A Framework for Trust Establishment and Assessment on the Web of Data

Qi Gao Geert-Jan Houben
Delft University of Technology,
PO Box 5031, 2600 GA Delft, the Netherlands
0031-1527874686
{q.gao, g.j.p.m.houben }@tudelft.nl

ABSTRACT
With the enormous and still growing amount of data and user interaction on the Web, it becomes more and more necessary for data consumers to be able to assess the trustworthiness of data on the Web. In this paper, we present a framework for trust establishment and assessment on the Web of Data. Different from many approaches that build trust metrics within networks of people, we propose a model to represent the trust in concrete pieces of web data for a specific consumer, in which also the context is considered. Further more, we provide three strategies for trust assessment based on the principles of Linked Data, to overcome the shortcomings of the conventional web of documents: i.e. the lack of semantics and interlinking.

Categories and Subject Descriptors
I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods

General Terms
Design.

Keywords
trust, web of data, linked data, semantic web

1. INTRODUCTION AND MOTIVATION
The volume of web data is sharply increasing in recent years mainly because of Web 2.0 technology and applications for social networking, with more and more user interaction, e.g. feedback and sharing, involved in the process of data exploitation on the Web. It becomes necessary for data consumers to be able to assess the trustworthiness of data on the Web. Trust can be helpful to determine the best quality data, to find accurate and relevant data, and to protect privacy [1].

Ideally, there would be a trust model that tells how trustworthy each piece of data is for a specific data consumer. However, due to the fact that on the conventional Web of documents, the "untyped hyperlinks" lead to a lack of semantic relationships between different pieces of data, the main barrier for this model is that if we would have obtained accurately the trustworthiness of a given piece of data for a data consumer, it is still very hard to assess and predict the trustworthiness for new pieces of data. Many approaches therefore choose to build the notion of trust and trust metrics within networks of people, agents or peers [5] that have trust relationships between them, based on which the trust in web data can be assessed and predicted indirectly. These approaches are often limited to a certain domain, in which the trust network is acquired.

However, based on the principles of Linked Data [4], we see the possibility to overcome the weakness of the conventional Web and build a trust relationship directly between web data and data consumer. Linked Data has billions of triples from various data sources, with contributions from organizations, companies, as well as individuals. Furthermore, it provides rich semantics of data and interlinking between data from different data sources. Some related work has been done for building trust in the Web of Data. Hartig [3] proposed a trust model to add a trust value to RDF statements and enable SPARQL to access the trust values. However, to the best of our knowledge, two critical tasks in this topic are still missing. The first one is the representation of trust in metadata in which both the context and consumer information are considered. A second one is the generation of trust in that representation, which is the foundation for further actions such as querying and retrieval. In this paper we present a framework for establishment and assessment of trust in the Web of Data. We address the problem based on two questions:

- How to represent a trust relationship between a data consumer and data in a specific context?
- How to assess the trust value in that trust relationship?

2. DEFINITIONS
Since in computer science there are various definitions of trust depending on the context and applications, we give our definition to claim the problem scope.

Definition 1: A *data consumer C* is the endpoint, an individual or some specific system, which consumes the data.

Definition 2: The *trustworthiness* of a piece of data *d* for a data consumer *C* is the measurable belief of *C* to represent the reliability of *d* within a specific context.

3. FRAMEWORK
In this section we propose a framework (Figure 1) for (1) establishment and (2) assessment of trust in metadata in the format of RDF, RDFa or microformats on the Web of Data. The establishment of trust focuses on how to represent the trust relationship. The assessment of trust relates to the problem of how to acquire the trustworthiness.

3.1 Trust Establishment
First of all, our framework requires a representation model of trust that defines and represents a trust relationship between a data consumer and a piece of metadata. There are three critical

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

elements to be investigated in our representation model: the trustworthiness of data (see Definition 2), consumer model, and context model. The consumer model specifies the preference and the existing trustworthiness of other pieces of metadata for this consumer. The context model mainly specifies the domain knowledge in which the data is consumed. Theses two models are formalized with RDF-based vocabulary. Currently we are working on developing a trust vocabulary for Web data that can be adopted to describe the context model and consumer model. Basically there are two solutions for providing the trustworthiness: manually provided by data consumer and generating them automatically by exploring the data. For the first one, the disadvantage is that in many cases it is not practical to push the users to give these trust ratings. Since the metadata is normally rich in semantic and Linked Data provides the possibility to investigate the interlinking between different pieces of data, we adopt the latter one as our solution to assess the trustworthiness. We propose three different assessment strategies for trust: vocabulary-based, triple-based and interlinking-based, which will be explained later. The main function of the trust computation component is to select one of the three strategies or to synthesize two or three of them according to the context model and consumer model, and then assign the trustworthiness to the piece of data.

