idMesh: Graph-Based Disambiguation of Linked Data

Philippe Cudré-Mauroux
CSAIL
MIT – USA
pcm@csail.mit.edu

Parisa Haghani
I&C
EPFL – Switzerland
parisa.haghani@epfl.ch

Michael Jost
I&C
EPFL – Switzerland
michael.jost@epfl.ch

Karl Aberer
I&C
EPFL – Switzerland
karl.aberer@epfl.ch

Hermann de Meer
U. Passau – Germany
demeer@fmi.uni-passau.de

ABSTRACT

We tackle the problem of disambiguating entities on the Web. We propose a user-driven scheme where graphs of entities – represented by globally identifiable declarative artifacts – self-organize in a dynamic and probabilistic manner. Our solution has the following two desirable properties: i) it lets end-users freely define associations between arbitrary entities and ii) it probabilistically infers entity relationships based on uncertain links using constraint-satisfaction mechanisms. We outline the interface between our scheme and the current data Web, and show how higher-layer applications can take advantage of our approach to enhance search and update of information relating to online entities. We describe a decentralized infrastructure supporting efficient and scalable entity disambiguation and demonstrate the practicability of our approach in a deployment over several hundreds of machines.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous; D.2.12.a [**Software Engineering**]: Interoperability—*Data Mapping*

General Terms

Algorithms, Design

Keywords

Entity Disambiguation, Linked Data, Emergent Semantics, Peer Data Management

1. INTRODUCTION

Until recently, the World Wide Web was a hierarchically organized space separating authoritative content providers from relatively passive information consumers. Today, the organization of the World Wide Web has flattened, empowering end-users with new roles. Publishing data on the Web is easier than ever with the advent of new declarative formats like XML, RDF, or Microformats allowing user-defined information to be encoded in machine-processable ways.

With an increasing amount of entities getting created online comes the pressing need to relate and integrate similar entities published by different end-users. Several initiatives, such as the Linked Data movement[1] recently suggested the use of various declarative links to connect semantically related online entities. As a result, a dynamic and decentralized Web of interlinked data is currently emerging on the Internet. We argue in the following that in this new context, entity disambiguation on the Web is drifting from local, pairwise data integration to large-scale, distributed and uncertain social data management.

Let us consider personal identity management as an example. Several formats encoding personal profiles in (semi) structured ways are today getting widely popular. FOAF [6] (an acronym of Friend of a Friend), as an example, is an RDF vocabulary specification describing persons, their activities, and their relations to other people and objects. XML vCard[2] and hCard [2] are two other examples of standards used to encode personal information in semi-structured ways. Structured profiles encoded in proprietary formats can also be found on an increasing list of Web portals or social communities such as DBLP, Wikipedia, LinkedIn, or Facebook. To add to the confusion, an ever increasing number of Web sites create new structured profiles by automatically combining or reformatting some of the aforementioned sources. DBpedia[3] and Spock[4] are two recent examples of this trend.

The result is a flurry of online, disparate, and machine-readable profiles. Relating these different profiles in a meaningful way would open the door to distributed, large-scale and automated personal information management. This remains however infeasible in practice, as these profiles often refer to different identifiers for the same identity, erroneously use a single identifier to refer to several different identities, or present fake identities altogether. As an example, an online survey aiming at retrieving online profiles related to Sir Tim Berners-Lee – the famous computer scientist – revealed the following in mid-2008: we found 109 different structured profiles related to Tim Berners-Lee (see Table 1). Three of them seemed to be created by Tim Berners-Lee himself. 53 profiles were managed by third-parties and another 53 profiles were generated automatically by combining several sources. Some contained legitimate and up-to-date information, while others were outdated or even fake. Information contained in these profiles vary from contact details to bibliographic records or project-related data.

A few of the profiles referred to his FOAF identity (*http://www.w3.org/People/Berners-Lee/card#i*), some to his

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

[1] *http://www.w3.org/DesignIssues/LinkedData.html*
[2] *http://www.xmpp.org/extensions/xep-0054.html*
[3] *http://dbpedia.org/*
[4] *http://www.spock.com/*

or [20] cannot discriminate uncertain links based on network dispersed evidences. They would thus either discard uncertain information from the matchers and spammers, or treat uncertain information on a uniform basis, generating poor results in both cases (low precision stemming respectively from low coverage and erroneous confidence values for the links). Graph partitioning and clustering methods such as [3] or [22] would suffer from the same fundamental problems, in the sense that they would not be able to disambiguate the links used to analyze the structure of the graph without the ground-truth related to the trust value of the sources.

In the end, the distinctiveness of our approach lies in the application of two core Emergent Semantics [9] principles that cannot be emulated by previous approaches: the analysis of the transitive closures of the probabilistic links, and the reinforcement of global information through network dispersed local evidences.

