Summary

Trusting Semi-structured Web Data

This thesis tackles the problem of trusting semi-structured data, in particular in the Web context. Different definitions of trust are in use in different areas. However, by constraining the context of application of our analyses, we also constrain the definition of trust that we adopt. In particular, we follow the definition of Castelfranchi and Falcone, the definition of Camp and the theory of O’Hara.

We focus on how to properly make use of metadata to estimate the trustworthiness of the corresponding data, given that this is the kind of information at our disposal and the knowledge or time is lacking to determine the trustworthiness of the content of these data. Moreover, we assume the existence of a correlation between metadata and the trustworthiness of the data themselves. For instance, we can estimate the trustworthiness of a piece of data by knowing who created it, and by observing part of his or her behavior. In fact, there exists a probabilistic relation between the user identity and the user trustworthiness, which is summarized by the user reputation. By using appropriate smoothing techniques, the user reputation can be reliably estimated based on a set of sample observations. The strength of this correlation is not the same for all metadata, but our evaluations confirm that this approach (possibly after combining data to strengthen the correlation) can be effective in estimating data trustworthiness.

Chapter 1 gives a first hint about the use of metadata enriched with Web data to assess the trustworthiness of museum annotations. That chapter presents an algorithm for assessing the trust of annotations, as well as an evaluation of it on a collection of bird specimen annotations from the Naturalis museum in Leiden in The Netherlands. The algorithm employs subjective logic, an uncertainty reasoning technique, to learn from a training set of evaluated annotations the relations between trust levels and annotations metadata, some of which are derived from Web sources. That chapter provides some useful insights. One is the fact that in order to make a trust assessment, one can distinguish two distinct phases: trustworthiness estimation and trust assessment (or decision strategy). Another insight is the fact that it shows how metadata are relevant features to be used to learn the trustworthiness of data, and how we can leverage the metadata at our disposal to enlarge their availability by means of Web sources, hence extending the information that we can use to accomplish our goal.

In this line of thinking, Chapter 2 shows that not all the metadata-based and Web-based heuristics that we can formulate to help in our trustworthiness predictions are
actually meaningful, and hence these need to be accurately selected. In that chapter we evaluate a series of heuristics used to estimate the confidence in a set of georeferenced records from specimen annotations from the Naturalis museum, and we show how the best combination of these heuristics is not the one that includes all of them. Also, the heuristics adopted in that case study present ample margins of improvement.

Both Chapters 1 and 2 share the use of statistical as well as evidential reasoning, in particular as a means to handle the uncertainty that is present in our computations. We use some information sources for our estimates, but we do not have any prior warranty, neither about their utility, nor about the representativity of the samples we face. We still use such information because it is the only handhold at our disposal, but we treat it prudently. Evidential reasoning allows us to take into account that the evidence that we did not observe about given metadata could have led us to different conclusions. So, we adopt evidential reasoning as a means to draw prudent conclusions about the data at our disposal, and in Chapter 3 we propose a series of case study applications where uncertainty reasoning is used to model piracy attacks recorded in a Web database. Also, we present a series of extensions of subjective logic which aim at enhancing the logic’s usability in the Web environment, thus allowing the logic to deal with semantic similarity, partial evidence observations and open world opinions. Chapter 4 shows, by means of a case study based on police open data, that uncertainty reasoning can also help to prudently extrapolate indications of reliability changes in the data at our disposal, even when it is not possible to have sure indications about this issue. By analyzing the data from different points of view (relative, absolute, etc.), we may be able to obtain insights about their reliability. If this is not possible, uncertainty reasoning can be employed again for aggregating such analyses and deriving weaker but still prudent conclusions about data reliability.

Having understood that statistical and evidential reasoning is particularly important to make trust assessments, we devote part of our research to the analysis of an important class of metadata, that is, provenance information. This leads to two different approaches: on the one hand we build Bayesian networks on top of small provenance graphs to determine the trustworthiness of naval messages, based on the trustworthiness of each single component used to build them (Chapter 5); on the other hand we run supervised learning algorithms on top of aggregated provenance graphs of media tags to predict the trustworthiness of these tags based on how they were produced (Chapter 6). Both approaches share the attempt to learn statistically some models that link information about how, when and by whom artifacts have been created as well as about their trustworthiness. Provenance information has a great potential for helping the trust assessment, and by using statistical techniques we are able to connect some feature combinations with the trustworthiness of the artifacts. This is a necessary step to rely on this class of information for our estimates, because provenance describes how artifacts have been produced, not how trustworthy they are. We first group provenance graphs in classes called “provenance stereotypes”, which represent user behaviors, and then use these stereotypes as a basis for trust assessments via machine learning. In this way, we obtain meaningful models, while overcoming possible problems due, for instance, to the complexity of the graphs. This
comes at a cost, since we do not focus deeply on the semantics of these graphs, and therefore we may have neglected useful knowledge. The results that we obtained are already satisfactory. We have not investigated deeper into the link between provenance semantics and trust, but we will in the future.

The last part of the thesis regards the use of semantic similarity measures in combination with uncertainty reasoning (as defined in Chapter 3) to make reputation-based (Chapter 7) and provenance-based (Chapter 8) trust assessments, in particular in the cultural heritage domain. Semantic similarity measures are useful at least for two reasons. Firstly, the uncertainty reasoning techniques we use rely on the availability of evidence, and the more evidence we get the better results we obtain. Semantic similarity measures allow us to enlarge the evidence set at our disposal, while keeping its relevance high. Secondly, semantic similarity measures allow us to reduce the computation complexity in our estimates by avoiding repetition of computations for pieces of evidence that are syntactically different but semantically similar. These are important achievements. Although they are currently limited to specific domains, like cultural heritage, in principle they may be adapted to other domains as well.