EVALUATING DATA RELIABILITY: AN EVIDENTIAL ANSWER WITH APPLICATION TO A WEB-ENABLED DATA WAREHOUSE

[1]K.Raja, [2]R.Saravanakumar

[1]Research Scholar, Department of Information Technology, Jayam College of Engineering and Technology, Dharmapuri

[2]Associate Professor/IT, Department of Information Technology, Jayam College of Engineering and Technology, Dharmapuri

[1]sakthivelvairam@gmail.com, [2]saravanakumar.surya@gmail.com

Abstract: We introduce a new correction scheme, which takes into account uncertain Meta knowledge on the source's relevance and truthfulness and that generalizes Shafer's discounting operation. We then show how to reinterpret all connectives of Boolean logic in terms of source behavior assumptions with respect to relevance and truthfulness. We are led to generalize the un normalized Dempster's rule to all Boolean connectives, while taking into account the uncertainties pertaining to assumptions concerning the behavior of sources. However, data reliability and confidence are essential components of a data warehousing system, as they influence subsequent retrieval and analysis. In this paper, we propose a generic method to assess data reliability from a set of criteria using the theory of belief functions. Customizable criteria and insightful decisions are provided. The chosen illustrative example comes from real-world data issued from the Sym'Previus predictive microbiology oriented data warehouse.

1. INTRODUCTION

The growth of the web and the emergence of dedicated data warehouses offer great opportunities to collect additional data, be it to build models or to make decisions. The reliability of these data depends on many different aspects and meta information: data source, experimental protocol,... Developing generic tools to evaluate this reliability represents a true challenge for the proper use of distributed data. In classical statistical procedures, a preprocessing step is generally done to remove outliers. In procedures using web facilities and data warehouses, this step is often omitted, implicit or simplistic. There are also very few works that propose a solution to evaluate data reliability. It is never- the less close to other notions that have received more attention, such as We propose a method to evaluate data reliability from Meta information. Several criteria are used, each one providing a piece of information about data reliability. These pieces are then aggregated into a global assessment that is sent back, after proper post-treatment, to the end user. In our opinion, such a method should

- deal with conflicting information, as different criteria may provide conflicting information about the reliability. For example, data may come from a reputed journal, but have been collected with rather unreliable instruments;
- be traceable, as it is important to be able to detect conflict and to provide insights about its origins, or in the absence of such conflicts, to know why such data have been declared poorly (or highly) reliable;
- be readable, both in its different input parameters and results, as the method and the system it is implemented in will be used mainly by non-computer scientists.

The method presented here answers these needs, by addressing two issues: first we propose a generic approach to evaluate global reliability from a set of criteria, second we consider the problem of ordering the reliability assessments so that they are presented in a useful manner to the end users. Indeed, the goal of the present work is to propose a partly automatic decision-support system to help in a data selection process.

As evaluating data reliability is subject to some uncertainties, we propose to model information by the means of evidence theory, for its capacity to model uncertainty and for its richness in fusion operators. Each criterion value is related to a reliability assessment by the means of fuzzy sets later transformed in basic belief assignments, for the use of fuzzy sets facilitates expert elicitation. Fusion is achieved by a compromise rule that both copes with conflicting information and provides insights about conflict origins. Finally, interval-valued evaluations based on lower and upper expectation notions are used to numerically summarize the results, for their capacity to reflect the imprecision (through interval width) in the final knowledge. As an application area, we focus on Life Sciences and on reliability evaluation of experimental data issued from arrays in electronic documents.

Section 2 explains what we understand by reliability and discusses related notions and works. Section 3 is dedicated to an analysis of the information available to infer data reliability (with a focus on experimental data). Section 4 describes the method used to model this information and to merge the different criteria using evidence theory. Section 5 addresses the question of data ordering by groups of decreasing reliability and subsequently the presentation of informative results to end users. Section 6 is devoted to the practical implementation of the approach to the case of the @Web data warehouse [2], [3]. It also presents a use case in the field of predictive microbiology.

2. RELATED NOTIONS AND WORKS

In this paper, a source (e.g., expert, sensor,...) is considered as reliable if its information can be used safely (i.e., be trusted), while the information of an unreliable source has to be used with caution (note that information coming from an unreliable source may be true, but nothing guarantees it).

