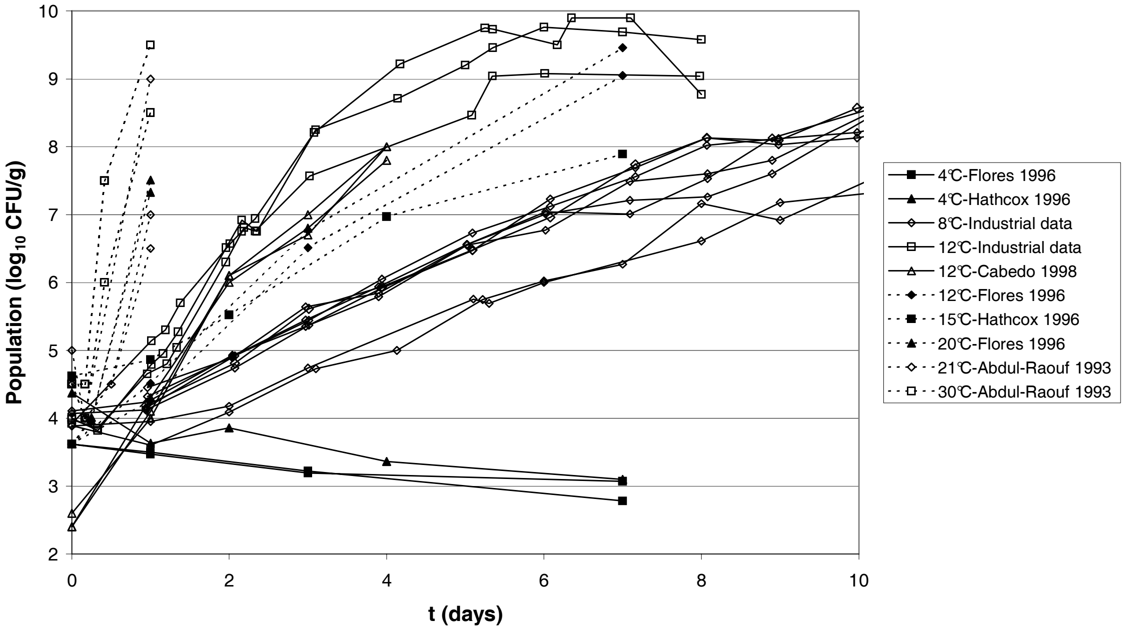
[1] Website http://www.symprevius.org/

Our aim is to provide a user querying the database with a compilation of data obtained from different sources so that he can assess the risks using the greatest quantity of available information about the micro-organism and the food product he is interested in.

Figure 1 presents an example of data which can be useful in a cold chain study concerning E. coli O157 on ground beef. Evidence from different sources can be used to draw simulations of growth of a micro-organism subjected to different temperature scenarios [Vialette, 2005].

Our main goal is to increase the quantity of data that a user can access when querying the database. The first solution we propose to increase the number of results for a query is an extended querying system, called MIEL, which allows a user to retrieve data that are close to his selection criteria, even if they do not exactly match the query [Buche, 2005]. This solution addresses the fact that not all combinations of micro-organisms, food products and controlled factors have been experimented. The second solution that we detail in this paper addresses another problem, when experiments have been conducted but the data was not stored in the database. Some data are not registered due to the cost of manually feeding the database, but there are also sources of data that are not considered when feeding the database, such as PhD theses, technical reports or course materials. We propose a way of semi-automatically building a data warehouse from documents found on the web, including the types of documents cited above, with as little human effort as possible.

Figure 1: growth kinetics of Escherichia coli O157 on ground beef at different temperature levels



This article first describes the architecture of our system, then defines a way of semantically annotating data from the web and gives experimental results.

