Extracting Trust Network Information from Scientific Web Portals

Alejandro Castañeda Paulo Pinheiro Da Silva

University of Texas at El Paso, El Paso TX 79968, USA

Abstract. An increased exchange of (scientific) information across organizations and disciplines is one of the long-term goals of the semantic web. In any such exchange of information, it is not difficult to identify one or more (scientific) communities responsible for the measurement, gathering and processing of scientific information. More challenging, however, is to understand the trust relations between members of these communities, whether the members are organizations or people. With a better understanding of trust relations, one may be able to compute trust recommendations for scientific information exchange, increasing in this way the acceptance of information by scientists. In this paper, we present CI-Learner, which is a systematic approach for extracting trust-related meta-information from scientific portals and related web sites. CI-Learner meta-information is organized as trust networks based on people, organizations, publications, and trust relations derived from publication co-authorship. Participation in a given trust network is restricted to organizations and people as identified by the CI-Learner information extraction process. The paper reports on the usefulness of the extracted trust network and related ranking as identified in a user study with subjects who are experts in the field of concern.

1 Introduction

A large-scale reuse of scientific data collected through funded projects is a major goal of funding agencies world-wide in support of new scientific discoveries [10] and a common concern with the semantic web. Reuse of data, however, requires the understanding of the data semantics so that scientists can integrate the data into their research, e.g., to other data, tools, documentation. The challenge of reusing data across disciplines, however, goes beyond the understanding of data since scientists would need to believe the data to accept and reuse the data in their scientific tasks. This means that the scientists would need to trust organizations and people from other disciplines and scientific communities.

In this paper, we investigate the challenge of learning about trust relations in a community and using these relations to establish a degree of trust on community members. This problem is particularly interesting assuming that one person who is not member of this community may be able to understand some important trust relations in the community without being a member. To facilitate the learning of trust relations, we introduce CI-Learner, a systematic approach for

extracting people and organizations trust information from scientific web portals supported by the community of interest. The extracted trust information is expected to support trust recommendations for scientific artifacts, e.g., maps, graphs, reports, created by the community. To illustrate the use of CI-Learner, this paper focuses in the creation of a trust network for the Earth Science community with the help of the scientific web portal maintained by the Incorporated Research Institutions for Seismology (IRIS) [7].

The paper also illustrates how CI-Learner can benefit the NSF-funded EarthScope Project in which many organizations such as the U.S. Geological Survey (USGS) along with hundreds of other scientific organizations worldwide exchange information to gain a better understanding of natural occurring hazards (e.g. volcano eruptions, earthquakes) supported by new information technology using CI initiatives. The EarthScope Project is particularly relevant for CI-Learner because the 3D model of the Earth will often rely on information coming from multiple sources from distinct organizations.

The rest of paper is organized as follows. Section 2 describes our use case scenario where we use trust networks to support the acceptance of scientific results. Section 3 describes some aspects of trust considered in this paper. Section 4 presents the CI-Learner approach used to extract trust networks from scientific web portals and Section 5 describes some agents used in the extraction process. Section 6 presents our user study. Section 7 presents related work. Finally, Section 8 summarizes the contribution of this paper.

2 Earth Science Community: A Use Case

The CyberSHARE Center is an NSF funded project that joins efforts of geophysicists, computer scientists, environmental scientists and computational mathematical experts developing new approaches for collection and analysis of geophysical and environmental data of the North American continent. For instance, of interest to CyberSHARE goals, is to develop a framework facilitating the integration of vast amount of data currently being collected from communities such as the NSF-funded Geosciences Network (GEON) and EarthScope, and apply it to understand and investigate the structure and evolution of the Earth through physical properties of our planet. An important step towards the creation of such an integrative data framework is the development of a region specific 3-D models of the lithosphere and upper mantle from existing information sources allowing a new global insight. To develop this 3-D model, it may be required the analysis of times series collected from two types of seismic experiments: source controlled and passive experiments. *Source* controlled experiments make use of explosions and high-frequency sensors, whereas *passive* experiments record natural occurring earthquakes. In any case, the 3-D model construction depends on the study of the time it takes for waves, produced by either source (e.g. explosion or earthquake), to travel through the body of the Earth. Since waves travel through Earth's interiors, the time it takes for a source-generated wave to reach a sensor is then measured [1]. This data is used for constructing a velocity model using

reverse slowness perturbations. A sample output of what could be this model along with an illustrative (but still to be implemented) trust recommendation is presented in Figure 1a.

