International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

The Survey Paper on Evaluating Data Reliability: An Evidential Answer with Application to a Web-Enabled Data Warehouse

Saloni Shah[1], Vina M. Lomte[2]

[1]Pune University, M.E., Department of Computer Engineering, RMD Sinhgad School of Engineering, Warje, Pune-58, India
[2]Pune University, Assistant Professor, Department of Computer Engineering, RMD Sinhgad School of Engineering, Warje, Pune-58, India

Abstract: ***There are many available methods to integrate information source reliability in an uncertainty representation, but there are only a few works focusing on the problem of evaluating this reliability. 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.***

Keywords: Belief functions, evidence, information fusion, confidence, maximal coherent subsets, trust, data quality, relevance

1. Introduction

There are many available methods to integrate information source reliability in an uncertainty representation, but there are only a few works focusing on the problem of evaluating this reliability. However, data reliability and confidence are essential components of a data warehousing system, as they influence subsequent retrieval and analysis. Aim of this project is a generic method to assess data reliability from a set of criteria using the theory of belief functions and domain ontology. Customizable criteria and insightful decisions are provided. The goal of the present work is to propose a partly automatic decision-support system to help in a data selection process.

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 nevertheless close to other notions that have received more attention, such as trust [1]. 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 are an ordered list of tables, from the most to the least reliable ones, together with an interval-valued evaluation. 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.

Licensed Under Creative Commons Attribution CC BY

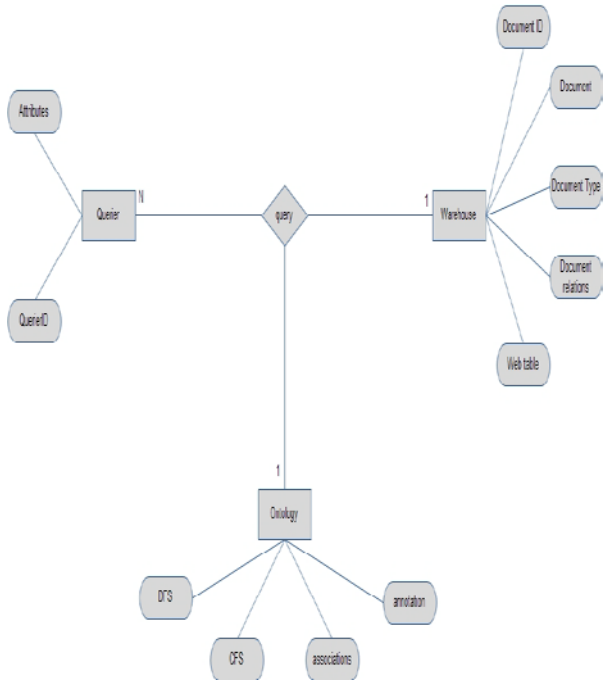

Figure 1: ER Diagram of the system

2. Literature Review

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 [2]. In spirit, the approach is similar to the comparison of source assessments with reference values (as done to evaluate experts in probabilistic [3] or possibilistic [4] 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 [5], 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.

Another paper [6] 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 Meta model 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 [7] 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. In [8], 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.

3. What Kind of Information Should We Considered to Evaluate the 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:

Table 1: Reliability Criteria

Source	Production	Statistics
Type	Protocol	Repetitions
Reputation	Material	Uncertainty quantification
Citation count		(variance, confidence interval)
Publication date		Experimental design

3.1. A first group concerns the data source itself. It contains features such as the source type (e.g., scientific publication, governmental report, webpage . . .), 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);

3.2. A second group is related to the means used to collect data. Information related to these criteria is

Licensed Under Creative Commons Attribution CC BY

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;

3.3. A third group is related to statistical procedures: presence of repetitions, uncertainty quantification, and 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. Required Notation and Information

We assume that reliability takes its value on a finite ordered space $\Theta = \{\theta_{1,\ldots,}\theta_N\}$
Where,
N is any odd number
$\exists\ \theta_i < \theta_j$ iff $i < j$
θ_1 corresponds to total unreliability
θ_N corresponds to total reliability
Natural Values Standing in middle of Θ
Denoted by $I_{a,b} = \{\ \theta_{\mathbf{a},\ldots,}\theta_{\mathbf{b}}\ \}$

Customized Criteria

The evaluation is based on reliability criteria and it is evaluated by source, production and statistics parameters

Let say,
S – is set of a criteria group
$S = \{A_{1,\ldots,}A_S\}$
$\forall\ A_i$, i = 1, ….., S
Where,
A_i is a finite space, representing individual criteria

