ISWC 2008

The 7th International Semantic Web Conference

John Breslin
Uldis Bojārs
Alexandre Passant
Sergio Fernández

Social Data on the Web (SDoW 2008)

October 27, 2008

Platinum Sponsors

Ontoprise

Gold Sponsors

BBN
eyeworkers
Microsoft
NeOn
SAP Research
Vulcan

Silver Sponsors

ACTIVE
ADUNA
Saltlux
SUPER
X-Media
Yahoo

Organizing Committee

General Chair
Tim Finin (University of Maryland, Baltimore County)

Local Chair
Rudi Studer (Universität Karlsruhe (TH), FZI Forschungszentrum Informatik)

Local Organizing Committee
Anne Eberhardt (Universität Karlsruhe)
Holger Lewen (Universität Karlsruhe)
York Sure (SAP Research Karlsruhe)

Program Chairs
Amit Sheth (Wright State University)
Steffen Staab (Universität Koblenz Landau)

Semantic Web in Use Chairs
Mike Dean (BBN)
Massimo Paolucci (DoCoMo Euro-labs)

Semantic Web Challenge Chairs
Jim Hendler (RPI, USA)
Peter Mika (Yahoo, ES)

Workshop chairs
Melliyal Annamalai (Oracle, USA)
Daniel Olmedilla (Leibniz Universität Hannover, DE)

Tutorial Chairs
Lalana Kagal (MIT)
David Martin (SRI)

Poster and Demos Chairs
Chris Bizer (Freie Universität Berlin)
Anupam Joshi (UMBC)

Doctoral Consortium Chairs
Diana Maynard (Sheffield)

Sponsor Chairs
John Domingue (The Open University)
Benjamin Grosof (Vulcan Inc.)

Metadata Chairs
Richard Cyganiak (DERI/Freie Universität Berlin)
Knud Möller (DERI)

Publicity Chair
Li Ding (RPI)

Proceedings Chair
Krishnaprasad Thirunarayan (Wright State University)

Fellowship Chair
Joel Sachs (UMBC)

Preface

The *1st Social Data on the Web workshop* (SDoW2008), co-located with the *7th International Semantic Web Conference* (ISWC2008), aims to bring together researchers, developers and practitioners involved in semantically-enhancing social media websites, as well as academics researching more formal aspect of these interactions between the Semantic Web and Social Media.

Since its first steps in 2001, many research issues have been tackled by the Semantic Web community such as data formalism for knowledge representation, data querying and scalability, or reasoning and inferencing. More recently, Web 2.0 offered new perspectives regarding information sharing, annotation, and social networking on the Web. It opens new research areas for the Semantic Web which has an important role to play to lead to the emergence of a Social Semantic Web that should provide novel services to end-users, combining the best of both Semantic Web and Web 2.0 worlds. To achieve this goal, various tasks and features are needed from data modeling and lightweight ontologies, to knowledge and social networks portability as well as ways to interlink data between Social Media websites, leveraging proprietary data silos to a Giant Global Graph.

This volume includes the papers presented at the *1st Social Data on the Web workshop* (SDoW2008), co-located with the *7th International Semantic Web Conference* (ISWC2008), in Karlsruhe, Germany, October 27th, 2008.

SDoW2008 chairs

Dr. John Breslin
Uldis Bojārs
Alexandre Passant
Sergio Fernández

Topics

- Creating RDF-based knowledge using social media services
- Data Portability and Social Network Portability
- Emerging semantic platforms for the Social Web
- Enriching Social Web with semantic data: RDFa, microformats and other approaches
- Linked Data on the Social Web: providing linked data from social media sites
- Ontologies for the Social Web: developing, using and extending lightweight ontologies for social media sites
- Querying and mining social semantic data
- Policies, authentication, security, and trust within collaborative scenarios
- Producing Semantic Web data from social software applications
- Reasoning for Social Web applications
- Semantic blogging, wikis and social networks
- Semantically-Interlinked Online Communities (SIOC)
- Social and semantic bookmarking, tagging and annotation
- Social Semantic Web: combining Web 2.0 and Semantic Web strategies and technologies

Workshop Organization

Program Chairs

- John Breslin, DERI, NUI Galway, Ireland
- Uldis Bojārs, DERI, NUI Galway, Ireland
- Alexandre Passant, LaLIC, Université Paris-Sorbonne, France
- Sergio Fernández, Fundación CTIC, Spain

Program Committee

- Benjamin Nowack, Appmosphere/Semsol, Germany
- Chris Bizer, Free University Berlin, Germany
- Christoph Görn, #B4mad.Net Network, Germany
- Dan Brickley, FOAF Project, World
- Denny Vrandecic, DFKI, University of Karlsruhe, Germany
- Diego Berrueta, Fundación CTIC, Spain
- Eyal Oren, VU Amsterdam, Netherlands
- Eric Prud'hommeaux, MIT/W3C, USA
- Fabien Gandon, INRIA, France
- Frederick Giasson, Zitgist, Canada
- Harry Halpin, University of Edinburgh, UK
- Ivan Herman, CWI/W3C, Netherlands
- Jie Bao, Rensselaer Polytechnic Institute, USA
- Jose E. Labra, University of Oviedo, Spain
- Li Ding, Rensselaer Polytechnic Institute, USA
- Martin Dzbor, KMi, Open University, UK
- Michael Hausenblas, Joanneum Research, Austria
- Paul Miller, Talis, UK
- Richard Cyganiak, DERI, NUI Galway, Ireland
- Sebastian Dietzold, University of Leipzig, Germany
- Sofia Angeletou, KMi, Open University, UK
- Sören Auer, University of Leipzig, Germany
- Susie M. Stephens, Eli Lilly and Company, USA
- Stefan Decker, DERI, NUI Galway, Ireland
- Steve Harris, Garlik, UK
- Tom Heath, Talis, UK

Additional Reviewers

- Jose María Álvarez, Fundación CTIC, Spain
- Michael Martin, University of Leipzig, Germany
- Emilio Rubiera, Fundación CTIC, Spain

Keynotes:

Beyond Walled Gardens: Open Standards for the Social Web

Harry Halpin, University of Edinburgh

Now that the Social Web has finally reached truly widespread adoption, the question remains: Why can't users have their data back? To flip the question, what could researchers discover and application-developers create if they had access to the masses of social data currently spread throughout the Web? Luckily, the building-blocks that allow us to open up these closed walls of data exist right now: all that is missing is a strategy for putting it together. As has been said before: You may not be interested in strategy, but strategy is interested in you.

One weak point is conceptual: Social data portability and privacy are usually viewed as opposing forces. Yet data portability and privacy are mutual benefits that a framework for a mature Social Web could bring users. Today, the walled garden of data fundamentally leads to less security and privacy for users. For example, the lack of portability does not imply a lack of privacy: the data of users may be data-mined and is all-too-portable, for it can be easily sold to parties unknown to the users without their explicit knowledge. Furthermore, the lack of portability has made it common practice for many social web services to ask users to give third-party services access e-mail inboxes, an insecure practice that has already easily led to identity theft. By forcing users to have their data spread throughout the Web under multiple accounts, users often just repeat passwords and user-names, leading to insecure transactions. Solving these real-world problems should not be rocket science, although it may be Web science.

Technically, solutions to all these problems already exist. SAML (Security Assertion Markup Language) in general, and OpenID, provide a usable framework for multi-site log-ins. OAuth can provide authenticated API access. For data portability, a host of incompatible APIs and data formats exist, ranging from OpenSocial to the Contact API, from the XFN microformat to FOAF. One large question that must be answered is how can all these different standards be harmonized, and on what level of abstraction? Given the large amount of work already put into these technologies, instead of asking for a single API to be adopted, a more sensible strategy would be move to an extensible and simple modeling framework, and rather shockingly the Semantic Web may very well be the best solution out there.

Yet, the Semantic Web has its own host of problems, and despite the years of research, very little research has gone into maturing the social side of the Semantic Web. This points out a fundamental flaw in the design of RDF as it stands today: While publishing triples in the wild may do for publicly-available Linked Data, this model of deployment will not work for Social Web. Instead, somehow privacy and data provenance must be built into the very core of the data-format, and be easily accessible. Furthermore, the work on trust needs to go beyond Goldbeck's famous Trust Ontology. How does the Semantic Web and identity providers interact? How can we support both identity integration and multiple profiles for different uses? And on a very basic level, what are the mappings between vCard, XFN, and FOAF? What precisely is the core of social data that should be standardized, and what other components should be left to develop in a decentralized manner? These questions are seemingly simple, and one is unlikely to get a dissertation working on them. Without concrete answers and running code to solve these practical questions, the Semantic Web vision is unlikely to take off.

Luckily, the World Wide Web Consortium provides just such a process where academia, industry, and developers can discuss the future of the Social Web, create a strategy, and then implement it. Furthermore, in a consensus-driven manner, various parties can get on board, without the fear of patent trolls due to the W3C's Royalty-Free Patent Policy. Lastly, the W3C process can help guarantee that various other parts of the Web, like the Mobile Web, can stay involved. What is needed from academia is that the research priorities of the Semantic Web move beyond its roots in classical artificial intelligence to the problem of creating a framework for collective intelligence on the Social Web.

Semantic Search and the Social Web

Peter Mika, Yahoo! Research Barcelona

We start our discussion on the role of search engines in the Social Web by introducing SearchMonkey, Yahoo!'s Semantic Web application platform. A part of Yahoo!'s Open Search strategy, SearchMonkey allows developers and content owners to exploit structured data to make Yahoo! Search results more useful and visually appealing, and drive more relevant traffic to their sites. SearchMonkey contributes to the transformation from the current web to a Semantic Web by creating an ecosystem of developers, publishers and end-users where all participants benefit from contributing and reusing structured data.

We will provide a quick overview of the social data gathered by the Yahoo! crawler and some of the applications built on top of this data. We close with a discussion of the role that (semantic) search engines can and should play in the growing universe of social data exposed in Semantic Web formats.

Contents

A State of the Art on Social Network Analysis and its Applications on a Semantic Web

Guillaume Erétéo[1], Michel Buffa[2], Fabien Gandon[3], Patrick Grohan[1], Mylène Leitzelman[4], Peter Sander[2]

[1] Orange Labs
{guillaume.ereteo, patrick.grohan}@orange-ftgroup.com
[2] KEWI, I3S, Université of Nice, France
buffa@unice.fr, sander@polytech.unice.fr
[3] EDELWEISS, INRIA Sophia-Antipolis, France
fabien.gandon@sophia.inria.fr
[4] Telecom ParisTech, Sophia Antipolis
mylene.leitzelman@telecom-paristech.fr

Abstract. The increasingly popular web 2.0 sites provide the largest social network ever analyzed - users are now considered as plain web resources. Some researchers apply classical methods of social network analysis to such networks; others provide models to leverage the semantics of their representation. We present a state of the art of these two approaches and propose an architecture to merge and exploit the best features of each.

Keywords: social network analysis, semantic web.

1 Introduction

Research conducted on large social networks has principally concerned interviews, enterprise human resources mining, or scientific publications references [17] [39] [51] [53]. However, since its birth in 1992, the web has provided many ways of interaction between people [9], revealing social network structures [54], a phenomenon amplified by the emergence of the web 2.0 [28]. Social networks have been extracted from email communications [52], hyperlink structure of home pages [1], co-occurrence of names [31] [39] [37] [30], and from web 2.0 applications [39]. Dedicated online platforms such as Facebook and Myspace now provide huge amounts of structured social network data to exploit.

In the first part of this paper we recall some classical work from Social Network Analysis (SNA), in particular we detail the popular models used by researchers for representing and visualizing social networks. Definitions of the features that characterize these networks will be presented as well as the corresponding algorithms. In a second part, we discuss the use of semantic web languages and technologies to represent social networks. Finally, we will show that these enhanced representations are a step forward to what we call the “semantic social network analysis” of online interactions.

2 Social Network Analysis

The first representations of social network were sociograms [38] where people are represented by points and relationships by lines connecting them. Much research has been conducted on SNA based on this graph-based view using graph theory [51] [53]. Among important results is the identification of sociometric features that characterize a network. The **density** indicates the cohesion of the network. The **centrality** highlights the most important actors of the network and three definitions have been proposed [19]. The **degree centrality** considers nodes with the higher degrees (number of adjacent edges). The **closeness centrality** is based on the average length of the paths (number of edges) linking a node to others and reveals the capacity of a node to be reached. The **betweenness centrality** focuses on the capacity of a node to be an intermediary between any two other nodes. A network is highly dependent on actors with high betweenness centrality and these actors have a strategic advantage due to their position as intermediaries and brokers [10] [29] [12]. Its exact computation is time consuming, several algorithms tackle this problem [20] [42] [35] [7] with a minimum time complexity of $O(n.m)$ - n is the number of vertices and m the number of edges. To deal with large networks, approximating algorithms [49] [8] [5] [22] and parallel algorithms [4] [50] have been proposed.

Community detection helps understanding the global structure of a network and the distribution of actors and activities [51]. Moreover, the community structure influences the way information is shared and the way actors behave [10] [11] [12]. Scott [51] gives three graph patterns that correspond to cohesive subgroups of actors playing an important role in community detection: **components (**isolated connected subgraphs), **cliques** (complete subgraphs), and **cycles** (paths returning to their point of departure)**.** Alternative definitions have also been proposed such as **n-clique, n-clan** and **k-plex** that extend these initial concepts. Community detection algorithms are decomposed into two categories, either hierarchical or based on heuristics [44] [24] [15]. Two strategies are used in hierarchical algorithms: the divisive algorithms consider the whole network and divide it iteratively into sub communities [23] [56] [21] [49] and the agglomerative algorithms group nodes into larger and larger communities [16] [58]. Other algorithms are based on heuristics such as random walk, analogies to electrical networks or formula optimization [45] [57] [48].

Social network graphs hold specific patterns that can be used to characterize them [43] and accelerate algorithms. According to the small world effect [40], the order of the shortest path between two actors in a social network of size n is $log(n)$. Social networks have an important clustering tendency and a community structure**,** furthermore, the degree distribution follows a **power law** [43].

These graph-based representations are only concerned with syntax – they all lack semantics, and have an especially poor exploitation of the types of relations. We will now see how recently online social networks started to be represented with rich structured data incorporating semantics.

3 Semantic Web Representation of Online Social Networks

Semantic web frameworks provide a graph model (RDF[1]), a query language (SPARQL[1]) and type and definition systems (RDFS[1] and OWL[1]) to represent and exchange knowledge online. These frameworks provide a whole new way of capturing social networks in much richer structures than raw graphs.