This section makes a short review of topics covered in this paper and of related notions, i.e., how to evaluate reliability, what are the notions related to reliability, and how reliability evaluations should be presented to the end user.

2.1 Reliability Evaluation

In practice, an information source is seldom always right or wrong, and evaluating/modeling the reliability of a source can be complex, especially if source information cannot be compared to a reference value.

In evidence theory, methods to evaluate reliability consist in choosing reliability scores that minimize an error function [4]. In spirit, the approach is similar to the comparison of source assessments with reference values (as done to evaluate experts in probabilistic [5] or possibility [6] methods). It requires the definition of an objective error function and a fair amount of data with a known reference value. This is hardly applicable in our case, as data are sparse and can be collected and stored for later use, i.e., not having a specific purpose in mind during collection. Other approaches rely on the analysis of conflict between source information [7], assuming that a source is more reliable when it agrees with the others. This comes down to make the assumption that the majority opinion is more reliable. If one accepts this assumption, then the results of such methods could possibly complement our approach.

2.2 Related Notions

Reliability has strong connections with other notions such as relevance, truthfulness, trust, and data quality ...All these related concepts are, however, either different from or less specific than the notion of reliability.

First, there is a difference between data reliability, i.e., the trust we can have in data values, and data relevance, termine trust in content provided by a web resource. Naturally these include source authority and direct experience. Among the remaining factors, one can find items like topic and criticality, which are somehow related to data relevance. The limitation of resources may play a role, as well as the incentive to provide good information or on the contrary to be biased or deceptive (elements that are related to the notion of truthfulness). Source agreement and user expertise also have an impact. Some factors that we considered as particularly important in the present work are highlighted, such as citations (through related resources or recommendations), or age/freshness, this last point being very domain-dependent. Another paper [12] advocates a multifaceted approach to trust models in internet environments. The authors point out the great number of terms and intertwined meanings of trust, and the difficulty to capture the wide range of subjective views of trust in single-faceted approaches. They propose an OWL-based ontology of trust related concepts, such as credibility, honesty, reliability, reputation or competency, as well as a metamodel of relationships between concepts. Through domain specific models of trust, they can propose personalized models suited to different needs. The idea is to provide internal trust management systems, i.e., the trust assessment being made inside the system, while using the annotation power of a user community to collect trust data.

Among methods proposing solutions to evaluate trust or data quality in web applications, the method presented in [13] for recommendation systems is close to our proposal, but uses possibility theory as a basis for evaluations rather than belief functions. Another difference between this approach and ours is that global information is not obtained by a fusion of multiple uncertainty models, but by the propagation of uncertain criteria through an aggregation function (e.g., a weighted mean). Each method has its pros and cons: it is easier to integrate criteria interactions in aggregation functions, while it is easier to retrieve explanations of the final result in our approach.

The problem with using available trust evaluation systems on the web is that they are often dedicated to a particular application. They propose relatively simple representation and aggregation tools, notably due to the fact that Semantic Web applications are confronted to scalability issues. Our situation is somehow different, since we aim for a general method applicable to situations where the number of items will seldom exceed tens of thousands, and will in fact be often limited to some dozens.

2.3 Output Post-treatment

In [14], the impact of data quality on decision making is explored, and an experimental study about the consequences of providing various kinds of information (none, two-point ordinal, and interval scale) regarding the quality of data is performed. They point out that the availability of the information is not enough, and that an important consideration is how data quality information is recorded and presented.

Decision tasks (apartment or restaurant selection) were experimented using two groups of subjects, one group performing the tasks without data quality information and the other one with a given quality format. Users were asked to explain their decision process, and issues like complacency, For simple tasks, the smallest level of complacency, corresponding to the greatest impact of data quality information, was observed when comparing groups with interval scaled data quality information with groups with no data quality information. For complex tasks, there seems to be an information overload effect, and no statistically significant conclusions appear. This is an important point in favor of giving a lot of attention in presenting readable results to end users. Interval scaled quality is used in the present paper, together with group ordering by decreasing reliability.

3. FROM WHAT INFORMATION SHOULD WE INFER RELIABILITY

In this section, we present the type of information we have considered to evaluate the reliability of experimental data in Life Science. These criteria are elements that are usually found in publications (reports, papers ...) reporting experimental results. Note that most of these criteria are not specific to Life Sciences, and can be used for any experimental data. The list of criteria is, of course, not exhaustive.