Fuzzy Set Theory

Fuzzy set is use on the criteria groups to calculate the reliability state in linguistic term by experts.
L – is a set of linguistic terms representing the state of data reliability
L = {very unreliable, slightly unreliable, neutral, slightly reliable, and very reliable}

$$\begin{cases} E_i = \{\theta \in \Theta \mid \mu(\theta) >= \alpha_i\} = A_{\alpha i} \\ m(E_i) = \alpha_i - \alpha_{i+1} \end{cases}$$

m – is a mass of E_i focal element

Merging Multiple Pieces of Information

Merging of all the masses to obtain global model.

$$\forall E \subseteq \Theta$$

$$m(E) = \sum_{\oplus_{i=1}^{S}(E_i)=E} \prod_{i=1}^{S} m_i(E_i),$$

Where,
F_i – is a focal elements of m_i

Presentation of Data

This step is used to provide readability to resultant data of merging.
Given Set, $D = \{e_1, \ldots, e_d\}$
Where,
d – is representing data
$\int : \Theta \rightarrow$ IR introduced by a bba m_g
The lower expectation - $IEg(\int) = \sum_{Ac\,\Theta} m(A) \min \int(\theta)$

Ordering the data by decreasing reliability

$$D_i = opt(IE, (\{e_{1,} \ldots, e_d\} \setminus \bigcup_{j=0}^{i-1} Dj))$$

5. Application to the Design of a Web-Enabled Data Warehouse

We present an application of the method to@Web, a web-enabled data warehouse. Indeed, the framework developed in this paper was originally motivated byte need to estimate the reliability of scientific experimental results collected in open data warehouses. To lighten the burden laid upon domain experts when selecting data for particular application, it is necessary to give them indicative reliability estimations. Formalizing reliability criteria will hopefully be a better asset for them to justify their choices and to capitalize knowledge than the use of an ad hoc estimation. For this application, numerical evaluations were chosen, the reason being that the initial ad hoc evaluation system proposed numerical evaluations hence users were more at ease with them. Tools development was carefully done using Semantic Web recommended languages, so that created tools would be generic and reusable in other data warehouses. This required an advanced design step, which is important to ensure modularity and to foresee future evolutions.

The current version of @Web has been implemented using the W3C recommended languages (http://www.w3.org/TR/): OWL to represent the domain ontology, RDF to annotate web tables and SPARQL to query annotated web tables. Nevertheless, to show the potential of these tools, we will illustrate the concepts with examples easy to understand. All belief function-related computations have been done thanks to the R [19] package belief [19].

We first describe the purpose and architecture of the data warehouse. We then focus on the extension developed to implement the reliability estimation. Finally, we provide a real-world use case in Food predictive microbiology. It is worthwhile to note that this application provides another good example of the difference between the notions of relevance and reliability: while relevant data are the answers returned by a query, some of these answers may be unreliable. Indeed, data base queries return relevant, but possibly unreliable, answers.

Licensed Under Creative Commons Attribution CC BY

@Web Presentation

@Web is a data warehouse opened on the web [2], [3]. Its current version is centered on the integration of heterogeneous data tables extracted from web documents. The focus has been put on web tables for two reasons: 1) experimental data are often summarized in tables and 2) data are already structured and easier to integrate in a data warehouse than, e.g., text or graphics.
The main steps of web table integration are given in Fig. 2. The key point of data integration in @Web is the central role played by the domain ontology. This ontology describes the concepts, their terminology, and the relationships 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. 2. Concepts found in a data table and semantic relations linking these concepts are automatically recognized and annotated, which allows interrogation and querying in a homogeneous way. The @Web instance used here is implemented in the Sym'Previus [13] 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).

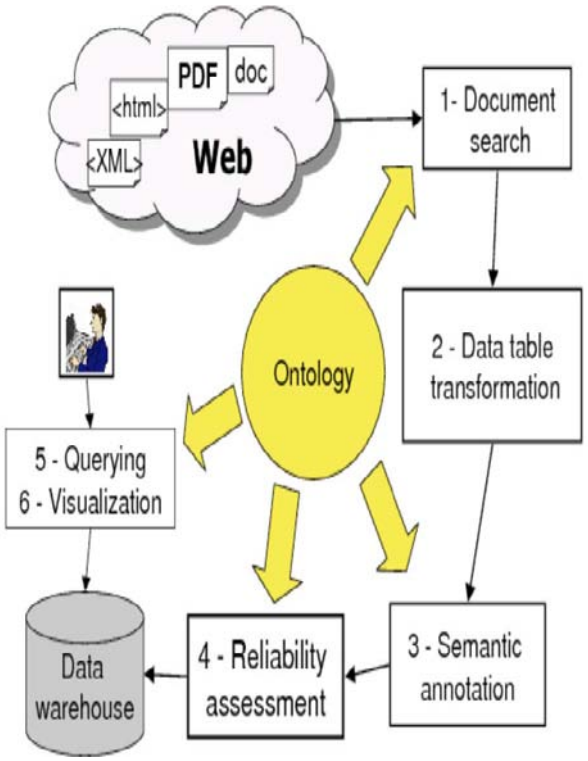

Figure 2: Main steps of the document work flow in @ web [14]

@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. 3 gives an excerpt of an OWL class organization for symbolic concepts.