Several ontologies can be used to represent social networks. The most popular is **FOAF**[2], used for describing people, their relationships and their activity. A large set of properties is dedicated to the definition of a user profile: "family name", "nick", "interest", etc. The “knows” property is used to connect people and to build a social network. Other properties are available to describe web usages: online account, weblog, memberships, etc. The properties defined in the **RELATIONSHIP**[3] ontology specialize the “knows” property of FOAF to type relationships in a social network more precisely (familial, friendship or professional relationships). For instance the relation “livesWith” specializes the relation “knows”. The primitives of the **SIOC**[4] ontology specialize “OnlineAccount” and “HasOnlineAccount” from FOAF in order to model the interactions and resources manipulated by social web applications; SIOC defines concepts such as posts in forums, blogs, etc. Researchers [6] have shown that SIOC and the other ontologies presented can be used and extended for linking reuse scenarios and data from web 2.0 community sites.

In parallel, web 2.0 applications made social tagging popular: users tag resources of the web (pictures, video, blog posts etc.) The set of tags forms a folksonomy that can be seen as a shared vocabulary that is both originated by, and familiar to, its primary users [39]. Ontologies have been designed to capture and exploit the activities of **social tagging** [27] [33] [46] while researchers have attempted to bridge folksonomies and ontologies to leverage the semantics of tags (see overview in [36]). Once they are typed and structured, the relations between the tags and between the tags and the users are also a new source of social networks.

A lighter way to add semantics to the representation of persons and usages of the web is to use microformats[5] [2] [32]. Some microformats can be used for describing user profiles, including resources and social networks. For example, hCard and hResume microformats describe a person (name, email, address, personal resume etc.) and XFN (XTML Friends Network) is useful for describing relationships.

Millions of FOAF profiles [26] are now published on the web, due to the adoption of this ontology by web 2.0 platforms with large audiences (www.livejournal.net, www.tribe.net). The acquaintance and expertise networks respectively formed by the properties "foaf:knows" and "foaf:interest" reflect real social networks [18]. As a consequence, researchers have applied classical SNA methods to FOAF [47] [25] [26]. Much as today there is only one community of email users (anyone can mail anyone), the adoption of standardized ontologies for non-specialist online social networks will lead to increasing interoperability between them and to the need for uniform tools to analyse and manage them.

[1] Semantic Web, W3C, http://www.w3.org/2001/sw/
[2] http://www.foaf-project.org/
[3] http://vocab.org/relationship/
[4] http://sioc-project.org/
[5] http://microformats.org/

4 Toward a Semantic Social Network Analysis

The online availability of social network data in different formats, the availability of associated semantic models and the graph structure of the RDF language are leading to a new way of analysing social networks. Current algorithms that are applied to SNA are based on graph pattern detection and use very little semantics. The semantics of sociometric patterns that are measured are never taken into account due to the lack of semantics of the representation of the analysed networks. As an example, community detection algorithms are based on graph structure characteristics of social networks but none is based on a sociological definition of community [55] and types of relations are under-exploited. Ontologies were designed to describe particular communities [41] and can be an interesting way to extend community detection among semantically described social networks.

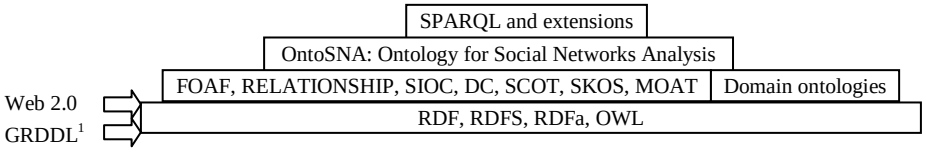

Fig 1: A semantic social network analysis architecture

We are designing an architecture (fig. 1) for a new tool to analyse online social networks. This tool explores RDF-based annotations describing profiles and interactions of users through social applications, using the conceptual vocabulary of previously mentioned ontologies and domain specific ontologies. An ontology, called **OntoSNA** (Ontology of Social Network Analysis), describes general sociometric features and their links to social RDF data. Recently, SPARQL extensions have been proposed for enhancing the RDF graph queries [3] [34] and have been implemented in the search engine CORESE [13] [14]. These extensions enable us to extract paths in RDF graphs by specifying multiple criteria such as the type of the properties involved in the path with regular expressions, or edge directions or constraints on the vertices that paths go through. We reuse these extensions and propose new ones dedicated to SNA in order to make easier the analysis of RDF-based representations of social networks. With such a tool, we can focus or parameterize the analysis specifying types of resources or properties to be considered, and extend classical algorithms with semantic features expressed in SPARQL and based on sociological definitions.

```
select count(?y) as ?cdegree
{
  {?y foaf:knows ?x}
  UNION
  {?x foaf:knows ?y}
}
group by ?x
```

```
select count(?y) as ?cdegree
{
  {?y relationship:worksWith ?x}
  UNION
  {?x relationship:worksWith ?y}
}
group by ?x
```

Table 1: SPARQL queries that extract the degree centrality of actors linked by the property foaf:knows and its specialization "relationship:worksWith".

5 Conclusion

We presented a state of the art on SNA and showed that while this research domain has been exploited for a long time, its application to the web opened new perspectives. The web is now a major medium of communication in our society and, as a consequence, an element of our socialization. The huge number of human interactions through web 2.0 platforms reveal real social networks, and understanding their life cycles is one of the challenges of knowledge sciences. Semantic models of these interactions are well developed and some are now massively integrated into online social applications. The semantic leverage of social data in a machine readable format opens a new way for SNA and the enhancement of online social experiences. We proposed an approach to go toward semantic-aware social network analysis.

References

1. Adamic, L. A., Adar, E.: Friends and Neighbors on the web. Social Networks 25: 211-230 (2003)
2. Adida, B.: hGRDDL: Bridging micorformats and RDFa. J. Web Sem. 6(1): 61-69 (2008)
3. Anyanwu, M., Maduko, A., Sheth, A.: SPARQL2L: Towards Support for Subgraph Extraction Queries in RDF Databases, Proc. WWW2007. (2007)
4. Bader, D. A., Madduri, K.: Parallel algorithms for evaluating centrality in real-world networks, ICPP2006 (2006).
5. Bader, D. A., Kintali, S., Madduri, K., Mihail, M.: Approximating betweenness centrality. WAW2007 (2007)
6. Bojars, U., Breslin, J.G, Finn, A., Decker, S.: "Using the Semantic Web for linking and reusing data across Web 2.0 communities", 2008, J. Web Sem. 6: 21-28.
7. Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Socio. 25(2): 163-177 (2001).
8. Brandes, U., Pich, C.: Centrality estimation in large networks. J. Bifurcation and Chaos in Applied Sciences and Engineering 17(7), 2303–2318 (2007).
9. Buffa, M.: Du Web aux wikis : une histoire des outils collaboratifs. http://interstices.info, online journal, issue of 23/05/08 (2008)
10. Burt, RS: Structural Holes. Cambridge University Press, New York (1992)
11. Burt, R.S.: Structural Holes versus Network Closure as Social Capital. Lin, N., Cook K., Burt, R.S.: *Social Capital: Theory and Research.* Aldine de Gruyter: 31-56 (2001)
12. Burt, R.S.: Structural Holes and Good Ideas. American J. of Sociology 100(2): 339-399 (2004)
13. Corby, C., Dieng-Kuntz, R., Faron-Zucker, C.: Querying the semantic web with the corese search engine. ECAI/PAIS2004 (2004)
14. Corby, C.: http://www-sop.inria.fr/edelweiss/software/corese/v2_4_0/manual/next.php (2008)
15. Danon, L., Diaz-Guilera, A., Duch, J., Arenas,A. Comparing community structure identification. J. Stat. Mech. P09008 (2005).
16. Donetti, L., Munoz, M. A.: Detecting communities: a new systematic and efficient algorithm. J. Stat. Mech. P10012 (2004).
17. Evans, J.A.: Electronic publishing and the narrowing of science and scholarship. Science 5887 (321). (2008)
18. Finin, T., Ding, L., Zou, L.: Social networking on the semantic web. J. Learning organization 5 (12): 418-435. (2005)
19. Freeman, L.C.: Centrality in Social Networks: I. Conceptual Clarification. Social Networks 1 (1979)
20. Freeman, L. C., Borgatti, S. P.: Centrality in valued graphs: A mesure of betweenness based on network flow. Social Networks 13: 141–154 (1991)
21. Fortunato, S., Latora, V., Marchiori, M.: Method to find community structures based on information centrality. Phy. Rev. E 70(5): 056104 (2004)
22. Geisberg, R., Sanders P., Scultes, D.: Better approximation of betweenness centrality. ALENEX08 (2008).

23. Girvan, M., Newman, M. E. J.: Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99(12) (2002)
24. Girvan, M., Newman, M. E. J.: Finding and evaluating community structure in networks. Phys. Rev. E, 69:026113 (2004)
25. Golbeck, J., Parsia, B., Hendler, J.: Trust network on the semantic web. CIA03(2003)
26. Goldbeck, J., Rothstein, M.: Linking social Networks on the web with FOAF AAA08 (2008)
27. Gruber, T.: Ontology of folksonomy: A mash-up of apples and oranges. MTSR2005 (2005)
28. Hendler, J., Goldbeck, J.: Metcalfe's law, web 2.0 and the Semantic Web. J. Web Sem. 6(1):14-20, 2008
29. Holme, P., Kim, B.J., Yoon, C.N., Han, S.K.: Attack vulnerability of complex networks. Phys. Rev. E 65: 056109 (2002).
30. Jin, Y., Matsuo, Y., Ishizuka, M.: Extracting a Social Network among Entities by Web mining. ESWC2007. (2007)
31. Kautz, H., Selman, B., Shah, M.: The hidden Web. AI magazine 18 (2) :27-35 (1997)
32. Khare, R., Celik, T.: Microformats: a pragmatic path to the Semantic Web. WWW2006 (2006)
33. Kim, H., Yang, S., Song, S., Breslin, J. G., Kim, H.: Tag Mediated Society with SCOT Ontology. ISWC2007. (2007).
34. Kochut, K.L., Janik, M.: SPARQLeR: Extended SPARQL for Semantic Association Discovery. Proc. ESWC2007 (2007)
35. Latora, V., Marchiori, M.: A measure of centrality based on the network efficiency. N. J. Phy 9 (6) : 188 (2007)
36. Limpens, F., Buffa, M., Gandon, F.: Rapprocher les ontologies et les folksonomies pour la gestion des connaissances partagées : un état de l'art. IC2008 (2008)
37. Matsuo, Y., Hamasaki, M., Takeda, H., Nishimura, T., Hasida, K., Ishizuka, M.: POLYPHONET: An advanced social network extraction system. WWW2006 (2006)
38. Moreno, J.L.: Emotions mapped by new geography. New York Times (1933)
39. Mika, P.: Social Networks and the Semantic Web. Springer (2007)
40. Milgram, S.: The Small World Problem. Psychology Today, 1(1): 61 – 67. (1967)
41. Mirbel, I.: Vers une ontologie pour les communautés de développement de logiciel libre. IC2009 (2008)
42. Newman, M. E. J.: Scientific collaboration networks. Shortests paths weighted networks, and centrality. Phys Rev E Stat Nonlin Soft Matter Phys 64: 016132 (2001)
43. Newman, M. E. J.: The structure and function of complex networks. SIAM Review 45, 167-256 (2003)
44. Newman, M. E. J.: Detecting community structure in networks. Eur. Phys. J. B 38:321-330. (2004).
45. Newman, M. E. J.: Fast algorithm for detecting community in networks. Phys. Rev. E 69, 066133 (2004)
46. Passant, A., Laublet, P.: Meaning Of A Tag: A Collaborative Approach to Bridge the Gap Between Tagging and Linked Data. LDOW2008 (2008)
47. Paolillo, J. C., Wright, E.: Social Network Analysis on the Semantic Web: Techniques and Challenges for Visualizing FOAF. Visualizing the semantic WebXml-based Internet And Information (2006).
48. Pons, P., Latapy, M.: Computing communities in large networks using random walks. ISCIS2005 (2005)
49. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proc Natl Acad Sci USA 101, 2658-2663 (2004)
50. Santos, E.E., Pan, L., Arendt, D., Pittkin, M.: An Effective Anytime Anywhere Parallel Approach for Centrality Measurements in Social Network Analysis. IEEE2006 (2006)
51. Scott 2000, J.: Social Network Analysis, a handbook, second edition. Sage (2000)
52. Tyler, J. R., Wilkinson, D. M., Huberman, B. A.: Email as spectroscopy: automated discovery of community structure within organizations. Proc. C&T2003 (2003)
53. Wasserman, S., Faust, K., Iacobucci, D., Granovetter, M.: Social Network Analysis: Methods and Applications. Cambridge University Press (1994)
54. Wellman, B.: Computer Networks As Social Networks. Science 293, 2031-34 (2001)
55. Wenger, E.: Communities of Practice: Learning as a Social System Thinker. (1998)
56. Wilkinson, D. M., Huberman, B. A. A method for finding communities of related genes. Proc. Natl. Acad. Sci. (2003)
57. Wu, F., Huberman, B. A.: Finding communities in linear time: a physics approach. HP Labs (2004)
58. Zhou, H., Lipowsky, R.: Network browniam motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. ICCS 2004 (2004).

Combining Social Music and Semantic Web for music-related recommender systems

Alexandre Passant[1], Yves Raimond[2]

[1] DERI, National University of Ireland, Galway,
IDA Business Park, Lower Dangan,
Galway, Ireland,
alexandre.passant@deri.org
[2] Center For Digital Music,
Queen Mary, University of London,
England,
yves.raimond@elec.qmul.ac.uk

Abstract. This paper introduces various ways to suggest music-related content on the Web thanks to Semantic Web technologies. Rather than focusing on features of musical signals or running statistical analysis over listening habits, we detail how social networking, user contributions, and other interlinked data published within the scope of the Linking Open Data initiative can be combined to provide *data-rich* recommendations.

Key words: Web 2.0, Music, FOAF, SIOC, MOAT, Linked Data, Recommendation systems

1 Introduction

Recent Web 2.0 trends introduced new paradigms regarding the way information is produced on the Web. Users were mainly consumers of content, but now tend to become producers by spontaneously publishing and exchanging data [9]. In particular, sharing musical tastes is a frequently used practice. For instance, Last.fm[3] allows its users to publish their listening habits on the Web and offers musical recommendation based on these behaviors. MySpace[4] allows its users to declare themselves as friends of artists that are member of the platform – since many of them are using it to interact with their audience or to promote their work – while Facebook allows one to declare himself as a "fan" of someone. Moreover, people can blog about the latest gig they have seen, tag their favourite band pages on del.icio.us[5] or maintain discographies of artists on Wikipedia[6]. Thus, while the term *social music* was coined by last.fm to mainly identify the act of sharing musical tastes, we will here refer to the publishing and sharing of

[3] http://last.fm/
[4] http://www.myspace.com/
[5] http://del.icio.us/
[6] http://wikipedia.org/

music-related data on the Web, whatever the format is: blog posts, wiki pages, community databases, mp3s or playlists.