For other popular cases such as touristic data or other applications of the Semantic Web, some criteria used here are universal enough to be valid, but they must be completed by other proper criteria. The approach itself remains generic. Table 1 summarizes the various criteria that can be considered in our applicative context:

- a first group concerns the data source itself. It contains features such as the source type (e.g., scientific publication, governmental report, web-page, ...), the source reputation (e.g., is the laboratory that has produced data known for its reliability), the number of times the source has been cited or the publication date (data freshness being important in Life Sciences, due to rapid evolution of measurement devices and experimental protocol);
- a second group is related to the means used to collect data. Information related to these criteria is typically included in a section called material and method in papers based on experiments in Life Science, which thoroughly describes the experimental protocol and material. Some methods may be known to be less accurate than others, but still be chosen for practical considerations;
- a third group is related to statistical procedures: presence of repetitions, uncertainty quantification, elaboration of an experimental design. These criteria can be reduced or enriched, according to the available information about the data and the relevant features to evaluate reliability

4. METHOD

This section describes the method we propose to evaluate and use reliability. We first describe how information is collected and modeled. Then, we briefly recall the basics of evidence theory needed in this paper.

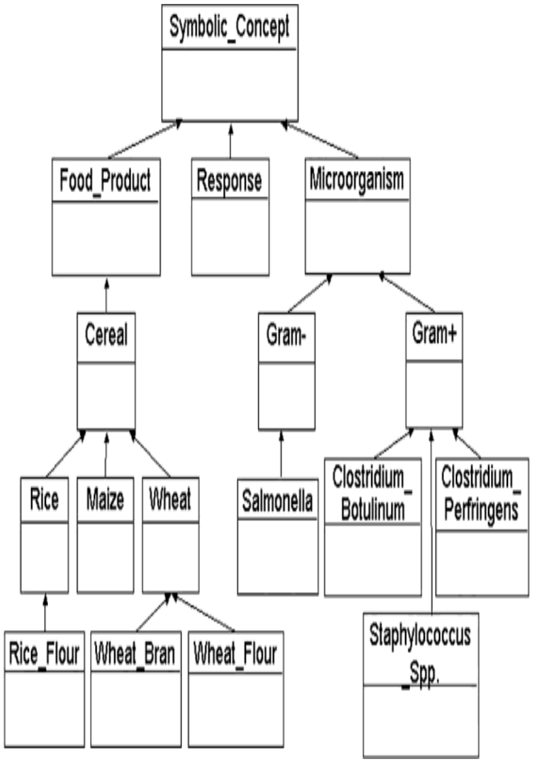

Figure 3: Main steps of the document workflow in @Web.

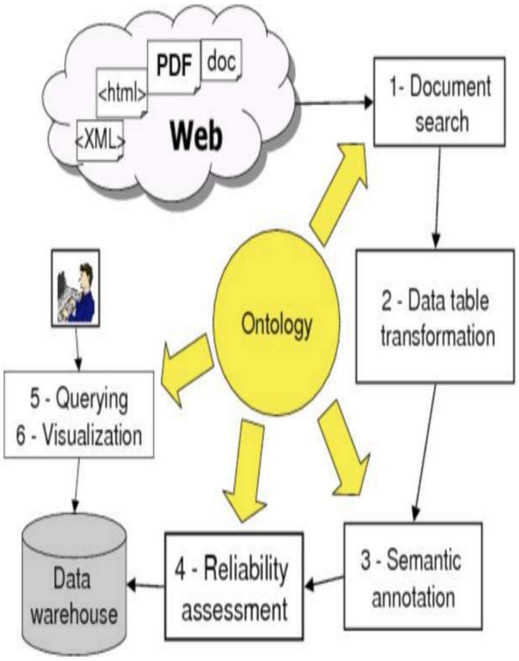

Figure 4: OWL class

Central role played by the domain ontology. This ontology describes the concepts, their terminology, and the relation- ships between concepts proper to a given application domain. Thanks to this feature, @Web can be instantiated for any application domain by defining the corresponding ontology including the domain knowledge. For instance, @Web has already been instantiated and tested in various domains such as food predictive microbiology, chemical risk in food, and aeronautics [3].