Of particular interest within CyberSHARE is the analysis of wave slowness data together with existing legacy applications required to produce 3-D models (e.g. velocity model) using and algorithm developed by Hole as mention in [1]. Hole's algorithm has been designed to make use of data produced by controlled experiments in order to derive a velocity model such as in Figure 1a. In CyberSHARE, we are interested in adapting Hole's algorithm to use existing collections of passive seismic data collected and stored in a central repository at Incorporated Research Institutions for Seismology (IRIS).

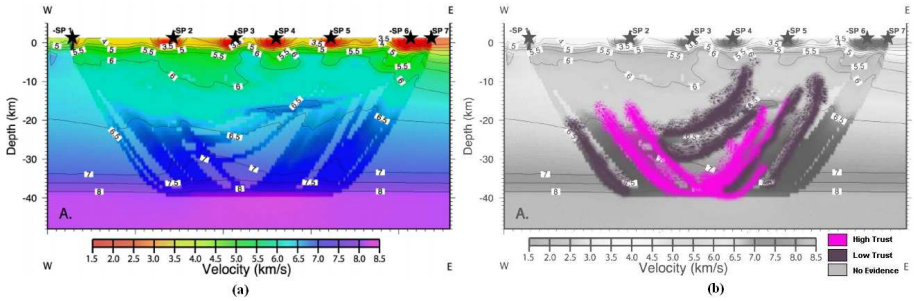

Fig. 1. Velocity model (a) and its corresponding trust layer (b).

3 Trust Networks

Trust scores are used to compare the degree of trust and distrust an agent may have in other agents in the network. This information can be used in the ranking mechanisms of the recommendation systems, e.g. by giving preference to recommendations from sources that are trusted more. To this aim, we introduce the trust score space as a model that allows to compare and preserve information about the provenance of trust scores. The available trust information should be thought of as a snapshot taken at a certain moment since trust scores can be updated. In this paper, we are particularly interested in a *probabilistic* trust based approach that can also accept unknown probability (as opposed to zero probability). Evidence of distrust is not considered in our work.

Scientists and organizations belonging to the communities directly and indirectly related to CyberSHARE scenario constitute a large trust network where exchange of information among members of this network is based on trust relations. Our interest is on the use of trust relations to aid in the problem of the acceptance of velocity models by scientists who are not necessarily involved in the creation of the models. The scientists may face velocity models derived

from passive seismic data, requiring the integration of data from multiple information sources (e.g. repositories, sensors) available from, and possibly managed by, different institutions and personnel involved in the collection of earthquake-generated seismic signals worldwide. The sources used to derive the models and trust relations between these sources are part of a social trust network that the CI-Learner approach builds. For instance, the scientists now know exactly how information sources contributed to the process of generating the velocity model, i.e., provenance information. In addition to the provenance information, the scientists know how much s/he can trust each one of these sources (i.e., degrees of trust) and also know how to generate trust maps from these degrees of trust. However, the scientists still have to aggregate and propagate the degree of trust originally assigned to the information sources. The aggregation and propagation of trust is outside the scope of this paper. Using the techniques described in [4], we can propagate and aggregate trust for the sources used to derive a velocity model, and this adds a layer of trust recommendation on top of a velocity map, creating in this way a trust map [1] such as in Figure 1b[1]. Using a trust map, scientists may be aided in the process of acceptance of quality products derived from scientific processes.

