Trusting Semi-structured Web Data

Davide Ceolin
supervised by Guus Schreiber*,* Wan Fokkink *and* Willem Robert van Hage

VU University, Amsterdam, The Netherlands

Abstract. The growth of the Web brings an uncountable amount of useful information to everybody who can access it. These data are often crowdsourced or provided by heterogenous or unknown sources, therefore they might be maliciously manipulated or unreliable. Moreover, because of their amount it is often impossible to extensively check them, and this gives rise to massive and ever growing trust issues. The research presented in this paper aims at investigating the use of data sources and reasoning techniques to address trust issues about Web data. In particular, these investigations include the use of trusted Web sources, of uncertainty reasoning, of semantic similarity measures and of provenance information as possible bases for trust estimation. The intended result of this thesis is a series of analyses and tools that allow to better understand and address the problem of trusting semi-structured Web data.

1 Research Questions

Trust is a crucial issue in the Web. The growth of the Web brings the impossibility to control and check every single piece of information we have to deal with. Moreover, the heterogeneity of data sources therein present makes the quality and the reliability of the data that these sources expose vary. Consequently, proper techniques need to be developed and proper analyses need to be performed to provide tools and indications to quantify the reliability of the data observed, so that users can properly handle them. This is the focus of the research described here, as summarized by the following overall problem statement.

> *How can the trustworthiness of semi-structured Web data be adequately estimated?*

I investigate about different aspects inherent to this problem: data, metadata and reasoning techniques useful to make adequate trust estimates.

Research Question 1 The first problem that I focus on is the usage of trusted semi-structured Web data to make trust evaluations of semi-structured data (not necessarily coming from Web sources). This gives a first insight into the possibility to use Web data for assessing the trustworthiness of data. Hence the first research question is:

> *Can Web data help the trust evaluation of semi-structured data?*

Research Question 2 Web data present peculiar characteristics that have to be taken into account when using them to make trust evaluations. For instance, they are often accessed incrementally (e.g. by crawling; so we do not always know how representative the data that we observe are), and also their reliability varies, and their source reputation is not always known. Proper reasoning techniques have to be employed to cope with this, and they will be investigated by addressing the following research question:

> *How can uncertainty reasoning be effectively used to estimate the trustworthiness of semi-structured data?*

Research Question 3 Also the Web as such can be exploited for the computation of meta-information that facilitates the estimation of trust values. Web-based semantic similarity measures can be used to weigh data and metadata at disposal of the uncertainty reasoning techniques adopted to estimate the trustworthiness of a given subject, hence the following research question:

> *Can semantic similarity measures improve the accuracy of trust estimates of semi-structured data based on uncertainty reasoning?*

Research Question 4 The Web offers also a meta-level of related information that is useful when dealing with trust, namely provenance information, that represents by whom and how data have been produced, manipulated and exposed. Reasoning over these data is important because this can provide indirect evidence about the reliability of a target object. Moreover, in general, this kind of data possibly enlarges our availability of reliable sources of evidence. This subject will be explored by addressing the following research question:

> *How can provenance information be used for making accurate trustworthiness estimations of semi-structured data?*

2 State of the Art

Trust is a widely explored topic in computer science, in the Web and Semantic Web. Sabater and Sierra [14], Golbeck [10] and Artz and Gil [1] present three comprehensive surveys of the fields. In particular the definition of trust that I make use of is the one of Castelfranchi and Falcone reported by Sabater and Sierra, that is “the decision that an agent x (trustor) takes to delegate a task to agent y (trustee) is based on a specific set of beliefs and goals, and this mental state is what we call trust”. Depending on the scenario where my case study locate, the trustors will vary and the goal of my research will be to build tools or models able to mimic their behavior given the constraints of the case. I do so by employing uncertainty reasoning, provenance analysis and semantic similarity measures. The link between provenance and trust, mentioned in the survey of Artz and Gil, has been explored by Golbeck [9] but, mainly for addressing socio-related issues, while my my focus is on the data trustworthiness estimation. Uncertainty reasoning techniques are often used to make trust assessments, like in the work of Fokoue et al. [8]. It is important to investigate further the possibility to represent these data by means of multiple layers of probabilities, because of their adequateness to deal with vast amounts of heterogenous data.