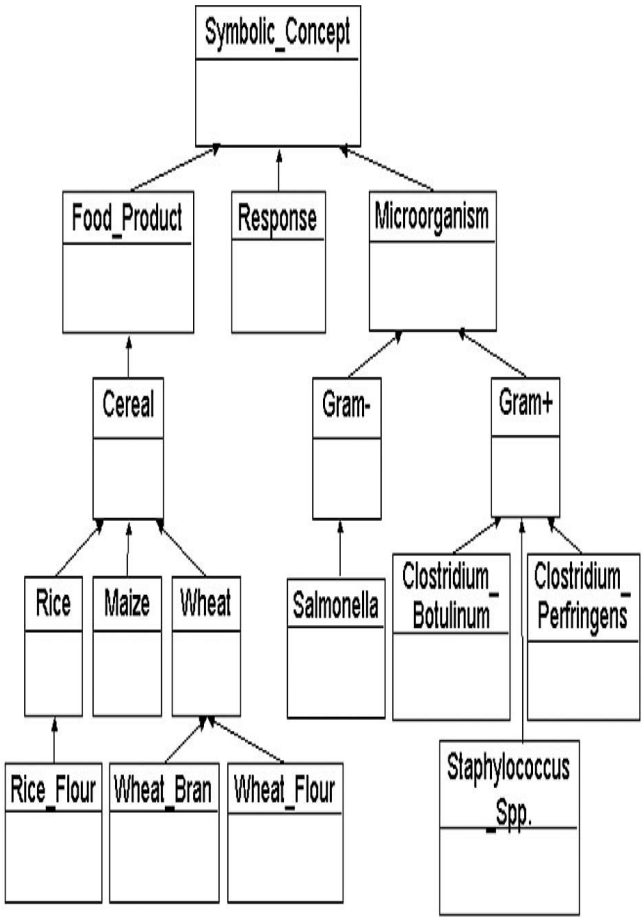

Figure 3: Excerpt of OWL class hierarchy for symbolic concepts in the microbial domain [14]

@Web Workflow

The first three steps of @Web workflow (see Fig. 2) are the following: the first task consists in retrieving relevant web 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

Licensed Under Creative Commons Attribution CC BY

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. This process generates RDF descriptions.

6. Conclusion

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 are an ordered list of tables, from the most to the least reliable ones, together with an interval-valued evaluation.

7. Future Scope

As future works, we see two main possible evolutions:

7.1. 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.

7.2. 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 SparqlQuerying 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 Belief 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 Truthfulness 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] 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, vol. 8, pp. 199-249, 1975.
[16] L. Zadeh, “Fuzzy Sets as a Basis for a Theory of Possibility,” Fuzzy Sets and Systems, vol. 1, pp. 3-28, 1978.
[17] G. Miller, “The Magical Number Seven, Plus or Minus Two,” The Psychological Rev., vol. 63, pp. 81-97, 1956.
[18] S. Destercke, D. Dubois, and E. Chojnacki, “Possibilistic Information Fusion Using Maximal Coherent Subsets,” IEEE Trans. Fuzzy Systems, vol. 17, no. 1, pp. 79-92, Feb. 2009.
[19] G. Shafer, A Mathematical Theory of Evidence. Princeton Univ. Press, 1976.

Author Profile

Vina M. Lomte received the B.E. and M.E. Degree in Computer engineering. She is now working with RMDSSOE, Warje, Pune as Asst. Professor. She has experiences of 10 yrs

Licensed Under Creative Commons Attribution CC BY

8 months and her Area of specialization - Web Security & S/W Engg.

Saloni Shah received the B.E. degree in computer engineering from Cummins College of engineering for women, Pune University in 2009. And also received the M.B.A. degree in IT/Systems from Dnyangana Institute of Career Empowerment & Research, Pune University in 2012. Now studying in RMD Sinhgad College of engineering, Pune University for Post graduation in M.E.

Licensed Under Creative Commons Attribution CC BY