Once the ontology is built, data integration in the warehouse is done according to the steps of Fig. 3. Concepts found in a data table and semantic relations linking these concepts are automatically recognized and annotated, which allows interrogation and querying in an homogeneous way.

The @Web instance used here is implemented in the Sym'Previus [32] decision support system which simulates the growth of a pathogenic microorganism in a food product. Semantic relations in this system include, for example, the Growth Rate linking a microorganism and a food product to the corresponding growth rate and its associated parameters. After semantic annotation, data retrieved from tables can be used for various tasks (e.g., estimate a model parameter).

4. WEB GENERIC ONTOLOGY

The current OWL ontology used in the @Web system is composed of two main parts: a generic part, the core ontology, which contains the structuring concepts of the web table integration task, and a specific part, the domain ontology, which contains the concepts specific to the domain of interest. The core ontology is composed of symbolic concepts, numeric concepts and relations between them. It is therefore separated from the definition of the concepts and relations specific to a given domain, the domain ontology. All the ontology concepts are materialized by OWL classes. For example, in the microbiological ontology, the symbolic concept Microorganism and the numeric concept pH are represented by OWL classes that are subclasses of the generic classes Symbolic Concept and Numeric Concept, respectively. Fig. 4 gives an excerpt of an OWL class organization for symbolic concepts.

4.1 Workflow

The first three steps of @Web workflow (see Fig. 3) are the following: the first task consists in retrieving relevant web Fig. 4. Excerpt of OWL class hierarchy for symbolic concepts in the microbial domain.

Documents (in html or pdf) for the application domain, using key words extracted from the domain ontology. It does so by defining queries executed by different crawlers; in the second task, data tables are extracted from the retrieved documents and are semi automatically translated into a generic XML format. The web tables are then represented in a classical and generic way—a table is a set of lines, each line being a set of cells; in the third task, the web tables are semantically annotated according to the domain ontology. This annotation consists in identifying what semantic relations of the domain ontology can be recognized in each row of the web table (see [3] for details). This process generates RDF descriptions.

Example 8. Table 3 is an example of a web table in which the semantic relation Growth Parameter Aw has been identified. The domain of this relation is a Micro organ- ism and its range is food product water activity (aw), a dimensionless value. For instance, the first row indicates that Clostridium water activity (aw) ranges from 0.943 to 0.97, and is known to be optimal in the range [0.95, 0.96].Some of the RDF descriptions associated with web tables by the semantic annotation process include values ex- pressed as fuzzy sets (e.g., aw values). Let us now introduce the use of fuzzy sets in @Web before illustrating it.

Uses of Fuzzy Sets in @Web We distinguish two kinds of fuzzy sets:

- discrete and
- Continuous. Each kind will be used in @Web for specific purposes.

Definition 1: A discrete fuzzy set , denoted by DFS in the RDF description, is a fuzzy set associated with a relation or a symbolic concept of the ontology. Its definition domain is the set of relations or the set of subclasses of the symbolic concept.

Table 3: Example of a Web Table

Organism	a_w min.	a_w optimum.	a_w max.
Clostridium	0.943	0.95_0.96	0.97
Staphylococcus	0.88	0.98	0.99
Salmonella	0.94	0.99	0.991

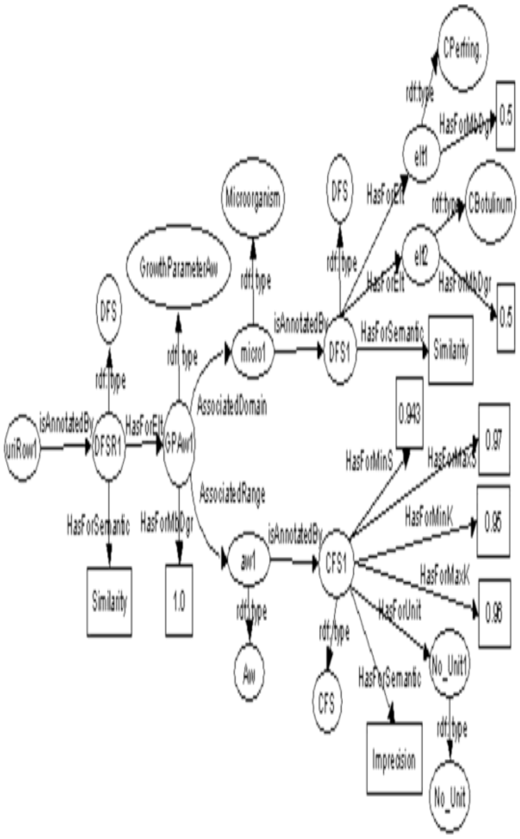

Figure 5: Example of RDF annotations generated from the web Table 3

We denote by fðx; yÞ;.. .g the fact that element x has membership degree y.

