ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

An Evaluating Enhanced Data Reliability to a Web-Enabled Data Warehouse

M.Vijayaganesh[1], Fungkha Narzary[2,] S.Rajanand[3], D.Dhayalan[4]

Final Year MCA Student, VelTech HighTech Engineering College, Chennai, India[1,2]

Assistant Professor, Department of MCA, VelTech HighTech Engineering College, Chennai, India[3,4]

ABSTRACT: Numerous available techniques 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: Data Reliability, Data Quality, Relevance, Data Warehousing, Maximal Coherent Subsets.

I. INTRODUCTION

The growth of the web and the emergence of dedicated data warehouses offer great opportunities to collect data and to be build to models or prototype or to make decisions. These data are used in further inferences. During collection, data reliability is mostly ensured by measurement device calibration, by adapted experimental design and by statistical repetition. However, full traceability is no longer ensured when data are reused at a later time by other scientists. This estimation is especially important in areas where data are scarce and difficult to obtain as it is the case, for example, in Life Sciences. i present an application of the method a web-enabled data warehouse. Indeed, the framework developed in this paper was originally motivated by the 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 a particular application, it is necessary to give them indicative reliability estimations.

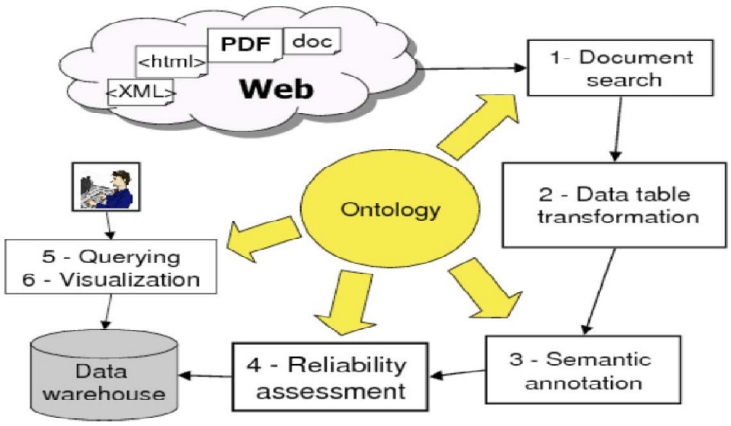

Figure1. Architecture Diagram for Web Presentation

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

II. LIMITATIONS FOR EXISTING SYSTEM

These data are used in further inferences. During collection, data reliability is mostly ensured by measurement device calibration, by adapted experimental design and by statistical repetition. However, full traceability is no longer ensured when data are reused at a later time by other scientists. This estimation is especially important in areas where data are scarce and difficult to obtain as it is the case, for example, in Life Sciences. 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.

- The conflicting information, as different criteria may provide conflicting information about the reliability.
- 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 in the final knowledge.
- Addresses the question of data ordering by groups of decreasing reliability and subsequently the presentation of informative results to end users.

III. IMPLEMENTATION IN PROPOSED SYSTEM

The evaluation of their reliability, it is natural to be interested in the reasons explaining why some particular data were assessed as (un)reliable. We now show how maximal coherent subsets of criteria, i.e., groups of agreeing criteria, may provide some insight as to which reasons have led to a particular assessment. We present an application of the method a web-enabled data warehouse. Indeed, the framework developed in this paper was originally motivated by the 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 a 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. 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.

Advantages

- This notion only makes sense if the source can be suspected of lying in order to gain some advantage, and is distinct from reliability.

- The differentiate between individual-level and system-level trust, the former concerning the trust one has in a particular agent, while the latter concerns the overall system and how it ensures that no one will be able to take advantage of the system.