The problem of encoding provenance information is assumed to be addressed elsewhere [16]. In the rest of the paper we address the problem of one learning how much s/he can trust the sources used to generate scientific products like the map shown in Figure 1b. The problem of generating trust maps is beyond the scope of this paper.

4 CI-Learner Approach

4.1 Provenance Elements as Trust Network Agents

The Proof Markup Language (PML) [15], which is based on OWL, is an ontology that includes a hierarchy of concepts used to characterize information sources. This hierarchy can serve as a schema for encoding information about agents used as actors in the trust network to be extracted.

Of interest to our approach are `Sources` of information, which can be either `Agents` or `Documents`. Agents are sources capable of stating new assertions, typically in the form of documents (e.g. Organizations, People). A relationship of interest to our approach is the `hasMember` relationship which denotes Agents being members of Organizations. `Documents` are artifacts created by agents, as identified by the `hasCreator` relationship (e.g. Publications, Ontologies).

In terms of trust, it is important to mention that if we are inclined to trust the creator of a document that means that we are inclined to believe the contents of the document. By using PML, we are able to identify objects in the real world that are potential sources of information in a given scientific field; therefore being able to provide trust related information captured by our approach using

[1] Currently being applied to Gravity data. Visit http://trust.utep.edu/probeit/trustmap

the Web as the source of information about sources, our *source meta-information* as in [16].

4.2 Trust Relations

Trust relations among information sources come on several flavors and are captured by a wide-range of evidences such as links between web pages, traces of exchanged e-mails, and co-authorship of documents. On top of these evidences, we may aggregate trust for each member of our network.

We consider scholarly literature as evidence of trust between two scientists. This is, co-authorship is a case of relationship that connects actors in our network. The CI-Learner approach requires access to a retrieval system for academic publications capable of: performing searches based on an author's name; allowing specification of subject field from science; and providing meta-information about the publication's title, year, and most importantly co-authors. Access to scholarly repositories is important in that we require information as rich as possible to facilitate and automate the process of capturing trust relationship information between our network's actors. Many of these repositories exist such as *Scopus* [2] and *Web of Science* [3] which are commercial repositories for scientific abstracts, references and citations. Even when these repositories provide complex functionalities to users, a subscription is required. Therefore we considered Google Scholar which is becoming a very popular resource for scientific publications and is open to public access. It provides basic functionalities similar to paid sources and satisfies the requirements of our approach.

4.3 CI-Learner Algorithm

The CI-Learner algorithm (Algorithm 1) is used to create trust networks from publications based on PML concepts. First, CI-Learner initializes sets O, P, D, L, G (line 1). The first major goal of the algorithm consists of capturing organizations that are affiliated to a scientific community's web page denoted by w_s. The process of capturing organizations consists of first finding potential pages containing organization's information about affiliates, and second to extract found information. Scientific organizations often present affiliates in listings through possibly one or more web pages. The step of $findOrgWebPages$ (Line 2) in the algorithm traverses the web domain of w_s using a web crawler in order to find web pages having hyper-links containing specific keywords as "member", "affiliate" and are potential candidates of having affiliates to the community in discourse storing each web page into W_o. The CI-Learner algorithm creates a context for extracting organizations by assuming that such keywords lead to pages containing organizations affiliated to the community supporting the portal after we verify them. The *extractAgents* step (Line 3) of the algorithm creates a web wrapper (further explained in Section 5) and extracts Organizations by labeling

[2] See http://www.info.scopus.com/
[3] See http://scientific.thomson.com/products/wos/

each of them as member of the community of discourse (applying `hasMember` relationship between the Source and Organization concepts in PML).