In this paper, we will see how *social music* data can be leveraged to the Semantic Web and how it can be used to let people find related musical-content regarding what they are producing or consuming. Rather than focusing on well-known practices for musical recommendations system (that will be described at the end of this paper), our approach consists in using relationships between various types of data (social networks, published content, tags, artist information, etc.) that have been modeled in RDF from those websites. The rest of the paper is organized as follows. In the first section, we describe current issues with music-related social networks and we see how FOAF and linked data can be used to provide a completely open and distributed *social graph*. We also see how SIOC can be used to model user contributions on the Web. In the second part, we overview the music-related part of the Linking Open Data project, providing machine-understandable and interlinked description of artists, bands, and so on. We also see how MOAT can help to move from simple tagging systems to semantic indexing using reference web identifiers. We then describe how the results of these different efforts can be used to suggest related content, whether it is music itself, blog posts, wiki pages... Finally, we go through an overview of existing musical recommendation systems and compare it with the ideas exposed in this paper. Then, we conclude the paper with overview of some future works.

2 Weaving social networks and music-related social data to the Semantic Web

2.1 From closed-world data silos to open social networks

One well-known feature of Web 2.0 websites is the ability to define and maintain social networks. These applications let their users add some other users in their network so that they can receive updates, send direct messages or access some semi-private data. Yet, most of the available social networking sites are isolated amongst each other. When subscribing to a new site, a user must invite his friends again, even if he already defined his network on another website. The process must be repeated each time a new website is joined, leading to what has been called *Social Network Fatigue*[7]. Another common limitation is that it is impossible to add as a friend on a particular service someone who is registered on another service.

Weaving these social networks into the Semantic Web can solve this problem. FOAF – Friend Of A Friend [5] –, a well-known vocabulary to describe agents and their relationships is especially adapted for this use-case. Various exporters for social websites have been written, as for Flickr [14], Last.fm[8] or MySpace[9], providing representations of social networks using FOAF and the

[7] `http://slashdot.org/articles/07/01/02/237223.shtml`
[8] `http://dbtune.org/last-fm/`
[9] `http://dbtune.org/myspace/`

`foaf:knows` relationship. Such networks can therefore be queried uniformally using the SPARQL query language, instead of relying on a specific API for each service. Moreover, various networks can be merged together thanks to resource consolidation, by explicitly defining `owl:sameAs` properties, or relying on the implicit use of `owl:InverseFunctionalProperties`. Such interlinking between various social networks provide a complete distributed and open *social graph*[10], that can be then queried and processed in an uniform way.

2.2 Representing Web 2.0 content on the Semantic Web

As for social networks, the content produced by users of Web 2.0 is held within closed systems, and not represented uniformally. There is a need for a shared semantics in order to represent in a common way user-generated content coming from multiple places. SIOC – Semantically-Interlinked Online Communities [4] – achieve this goal, by offering a model to represent activities of online communities and their contributions. More than 40 applications are now available for SIOC, both for creating or browsing and querying data, from PHP APIs to dedicated SIOC browsers[11].

Furthermore, as Web 2.0 content is not limited to textual data, SIOC defines a Type module[12], so that new types of user-generated content can be easily specified in data exporters, sometimes by reusing terms from other domain-specific ontologies (for example, `dcmi:MovingImage` or `mo:Playlist`). The following example represents a playlist using SIOC and Music Ontology terms.

```
:myRadio a mo:Playlist ;
  mo:track :song1 ;
  sioc:has_creator :me ;
  sioc:site <http://lastfm.com> ;
  dc:title ‘‘Alex’s last.fm playlist’’ .
:song1 a mo:Track ;
  dc:title ‘‘Monkey Man’’ ;
  foaf:maker dbpedia:The_Specials .
```

3 Music and the Linking Open Data project

The Linking Open Data community project [3] aims at publishing and interlinking a wide range of open data sources, by following the four principles outlined by Berners-Lee in [2]: (1) Use URIs as names for things, (2) Use HTTP URIs so that people can look up those names, (3) When someone looks up a URI, provide useful information and (4) Include links to other URIs so that they can discover more things. For example, the Flickr exporter mentioned above provides

[10] `http://www.bradfitz.com/social-graph-problem/`
[11] `http://sioc-project.org/applications`
[12] `http://rdfs.org/sioc/types`

the geolocation of the user using a link to a Geonames resource, allowing an user agent to jump from one dataset to another.

A wide range of music-related data sources have been interlinked within the Linking Open Data initiative [17]. For example, the DBTune project[13] exports the datasets depicted in Fig. 1 in RDF. These datasets encompass detailed editorial information, geolocation of artists, social networking information amongst artists and listeners, listening habits, Creative Commons content, public broadcasting information, and content-based data (eg. features extracted from the audio signal characterising structure, harmony, melody, rhythm or timbre, and content-based similarity measures derived from these). These datasets are linked to other ones. For example, Jamendo is linked to Geonames, therefore providing an easy to build geolocation-based mash-up for musical data. Artists within Musicbrainz are linked to DBPedia artists, MySpace artists, and artists within the BBC Playcount data.

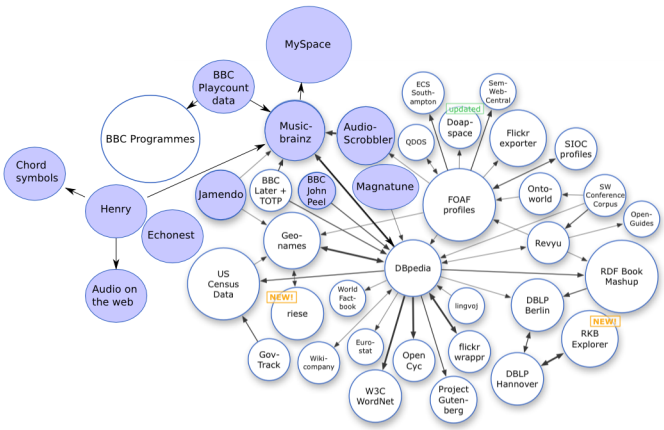

Fig. 1. Depiction of different music-related datasets and their interlinks. The coloured circles correspond to dataset made available within the DBTune project.

4 Using distributed RDF data for musical recommendation

We now have, to the extent of our knowledge, the largest publicly available distributed database of music-related information. We detail in the following the usefulness of such a database for musical recommendations.

[13] `http://dbtune.org/`

4.1 Social-networking and music recommendations

Using social-networks is a first method to suggest musical recommendations, based on listening habits in the network. The Music Ontology (MO [16]) and the related Event Ontology can be used to represent listening habits, as done on the last.fm exporter. Another way is to rely on the `foaf:topic_interest` to provide a direct link between someone and his interests, as artists or bands URIs. An inference rule can be defined to go from the first representation to the second one, represented as follows (using the SPARQL `CONSTRUCT` pattern):

```
CONSTRUCT { ?user foaf:topic_interest ?artist }
WHERE {
  ?ev a event:Event ;
     event:agent ?user ;
     event:factor ?track .
  ?track foaf:maker ?artist .
}
```

Once such `foaf:topic_interest` links have been defined, a single SPARQL query can be used to find listening habits of one's friends, so that they can be suggested to him:

```
SELECT ?artist
WHERE {
  <$uri> foaf:knows [
    foaf:foaf_interest ?artist .
  ] .
}
```

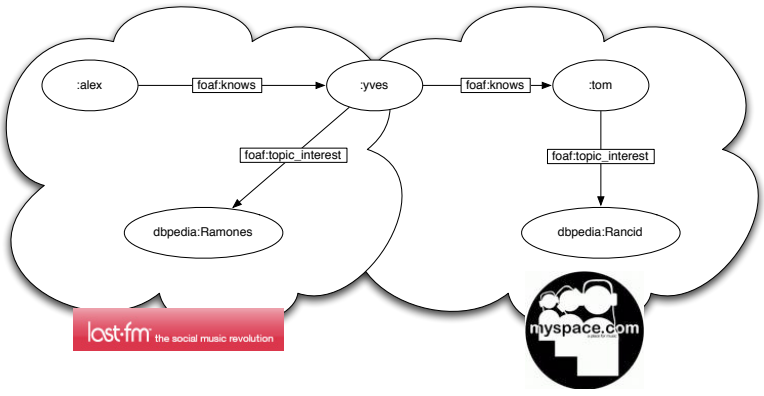

Fig. 2. Using distributed social-networks in recommendation systems

This query therefore implements a really simple collaborative filtering algorithm, that already exist in most of *social music* services. Going further, we can extend this relationships path, for instance to suggest tastes of friends of a friend, and so on. Yet, the most interesting aspect here is that, thanks to the distributed social graph model detailed in Sec. 2.1, the relationships paths can be distributed in various social networks. A recommender system is therefore not limited to a single network, but can combine the aggregation of various social networks and listening habits (or interests), unified using Semantic Web technologies, as outlined in Fig. 2.

4.2 Tag-based recommendations

Tagged content, by using the MOAT framework [15], can also be used to drive music-related recommendations. MOAT allows people to tag their content with URIs, rather than simple keywords, and can be used on existing tagged content, thanks to the LODr application[14].

Once people have tagged their data using URIs, we can rely on relationships between those URIs to suggest related data. For instance, a recommender system for music-related content might suggest to browse a picture tagged with the URI of Joe Strummer on Flickr when browsing a blog post about The Clash, since there is a relationship between both defined in DBpedia, as seen in Fig. 3. By leveraging existing Web 2.0 content on the Semantic Web, as for social networks, we break the bareers between applications and can use data combined, mashed and eventually interlinked through various paths and datasets, that are completely disconnected at a usual "Web of documents" level.

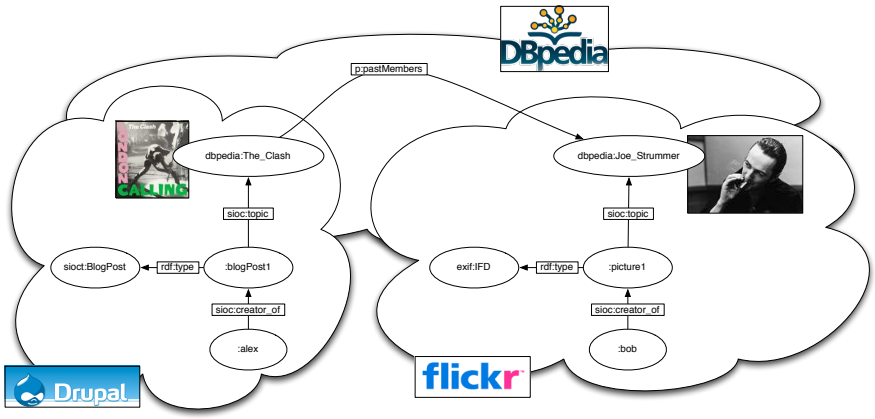

Fig. 3. Interlinking distributed tagged content

[14] http://lodr.info

4.3 Using LOD to find relevant musical content

Using RDF data available from the LOD cloud, various strategies can be adapted to find related content, especially bands or artists. While we can consider direct relationships, as suggesting someone to look for solo members when browsing a band's page, one of our interest is to exploit the structure of linked data and provide recommendations based on paths between artists and bands, for instance, because two bands played at the Cavern in Liverpool or at the CBGB in New York City. Here, one challenge is to find relevant paths between concepts, especially since a lot of various paths exist. As a first step, we analyzed 400 random bands and artist pages from DBpedia to find which are the most used properties, and see how they could be used for music-based recommendations. The 20 most popular are described in the table below (we voluntary excluded plain-text properties as using plain-text relationships may lead to noise and lack of precision):

Rank	Property	Number of relationships
1	`skos:subject`	1930
2	`rdf:type`	882
3	`dbpedia:reference`	847
4	`dbpedia:genre`	450
5	`dbpedia:page`	400
6	`dbpedia:hasPhotoCollection`	400
7	`dbpedia:origin`	355
8	`dbpedia:wikiPageUsesTemplate`	333
9	`dbpedia:label`	265
10	`dbpedia:wordnet_type`	194
11	`dbpedia:associatedActs`	189
12	`foaf:homepage`	178
13	`dbpedia:currentMembers`	151
14	`dbpedia:url`	114
15	`dbpedia:pastMembers`	108
16	`dbpedia:occupation`	97
17	`owl:sameAs`	95
18	`foaf:depiction`	89
19	`foaf:img`	89
20	`dpbedia:wikipage-de`	85

From this list of property, we voluntary excluded those leading to nonrelevant results (as `dbpedia:wikiPageUsesTemplate`), and designed a lightweight facet browser that suggest related artists based on the other properties (and their values) (Fig. 4). We may in the future rely on automatic extraction [12] so that the browsing interface could be automatically adapted to the changes in the Wikipedia structure (and so on the DBpedia one), and it would help to exclude property / values tuples leading to too many results. Another challenge regarding facets definitions is also to automatically find which paths might be

Fig. 4. Finding related artists using Linked Data

relevant to musical recommendations. Especially, future work includes the automatic generation of relevant recommendation paths in linked data. So far, our prototype is closer to a faceted browsing interface than a recommender system, although it demonstrates the richness of the available music-related linked data

As Fig 5 shows, there is a similar path between artists related because of their music genre of vegetarian habits. Yet, on some case this is not so trivial and may also depend on the interest of the users and musical styles. For instance, some might be interested when listening to The Ramones by being suggested to Listen to the Sex Pistols not because they are somehow musically-related by 3-chords songs, but because graph browsing would have tell us that members of both of them died from a drug overdose in a NYC hotel.

5 Comparison with existing approaches

Current music recommendation systems can be clustered in two main categories: collaborative filtering and content-based.

Collaborative filtering consists in recommending items to an user based on the stated tastes of other related users. For example, a user u_2 might like a song s_2 if he likes a song s_1, and that a user u_1 likes both s_1 and s_2. Usually, music recommendations service based on such a methodology use a closed set of information, gathered through a particular service, eg. listening habbits for last.fm

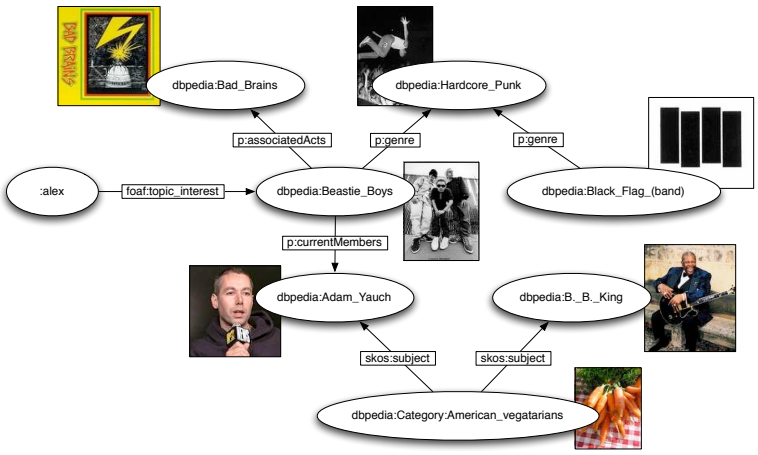

Fig. 5. Exploring different paths between artists

or consumer behaviors for Amazon. Some systems adapt similar ideas in open information environments, such as in [8], where structured taste information is extracted from web documents. Collaborative filtering can also be enhanced by integrating contextual data, reflecting the user's interests at a particular moment [10]. However, collaborative filtering does not tackle the problem of finding items within the *long tail* of music production — those for which the amount of taste data is limited. This approach is only suitable when a large number of users have heard the music and expressed their interest in it.