Building the data warehouse

In the MIEL system, data about food microbiology are stored in a relational database using a predefined vocabulary to describe the experiments. We call this vocabulary the ontology of the system. It consists of three lists of terms (the food products, the micro-organisms and the controlled factors) that are structured using the "is a kind of" relation. For example, "ground beef" is a kind of "fresh meat". The users can then query the relational database using this ontology. In order to get more answers to one user query, the "is a kind of" links are used to generalize the query and find approximate answers. Documents found on the web that we want to integrate in our system must be annotated using the same ontology. Then the user can keep on querying the database as usual using the ontology, and, on top of the data manually fed into the database, his queries can be extended to the documents found on the web.

Figure 2: the XML data warehouse and its integration into the MIEL++ system

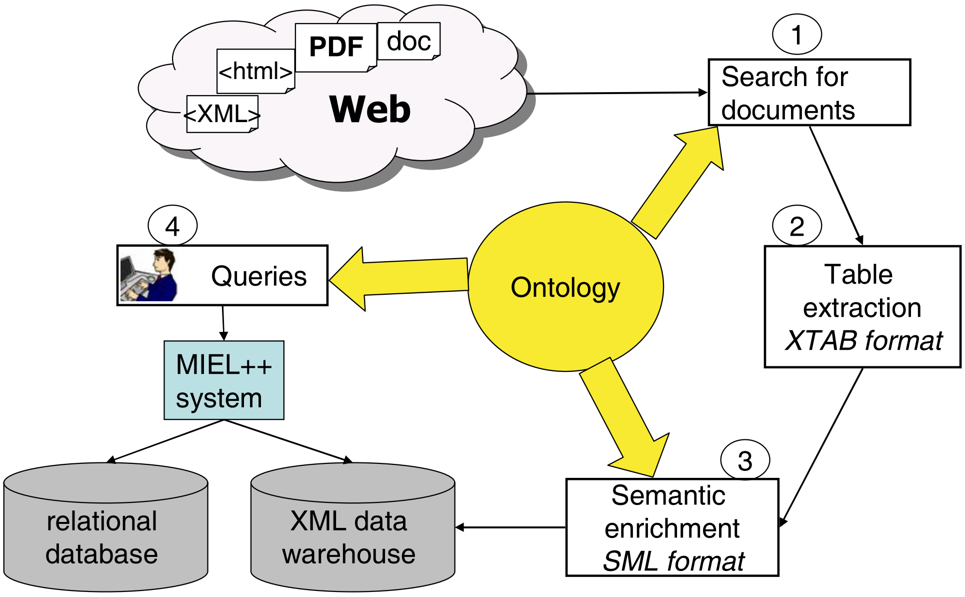


Figure 2 illustrates the steps we follow when building the data warehouse containing the data found on the web.

Terms of the ontology are used as keywords in a web crawler to find documents relevant to the food microbiology domain. We only retain documents in PDF format containing keywords such as "abstract" and "references", which helps in finding scientific papers.

The data tables are automatically extracted from the documents. We concentrate our work on tables because they usually contain the synthesis of experimental data. The tables are extracted into XTab format, a format based on XML to represent tables [Saïs, 2005]. We chose the XML format because it is a standard interchange format and it allows great flexibility, which is important because the structure of data found on the web is very heterogeneous.

Those tables are semantically enriched, original terms from the table being annotated with their equivalent in the ontology. The XTab format with added annotations is the SML (Semantic Markup Language) format [Saïs, 2005], still based on XML. The tables in SML are then stored in a data warehouse.

The MIEL system extended to the XML data warehouse is called MIEL++. When users query the MIEL++ system using the ontology, the relational database and the data warehouse are queried simultaneously. The answers to the query, coming from the relational database (manually fed from published papers or industrial data) and from the data warehouse (semi-automatically fed from the internet), are presented in two separate sets which are ordered according to their relevance to the query.

Fuzzy Semantic Annotation

For our data warehouse to be usable in the MIEL++ system, we want to annotate terms from the web with their equivalent in the ontology. For this purpose, words composing a term are separated, and each word is lemmatized. For example “carrot cuts” and “cut carrots” become the same set {“carrot”, “cut”}. Then each term of the original table extracted from the web is compared word by word with each term of the ontology. We keep terms from the ontology which have at least one word in common with the term from the web as candidates for annotating this term. The challenge is then to discriminate among several candidates. Consider for example the term “minced beef” found on the web. It does not correspond exactly to any term of the ontology, but word by word comparison matches with “minced poultry” and “ground beef”. Therefore we designed a fuzzy semantic annotation, in which terms from the web are annotated with several possible terms from the ontology, ordered by their similarity to the original term.