Algorithm 1 CI-Learner Algorithm

input: w_s; **output**: O, P, D, L, G

1: $O, P, D, W_o, L, G \leftarrow \emptyset$;
2: $W_o \leftarrow findOrgWebPages(w_s)$;
3: $O \leftarrow extractAgents(W_o)$;
4: **for all** $(o_i \in O)$ **do**
5: $\quad W_{poi}, P_{woi} \leftarrow \emptyset$;
6: $\quad w \leftarrow getWebPage(o_i)$;
7: $\quad W_{poi} \leftarrow findPersonWebPages(w)$;
8: $\quad P_{woi} \leftarrow extractAgents(W_{poi})$;
9: $\quad P \leftarrow P \cup P_{woi}$;
10: **end for**
11: **for all** $(p_i \in P)$ **do**
12: $\quad D_{pi} \leftarrow \emptyset$;
13: $\quad D_{pi} \leftarrow getPublications(p_i)$;
14: $\quad D \leftarrow D \cup D_{pi}$;
15: **end for**
16: $L \leftarrow ExtractTrustRelations(P, D)$
17: $G \leftarrow ComputeAggregateTrust(L)$

The CI-Learner algorithm imposes an order in which the agents are extracted (i.e., organizations before people); therefore, making sure the algorithm does not relate an organization being part of a person. However, the algorithm does not capture information about sub-organizations leaving this as future work.

After extracting organizations, CI-Learner proceeds to find and extract people (i.e. scientists). The approach is similar to that of capturing organizations as in Lines 4-10. For each organization $o_i \in O$, the algorithm explores o_i's web page to find listing of faculty, research specialists. We assume that each o_i's web page points to the organization's scientific group (e.g. geology faculty page for geosciences community). Again, CI-Learner instruments its web crawler to use keywords such as "people", "members", "faculty", "about us", and common variations of these (e.g. "ppl" for people), exploring in this way each page to find web pages within their web domain containing meta-information about scientists. These web pages are then stored in W_{poi} ($findPersonWebPages$ Line 7).

The step *extractAgents* from Line 8 extracts people for each particular o_i and assigns each person its respective affiliation w_o (i.e. o_i's id_o). Lines 11-15 from Algorithm 1 retrieves the set of publications D_{pi} for person $p_i \in P$. A call to *getPublications* for p_i is at line 13. The CI-Learner approach instruments a software component to interface with Google Scholar. This software component sets the subject area to `Physics, Astronomy, and Planetary Science`,

restricts number of results to 100 and import citation format to BibTex. Finally, a wrapper is used to extract title, year, and co-authors for each publication.

In order to *Extract Trust Relations* we first collect publications having two or more *known* co-authors. By *known* co-authors we refer to authors from $d_i \in D$ that belong to previously extracted people (i.e. each d_i's co-author $c_i \in P$); thus, setting the boundaries of our network. We construct tuples representing the possible combinations of known co-authors. This is, the approach considers co-authorship between two scientists as being bi-directional. If a combination of two co-authors exists, we proceed to increment the co-authorship number between scientists; thus, updating both relations in both directions.

Computing the aggregate trust values for the identified network (sets P, O, D and L, is straightforward using EigenTrust [8], we need to construct a matrix C^T using local trust relations (i.e. set L) where C_{ij} represents the number of joint publications between $person_i$ and $person_j$; and an m-vector $\boldsymbol{e}$ initialized to $1/m$, where m is the number of people in our network. The resulting vector $\boldsymbol{t}$ contains the global trust values for each member in our network.

At this point, the CI-Learner algorithm has formally constructed a network $N(w_s, O, P, L, G)$ consisting of people P affiliated to w_s through organization $o_i \in O$. Relationship (e.g. edge) between people in P is determined by tuples $l_t\langle t_r, t_e, n\rangle \in L$ being both $t_r, t_e \in P$ and having weight the co-authorship n. The degree of trust for each $p_i \in P$ can be retrieved from tuples $g_t\langle p_i, t\rangle \in G$.

5 CI-Learner Agents

The CI Learner approach makes use of agents to extract information about sources of interest. Web wrappers are used to deal with collections of data sharing a common structure and that are usually encoded as HTML tables. Web crawlers are used to explore web page links to find information of interest (e.g. people, organization names).