A recent trend for music recommendation is to use automated analysis of the musical items to drive recommendations, by modeling musical audio similarity [13]. For example, a particular song might be recommended to an user because its rhythm is similar to a song mentioned in his listening habbits. A system based on such a methodology will therefore be able to recommend items within the long tail [7] — it will be able to recommend unknown artists, as the system has no notion of 'popularity'. Current content-based recommendation systems focus on low-level representations of the musical audio content, relative to one of its facet (timbre, structure, melody, harmony or rhythm). However, higher-level representations may lead towards a more widespread adoption of such music recommenders.

Several works combine these two methodologies (collaborative filtering and content-based) to achieve better recommendation [1,18]. Indeed, content-based recommendation can help for the *cold start* problem of collaborative filtering — if the user made no statements about his tastes at first, it will be impossible to provide him with an acceptable recommendation and therefore obtaining taste data from him. The Foafing-the-music project [6] is particularly related to our

work, as it uses distributed social networks using FOAF as well as content-based data available in RDF.

We believe our approach, using not only these two methodologies (collaborative filtering and content-based similarities), but also a wide range of interlinked data in multiple domains, allows the user to get much more justified recommendations [11], such as the ones detailed in sec 4.3. We therefore move towards *data-rich* musical recommendations.

6 Conclusion and future works

In this paper, we outlined various ways to extract RDF data from social music websites and see how it can be used in recommendation systems. While we defined various strategies (based respectively on social networks, tagged data and data published and interlinked within the Linking Open Data community project), they can be combined together for advanced querying and suggesting data, especially as the underlying models are also interlinked (for instance, SIOC and MO reuses the FOAF ontology and MOAT allows to link SIOC data to LOD URIs). Using this global data integration, a query could be used to suggest all bands that will play in your town next month for less than 15euros, that one of your friend blogged about and that play cover songs from a band that you listened to more than 10 times on the last week. Going further, wen can also combine content-based similarity information and interest data as in the following query. The Henry[15] dataset within DBTune indeed publishes `mo:similar_to` statement between musical audio items, based on timbre similarity, so that we could also imagine SPARQL queries using extracted keys and rhythms published by Henry to generate a playlist of recommendations with smooth transitions.

```
SELECT ?track1 ?track2
WHERE {
  <$me> foaf:topic_interest ?artist .
  ?artist foaf:made ?track1 .
  ?track1 mo:available_as ?audio .
  ?track2 mo:available_as ?audio2 .
  ?audio mo:similar_to ?audio2 .
}
```

Regarding those kind of query, relying on highly-distributed data, while not directly related to our work, some efforts should be made regarding dedicated crawlers or distributed querying systems as well as using voiD[16] to find the relevant datasets.

Also, a lot of future work still needs to be done to find the best way to browse the graph of linked data to provide good recommendations. Using attention and taste data available in a user's profile is a first step towards it, but we could also

[15] `http://dbtune.org/henry/`
[16] `http://community.linkeddata.org/MediaWiki/index.php?MetaLOD`

imagine using lots of other data, such as political interests, interests in other arts, etc. While we worked on a first implementation based on data available from DBpedia, we plan to extend it to re-use other datasets and make it publicly available. Especially, future work includes the evaluation of a recommendation system based on such approach, and compare it to existing approaches, to see how it can augment the user experience of discovering musical content.

Acknowledgments

This material is based (in part) upon works supported by the Science Foundation Ireland under Grant No. SFI/02/CE1/I131.

References

1. Marko Balabanovi and Yoav Shoham. Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, March 1997.
2. Tim Berners-Lee. Linked data. World wide web design issues, July 2006.
3. Chris Bizer, Tom Heath, Danny Ayers, and Yves Raimond. Interlinking open data on the web. In *Poster, 4th Annual European Semantic Web Conference (ESWC2007), Innsbruck,Austria*, 2007.
4. J.G. Breslin, A. Harth, U. Bojars, and S. Decker. Towards Semantically-Interlinked Online Communities. *2nd European Semantic Web Conference*, May 2005.
5. Dan Brickley and Libby Miller. FOAF Vocabulary Specification. Namespace Document 2 Sept 2004, FOAF Project, 2004. http://xmlns.com/foaf/0.1/.
6. O. Celma, M. Ramirez, and P. Herrera. Foafing the music: A music recommendation system based on rss feeds and user preferences. In *Proceedings of the International Conference on Music Information Retrieval*, 2005.
7. Oscar Celma and Pedro Cano. From hits to niches? or how popular artists can bias music recommendation and discovery. In *2nd Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition (ACM KDD)*, Las Vegas, USA, August 2008.
8. William W. Cohen and Wei Fan. Web-collaborative filtering: recommending music by crawling the web. *Computer Networks*, 33(1-6):685–698, June 2000.
9. Dan Gillmor. *We the Media: Grassroots Journalism by the People, for the People*. O'Reilly Media , Inc.,, Sebastopol, CA, August 2004.
10. C. Hayes. *Smart Radio: Building community-Based Internet Music Radio*. PhD thesis, Trinity College Dublin, October 2003.
11. Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *CSCW '00: Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250, New York, NY, USA, 2000. ACM.
12. Eyal Oren, Renaud Delbru, Knud Möller, Max Völkel, and Siegfried Handschuh. Annotation and navigation in semantic wikis. In Max Völkel, editor, *Proceedings of the First Workshop on Semantic Wikis – From Wiki*, JUN 2006.
13. Elias Pampalk. *Computational Models of Music Similarity and their Application in Music Information Retrieval*. PhD thesis, Vienna University of Technology, 2006.

14. Alexandre Passant. :me owl:sameAs flickr:33669349@N00. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China*, Apr 2008.
15. Alexandre Passant and Philippe Laublet. Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China*, Apr 2008.
16. Yves Raimond, Samer Abdallah, Mark Sandler, and Frederick Giasson. The music ontology. In *Proceedings of the International Conference on Music Information Retrieval*, pages 417–422, September 2007.
17. Yves Raimond and Mark Sandler. A web of musical information. In *Proceedings of the International Conference on Music Information Retrieval*, Philadelphia, USA, 2008.
18. Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In *Proceedings of the International Conference on Music Information Retrieval*, 2006.

Expressing Argumentative Discussions in Social Media Sites

Christoph Lange[1,2], Uldis Bojārs[2], Tudor Groza[2], John G. Breslin[2], and Siegfried Handschuh[2]

[1] Computer Science, Jacobs University Bremen
ch.lange@jacobs-university.de
[2] DERI, National University of Ireland, Galway,
IDA Business Park, Lower Dangan, Galway, Ireland
{uldis.bojars, tudor.groza, john.breslin, siegfried.handschuh}@deri.org
http://www.deri.ie/

Abstract. Among the activities that people participate in on the Social Web are argumentative discussions and decision making. This paper analyzes a series of use-cases (from the perspective of social media sites) that share the presence of such argumentative discussions and where the structure of online discussions can be represented in SIOC. Our goal is to externalize implicit argumentation structures hidden in the user-generated content. For capturing it and making it explicit, we propose a SIOC Argumentation ontology module as a formal representation.

1 Introduction

Argumentation can be found and captured in a variety of fields ranging from scientific publications to ontology engineering or agent interaction. Social media sites, which represent the hype of the moment, also host argumentative discussions between their members. Such an interactive argumentative discussion usually starts with an initial proposition stated by a single creator. This is then followed by supporting propositions or counter-propositions from other contributors. The actual semantics, both of the interactivity and the argumentation side of the discussion, is hidden in the structure and content created by the participants, and therefore it is difficult to leverage for use by machines.

A possible solution for the first part of the problem is represented by the SIOC initiative (Semantically Interlinked Online Communities) [3]. SIOC aims at integrating online community information, by representing rich data from the social web in RDF. Lately, SIOC became a standard way for expressing user-generated content from social media sites, thus being able to capture their dynamic aspect (interactivity), by modeling the underlying structure of the content. In addition, when complemented with other commonly used vocabularies (like FOAF[3]), SIOC enables innovative ways of expressing personal profiles and social networking information.

[3] http://www.foaf-project.org/

Unfortunately, the second part of the problem, i. e. capturing the semantics of the argumentative discussions, is still open. SIOC provides the means for modeling the structure of the discussions, but it needs a complementary and more precise way to acquire the actual argumentation present in them. There is a relevant number of argumentation models, most of them following the direction given by the IBIS methodology [11]. One of the main issues with many of these models is the focus on a particular knowledge domain, limiting the view of the argumentation to the scope of that domain only, and enabling only partial re-use.

In this paper, we make the first steps towards building an argumentation module for SIOC. We performed a thorough analysis of the existing work done in the argumentation area, and step by step we created our own model that has the specific target of social media sites. By taking into account models like IBIS [11], or DILIGENT [17], our tendency was more towards building upon concepts from these models and adapting them for our own needs.

In sect. 2, we describe background research performed in the SIOC initiative. In sect. 3 we present use-cases from which we extracted the need for an argumentation model for social media sites. Sect. 4 details our proposal, and sect. 5 outlines our plans for deploying our model. In sect. 6 we provide a comprehensive overview of the related work. Sect. 7 presents our future work and conclusions.

2 SIOC Ontology

The SIOC initiative (Semantically Interlinked Online Communities) [3][4] aims to enable the integration of online community information by providing an ontology for representing rich data from social web sites in RDF. It has recently achieved significant adoption through its usage in a variety of commercial and open-source software applications, and is commonly used in conjunction with the FOAF vocabulary for expressing personal profile and social networking information. The SIOC ontology has been published as a W3C Member Submission[5].

The ontology consists of the SIOC Core ontology[6] (consisting of 11 classes and 53 properties) and three complementary ontology modules: SIOC Access, SIOC Services and SIOC Types.

The SIOC Core ontology defines the main concepts and properties required to describe information from online communities on the semantic web. The main terms in the SIOC Core ontology are shown in fig. 1. The SIOC Core ontology was created with the terms used to describe web-based discussion areas such as blogs and message boards: namely *Site*, *Forum* and *Post*. Users create *Post*s organized in *Forum*s which are hosted on *Site*s. *Post*s can reply to other *Post*s. Higher level concepts (data spaces, containers and content items) were added to SIOC as it evolved. By using these classes and related properties, SIOC allows us to structure the information in online community sites and distinguish between different kinds of social web objects.

[4] `http://sioc-project.org`
[5] `http://www.w3.org/Submission/2007/02/`
[6] `http://rdfs.org/sioc/spec`

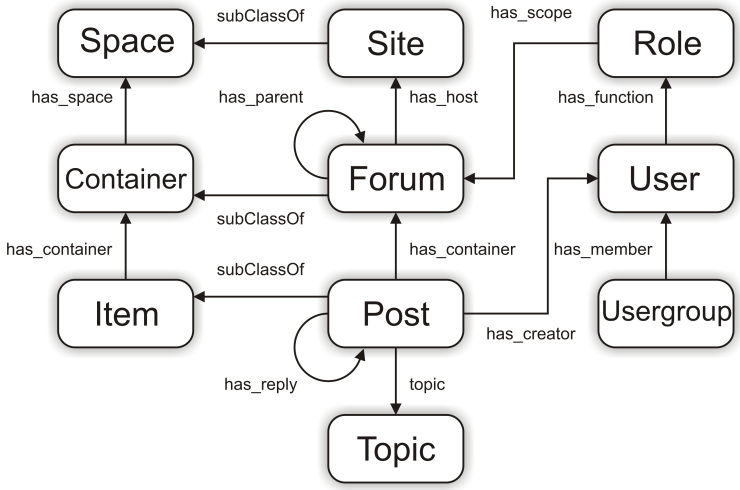

Fig. 1. Main classes and properties in the SIOC Core ontology.

Modules of the SIOC ontology were created as the core ontology was growing. They contain classes and properties that are too specific for the core ontology or cover a particular use case. E. g., the SIOC Types module defines subclasses of SIOC concepts needed for more precise representation of various elements of online community sites (e. g. *sioc_t:MessageBoard* is a subclass of *sioc:Forum*), and introduces new subclasses for describing different kinds of social web objects in SIOC. For example, as a subclass of *sioc:Forum*, one can use *sioc_t:ArgumentativeDiscussion*. With the SIOC Types module we have envisioned certain use cases and provided specific subclasses for them, but we have not further elaborated on supporting these use cases with SIOC. In this paper we are going to elaborate on specific support for argumentation.

3 Use Cases

This section describes use cases for argumentation as used on the social web.

3.1 Forum and Blog Discussions

Forums and blog posts are among the most popular ways of online discussions. Such discussions are a natural place where argumentation and decision making may take place. For example, a group of software developers may use a forum for deciding on the place for their next face-to-face meeting (or decide on details how a particular software function should work). We want to be able to formalize the argumentative structure of these online conversations.

A simple example is a blog post A with a number of replies (A-R_1, ..., A-R_n). The blog post may express an opinion (a position) about something, and

comments are agreeing or disagreeing with it. However, the two-level structure of blogs (post + comments) only allows for simple argumentation and may not be sufficient for "full scale" decision making. If we consider the whole blogosphere and conversations across blogs, argumentation is more interesting. E. g., when a blogger Bob makes a statement on his blog, another blogger Carl can refer to this from his own blog, e. g. with an argument why he thinks Bob is wrong.

Forums and bulletin boards usually have larger communities than blogs do and have a richer conversation structure which can be used for argumentative discussions. Forum sites usually have a number of discussion rooms or forums where each forum is used for conversation on a particular topic or subject area. Each forum consists of conversation threads. A conversation thread is the place where a particular discussion about the topic of this conversation takes place.

The first message starts a thread and is the root of all other conversations and decision making that is taking place in this thread. As such it defines what a thread is about. Imagine a thread which starts with a message "It is time for our monthly off-site meeting. Where shall we go?". This message defines an issue and a goal for this conversation (decision about the meeting venue).

Colleagues respond to this message by posting replies in the same thread. Some of these replies may offer suggestions about the venue (proposed solutions for the issue) and propose the following locations, each with a justification:

1. Aran Islands ("a relaxing place where we can be away from all the noise")
2. London ("a prime business location")
3. the local pub ("it's just across the road!")

Other messages express their support or disagreement with one of the proposals and, finally, the group will come up with a decision, e. g., "local pub it is!"

As can be seen from this example, forums can have a rich conversation structure and posts inside a thread may have different roles in the decision making process.