The link between trust and semantic similarity measures has already been explored, for instance by Ibrahim et al. [11] and by Sensoy et al. [15]. This link can be further explored by considering the relation between different kinds of semantic similarity measures (e.g. deterministic or probabilistic ones) and evidential reasoning. Also, the trust evaluations obtained by means of semantic similarity measures may be effectively integrated with those based on provenance.

3 Proposed Approach

I propose the following approaches to tackle each research question.

Research Question 1 I propose a quantitative empirical approach for this research question, by using uncertainty reasoning to make sense of Web data to trust unknown data. This has merely explorative goals (proving the possibility to use Web data to make trust assessments), and its novelty resides in the use of evidential reasoning in combination with Web data for making trust assessments.

Research Question 2 The approach proposed for this question is quantitative and empirical, and aims at producing a description of how categorical Web data fit higher-order probability distributions. This approach is novel as it provides a first description of Web data in terms of higher-order probabilities.

Research Question 3 I employ a quantitative approach to determine whether I can improve the accuracy of trust values by into account semantic similarity measures. I adopt a theoretical approach to incorporate semantic similarity measures in uncertainty reasoning techniques, which is yet another novel result.

Research Question 4 This research question is tackled empirically. By obtaining an analysis of the use of provenance for trust estimation using statistical techniques, I obtain a novel application.

4 Methodology

Here I introduce the methodologies chosen to implement the above approaches.

Research Question 1 The Naturalis Museum in The Netherlands holds a collection of annotated bird specimen, which includes information like the species these specimens belong to, and the authors of the annotations. These annotations are not fully trustworthy, either because of their inaccuracy or because of the obsolescence of the taxonomy. I map these annotations to trusted Semantic Web sources to check them and, based on a gold standard, I estimate their trustworthiness using a probabilistic logic, named subjective logic [12], that allows to cope with uncertainty about the representativity of the sample observed. I use these trust values with range of decision strategies to decide whether to trust the annotations and I measure the accuracy of the algorithm.

Research Question 2 I investigate further about the statistical foundations of subjective logic, and I use second-order probability distributions and stochastic processes to model the data contained in the Linked Open Piracy dataset [17], which contains a partial collection of piracy attacks descriptions. I focus on categorical data, which are among the most popular kind of data on the Web (URI). I model the data by means of Dirichlet-multinomial distributions and Dirichlet Processes, high-order probabilistic models for categorical data and I compare their ability to cope with the lack of a full view on the data with multinomial probability distributions based on the evidence at my disposal.

Research Question 3 Semantic similarity measures (e.g. the Wu & Palmer similarity [19]) are used to improve the precision of the uncertainty reasoning techniques adopted for trust estimation. I incorporate semantic similarity measures in the uncertainty reasoning techniques, in particular in subjective logic, proving theoretically whether they can be used as a "discounting" factor for probabilities in subjective logic. I compare the precision and the accuracy of trust values of tags of the Steve Museum [16] dataset (which annotate cultural heritage artworks) when semantic similarity weighing is used and when it is not.

Research Question 4 First, I build a bayesian network using subjective logic on top of provenance graphs, to derive a trust value for a data artifact from the analysis of how it has been produced. This is validated over a set of messages (AIS) sent by ships to coast guard authorities to communicate mandatory information (e.g. their nationality). The validation focuses on the feasibility of the approach, by proving the possibility to build an algorithm that provides such a network. Second, I use machine learning methods to make trust predictions based on the provenance graph of the target objects. In particular, I predict the trustworthiness of a collection of video tags provided by the gaming platform *Waisda?* [13]. Accuracy, precision and recall of the predictions are computed.

5 Results

Here I report the results obtained by addressing the research questions above.

Research Question 1 An algorithm based on subjective logic that uses Web data to assess trust values about the dataset of 65,600 bird specimen annotations of the Naturalis Museum (30% of which serve as training set) [6].