8. CONCLUSIONS

As the data Web develops, managing heterogeneous online entities is becoming a key problem impeding online data processing and information reuse. Current approaches mostly focus on matching pairs of entities, either by asking the help of end-users or by creating automatic matchers. We proposed in this paper a different approach, based on an analysis of graphs of interlinked entities. Our method complements previous approaches, and could be used in combination with them (in fact, the OKKAM project is investigating such possibilities). Our approach leverages entity relationships to identify constraints and to resolve conflicts by handling trust metrics attached to the sources declaring the relationships.

The technique we presented can be extended in many ways. One compelling extension would be to generalize the constructs we defined to answer other classes of queries. An interesting example would be the *relatedness* relationship. The semantics of this relationship are not well defined in general, but could be expressed in many specific contexts, for example for several variations of rel tags or FOAF links.

9. ACKNOWLEDGEMENT

The work presented in this paper was partially supported by the European project OKKAM No. 215032.

10. REFERENCES

[1] K. Aberer and Z. Despotovic. Managing Trust in a Peer-2-Peer Information System. In *International Conference on Information and Knowledge Management (CIKM)*, 2001.

[2] J. Allsopp. *Microformats: Empowering Your Markup for Web 2.0*. Friends of ED Publisher, 2007.

[3] R. Bekkermand and A. McCallum. Disambiguating Web Appearances of People in a Social Network. In *International World Wide Web Conference (WWW)*, 2005.

[4] G. Bianconi and M. Marsili. Loops of any size and hamilton cycles in random scale-free networks. *Journal of Statistical Mechanics: Theory and Experiments*, P06005, 2005.

[5] P. Bouquet, H. Stoermer, C. Niederee, and A. Mana. Entity Name System: The Backbone of an Open and Scalable Web of Data. In *IEEE International Conference on Semantic Computing (ICSC)*, 2008.

[6] D. Brickley and L. Miller. FOAF Vocabulary Specification 0.91. http://xmlns.com/foaf/spec/.

[7] T. Celic, M. Mullenweg, and E. Meyer. Xhtml Friends Network Relationships Meta Data Profile 1.1. http://gmpg.org/xfn/11.

[8] H. Choi, S. Kruk, S. Grzonkowski, K. Stankiewicz, B. Davis, and J. Breslin. Trust Models for Community-Aware Identity Management. *Identity, Reference, and the Web Workshop at the World Wide Web Conference (WWW)*, 2006.

[9] P. Cudré-Mauroux. *Emergent Semantics*. EPFL & CRC Press, 2008.

[10] P. Cudré-Mauroux, K. Aberer, and A. Feher. Probabilistic Message Passing in Peer Data Management Systems. In *International Conference on Data Engineering (ICDE)*, 2006.

[11] P. Cudré-Mauroux, S. Agarwal, and K. Aberer. Gridvine: An infrastructure for peer information management. *IEEE Internet Computing*, 11(5), 2007.

[12] X. Dong, A. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In *SIGMOD International Conference on Management of Data*, 2005.

[13] R. Hoelzer, B. Malin, and L. Sweeney. Email Alias Detection Using Social Network Analysis. In *International Workshop on Link Analysis (LinkKDD)*, 2005.

[14] A. Hogan, A. Harth, and S. Decker. Performing object consolidation on the semantic web data graph. In *I^3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web workshop at the World Wide Web conference (WWW)*, 2007.

[15] E. Ioannou, C. Niederée, and W. Nejdl. Probabilistic Entity Linkage for Heterogeneous Information Sources. In *International Conference on Advanced Information Systems Engineering (CAiSE)*, 2008.

[16] A. Jaffri, H. Glaser, and I. Millard. URI Disambiguation in the Context of Linked Data. In *Workshop on Linked Data on the Web (LDOW)*, 2008.

[17] N. Koudas, S. Sarawagi, and D. Srivastava. Record linkage: similarity measures and algorithms. In *ACM SIGMOD international conference on Management of data*, 2006.

[18] N. Koudas and D. Srivastava. Approximate Joins: Concepts and Techniques. In *International Conference on Very Large Data Bases (VLDB)*, 2005.

[19] F. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2), 2001.

[20] E. Minkov, W. Cohen, and A. Ng. Contextual search and name disambiguation in email using graphs. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, 2006.

[21] K. M. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Uncertainty in Artificial Intelligence (UAI)*, 1999.

[22] B. On, E. Elmacioglu, D. Lee, J. Kang, and J. Pei. Improving Grouped-Entity Resolution using Quasi-Cliques. In *IEEE International Conference on Data Mining (ICDM)*, 2006.

[23] Y. Raimond, C. Sutton, and M. Sandler. Automatic Interlinking of Music Datasets. In *International Workshop on Linked Data on the Web (LDOW)*, 2008.

[24] W. Roush. People-powered search. *MIT Technology Review*, May-June, 2007. https://www.technologyreview.com/Infotech/18655/.

[25] G. Shafer. *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton Univ. Press, 1976.

[26] W. Shen, X. Li, and A. Doan. Constraint-based entity matching. In *National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference (AAAI)*, 2005.