3.2 Wiki Discussions and Bug Tracking

Wiki discussion pages and comments in bug tracking systems have in common that they usually contain discussions about artifacts of domain knowledge: In wikis that are used as knowledge collections, one article page usually holds knowledge about a distinct subject of interest, and on the corresponding discussion page, people can discuss *about* that subject, or about the way that subject is presented on the article page. On Wikipedia, for example, the former type of discussion is discouraged, whereas the latter prevails [21]. It is common to report issues with the corresponding article (e. g. that the article is found to violate a community policy such as taking a neutral point of view), coming up with ideas on how to solve this problem, and finally voting on these ideas. It is then up to an experienced member of the community to identify the best solution and put it into practice by revising the article [12].

In bug tracking systems, users or developers of a software system report issues with that system (see, e. g., baetle [2]). While unexperienced users often report

issues with the system in general, developers can usually narrow them down to issues with a particular component of the system. Follow-up comments giving elaborations on the issue description or proposing solutions can be given. Some systems support voting on the importance of bugs. In the end a developer takes a decision and changes the affected source code, i. e. fixes the bug. Links from bug reports to the affected software artifacts are shown in some bug trackers which are closely integrated with (source code) revision management systems, such as Trac with Subversion[7]. Similar patterns (discussion of changes and voting or decisions on their acceptance) are present in source code review systems[8].

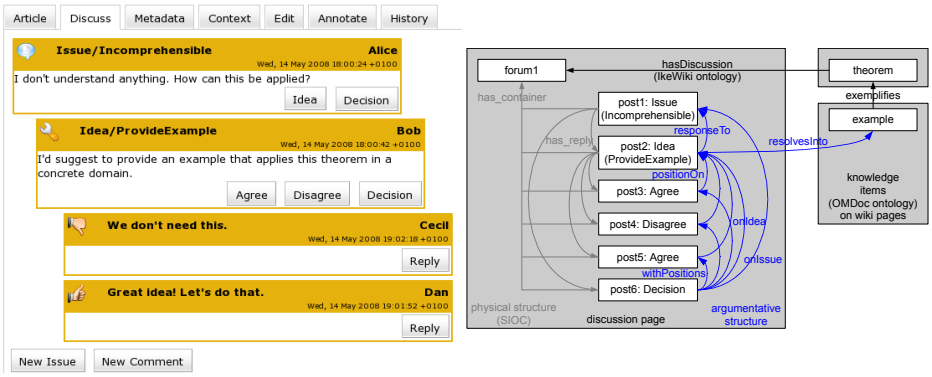

Fig. 2. A structured discussion about mathematical knowledge (left: user interface while discussing; right: full RDF graph)

In previous work [12], we have extended SWiM, a semantic wiki for mathematics, by discussion pages using SIOC for the infrastructure of threaded discussion pages and the DILIGENT argumentation ontology [17] for argumentative structures. The Wiki pages contain artifacts of domain knowledge such as definitions of symbols or theorems, and problems with their conceptualization or formalization can be discussed. In any step of the discussion, the system not just offers to post a "reply", but it displays a button for every type of argumentative primitive that can follow up on the type of the current post, as specified by the DILIGENT ontology. We extended the argumentation ontology by domain-specific subclasses of DILIGENT's *Issue* and *Idea* classes, which allows for arguing about common problems in a more directed way, and for offering semi-automatic software assistance in solving problems[9]. For example, a particular type of issue with a mathematical theorem could be that it is hard to understand, and an idea to

[7] `http://trac.edgewall.org/wiki/TracSubversion`

[8] `http://google-code-updates.blogspot.com/2008/07/looks-good-to-me-source-code-review.html`

[9] A public prototype of the system is currently used at `http://wiki.openmath.org` by domain experts who are revising the OpenMath Content Dictionaries, a lightweight ontology of mathematical symbols.

solve this could be to add an example to the Wiki, which applies the theorem in a practical setting.

This previous work on extending a semantic wiki with argumentation proves the usefulness of combining SIOC and argumentation models. Nevertheless, as we detail in sect. 4, it is not abstract enough to fit *all* the use-cases that generate and manage social media content. At the same time, the focus of DILIGENT on ontology engineering raises, from the social media sites perspective, differences in the interpretation of the semantics of the argumentation concepts. These differences in semantics constitute our main motivation in building a specific argumentation module for SIOC.

4 Approach

4.1 SIOC Argumentation Module

We have identified common cases of argumentative discussions on social media sites (cf. sect. 3) and developed a module for expressing argumentation in SIOC[10].

The SIOC Types module already contains a *sioc_t:ArgumentativeDiscussion*, a subclass of *sioc:Forum* and represents a "placeholder" for expressing that argumentative discussions are taking place in this discussion area (i. e. *sioc:Forum*). Nevertheless, in order to be able to provide a rich and comprehensive argumentation structure, we opted for creating an individual module, that captures the main argumentation concepts we identified as being relevant for our use cases.

The minimum needed for argumentation in SIOC is having a class that can be assigned to any resource in addition to *sioc:Item* or *sioc:Post*, stating that this post has the role of an argumentative *statement*. A post of type *sioc_arg:Statement* represents the root of the argumentative discussion, as it can be followed by a replying post of the same type, modeled by *sioc:has_reply* in core SIOC (thus one statement *refers to* another statement).

The way in which we modeled this relation, was by introducing *sioc_arg :refers_to* as a sub-property of *sioc:has_reply*. Starting from this, we specify additional classes and properties for arguments, all subclasses of *Statement*, or sub-properties of *refers_to*. The reason behind our design was to provide both developers and users with the flexibility of choosing their own way for identifying the argumentation (statement) types for their posts.

From the use cases described in sect. 3 we observe that discussions usually start with an issue or an idea. An *Issue* is a problem to be discussed, a decision on a solution being expected as the result of the discussion. An *Idea* can refer to an *Issue*, then taking the role of a solution proposed for that issue, or it can stand on its own. In this last case, the *Idea* can either be a general idea, not proposing to solve any *particular* issue, or it is a proposed solution for an implicit issue that is not addressed in a discussion post of its own.

On the other hand, *Issue*s can also follow up on *Idea*s – particularly when a discussion was initialized by an *Idea* and then the idea turns out to be problematic.

[10] `http://rdfs.org/sioc/argument`

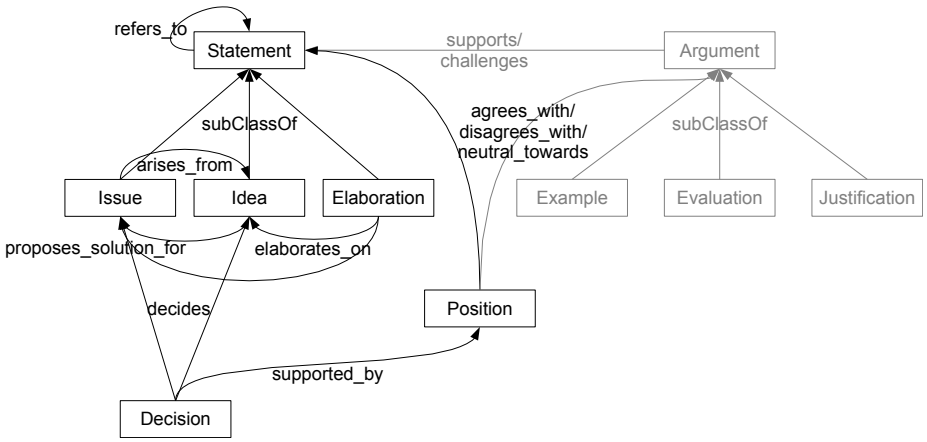

Fig. 3. The argumentation module for SIOC

Most of our concepts (as depicted in fig. 3) have their roots in the DILIGENT argumentation ontology [17], but have a slightly different semantics. A DILIGENT *Issue* states a requirement for the ontology to be designed, and an *Idea* would propose a concrete conceptualization or formalization; ideas cannot represent roots of argumentation threads. Both *Issue*s and *Idea*s can be followed up by *Elaboration*s, which continue the line given by the parent statement, and thus enrich the argumentation model of the discussion.

Users can reply to *Issue*s, *Idea*s, and *Elaboration*s on the former, with *Argument*s, which can be justifications or challenges. An *Argument* tries to argue objectively; it is distinct from a *Position* (see below), which rather conveys the personal opinion of a user. On the other hand, depending on the particular use-case, the presence of the *Argument* concept might not be needed (this being the reason for the different way of representing it in fig. 3). In the Blogosphere, every opinion can be seen as a personal interpretation of the reality, while in a bug tracking system, such opinions are supported by real issues, thus having the circumstance of being considered objective. In addition, the role of an *Argument* can be resumed to: (i) an expression that states if an *Issue* is considered legitimate and worth discussing, and (ii) an expression that shows if an *Idea* can be considered a good solution. Subclasses of *Argument* comprise: *Example*, *Evaluation*, and *Justification*, which can be attached to their parent post by one of the properties *supports* or *challenges*.

In this case, our design was motivated by the Cicero system [7] and allows the retrieval of supporting or challenging arguments with one query step less than a model with positive and negative argument classes and just one property. Also, we opted for only this small set of subclasses for the *Argument* concept, as earlier studies in argumentation have shown that a restricted space of argument types helps to keep a discussion more focused [16].

In a more subjective manner, users can express their *Position*s on a statement – either agreeing or disagreeing. The relation to the statement is represented by one of the properties *agrees_with*, *disagrees_with*, *neutral_towards*. While most argumentation ontologies do not allow the representation of neutral positions in order to force the argumentation towards solutions, they are nevertheless quite common in online discussions. In fact, they are different from the absence of the position in that they express "I do care about this statement, I'm just not decided whether to support it or not." For a minimum working model, it is sufficient to give *Position*s on *Idea*s, but in a more elaborate model *Position*s on *Issues*, *Elaborations*, and even *Arguments* could make sense.

At the end of an argumentative discussion a decision can be taken. It can be documented by replying to the post that started the discussion (either an *Issue* or an *Idea*) with a *Decision*. In the case of making a decision on an issue, one can also link the *Decision* to the winning *Idea*. A *Decision* should be backed by linking to the positions that were in favor of the action decided.

4.2 Overall Recommendations

We would like to leave to the developers of social applications the decision of how much of the SIOC Argumentation Module to support. As shown in sect. 3, the list of use cases is diverse, and thus the need for argumentation support is present. Nevertheless, we do recommend that applications restrict the statement types with which the user can reply to a post to exactly those that are allowed by the schema.

One aspect that our model currently does not capture is a voting scheme. The developer should make the choice of implementing positions as proper posts, or by introducing a vote mechanism on statements. There exist several possibilities to model voting: (i) Collaborative Protégé, for example, allows for either "5-star" or "yes/no" voting [19], whereas (ii) Cicero allows for "yes/no" voting either on individual ideas or in a multiple choice way [7].

When using voting in problem solving, the process can be made more efficient by separating it into two stages: setting a deadline until which all argumentation (such as coming up with ideas and arguing on them) has to be finished, and then allowing the community to vote, as to prepare a final decision. This has been investigated in the Cicero system (cf. sect. 6.3).

A final recommendation would be to close an argumentative thread with a decision, with no more possibility to submit posts. In some applications, such as bug tracking systems, however, the possibility to reopen a discussion should be offered. In a small web of trust it may be feasible to let every user make decisions, whereas in larger social networks we recommend this to be restricted to moderators.

5 Deploying Argumentation on Social Media Sites

The actual process of deploying the argumentation module for SIOC to social media sites is twofold: The software needs to support it, and the users should make

use of it. Concerning software support, the key difference from the deployment of the SIOC Core ontology is that the main SIOC concepts, such as forums, posts, and users, have always been present in software systems running social media sites. Each system had its own internal, idiosyncratic data model with notions of these concepts, so they just had to be externalized using SIOC as a common data model. The argumentation functionality is different in that only very few social media systems already have a model for it. In order to support the SIOC argumentation module, the data model of a system would have to be extended by new components.

Once the software supports the SIOC argumentation in principle, the next challenge is acquiring information about argumentative structures in discussions. This could be done automatically, or by letting the users annotate their posts manually. Automatic annotation would most likely be done using natural language processing techniques, whereas manual annotation needs to be encouraged by a simple and intuitive user interface. We believe that a good approach would offer both automated suggestions, which the user can approve manually, and a user interface that reflects the primitives of our model of argumentation. An example user interface is shown in fig. 2, where an appropriate set of reply buttons is displayed for each post.

6 Related Work

The background and previous research performed in argumentation covers, in general, an important number of related directions. In this section we will focus on three main aspects: (i) background theories and models, which were used as an inspiration by the majority of the currently existing argumentation models, (ii) existing argumentation models, having similar goals with our approach, and (iii) specific implementations of such models in social applications.

6.1 Background Theories

In terms of background theories and models, we found three of them as being relevant for covering the directions from which most of the argumentation models were inspired. The first of them is IBIS (Issue Based Information Systems) [11]. IBIS introduced a methodology for argumentation-based decision making in information systems, adopted by most of the current ontology-based argumentation frameworks, like DILIGENT [17], the Compendium methodology [14] or SALT [10]. A second important background theory is the Speech Acts Theory [5] that models the language aspects of speech acts and their planning in human communication. Although not directly (re)used in argumentation models (one of the main application areas being e-mail workflow modeling), this theory represented groundbreaking research that later led, for example, to the third important theory, i. e. the Dialogue Games Theory. The Dialogue Games theory [4] proposed a novel direction for the general Game Theory by considering discourse analysis and the logics and rhetorics of the human communication. This approach can be found as inspiration in most of the agent-based argumentation models.

6.2 Argumentation Models

One of the early argumentation models was the one of Conklin et al., i. e. gIBIS [6]. This was following closely the original IBIS model and applied its methodology in team-based deliberation. gIBIS served as inspiration for later models like: (i) DILIGENT [17], which applies argumentation in ontology engineering, (ii) Compendium [13], that follows a semiotic [15] approach for dealing with knowledge visualization and design rationale, while complementing argumentation with Cognitive Coherence Relations [14], or (iii) The Zeno argumentation framework [9] applied in mediation systems. Other relevant argumentation models include the one proposed by Torroni et al. in [18] for dealing with agent-based argumentation in the semantic web, in the case of communities of web services the one introduced by Bentahar et al. in [1], or a more lightweight text-based argumentation syntax, as the one proposed by Völkel[11].

6.3 Social Applications

Cicero is a Semantic MediaWiki extension for DILIGENT-like argumentation [7]. In contrast to the SWiM system introduced in sect. 3.2, Cicero is not made for arguing *about* knowledge items, but for solving problems in projects in general. One Wiki page corresponds to one project, issue, or solution proposal (= idea). Arguments are represented as subsections of a solution proposal page. Cicero offers versatile options for voting and deciding. The ontology is DILIGENT-like but slightly different. It is only available in the Wiki; no external implementation is known. For the non-argumentative infrastructure, no ontology (such as, e. g., SIOC) is used.