Research Question 2 An analysis of the effectiveness of second-order probability distributions in representing Web data, tested over 2,309 LOP piracy attacks [7]. A first extension of subjective logic to handle higher-order probabilities, which I demonstrate theoretically [5].

Research Question 3 An extension of subjective logic to incorporate semantic similarity measures as a means to weigh evidence within the logic, which I prove theoretically [5], and a first algorithm that employs this extension for computing trust estimates over samples from the 45,860 tags from the Steve.Museum dataset [4], which has been validated by means of a statistical hypothesis test.

Research Question 4 An algorithm that builds a subjective logic-based bayesian network over a provenance graph, compliant to AIS messages [2]; an algorithm that estimates trust based on provenance graphs of 37,850 *Waisda?* tags (training set 70%, test set 30%), by using machine learning classifiers [3].

6 Remaining Work

In this section I describe the remaining work and I indicate a time plan.

Research Question 2-3 Additional extensions of subjective logic incorporating semantic similarity measures and higher-order probabilities; 2 months.

An algorithm for trust computation using uncertainty reasoning combined with semantic similarity measures and provenance metadata; 2 months.

Research Question 4 An algorithm for trust computation based on the semantics of the PROV-O ontology [18]; 3 months. **Thesis writing** 4 months.

References

1. D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Journal of Semantic Web*, 5(2):131–197, 2007.
2. D. Ceolin, P. Groth, and W. R. van Hage. Calculating the trust of event descriptions using provenance. In *SWPM*, volume 670, pages 11–16. ceur-ws.org, 2010.
3. D. Ceolin, P. Groth, W. R. van Hage, A. Nottamkandath, and W. Fokkink. Trust evaluation through user reputation and provenance analysis. In *URSW*, volume 900, pages 15–26. ceur-ws.org, 2012.
4. D. Ceolin, A. Nottamkandath, and W. Fokkink. Automated evaluation of annotators for museum collections using subjective logic. In *IFIPTM*, volume 374, pages 232–239. Springer, 2012.
5. D. Ceolin, A. Nottamkandath, and W. Fokkink. Subjective logic extensions for the semantic web. In *URSW*, volume 900, pages 27–38. ceur-ws.org, 2012.
6. D. Ceolin, W. R. van Hage, and W. Fokkink. A Trust Model to Estimate the Quality of Annotations Using the Web. In *WebSci*. Web Science Repository, 2010.
7. D. Ceolin, W. R. van Hage, W. Fokkink, and G. Schreiber. Estimating uncertainty of categorical web data. In *URSW*, volume 778, pages 15–26. ceur-ws.org, 2011.
8. A. Fokoue, M. Srivatsa, and R. Young. Assessing trust in uncertain information. In *ISWC*, pages 209–224, 2010.
9. J. Golbeck. Combining provenance with trust in social networks for semantic web content filtering. In *IPAW*, pages 101–108. Springer, 2006.
10. J. Golbeck. Trust on the World Wide Web: A Survey. *Foundations and Trends in Web Science*, 1(2):131–197, 2006.
11. H. Ibrahim, P. K. Atrey, and A. El Saddik. Semantic similarity based trust computation in websites. In *MS*, pages 65–72. ACM, 2007.
12. A. Jøsang. A Logic for Uncertain Probabilities. *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–212, 2001.
13. Netherlands Inst. for Sound and Vision. Waisda? http://wasida.nl, Aug. 2012.
14. J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artificial Intelligence Review*, 24:33–60, 2005.
15. M. Sensoy, J. Z. Pan, A. Fokoue, M. Srivatsa, and F. Meneguzzi. Using subjective logic to handle uncertainty and conflicts. In *TrustCom*, pages 1323–1326. IEEE Computer Society, 2012.
16. Steve. The Museum Social Tagging Project. http://www.steve.museum, Jan. 2013.
17. W. R. van Hage, V. Malaisé, and M. van Erp. Linked Open Piracy. http://semanticweb.cs.vu.nl/lop/, Nov. 2012.
18. W3C. PROV-O. http://www.w3.org/TR/prov-o//, June 2012.
19. Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *ACL*, pages 133–138. ACL, 1994.