Definition 2: A continuous fuzzy set, denoted by CFS in the RDF description, is a trapezoidal fuzzy set associated with a numeric concept of the ontology. A trapezoidal fuzzy set is defined by its four characteristic points ½a; b; c; d which correspond to its support ½a; d and its kernel ½b; c (with a linear interpolation between a; b and c; d). Its definition domain is the interval of possible values for the concept.

The fuzzy values used to annotate web tables may express two of the three classical semantics of fuzzy sets (see [33]): similarity or imprecision. In the @Web system, similarity interpretation is used to recognize symbolic concepts and relations inside the table, while imprecision interpretation is used when modeling some ill-known values about some particular instances of numerical concepts.

Example 9: Fig. 5 gives the main part of the RDF description corresponding to the recognition of the relation Growth- ParameterAw in the first row of the web Table 3, denoted by uriRow1 in Fig. 5. Starting from the left part of the figure, the row is annotated by a discrete fuzzy set DFSR1. A list of closest relations is extracted from within the ontology, in the present case a single element corresponding to the relation Growth ParameterAw (GPaw1). The membership degree (1 here) is a certainty score, denoted ps, expres- sing the degree of certainty associated with the relation recognition. In the top right part of the figure, the domain of the relation Growth ParameterAw, an instance of the symbolic concept Microorganism, is annotated by a discrete fuzzy set.

This fuzzy set, typed by the OWL class DFS, has semantics of similarity and gives the list of closest ontology concepts compared to Clostridium (First row of Table 3). Starting from the end, we see two instances of symbolic concepts, CPerfring for Clostridium Perfringens, and CBotulinum for Clostridium Botulinum, with membership degree equal to 0.5 for each of them.

Finally one can see the use of continuous fuzzy sets, like CFS1 on the bottom right, to express numerical values associated with the range of the relation Growth- ParameterAw.

4.2 SPARQL Querying of RDF Graphs

In the XML/RDF data warehouse, the querying is done through MIEL++ queries. We briefly recall how MIEL++ queries are executed in the current version of @Web (see [2] for details). A MIEL++ query is asked in a view which corresponds to a given relation of the ontology (e.g., the relation Growth ParameterAw of example 8). A MIEL++ query is an instantiation of a given view by the end user, specifying among the set of queryable attributes of the view, which are the selection attributes (i.e., the one used to select relevant answers) and their corresponding searched values, and which are the projection attributes (i.e., the one displayed in the answers).

In such MIEL++ queries, fuzzy sets allow representing end-user preferences (the third semantic [33] of fuzzy sets) and are used to retrieve not only exact answers but also answers which are semantically close (kernel matching versus support matching). Since the XML/RDF data ware- house contains fuzzy values generated by the annotation process, the query processing has 1) to consider the certainty score associated with the semantic relations identified in web tables and 2) to compare a fuzzy set expressing querying preferences to a fuzzy set, generated by the annotation process, having a semantics of similarity or imprecision.

Example 10. Let us define a MIEL++ query Q expressed in the view Growth ParameterAw as follows:
Q¼fMicroorganism;awjðGrowthParameterAwðMicroorganism; awÞ
^ ðMicroorganism MicroPreferencesÞ
^ ðaw awPreferencesÞ^ ðps 0:5Þg:

In Q, the projection attributes are Microorganism and aw, while the second part describes selection attributes. The discrete fuzzy set Micro Preferences, The discrete fuzzy

5. CONCLUSION AND PERSPECTIVES

We proposed a generic method to evaluate the reliability of data automatically retrieved from the web or from electronic documents. Even if the method is generic, we were more specifically interested in scientific experimental data.