Fraser et al. have developed an argumentation ontology for e-mails [8]. They shallowly annotate on the top level of every e-mail to keep the annotation easy for users. That means, however, that if an e-mail agrees with some statements of another e-mail but disagrees with others, the value of the argumentative annotation is limited. This issue can also be present in our use cases, and that is why we intend to solve it in the near future, by allowing the representation of fine-grained structures within posts.

7 Conclusion

In this paper we presented the first steps that we have made towards creating an Argumentation Module for SIOC. We started with a series of use-cases that have two facts in common: (i) their structure can be represented semantically with SIOC, and (ii) part of the content created by the users has an implicit argumentative structure. Our goals were to externalize these argumentative discussions and make them explicit via models that are machine-understandable. The model that we have proposed is in its initial stage, and thus we are looking forward to improving it based on the community's feedback.

[11] `http://xam.de/2006/02-ibaw.html`

Most of the use cases presented here deal with problem solving, but we believe that another important benefit of making argumentative structures on social media sites explicit will be a precise *documentation* of discourses that led to earlier decisions. This strengthens the collective memory of a community and will allow new members to retrace and understand the steps of their "ancestors".

For future work, we consider unleashing the potential of SIOC in representing distributed conversation and interlinking argumentations across multiple social media sites. An analysis of the RDF graphs of the argumentations on a single site enables the identification of the merited members of one community, e. g. by counting how many of their ideas have received positive feedback (by *Argument*s or *Position*s) and finally got accepted (by *Decision*s). Then, by making the data of several SIOC-enabled social media sites available to a linked data crawler such as Sindice [20], we can identify traces of the same users in other communities. Such merited users could then automatically be promoted to moderators that are allowed to take decisions. Argumentation in distributed blog conversations can also be an interesting topic to explore in this way.

A second direction we want to follow is to model and enable the representation of fine-grained structured for argumentation in social media sites. Some of the main challenges here are: the creation of the appropriate underlying structures and their links to the SIOC concepts, proper identification of such structures for building the argumentation model, and how to make users willing to split their discourse and to describe its rhetorical structure, all without disrupting their normal flow of work.

In terms of deployment, an interesting direction would be enhancing the existing wiki talk pages (e. g. as used on Wikipedia for discussions and issue solving [12]) with a structured argumentation module as described here. Benefits of doing so can be a more efficient workflow for improving wiki content.

Acknowledgments

The work presented in this paper was supported (in part) by the European project NEPOMUK No FP6-027705 and (in part) by the Líon project supported by Science Foundation Ireland under Grant No. SFI/02/CE1/I131. The authors would like to thank Tuukka Hastrup, Thomas Schandl, Christoph Tempich, Max Völkel, and Stefan Decker for their support and fruitful discussions.

References

1. J. Bentahar, Z. Maamar, D. Benslimane, and P. Thiran. An Argumentation Framework for Communities of Web Services. *IEEE Intelligent Systems*, 22(6), 2007.
2. A. Bernstein, J. Tappolet, H. Story, et al. baetle – bug and enhancement tracking language, seen August 2008. `http://code.google.com/p/baetle`.
3. U. Bojārs, J. G. Breslin, V. Peristeras, G. Tummarello, and S. Decker. Interlinking the Social Web with Semantics. *IEEE Intelligent Systems*, 23(3), May/June 2008.

4. L. M. Carlson. *Dialogue Games: An Approach to Discourse Analysis*. D. Reidel Publishing Company, 1983.
5. P. R. Cohen and C. R. Perrault. Elements of a Plan-Based Theory of Speech Acts. *Cognitive Science*, 3:177–212, 1979.
6. J. Conklin and M. L. Begeman. gIBIS: A Hypertext Tool for Team Design Deliberation. In *ACM Hypertext*, pages 247–251. ACM Press, 1987.
7. K. Dellschaft, A. Gangemi, J. M. Gomez, H. Lewen, V. Presutti, and M. Sini. Practical methods to support collaborative ontology design. `http://www.neon-project.org/web-content/images/Publications/neon_2008_d%2.3.1.pdf`, Feb. 2008. NEON EU-IST-2005-027595 Deliverable D2.3.1, `http://www.neon-project.org/web-content/images/Publications/neon_2008_d%2.3.1.pdf`.
8. C. Fraser, H. Halpin, and K. E. Thomas. Developing an argumentation ontology for mailing lists. In J. Euzenat and J. Domingue, editors, *AIMSA*, volume 4183 of *LNAI*, pages 150–161. Springer, 2006.
9. T. Gordon. The Zeno Argumentation Framework. In *International Conference on Artificial Intelligence and Law*, 1997.
10. T. Groza, S. Handschuh, K. Möller, and S. Decker. SALT – semantically annotated LaTeX for scientific publications. In E. Franconi, M. Kifer, and W. May, editors, *ESWC*, volume 4519 of *LNCS*, pages 518–532. Springer, 2007.
11. W. Kunz and H. Rittel. Issues as elements of information system. Working paper 131, Institute of Urban and Regional Development, University of California, 1970.
12. C. Lange, T. Hastrup, and S. Corlosquet. Arguing on issues with mathematical knowledge items in a semantic wiki. In J. Baumeister and M. Atzmüller, editors, *LWA*, 2008.
13. C. Mancini, D. Scott, and S. B. Shum. Visualising Discourse Coherence in Non-Linear Documents. *Traitement Automatique des Langues*, 47(1):101–108, 2006.
14. C. Mancini and S. B. Shum. Modelling discourse in contested domains: a semiotic and cognitive framework. *Int. J. Hum.-Comput. Stud.*, 64(11):1154–1171, 2006.
15. C. K. Ogden and I. A. Richards. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Magdalene College, University of Cambridge, 1923.
16. H. S. Pinto, S. Staab, and C. Tempich. DILIGENT: Towards a fine-grained methodology for Distributed, Loosely-controlled and evolving Engineering of oNTologies. In R. L. de Mántaras and L. Saitta, editors, *ECAI*, pages 393–397. IOS Press, 2004.
17. C. Tempich, H. S. Pinto, Y. Sure, and S. Staab. An Argumentation Ontology for Distributed, Loosely-controlled and evolvInG Engineering processes of oNTologies (DILIGENT). In *ESWC 2005*, pages 241–256, 2005.
18. P. Torroni, M. Gavanelli, and F. Chesani. Argumentation in the Semantic Web. *IEEE Intelligent Systems*, 22(6), 2007.
19. T. Tudorache and N. Noy. Collaborative Protégé. In C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, editors, *WWW*. ACM, 2007.
20. G. Tummarello, R. Delbru, and E. Oren. Sindice.com: Weaving the open linked data. In K. Aberer, K.-S. Choi, N. F. Noy, D. Allemang, K.-I. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, editors, *ISWC/ASWC*, volume 4825 of *LNCS*, pages 552–565. Springer, 2007.
21. Wikipedia: Talk page guidelines. `http://en.wikipedia.org/w/index.php?title=Wikipedia:Talk_page_guideline%s&oldid=227197584`, July 2008.

Getting to Me – Exporting Semantic Social Network Information from Facebook

Matthew Rowe, Fabio Ciravegna

Web Intelligence Technologies Lab
Department of Computer Science
University of Sheffield, UK
{m.rowe,f.ciravegna}@dcs.shef.ac.uk

Abstract: Information sharing forms a large component of the Social Web as web users become citizens within the web sphere. Within 'walled garden' services they interact socially with their peers through blog posts, image sharing and writing on one another's 'walls'. Porting this data for reuse opens up the social graph associated with the user. In this paper we present work to export Semantic information from the social networking site Facebook using the FOAF ontology by mashing up several Web 2.0 services. We explain the details of our approach and how it supports the data portability movement.

Keywords: data portability, identity, social networks, social networking sites

1 Introduction

The uptake of Social Web and Web 2.0 sites and services has seen a staggering number of people organising their lives online, and moving their offline social space into the virtual realm. A blog post is the new public ego-centric diary entry, writing on someone's virtual 'wall' is the new email, and planning an event can be done within minutes, including sending all invitees their invitations. The general public has embraced the Social Web with open arms, willingly divulging their personal information, their email address, their home address, who they are friends with, and what they like to do. This abundance of information forms a rich social sphere, enabling intelligent reuse for advertisers and consumer groups. However, one may pose the question, who owns this information? A common misconception is that the

user, essentially the author, owns it. It is also a misconception that companies will tell you everything they have logged about you using their site by simply asking for it as the general public, in the UK, believe they have a right to such information based on the data protection act[1]. Therefore surely there is a need for exporting, or claiming back, our information and what belongs to us.

The social web has grown to 9 million unique Facebook users, 5 million unique MySpace users and 4.1 million unique Bebo users [18] in the UK alone, this sharp rise in users and the accompanying habit of information divulgence has begun to plateau: In February of 2008, Facebook saw it's first slump in UK numbers [18], and the movement to make data portable from the social web, namely from social networking sites (SNS), saw the foundation of the Data Portability group[2] to oversee data portability standards and roadmap work within this area. It has become apparent that web users are beginning to wonder what happens to their data.

The data portability movement focuses on opening up the social graph, enabling social information to become more accessible. In 2007 work by [17] explained a 'Bill of Rights for Users of the Social Web' detailing the role of web users as the sole owners of their information, they have the ability to control the distribution of their information, and have the freedom to allow or deny access to their information to any requesting site or service. Such rights allow for a user-centric identity allowing a web user to take their identity with them, similar to the real world, as they move through the social web. The reputation they build up on Ebay, for example, would be transferable to other similar auction sites deeming the level of trust associated with the buyer or seller. The idea that an individual is to be treated as a citizen and not as a consumer was presented in early work by [13] in which virtuality is merged with reality and people in the real world can be held accountable for their online actions.

In this paper we present our work contributing to the data portability movement. We have provided an online application called the Facebook FOAF Generator[3] allowing any user of Facebook to extract their identity and the associated relationships. We use Semantic Web technologies to formalise the information we export into a Semantic format, enabling machine readability. At present generating reusable FOAF is a laborious process, either building it by hand, or using an existing generation tool does not explicitly capture identity. Instead using existing content from a Web 2.0 site bootstraps this process, and provides a rich source of FOAF content. FOAF provides a useful specification for describing information found on SNS, which can be later reused by FOAF enabled sites and services.

Following this section we explain existing methods for formalising identity information using Semantic Web technologies. Section three explains existing techniques for the extraction of social information both from the wider Web and the Social Web. Section four then explains our approach to exporting identity information from Facebook. Section five presents the response we have had to the application, and section six describes our conclusions from the work that we have carried out.

[1] www.opsi.gov.uk/Acts/Acts1998/ukpga_19980029_en_1

[2] http://www.dataportability.org/

[3] http://ext.dcs.shef.ac.uk/~u0057/FoafGenerator

2 Formalising Identity

Before we move on to describing the state of the art within the sensitive area of exporting *my* data from a SNS, it is crucial that we explain existing specifications for describing who a given person is, and who they are associated with. Therefore we describe two existing semantic technologies and their application in the formalisation of personal information, we then move on to present the use of small-level semantics to express implicit knowledge within web pages.

FOAF

In order to capture knowledge depicting a given person the Friend of a Friend (FOAF) specification was established. Work in [3] describes an ontology containing classes and properties designed to encapsulate existing identity knowledge available on the Internet at the time. A person begins by describing himself or herself using the *foaf:Person* class, listing key identity attributes such as name, gender, and resources relating to them. They can also list their interests, and each person is uniquely identified by using the *foaf:mbox* property containing their email address (it is still useful to consider how this is normally the primary key for online 'signup' forms). An alternative identification property is *foaf:openid* conforming to the OpenID[4] perspective of using a unique single URI to establish the identity. The person in question then moves on to describing their friends, each friend is an instance of the *foaf:Person* class. FOAF is both machine-readable, and human-readable, and was adopted by LiveJournal[5], the blogging site, to offer the facility for each user to export their personal information. FOAF has seen a slow adoption by Web 2.0 sites and services, this could be due to the lack of interest in the exportation of social information from one rival site to another. It is, however, the most widely used specification for expressing personal and relationship information within the Semantic Web community.

SIOC

The creation and development of online communities has provided web users with a virtual realm where they can express their thoughts, gain feedback and critique, and interact with individuals harbouring similar interests and beliefs. Most modern web users participate in a web community in some form, albeit forums, chat rooms, newsgroups and SNS. Each community can be interpreted as a single island of rich knowledge, unlinked and unique. The SIOC project focuses on ways to integrate and merge these islands, providing bridges between the knowledge that exists there. As an extension of the FOAF ontology, work presented in [4] outlines an ontology capable of capturing knowledge associated with online communities, and offers formalisms for establishing links between communities.

[4] http://openid.net/

[5] http://www.livejournal.com

The SIOC ontology uses existing specifications such as FOAF and RSS to define classes and their properties. The SIOC ontology contains six main classes of knowledge associated with a community: Site, forum, post, role, usergroup and user. These abstract class definitions offer a flexible range for the capture of knowledge existing in online communities. The concept *sioc:User* can be thought of as the central point for the ontology, s*ioc:User* is a sub class of the *foaf:OnlineAccount* class. The SIOC ontology also defines properties that relate each of the classes to the user in a similar manner to how a user interacts within an online community. Such as *has_moderator* to describe the relationship a forum has with the person monitoring it, *has_administrator* to describe a site's relationship with the user who runs it, and so on. For data portability purposes, SIOC provides a vital specification when exporting data from sites containing groups or forums, or any services where messaging plays a crucial role. SIOC allows the capture of interactions between individuals, and is capable of expressing the role an individual plays in an online social space.

Implicit Social Information

Inclusion of social semantics within web pages has achieved wide spread adoption, particularly with the move by search engines to incorporate advanced knowledge retrieval in their results. New initiatives such as Search Monkey[6] and Google's Social Graph[7] readily use small snippets of semantics within web page code to either return search results containing knowledge of people (Search Monkey), or to construct a linked social graph (Social Graph). This subsection presents two technologies used to code the semantics of social information within a given web page. Each technology uses small-level semantics in order to achieve knowledge descriptions, complying with Tim Berners-Lee's ideology:

"The trick.... is to make sure that each limited mechanical part of the Web, each application, is within itself composed of simple parts that will never get too powerful." [2]

Microformats

The lowercase Semantic Web as paraphrased by [5] involves small-level semantics embedded in XHTML code. Microformats[8] contain useful representations of everyday knowledge commonly found in web pages. The web designer responsible for a given web page is able to add additional machine-readable information within the page content, invisible to the human reader as it is rendered within the browser. According to [5], Microformats make Semantic annotations possible for people and organisations, calendars and events, opinions, ratings and reviews, and tags and keywords.

[6] http://developer.yahoo.com/searchmonkey/
[7] http://code.google.com/apis/socialgraph/
[8] http://microformats.org/

Microformats can also be used to describe relationships between people through link structures[9] using the XFN Microformat. Small level semantics are included within XHMTL code by web designers to support semantic agents to access valid relationship metadata. Examples of this application appear in Blogrolls and link pages. XFN allows the web designers to include details about a relationship such as the type of relationship, whether it is a friend, a colleague or a member of the person's family, the origin of the relationship and the geographical semantics of the relationship.