5.1 Scientific Web Portal Crawling Agent

Exploring web pages that contain information of interest is important in establishing a context for the extraction of Agents, as described in Algorithm 1 Lines 2,7 (see Section 4.3). The IRIS web site provides a listing of more than 190[4] organizations affiliated to it, making the portal a rich information source for our network buildup process.

By observing that many of the web sites we found of interest had a commonality in their anchors and URLs. Most of these sites had keywords such as "membership", "affiliates", "institutions" as part of their hyper-links. A simple version of a crawler using URL analysis [3] was used assuming that an anchor or URL that has any previously listed keywords receives highest score and nothing else is considered.

[4] This figure was at time of extraction. Currently the IRIS portal increased affiliated institutions to more than 230 organizations

Moreover, a *breadth-first* crawling strategy was used to provide more coverage of the web domain of IRIS. In the case of people, we used "people", "directory", "about us", "members", "faculty" for instance, and many other variations of these words (e.g. "ppl" for people). In few instances, URLs for an organization pointed to a home page of a university; thus, we used our web crawler to find the geosciences department using keywords such as "geosciences","geology" and variations (e.g. "geo"). A criticism for this approach becomes apparent when dealing with information of foreign institutions where a translation of keywords may be required. In both cases, the crawler restricts its search to the web domain of the page it was processing at the time. Once a collection of pages containing people names is found, we proceed to extracting agent's meta-information.

5.2 Source Extraction Agents

Extracting agents is another important step in the CI-Learner algorithm (refer to Section 4.3, *extractAgents* at Lines 4,8). In order to facilitate extraction processes of organizations and people, we based our extraction strategy on XWRAP [11] and followed most guidelines from ANDES [14]. CI-Learner uses HtmlUnit [2] as an API to interact with web pages and obtain an XML representation of the HTML source of the web page. Having this XML transformation, we construct an XPath pointing to a location in the XML where the collection of records of interest begins. Then an instrumented web wrapper extracts the desired meta-information according to the enclosing HTML type for the record (e.g. table, division). Using this technique, one can device a wrapper for candidate pages and supervise the extraction process by manually instructing the wrapper to filter out data of no concern to our approach.

5.3 Trust Relation Extraction Agent

An engine that uses HtmlUnit to interact with Google Scholar has been developed, which retrieves the first 100 publications for each person name previously extracted. This step is represented in lines 11-14 in Algorithm 1 (see Section 4.3). A web wrapper was created to extract meta-information of interest from publication such as title, co-authors, year, and BibTeX citation file using approach similar to Section 4.3. We reduce the possibility of introducing noise into our network by requesting 100 in that further publications may not belong to the scientist in particular. Other problems were the disambiguation of names.

A strategy to overcome these difficulties was to query for the publications of all people in P. Each publication d_i was stored an the person p_i, used to reach d_i, is added to the set of d_i's co-authors. In the case that another person has a publication already in our repository, we add this new person to the set of its co-authors. This way, capturing the co-authors of a given publication plus other meta-information is trivial.

6 Evaluation

The goal of this study is to determine how useful the extracted network is as perceived by members of the community. In our context, usefulness is measured by the completeness of the extracted trust network and the accuracy of the rank.

User Study: Test subjects were asked to complete an online user study [5] The experiment captures personal information from subjects such as name, e-mail, affiliation, area of expertise among other. The experiment consisted of two sections assessing two aspects of our network: Completeness of the network and rank accuracy. In the *Completeness of the network* section subjects were individually asked to provide the names of five individuals they considered to be part of the network. A tool presents possible matches and aids scientists to confirm names provided by using extracted meta-information (i.e. personal web page, co-authorship, publications) to disambiguate names provided. The *Rank accuracy* presents the user a ranking of the names input to the tool by using trust scores from ScienceTrust. Subjects, finally, provide a degree of acceptance of this ranking using a 1 (low) to 5 (high) scale. The tool presents a separate list of scientists for which we cannot provide ranking (i.e. unable to capture evidence of trust) After completing our online study, subjects were invited to visit the ScienceTrust web portal to browse and provide more feedback about the information we captured for the trust networks presented in the web portal.