At first, this similarity was computed according to the number of words in common between the term from the web and the term from the ontology. But such a similarity does not discriminate between “minced poultry” and “ground beef” to annotate “minced beef”, even though “ground beef” is better. For this reason, we developed the notion of word weights. For each term in the ontology, each word is assigned a weight according to its importance in the meaning of the term. For example, in “minced poultry”, we consider that the most important fact is that the meat considered is “poultry”, and that “minced” is a minor word. Weights between 0 and 1 are given to each word in a term. To simplify the manual definition of weights of words in the ontology, we restrict the weights to three values:

- “empty” words (articles, conjunctions…) are predefined in a list and given the weight 0;
- “major” words which are essential to the semantic of the term are given the weight 1;
- “minor” words which only slightly modify the meaning of the term, are given the weight 0.2 (this value was chosen after a small preliminary experiment).

For example, weight(minced poultry, minced)=0.2 and weight(minced poultry, poultry)=1. As for the terms from the web, all words are given a weight of 1 because those terms, contrary to the ones from the ontology, are automatically extracted so we do not know their real meaning.

Given those weights, we define the similarity measure between a term W={w1, …wi} from the web and a term O={o1, …oj} from the ontology as the ratio between the sum of weights of words in common in both terms and the total weights of words in both terms:

$$sim(W,O)=\frac{\sum_{c\in C} weight(W,c)+weight(O,c)}{\sum_{k=1}^{i} weight(W,w_k)+\sum_{k=1}^{j} weight(O,o_k)}$$

with C the set of words in common in both terms.

For example, the similarity between “minced beef” and “minced poultry” is

$$\frac{weight(\text{minced\_beef}, \text{minced}) + weight(\text{minced\_poultry}, \text{minced})}{totalWeights(\text{minced\_beef}) + totalWeights(\text{minced\_poultry})} = \frac{1+0.2}{2+1.2} = 0.375$$

whereas the similarity between “minced beef” and “ground beef” is

$$\frac{weight(\text{minced\_beef}, \text{beef}) + weight(\text{ground\_beef}, \text{beef})}{totalWeights(\text{minced\_beef}) + totalWeights(\text{ground\_beef})} = \frac{1+1}{2+1.2} = 0.625$$

So, a fuzzy annotation would show that we have several matches, but that we trust better “ground beef” than “minced poultry” to annotate “minced beef”. In the MIEL++ system, a user querying on the term “ground beef” would also be given data about minced beef. However, answers about minced beef would be presented after the ones about ground beef, according to the similarity to the query.

Experimental results

To assess our annotation system, we have extracted tables from scientific papers on food microbiology and listed the 185 distinct food products occurring in those tables. We have used two different food ontologies: the food product part of the ontology used in the MIEL system, and the Codex Alimentarius (used by the World Health Organization). Both ontologies were manually reviewed to give weights to each word in each term. Then each term from the web was manually compared to each ontology: the term chosen in each ontology to represent the term from the web is called the “best match”.

Each term from the web was then automatically annotated, twice with both ontologies: once with the weights on words of the ontology, once with weights on minor and major words set to 1 (as if all words were equally important, except for empty words, the weight of which is still 0), in order to assess the gain obtained by assigning weights. Automatic annotation was then compared with the manual definition of the “best match”.

Table 1: Experimental results

Terms from the web for which:	best match has the best similarity score	best match is within the five best similarity scores	best match is within the annotations
using the Codex Alimentarius ontology			
weighted words	34%	52%	60%
weights set to 1	30%	46%	
using the MIEL ontology			
weighted words	46%	65%	78%
weights set to 1	45%	61%	

Table 1 presents the results of our experiments. The terms of the ontology that are used for annotation are chosen based on the fact that they have a word in common with the term from the web. Using weights on the words in the ontology does not change the terms chosen for annotation, but it changes the ordering of the annotations. In all cases, using weights on words in the ontology allows a better annotation, in which the best match comes in a better place according to the similarity measure. Using word weights on the MIEL ontology allows up to 65% of the terms from the web to be recognized and have their best match within the first five positions. The MIEL ontology allows better results than the Codex Alimentarius.