RDFa

One of the most widely adopted specifications for the semantic description of knowledge is the Resource Description Framework (RDF) [6]. RDF offers a useful formalisation for knowledge capture and has been adapted by [1] to allow inclusion within XHTML through lightweight semantics. Unlike Microformats and particularly XFN, the web designer incorporating RDFa is able to include references to classes that appear in any ontology on the web, therefore supporting ontology reuse. In essence RDFa allows a much more expressive formalisation of semantic information, capable of encapsulating any concept from any ontology providing its namespace exists. Microformats and XFN offer a much lighter approach, one that a non-Semantic Web-savvy web designer would be more comfortable using due to the exclusion of ontology references. All three lightweight formalisms presented in this section offer useful means for the inclusion of personal semantic information within existing web pages, facilitating the extraction and portability of social information.

3 Extracting Social Information

This section details the current state of the art associated with extracting social information. Our work focuses on the portability of an individual's identity, of which we regard social network information to be a crucial component. As described in [9] there are 3 tiers of identity that exist in the modern realm of the Social Web:

- My Identity
- Shared Identity
- Abstracted Identity

The top tier, My Identity, contains information that is persistent about a given person. This information is rarely altered and is unique to the person, aiding their disambiguation from others. The second tier depicts identity information that is more temporal, and contains features and relations that help to bind a given person with others; i.e. relationships. One of the main challenges that the social network mining community has faced is how to ascertain knowledge from this tier. We now present relevant work in social network mining, and explain existing technologies designed to aid with Shared Identity knowledge capture. We conclude by describing existing approaches for porting data from Web 2.0 services into RDF.

[9] http://gmpg.org/xfn/

Social Network Mining

Work by [14] and [11] presents a three-step approach by mining the web for social network information to isolate links between two people, these links are monitored for interactions both in the real world and on the web. [15] presents a two-part process by mining the web and retrieving semantic documents containing FOAF. Relationship strengths between two people are derived using name co-occurrence from web queries. A methodology in [10] crawls the FOAF-web, extracting information from each FOAF file and aggregating with information from other FOAF files. Assertions are made about discovered individuals using the supplied semantic information. Work by both [12] and [16] identifies labels for relations between two people, not only providing a social network, but also providing extra information for each edge in the social graph.

Exporting Social Networks from the Social Web

Flickr Exporter

Work by [7] has created a web application designed to export a user's personal and community information from Flickr[10]. The 'Flickr Exporter'[11] works using a simple user interface requesting the username of the user to export data for. The exporter generates RDF using both the FOAF and SIOC specifications [8]. The user's personal information is described using FOAF, detailing their name, and location, along with their friends on Flickr. The *foaf:knows* relation is used to depict those relationships. User data is interlinked by describing the geographical location of the user using Geonames[12], giving a unique resolvable URI to the location. The SIOC specification is used to describe additional properties of the user detailing the location of the user's photo gallery and their profile image. The exporter also details any communities that the user has joined within Flickr, as this forms an integral part of the site allowing users to share and discuss specific photo topics. The SIOC specification is well suited to this task, as it facilitates the expression of forums and groups, and whether a given user is a member, and what their role is.

Twitter Exporter

Similarly, an exporter of RDF[13] has also been written for the micro-blogging service Twitter. This exporter uses both the FOAF and SIOC ontologies to describe the friends a user is 'following' on Twitter, i.e. their social network. Each friend is described as an instance of *foaf:Agent* containing FOAF semantics such as *foaf:name*, *foaf:nick* and *foaf:homepage*. A blog post made by a user is expressed as an instance

[10] http://www.flickr.com
[11] http://apassant.net/home/2007/12/flickrdf/
[12] http://www.geonames.org/
[13] http://tools.opiumfield.com/twitter/mattroweshow/rdf

of *rdf:Description* containing *sioc:content* to express the content of the post, with *foaf:maker* linking back to the namespace of the poster.

4 Exporting My Identity

Porting data from an existing web service is the main focus of this paper. We present in this section, a successful approach to perform this task involving a social networking site containing information hidden within a 'walled garden'. Our approach makes use of Semantic Web technologies to support reuse of the produced RDF as part of an initiative to make social network information portable from all of the major social networking and social web sites.

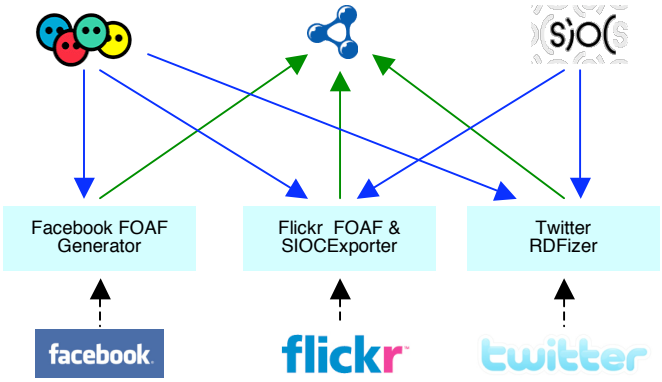

Figure 1 – Current status of data portability from Web 2.0

As figure 1 demonstrates the current model of data portability from social web sites employs various existing Semantic Web ontologies and specifications to generate reusable RDF. Each exporter or generator shown within the middle tier operates by producing RDF according to either the FOAF or SIOC ontologies. In essence the middle tier provides a mapping between the implicit knowledge offered by the service and Semantic Web ontologies, the mapping is commonly between an XML schema or API specification and several ontology concepts. Should a service in the lower tier alter, for example by adding new features, the exporter can be easily adapted to capture this new knowledge. By separating the model as such, this allows for greater flexibility, important when considering the rate of growth of the social web, and its susceptibility to change.

Each exporter and generator within the middle tier can be regarded as a wrapper, purporting knowledge reuse through the generated RDF. In the remainder of this section we present our approach to aiding data portability for social networks through the generation of RDF content according to the FOAF specification from the social networking site Facebook using our application; the Facebook FOAF Generator. In the context of Figure 1, the Facebook FOAF Generator only uses the FOAF

specification for RDF generation; this is due to the lack of community data accessible from the site. Should this data become available, the incorporation of knowledge according to SIOC ontology would also be included. At present only the porting of social network data is supported.

Facebook FOAF Generator

This subsection presents the methodology and processes involved in exporting information from Facebook. The actual exporting of personal information involves mapping the existing Facebook XML data schema with concepts from the FOAF specification to enable sufficient knowledge representation. Figure 2 presents an overview of the process involved when generating FOAF from Facebook.

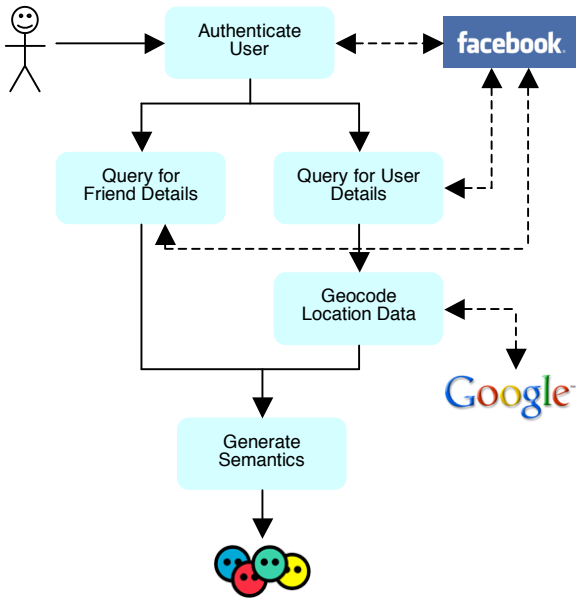

Figure 2 - Overview of the process of FOAF generation from Facebook

The process of FOAF generation begins by authenticating the user through Facebook's API. Upon authentication of the user the process can then begin by querying the API to retrieve the user's personal information and a list of the user's friends together with their details. First the user details are converted into FOAF, the *foaf:name* class contains the name of the user, *foaf:gender* describes the sex of the user, and *foaf:img* is used to contain a URL depicting an image the user.

In order to uniquely identify this FOAF file from others containing similar properties, the *foaf:mbox* or *foaf:mbox_sha1sum* classes are normally used. Facebook does not allow the exporting of email addresses, they are not offered by the API. Instead, our approach uses the user identification number assigned by Facebook to the

user. This is a simple incremental integer that is often seen in browser query strings when using the site. Therefore we used the *foaf:holdsAccount* property, containing the details of the user's account with Facebook. The account is described using the class; *foaf:OnlineAccount* within which the account service, in this case Facebook, is identified using *foaf:accountServiceHomepage*, and the *foaf:accountName* property contains the user identification number. The user's interests are also expressed using the *foaf:interest* property.

The next stage in the process is to create a geocoded representation of the given user's location. In order to do this we took the location information returned by the API and queried the Google Map[14] web service. The service returned a geocoded object containing the longitude and latitude of the given location. Using the *foaf:based_near* property to express a reference to the location, we included *geo:Point* class from the geocode ontology[15] and populated the *geo:lat* and *geo:long* attributes with the latitude and longitude respectively.

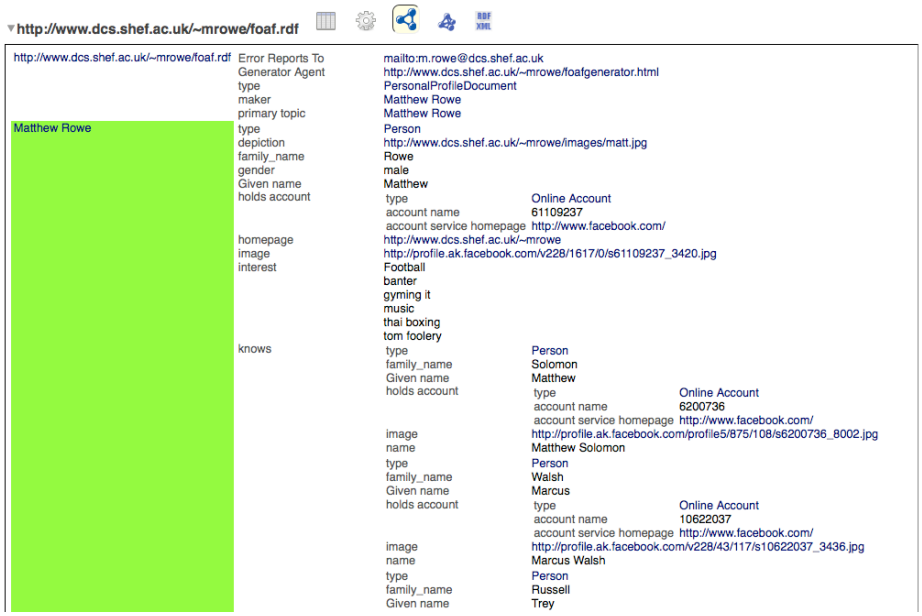

Figure 3 - Example FOAF produced by the Facebook FOAF Generator. Visualisation is achieved using the Tabulator[16] Addon for Mozilla Firefox.

The following stage involves capturing details of each of the user's friends in their social network, essential as we are using a 'Friend of a Friend' specification! The FOAF ontology contains the *foaf:knows* property to specify a relationship between

[14] http://code.google.com/apis/maps/
[15] http://www.w3.org/2003/01/geo/
[16] http://www.w3.org/2005/ajar/tab

two people. We use this property for each of the user's friends, and include the *foaf:Person* class to express the existence of a person. Within this class we include the *foaf:name* property to identity the friend, and the *foaf:img* property to contain an image of the friend. In order to uniquely identify the individual we use the same technique as mentioned previously through the user identification number provided by Facebook and store this value in the *foaf:accountName* property, within the *foaf:OnlineAccount* class.

The completion of this process generates a FOAF file that can then be downloaded by the user and placed on the web for usage by Semantic web agents and crawlers, to open to up the user's social graph enabling social network analysis and knowledge reuse.

5 Response

In this section we present the response given by users of the FOAF generator. Evaluating for usage is a fairly trivial process given that the user has a minimal amount of input, clicking a button, and the process of generating the RDF is fairly straightforward. Therefore we present results from RDF validators on 10 randomly selected FOAF files generated by the application, and the usage statistics since the applications launch in April 2008.

We tested 10 randomly selected FOAF files from the generator. We ran each file through the RDF:About Validator[17], and the W3C RDF Validation Service[18]. All the FOAF files that we tested using both validators were validated as being correct, and no errors were found. Figure 4 presents the usage statistics so far, with time displayed along the x-axis and user numbers along the y-axis. To date the FOAF generator has had 639 users, with an average of 6 active daily users, and 11 people have expressed their satisfaction and admiration for the application by becoming a 'fan' (feature in Facebook). Of those users we are unsure how many are Semantic Web 'enthusiasts'.

Figure 4 - FOAF Generator Usage

[17] http://www.rdfabout.com/demo/validator/
[18] http://www.w3.org/RDF/Validator/

6 Conclusions

In this paper we presented our approach to making social information portable from the social networking site Facebook. Our service has been widely adopted and appreciated by the web community based on the response we have had. The approach provides valid RDF that can be reused across the web, following the Bill of Rights for Users of the Social Web. Using FOAF to provide the semantics of the identity to be exported allows for the expression of relationships and the properties of those individuals. At the time of developing the application there was no need to incorporate the SIOC ontology as the Facebook API did not provide sufficient access to groups and communities the user may have joined and participated in, like Flickr allows. Future work would include the SIOC ontology for describing such community information should this information become available through the API.

One issue that needs to be addressed in the future is the usage of *foaf:accountName* property within the *foaf:OnlineAccount* class, which presents the unique identifier for the exported user and each of the user's friends within the RDF. In a similar manner to *foaf:mbox* we propose that this value should also have a hashed alternative able to protect the web presence of the user if they requested. However, such a process could be leveraged by the adoption of *foaf:openid* as an alternative identification property, by resolving a given profile URL from Facebook as the OpenID URI. Other future work will include additional geographical semantics similar to the technique adopted in [7] to enable interlinking of data. Place names will be resolved to specific geographical concepts using the Geonames service.

The walled garden social networking model inhibits the portability of social network information, forcing researchers to find alternatives through the creation of mashups and web applications to perform the exporting process. By offering a service to export RDF according to the FOAF specification, the social graph linked to a given person can then be reused by other service aiding such functions as relationship derivation, and social network analysis. Once we had developed the application issues arose when attempting to enlist the application with Facebook's application directory. Facebook were unhappy with the exportation of what they believed to be their information onto a separate server for reuse. The application was not blacklisted however; it was simply not listed to inhibit its distribution.