The ScienceTrust[6] portal's purpose is to provide public access to trust networks captured by using the CI-Learner approach. As of the writing of this publication, the network has 4538 people, 208 organizations, and 38053 publications. From these numbers, we were able to select 5798 co-authorship relationships of interest. This information was extracted from about an 85% of the 208 organizations [4] affiliated to IRIS.

Results: A larger experiment is still in progress and a smaller subset of an intended study is reported in this paper. We are not able to derive and support any conclusion statistically. We have received very interesting feedback from participant subjects and are presented in the aspects of completeness of our network and rank feedback. Despite the fact that CI-Learner covered around 85% of all institutions listed under IRIS web portal, test subjects expressed a satisfactory coverage of scientists they have had collaborations with as well as in the number of publications and trust relations we extracted for each of them. Other related issues to completeness of our network were: missing agents out of the scope of the extraction process of CI-Learner and introduction of noise to our network. After subjects inspected the rankings presented in ScienceTrust, most of them *agreed on the majority of the ranking* overall and there were few instances in which they disagreed with it. Still we are waiting for this experiment to complete and therefore be able to present more interesting results as to what

[5] http://iw.cs.utep.edu/:8081/CI_Learner/evaluation
[6] http://iw.cs.utep.edu:8081/CI_Learner

degree test subjects agree on the ranking provided for any random subset of people from the trust network.

7 Related work

IE and IR have been used to understand complex relationships and structures of social networks and their construction [6]. Interactive systems such as REFFERAL WEB [9] require the explicit participation of individuals members of the networks through the use of personal referrals. This approach is similar to that of the Friend of a Friend (FOAF) project. Another trend in social network construction are POLYPHONET [12], and Flink [13].

POLYPHONET is intended for extracting relations, groups, keywords for an individual. POLYPHONET builds social network out of information obtained through querying Google Scholar; however, focusing more on a generic social network construction. In this case, POLYPHONET has a single source of information for its network components (e.g. individuals, relationships, keywords), requiring an initial set of names later tied by using co-occurrence in web pages; thus, relying on the capabilities of Scholar.

Flink extracts and presents connectivity of Semantic Web researchers limiting its network to researchers who submitted a publication or played a role in the International Semantic Web Conference (ISWC) and Semantic Web Working Symposiums (SWWS). Flink, similar to previous systems, has a set of starting names and includes other extraction mechanisms such as Google using co-occurrence of a name plus the phrase "Semantic Web OR Ontology", FOAF profiles, e-mails, and publications. This project is more topic-centric in the sense it focuses and sets a boundary to researchers related to the sematic web community. Moreover, Flink continues to expand its network through explicit evidence found in FOAF profiles, e-mails, and publication co-authors; however, sets the network boundaries by building an ontology of the topic of interest from the extracted members. Therefore, individuals are related by topic as by means of other relationships (e.g co-authorship).

Similarly there exist systems which incorporate social networks with trust. FilmTrust[7], for instance, is a recommender system which integrates Semantic Web-based social networking enhanced with trust to create a system for movie recommendations[5].

The CI-Learner approach presented in this paper has some similarities and differences to previous systems mentioned. Our approach is topic-centric which focuses on affiliated institutions to IRIS. CI-Learner crawls a scientific portal to find, in our case, organization names that could lead us to discover individuals rather than having such names beforehand.

Similarly, our approach makes use of publications to find the a relationship of interest to our network (e.g. co-authorship). CI-Learner, on the other hand, does not continue to extract meta-information for people that are unknown to us

[7] http://trust.mindswap.org/FilmTrust/

(e.g. not found through the extraction process); therefore, setting the boundary for the extracted network. Moreover, we make use of Google Scholar to extract publications and co-authorship relationships between our network's actors and, therefore, explicitly showing the trust relations among the actors. Consequently, our approach relies on the accuracy and capabilities of Google Scholar.