The method evaluates data reliability from a set of common sense (and general) criteria. It relies on the use of basic probabilistic assignments and of induced belief functions, since they offer a good compromise between flexibility and computational tractability. To handle conflicting information while keeping a maximal amount of it, the information merging follows a maximal coherent subset approach. Finally, reliability evaluations and ordering of data tables are achieved by using lower/upper expectations, allowing us to reflect uncertainty in the evaluation. The results displayed to end users is an ordered list of tables, from the most to the least reliable ones, together with an interval-valued evaluation.

We have demonstrated the applicability of the method by its integration in the @Web system, and its use on the Sym’Previus data warehouse. As future works, we see two main possible evolutions:

- complementing the current method with useful additional features: the possibility to cope with multiple experts, with criteria of non-equal importance and with uncertainly known criteria;
- Combining the current approach with other notions or sources of information: relevance, in particular, appears to be equally important to characterize experimental data. Also, we may consider adding user feedback as an additional (and parallel) source

of information about reliability or relevance, as it is done in web applications.

REFERENCES

[1] S.Ramchurn, D. Huynh, and N. Jennings, "Trust in Multi-Agent Systems," The Knowledge Eng. Rev., vol. 19, pp. 1-25, 2004.

[2] P.Buche,.J. Dibie-Barthe´ lemy, and H. Chebil, "Flexible Sparql Querying of Web Data Tables Driven by an Ontology," Proc. Eighth Int'l Conf. Flexible Query Answering Systems (FQAS), pp. 345-357, 2009.

[3] G. Hignette, P. Buche, J. Dibie-Barthe´ lemy, and O. Haemmerle´ , "Fuzzy Annotation of Web Data Tables Driven by a Domain Ontology," Proc. Sixth European Semantic Web Conf. The Semantic Web: Research and Applications (ESWC), pp. 638-653, 2009.

[4] D. Mercier, B. Quost, and T. Denoeux, "Refined Modeling of Sensor Reliability in the Bellief Function Framework Using Contextual Discounting," Information Fusion, vol. 9,pp. 246-258,2008.

[5] R. Cooke, Experts in Uncertainty. Oxford Univ. Press, 1991.

[6] S. Sandri, D. Dubois, and H. Kalfsbeek, "Elicitation, Assessment and Pooling of Expert Judgments Using Possibility Theory," IEEE Trans. Fuzzy Systems, vol. 3, no. 3, pp. 313-335, Aug. 1995.

[7] F. Delmotte and P. Borne, "Modeling of Reliability with Possibility Theory," IEEE Trans. Systems, Man, and Cybernetics A, vol. 28, no. 1, pp. 78-88, 1998.

[8] F. Pichon, D. Dubois, and T. Denoeux, "Relevance and Truthful- ness in Information Correction and Fusion," Int'l J. Approximate Reasoning, vol. 53, pp. 159-175, 2011.

[9] J. Sabater and S. Sierra, "Review on Computational Trust and Reputation Models," Artificial Intelligence Rev., vol. 24, pp. 33-60,2005

[10] J. Golbeck and J. Hendler, "Inferring Reputation on the Semantic Web," Proc. 13th Int'l World Wide Web Conf., 2004.

[11] Y. Gil and D. Artz, "Towards Content Trust of Web Resources,"Proc. 15th Int'l Conf. World Wide Web (WWW '06), pp. 565-574, 2006.

[12] K. Quinn, D. Lewis, D. O'Sullivan, and V. Wade, "An Analysis of Accuracy Experiments Carried Out over a Multi-Faceted Model of Trust," Int'l J. Information Security, vol. 8, pp. 103-119, 2009.

[13] A. Denguir-Rekik, J. Montmain, and G. Mauris, "A Possibilistic- Valued Multi-Criteria Decision-Making Support for Marketing Activities in E-Commerce: Feedback Based Diagnosis System," European J. Operational Research, vol. 195, no. 3, pp. 876-888, 2009.

[14] I.N. Chengalur-Smith, D.P. Ballou, and H.L. Pazer, "The Impact of Data Quality Information on Decision Making: An Exploratory Analysis," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 853-864, Nov./Dec. 1999.

[15] L. Zadeh, "The Concept of a Linguistic Variable and Its Application to Approximate Reasoning-i," Information Sciences,