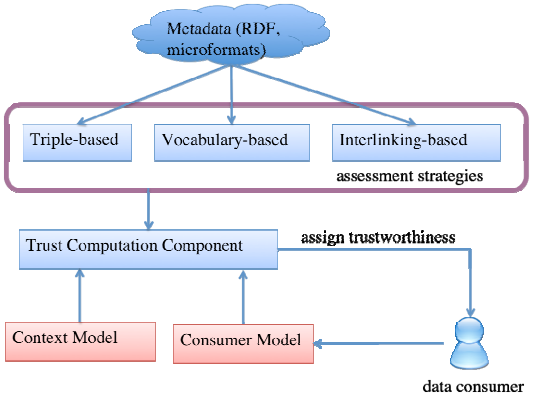

Figure 1. A framework for trust establishment and assessment on the Web of Data

3.2 Trust Assessment Strategies

Here we explain three strategies of trust assessment in detail.

(1) Vocabulary-based strategy: We can assess the trustworthiness of data based on some vocabularies used in this piece of data and the occurrence of these vocabularies. Here we consider the vocabularies that can represent the provenance information. In paper [2] author has listed a set of vocabularies that can be used to obtain the provenance information of data. The helpful vocabularies include Dublin Core Metadata Terms (http://dublincore.org/documents/dcmi-terms/, dcterms), the Friend of A Friend vocabulary (FOAF, http://www.foaf-project.org), and the Semantically-Interlinked Online Communities (SIOC, http://rdfs.org/sioc/spec/). The dcterms vocabulary is capable to represent the creation information and related resources from which the data is derived. The SIOC vocabulary can represent the information of creators or owners within online communities, e.g. blog, posts or comments. Besides these highly used vocabularies, some other vocabularies are also useful, for example, the Web Of Trust (WOT, http://xmlns.com/wot/0.1) schema that utilizes Public Key Cryptography tools such as PGP or GPG to sign RDF documents. However, the weak point of these vocabularies is that they are used only by a few pieces of data. For example, if we use Sindice (http://sindice.com/), a semantic web index tool, to search the metadata that includes the WOT vocabulary, we only get two hits.

(2) Triple-based strategy: The consumer model includes the trustworthiness already acquired before, for a given data consumer and a set of data. This strategy assesses the trustworthiness by calculating the similarity between the triples in the piece of unknown data and in the data of which we have already got trustworthiness. This similarity can be obtained by techniques for string matching or semantic matching.

(3) Interlinking-based: Linked Data extends the Web by creating typed links between data from different data sources. Instead of having "*untyped hyperlinks*" between documents, RDF properties can be interpreted as "semantic hyperlinks" [4]. Whilst the first two strategies focus on the explicit information obtained from the metadata itself, the interlinking-based strategy investigates the links between different pieces of data. Then we assess the trustworthiness of data by matching the semantics of these links and concepts in the context model.

4. CASE STUDY

In order to demonstrate the capability of our framework, we perform a case study on how the trust metric that we propose in this paper helps the semantic-featured personalization data exploitation over two data sources: DBpedia and Freebase. An individual user as a data consumer wants to search data about traveling in these three data sources. The consumer model is build by utilizing the personal information that is collected from existing data such as rating data on restaurant, city etc. The context model includes the domain knowledge related to traveling such as geographical concepts. Using the assessment strategies proposed previously we give a trusted-based ranking for the search results.

5. CONLUSION AND FUTURE WORK

In this paper, we present a framework for building trust in the Web of Data. Our approach stresses two specific tasks: the creation of a representation model of trust in web data for specific data consumers, and the support for three different strategies for data consumers to assess the trustworthiness of data that they consume. For future work, we are going to validate our framework from two aspects: (1) the flexibility and compatibility of our framework when it is applied to existing datasets and (2) its capability to support query language, e.g. SPARQL.

6. REFERENCES

[1] Artz, D., Gil, Y.: A Survey of Trust in Computer Science and the Semantic Web. J. Web Semantics. 5(2): 58–71(2007)

[2] Hartig, O.: Provenance Information in the Web of Data. In: Lined Data on the Web Workshop at WWW'09 (2009)

[3] Hartig, O.: Querying Trust in RDF Data with tSparql. In: ESWC'09: 5–20 (2009)

[4] Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The Story So Far. Journal on Semantic Web and Information Systems: 1-22 (2009)

[5] Golbeck, J.A.: Computing and Applying Trust in Web-based Social Networks. PhD thesis, University of Maryland (2005)