8 Conclusions

The constant increase of information exchange poses a problem for scientific information acceptance possibly due to the many sources involved in the creation of scientific information and to the fact that scientists of one domain may have no opinion about scientists in another domain. The CI-Learner approach was presented as a possible solution to the problem of extracting a social trust network from information available on scientific portals or the web. The approach consists and considers PML ontology as a characterization mean for the sources in our network; co-authorship from publications are used as evidence of trust among scientists; Google Scholar as the repository for publications; and the CI-Learner algorithm to extract and construct trust networks from scientific web portals. The paper presents the application of the CI-Learner approach to the IRIS web portal to extract a trust network for the Geosciences community. Feedback from test subjects, at this time, expresses an overall satisfaction with the so far extracted network as well as with the ranking presented in ScienceTrust.

Acknowledgments

This work was supported in part by NSF grant HRD-0734825 and by DHS grant 2008-ST-062-000007.

References

1. M. G. Averill. *A Lithospheric Investigation of the Southern Rio Grande Rift*. PhD thesis, 2007.
2. M. Bowler. Htmlunit - A unit testing framework written in Java. http://htmlunit.sourceforge.net/.
3. J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1-7):161–172, 1998.
4. M. D. Cock and P. Pinheiro da Silva. A Many Valued Representation and Propagation of Trust and Distrust. In *In Proceedings of International Workshop on Fuzzy Logic and Applications (WILF2005)*, pages 108–113, Crema, Italy, 2006. Springer.
5. J. Golbeck and J. Hendler. Filmtrust: Movie recommendations using trust in web-based social networks. *Proc. IEEE Consumer Communications and Networking Conference*, 2006.
6. C. Haythornthwaite. Social network analysis: An approach and technique for the study of information exchange. *Library and Information Science Research*, 18(4):323–342, 1996.

7. Incorporated Research Institutions for Seismology (IRIS). www.iris.edu.
8. S. Kamvar, M. Schlosser, and H. Garcia-Molina. The Eigentrust algorithm for reputation management in P2P networks. *Proceedings of the 12th international conference on World Wide Web*, pages 640–651, 2003.
9. H. Kautz, B. Selman, and M. Shah. The Hidden Web. *AI Magazine*, 18(2):27–36, 1997.
10. G. Keller, T. Hildenbrand, R. Kucks, M. Webring, A. Briesacher, K. Rujawitz, R. Torres, A. Gates, and V. Kreinovich. A Community Effort to Construct a Gravity Database for the United States and an Associated Web Portal. In *Special Paper on Geoinformatics*. Geological Society of America, 2004.
11. L. Liu, C. Pu, and W. Han. XWRAP: an XML-enabled wrapper construction system for Webinformation sources. *Data Engineering, 2000. Proceedings. 16th International Conference on*, pages 611–621, 2000.
12. Y. Matsuo, J. Mori, M. Hamasaki, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka. POLYPHONET: An advanced social network extraction system from the Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2007.
13. P. Mika. Flink: Semantic Web technology for the extraction and analysis of social networks. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):211–223, 2005.
14. J. Myllymaki. Effective Web data extraction with standard XML technologies. *Computer Networks*, 39(5):635–644, 2002.
15. P. Pinheiro da Silva, D. L. McGuinness, and R. Fikes. A Proof Markup Language for Semantic Web Services. *Information Systems*, 31(4-5):381–395, 2006.
16. N. D. Rio, P. Pinheiro da Silva, A. Q. Gates, and L. Salayandia. Semantic Annotation of Maps Through Knowledge Provenance. In *Proceedings of the Second International Conference on Geospatial Semantics (GeoS 2007)*, Mexico City, Mexico, November 29-30 2007.