IV.LITERATURE SURVEY

Towards Content Trust of Web Resources: This paper defines content trust and discusses it in the context of other trust measures that have been previously studied. We introduced several factors that users consider in deciding whether to trust the content provided by a Web resource. Our goal is to discern which of these factors could be captured in practice with minimal user interaction in order to maximize the quality of the system's trust estimates. A study to determine which factors were more important to capture, and describe a simulation environment that i have designed to study alternative models of content trust. In the proposed work is needed on mechanisms to capture how much trust users ultimately assign to open Web sources, while balancing the burden from eliciting feedback during regular use of the Web. There may be very transparent mechanisms based on studying regular browsing and downloading habits. Users will not be the only ones making trust decisions on the Semantic Web. Reasons, gents, and other automated systems will be making trust judgments as well, deciding which sources to use when faces with alternatives. Semantic representations of Web content should also enable the detection of related statements and whether they are contradictory. Further research is needed on how to discern which source a reasons should trust in case of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

contradictions or missing information. Modeling of Reliability with Possibility Theory**:** Possibility theory aims at representing and handling uncertain information. An important property of this theory is the ability to merge different data sources in order to increase the quality of the information. Different fusion rules have been defined in the literature, each with its own advantages and drawbacks. However, these rules usually cannot deal rigorously with contradictory data. This paper proposes a new approach, based on a fusion rule using a vector expressing the reliability of the data sources. Comparisons are made with classical fusion rules. An algorithm assessing the indexes of reliability of the data used by the new rule is given, and moreover an index of the quality of the result is proposed. These three tools define a new method dealing with the reliability of the data in the fusion field, and enable a clear distinction between the data and their quality. Relevance and Truthfulness in Information Correction and Fusion: A general approach to information correction and fusion for belief functions is not only may the information items be irrelevant, but sources may lie as well. Correction scheme, which takes into account uncertain meta knowledge on the source's relevance and truthfulness and that generalizes Shafer's discounting operation. It shows how to reinterpret all connectives of Boolean logic in terms of source behavior assumptions with respect to relevance and truthfulness. While taking into account the uncertainties pertaining to assumptions concerning the behavior of sources. The Proposed work is to extend the framework to the case of sources reporting to the agent what other sources reported to them. In other words, instead of considering several parallel testimonies, one may consider a series of agents, each reporting to the next one what the previous agent reported. There are then several uncertain information distortion steps in a row by sources having uncertain behavior. Interestingly, this alternative line of research is already present in the entry "Probabilit´e" in D'Alembert and Diderot Encyclopedia. Recently, Cholvy also investigated this issue. Eventually one may consider the case of series-parallel networks of more or less reliable sources with uncertain information flows. Another interesting perspective is the possibility to learn the behavior of sources by comparing the pieces of information provided by those sources with the ground truth, as done in a simple framework for discounting and contextual discounting. Possibilistic Information Fusion Using Maximal Coherent Subsets: Maximal Coherent Subsets use possibility theory and the notion of maximal coherent subsets, often used in logic-based representations, to build a fuzzy belief structure that will be instrumental both for extracting useful information about various features of the information conveyed by the sources and for compressing this information into a unique possibility distribution. The proposed work will be to integrate in a meaningful way such information into the fusion process, while keeping the idea of using maximal coherent subsets. Let us notice that, instead of considering the set of fuzzy focal elements generated by our method, one could consider a continuous belief function with a uniform density of weights distributed over cuts F_k . An interesting work would be to compare the method presented here with calculi made with this density. It is also necessary to compare the results of this method with other ones designed to aggregate conflicting possibility distributions. Conjunctive and Disjunctive Combination of Belief Functions Induced By Non Distinct Bodies of Evidence: Dempster's rule plays a central role in the theory of belief functions. However, it assumes the combined bodies of evidence to be distinct, an assumption which is not always verified in practice.In this paper a new operator, the cautious rule of combination, is introduced. This operator is commutative, associative and idempotent. This latter property makes it suitable to combine belief functions induced by reliable, but possibly overlapping bodies of evidence. A dual operator, the bold disjunctive rule, is also introduced. This operator is also commutative, associative and idempotent, and can be used to combine belief functions issues from possibly overlapping and unreliable sources. Finally, the cautious and bold rules are shown to be particular members of infinite families of conjunctive and disjunctive combination rules based on triangular norms and conforms.

V. 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 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. 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

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 Truthfulness in Information Correction and Fusion," Int'l J. Approximate Reasoning, vol. 53, pp. 159-175, 2011.