This can be explained: the MIEL ontology was built on purpose for food microbiology and contains names of transformed products, whereas the Codex Alimentarius is focused on the origin of food (for example, we would not find “butter” but “cow milk fat”). Our experimental results also show why we decided to keep a fuzzy annotation instead of keeping only the term of the ontology with the best similarity to the original term: often the best match does not have the best similarity but is among the first annotations proposed by the system. It is then possible to present a short-list to a human annotator, who can chose the right term of the ontology to annotate the term from the web.

We also conducted experiments with two other well-known definitions of similarity, the Dice coefficient and the cosine similarity [Lin, 1998], and we found similar results.

Conclusion

We have designed a system to semi-automatically annotate terms from the web with terms predefined in an ontology. This will allow the capture of much more experimental data for food microbiology than is possible with a manually fed database.

We are currently working on recognizing not only the food products and micro-organisms presented in a table, but also what kind of semantic relations are given in the table (for example, growth rate of a micro-organism at different temperatures, or minimal and maximal growth pH, etc.) We would then be able to present results such as in figure 1, in which the quantitative results are recognized and presented as a graph.

Another research axis for our team is a better integration between the database that contains experimental results entirely entered by a human, and the data warehouse that contains data which are automatically annotated. We need to take into account the fact that we do not have the same confidence in both kinds of data, unless the automatic annotations are manually validated.

References

[Abdul-Raouf, 1993] Abdul-Raouf, U. M., Beuchat, L. R., Ammar, M. S. (1993). *Survival and growth of Escherichia coli O157:H7 in ground, roasted beef as affected by pH, acidulants, and temperature.* Applied and Environmental Microbiology, 59, 2364-8

[Buche, 2005] Buche P., Dervin C., Haemmerlé O., Thomopoulos R., (2005) *Fuzzy querying on incomplete, imprecise and heterogeneously structured data in the relational model using ontologies and rules.* IEEE Transactions on fuzzy systems, 13 (3): 373-383.

[Cabedo, 1998]Cabedo, L., Sofos, J. N., Smith, G. C. (1998). *Bacterial growth in ground beef patties made with meat from animals fed diets without or with supplemental vitamin E.* Journal of Food Protection, 61, 36-40.

[Flores, 1996] Flores, L. M., Sumner, S. S., Peters, D. L., Mandigo, R. (1996). *Evaluation of a phosphate to control pathogen growth in fresh and processed meat products.* Journal of Food Protection, 59, 356-359

[Hathcox, 1996]Hathcox, A. K., Beuchat, L. R. (1996). *Inhibitory effects of sucrose fatty acid esters, alone and in combination with ethylenediaminetetraacetic acid and other organic acids, on viability of Escherichia coli O157:H7.* Food Microbiology, 13, 213-225.

IUFoST 2006
DOI: 10.1051/IUFoST:20060659

[Lin, 1998] Lin, D. (1998). *An information-theoretic definition of similarity.* In ICML ’98: Proceedings of the Fifteenth International Conference on Machine Learning (pp. 296–304). San Francisco, CA, USA

[Saïs, 2005] Saïs, F., Gagliardi, H., Haemmerlé, O., & Pernelle, N. (2005). *Enrichissement sémantique de documents SML représentant des tableaux.* In Extraction et Gestion des Connaissances (EGC’2005) volume II - RNTI-E-3 (pp. 407–418).

[Vialette, 2005] Vialette M., Pinon A., Leporq B., Dervin C., Membre J-M. (2005). *Meta analysis of food safety information based on a combination of a relational database and a predictive modelling tool.* Risk Analysis, 25 : 75-83.