The future of access to social information really depends on the adoption of open standards capable of expressing identity implicitly. As mentioned previously, formalisms such as MicroFormts, XFN and RDFa go some way to allowing the expression of identity features within XHTML. With the creation of standards such as Google's OpenSocial[19] and the steady uptake of shared standards the move towards a more portable Social Web is becoming a reality, with a focus of user-centric identity at the forefront.

[19] http://code.google.com/apis/opensocial/

References

1. Abida. B., Birbeck. M.: RDFa Primer: Bridging the Human and Data Webs. http://www.w3.org/TR/xhtml-rdfa-primer/ (2008)
2. Berners-Lee. T.: Weaving the Web. Harper, San Francisco. (1999).
3. Brickley. D., Miller. L.: FOAF Vocabulary Specification. (2004).
4. Breslin. J., Harth. A., Bojars. U., Decker. S.: Towards Semantically Interlinked Online Communities. The Semantic Web: Research and Applications. (2005).
5. Khare. R.; Microformats: the next (small) thing on the Semantic Web? Internet Computing, IEEE. Vol. 10. Issue 1. Pp 68-75. (2006).
6. Lasilla. O., Swick. R.: Resource Description Framework (RDF) Model and Syntax Specification. W3C Working Draft WD-rdf-syntax-19981008. (1998).
7. Passant. A.: RDF Export of Flickr Profiles with FOAF and SIOC. http://apassant.net/blog/2007/12/18/rdf-export-of-flickr-profiles-with-foaf-and-sioc/ (2007).
8. Passant. A.: me owl:sameAs flickr:33669349@N00. Linked Data on the Web Workshop. WWW08, Beijing, China. (2008).
9. Windley. P. J.: Digital Identity. O'Reilly Media (2005).
10. Finin. T., Ding. L., Zhou. L., Joshi. A.: Social Networking on the Semantic Web. The Learning Organisation, vol. 1 , no. 5, pp. 418-435 (2005).
11. Hamasaki. M., Matsuo. Y., Ishida. K., Nakamura. Y., Nishimura. Y., Takeda. H.: Community Focused Social Network Extraction. Proceedings of 2006 Asian Semantic Web Conference (2006).
12. Jin. Y., Matsuo. Y., Ishizuka. M.: Extracting Social Networks among Various Entities on Web. The Semantic Web. International Semantic Web Conference 2006. pp. 487-500 (2006).
13. K Jordan, J Hauser, and S Foster. The Augmented Social Network: Building Identity and Trust into the Next-Generation Internet. (2003).
14. Matsuo. Y., Hamasaki. M., Nakamura. Y.: Spinning Multiple Social Networks for the Semantic Web. Proceedings of the 2006 Asian Artificial Intelligence Conference (2006).
15. Mika. P.: Bootstrapping the FOAF-Web: An Experiment in Social Network Mining. 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web, Galway, Ireland (2004).
16. Mori. J., Tsujishita. T., Matsuo. Y., Ishizuka. M.: Extracting Relations in Social Networks from Web using Similarity between Collective Contexts. International Semantic Web Conference (2006).
17. Smarr. J.: Bill of Rights for Users of the Social Web. http://opensocialweb.org/2007/09/05/bill-of-rights/ . (2007)
18. Sweeney. M.: Facebook sees first dip in UK users. Guardian Newspaper. 21st Feb. http://www.guardian.co.uk/media/2008/feb/21/facebook.digitalmedia. (2008).

LODr – A Linking Open Data Tagging System

Alexandre Passant

DERI, National University of Ireland, Galway,
IDA Business Park, Lower Dangan,
Galway, Ireland,
alexandre.passant@deri.org

Abstract. This demo paper introduces LODr, a service providing semantic-enrichment features for existing tagged content from various Web 2.0 services, based on the MOAT and Linked Data principles.

Key words: Web 2.0, MOAT, Linked Data, SIOC, Tagging

1 A proposal for semantically-enhanced tagging

While tagging is widely deployed on Web 2.0 websites, it raises various issues which have been largely studied and mainly consist in tags ambiguity and heterogeneity, as well as the lack of organisation between them [3]. While it may not be a problem regarding personal tagging, it becomes relevant when trying to discover and retrieve content that have been tagged by others. Our approach to solve these issues consists in letting people give meaning to their tags using URIs of Semantic Web resources, especially *reference* URIs from the Linking Open Data initiative [1]. This means modeling facts as: *"When I tag this picture 'apple', I mean http://dbpedia.org/resource/Apple_Records, i.e. the record label, not the fruit"*.

Such vision of semantically-enhanced tagging has been recently published through MOAT [5], which consists in (1) an ontology to represent relationships between tags and resources URIs, extending the Tag Ontology [4] and (2) an open-source and collaborative framework to define and share those relationships within a community and help people to bridge the gap between tagging and semantic indexing, without directly facing RDF modeling[1]. This way of tagging content with URIs offer various advantages, as solving ambiguity and heterogenity issues by dealing with machine-understandable URIs rather than words. Most important, it makes tagged data enters the Semantic Web, being interlinked with other resources (DBpedia concepts, FOAF URIs ...), that can be used to retrieve and browse related content.

[1] http://moat-project.org

2 Introducing LODr

While our first experiments with MOAT have been done in a corporate context[2], we decided to extend the approach and implemented LODr – `http://lodr.info` –, a personal open-source application that allows one to re-tag his existing Web 2.0 content and weave it into the Semantic Web thanks to the previous principles.

LODr is thus not yet another tagging service, but a system that provide users a way to semantically enrich existing tagged data that have been created thanks to their favourite tools. The system is based on a set of wrappers (currently available for 5 different services including Flickr and Slideshare), that parse the RSS feeds of user's data, extract items and related tags and translate it to RDF using SIOC [2] and the Tagging Ontology. The data is then stored into a local triple-store and for each tagged item, the user can browse it and give meaning to its tags, using relationships that have been defined by the community, as depicted in Fig. 1. To ease the process of chosing the right meaning, human-readable labels can be displayed instead of URIs. When no URI have been previously defined or when existing ones do not correspond to the meaning of the tag in the current context, a new URI can be added, directly or using the Sindice search widget[3].

As LODr is based on the MOAT principles, it requires interaction with a dedicated tag server that stores the relationships between tags and URIs for the community that uses it. While a default public server is available, a community can use the tool with its own tag server which might be useful, for instance, in a company. LODr is completely Semantic-Web based, and its RDF backend is powered by ARC2[4]. It features tagcloud and conceptcloud interfaces, different ways to browse items (all items, re-tagged items ...) and use Exhibit[5] for faceted browsing. The whole template uses RDFa[6], so that content can be easily discovered and crawled by dedicated semantic search engine. Moreover, as we wanted to offer value-added services to end-users, we wrote a dedicated Ubiquity[7] command to discover tagged items when browsing the Web, e.g. to find all items linked to a DBpedia URI when browsing the related Wikipedia page.

Finally, as various *augmented-tagging* applications have been recently published, we think the originality of LODr resides in: (1) its way of linking tags and tagged data to existing Semantic Web resources, and not only relating tags together as in Gnizr[8], which makes the application live in its own closed-world, (2) its ability to use any URI (e.g. FOAF profiles, Semantic Web conference corpus URIs) and not only DBpedia ones as in Faviki[9], (3) its integration with existing Web 2.0 content, which does not require to subscribe to a new an independant tagging application, avoiding *social network fatigue* and (4) its complete

[2] `http://www.w3.org/2001/sw/sweo/public/UseCases/EDF/`
[3] `http://sindice.com/developers/widget`
[4] `http://arc.semsol.org`
[5] `http://simile.mit.edu/exhibit/`
[6] `http://rdfa.info`
[7] `https://wiki.mozilla.org/Labs/Ubiquity`
[8] `http://gnizr.com`
[9] `http://faviki.com`

Fig. 1. Re-tagging a Slideshare item with LODr

Semantic-Web based interface and especially its RDFa output and SPARQL endpoint, which makes easy to integrate its data into other applications.

Acknowledgments

This material is based upon works supported by the Science Foundation Ireland under Grant No. SFI/02/CE1/I131.

References

1. Chris Bizer, Tom Heath, Danny Ayers, and Yves Raimond. Interlinking open data on the web. In *Poster, 4th Annual European Semantic Web Conference (ESWC2007), Innsbruck,Austria*, 2007.
2. John G. Breslin, Andreas Harth, Uldis Bojars, and Stefan Decker. Towards Semantically-Interlinked Online Communities. *2nd European Semantic Web Conference*, May 2005.
3. Adam Mathes. Folksonomies - cooperative classification and communication through shared metadata, December 2004.
4. Richard Newman, Danny Ayers, and Seth Russell. Tag ontology, December 2005.
5. Alexandre Passant and Philippe Laublet. Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China*, Apr 2008.

Modeling Online Presence

Milan Stankovic
Université Paris-Sud XI
91405 Orsay cedex, France
milan.stankovic@gmail.com

Abstract. In this paper, we introduce the notion of Online Presence, a concept related to user's presence on online services. We identify interoperability issues in the field of exchange of the online presence data and propose a solution in building a common model for semantic representation of online presence data. We present the Online Presence Ontology (OPO) together with benefits such an ontology could bring. Finally we outline some directions for future work on this matter of emerging importance.

Keywords: Online Presence, Ontology, Social Web

1 Introduction

With the appearance of instant messaging (IM) tools and Social Web sites, most notably Social Networks, Internet faced a proliferation of social activities among users. On a typical service that offers some form of social interactions, users present themselves to their contacts by maintaining user profiles. Services that favor direct and frequent communication tend to include descriptions of user's temporary state in the profile. By the elements of temporary state, we mean primarily custom messages on IM platforms and social networks, as well as description of availability/willingness to chat. Often, visual representations known as avatars are used to depict user's online persona. In fact, the activity of maintaining this kind of user profiles is no more than creating an image of one self's presence in the online world, a representation how one wishes to be seen by his/her contacts. The use of custom messages, IM statuses and avatars became a common way for users to make known the character of their presence on some online service and in the online world in general.

The variety of different purpose social applications and the fact that one's friends can be spread over various services for the same purpose, motivate users to maintain their profiles on many services, often just copying custom messages and other data related to their presence online. Different formats used by those services to represent semantically identical data, stands in the way of a user's ability to seamlessly transfer the data among services.

In this paper we propose an ontology-based approach for modeling the semantics of the aforementioned aspects of a user's appearance in the online world, with the final aim of enabling interoperability among services that collect and use online presence data. This is especially significant in the domain of exchanging IM statuses from different IM status scales – a use of emerging importance by the recent proliferation of inter-platform IM.

In Section 2 we give an insight into the problem's nature, and illustrate it with scenarios of use where difficulties for users, influenced by this problem, can be easily spotted. In Section 3 we explain our ontology-based approach for facing the outlined interoperability

problem, and go into details of the Ontology design. In Section 4 we take a look at related work while Section 5 is reserved for conclusions and plans for future work.

2 Problem Specification

Let us first consider some sources of online presence data to assure better understanding of later discussions. First of all, those are IM platforms that publish information like custom messages, avatars and statuses of availability for chat. We will call this kind of information Online Status. Then, there are Online Social Networks with custom messages and profile pictures, as well as services that publish short Online Statuses (like Twitter). There is also a large number of Social Web sites (e.g., Digg, Technorati, Flickr, etc.) and online communities (e.g., Web Forums) publishing similar data.

The following two scenarios demonstrate the essence of the interoperability problem. Let us first consider an example user, Harry, who uses an IM platform and a variety of different purpose social networks in order to keep up with his friends who are spread over several social networks. Besides that, he likes all those networks for different functionalities they offer. When Harry wants to define a custom message to share an insight about his current state (e.g., "going to New York") he has to do it on every particular service i.e., on the IM platform and every social network used. It would be a lot easier for Harry to define his custom message only once, and rely on some kind of exchange mechanism between services. A part from the lack of actual collaboration between the services in question, the lack of unified data exchange format that would define the shared semantics of the domain also presents a significant obstacle for solving the user's problem.

In the second scenario we consider the problem of another example user, Sally, who uses two different IM platforms (e.g., Skype and GoogleTalk) for the same reason of keeping in touch with friends using different platforms. When Sally is doing something important and does not want to be disturbed she has to choose either 'Busy' or 'Do Not Disturb' status on each IM platform. Sally's difficulty is even a greater challenge than the one faced by Harry, because in the case of exchanging IM statuses between different IM platforms the problem of mappings between different IM status scales appears as an additional obstacle.

3 Modeling Online Presence

In order to meet the interoperability challenges introduced in the previous section, we created a model that enables semantic representation of all the data that are the subject of exchange. Using OWL-DL we formalized the model into the Online Presence Ontology (OPO). In further sections we consider the specifics of the solution and their practical implications.

3.1 The General Idea

Having observed that the creation of custom messages, IM statuses, avatars, etc. represents a part of users effort to publish their presence online, we decided to gather all that data under a common roof of the notion of Online Presence. This notion congregates all the data representing temporary aspects of a user's online presence, thus complementing his/her more stable online profile data defined, for example, by the FOAF vocabulary [1].

In order to develop a comprehensive model, we have analyzed the major sources of online presence data i.e., IM platforms, social networks and other social applications. As a result of that analysis we created a list of aspects that determine online presence in the sense of today's applications. The list contains avatars, custom messages, IM statuses, but also some complex aspects like the possibility for other users to find the user's contact details in public listings as well as the willingness of a user to receive notifications by applications.

While designing the model, we had in mind the dynamic nature of social applications, and their ever increasing functionalities. Thus we favored the flexibility and extensibility in our design in order for it to be able to support further changes in the way people present themselves online.

3.2 Ontology Design

OnlinePresence, the core class in OPO, represents a placeholder for all the aspects of a user's presence in the online world. Having in mind possible development of new, currently unpredictable, aspects of presence in the online world, we defined a class, OnlinePresenceComponent, to represent an abstract component of the OnlinePresence. This design decision introduces flexibility in modeling the building blocks of OnlinePresence.

Relying on the current state of practice in the area of online social interactions we have defined three components of Online Presence: Online Status, Notifiability and Findability (Figure 1). These are modeled as subclasses of the OnlinePresenceComponent class.

First we perceived the need to distinguish the attitude towards the possibility of interaction with humans (represented with Online Status) from the attitude towards the possibility of being contacted/interrupted by a machine. By a contact from a machine we mean the practice of IM programs to pop-up notifications. Many IM programs allow users to specify whether to allow this type of disturbance or not. This particularity is modeled with the Notifiability component, by assigning one of the different Notifiability instances (e.g., AllNotificationsPass, NotificationsProhibited) to the Online Presence.

Findability is a component meant to describe the possibility of other users to access a person's contact details and online presence data. In most systems this property is defined by users in some form of settings. The approach for defining Findability is the same as with Notifiability. Different predefined instances are used to denote various states of Findability (e.g., PubliclyFindable and ConstrainedFindability).

Finally, Online Status represents what one may call availability for chat – the status used by IM platforms. While analyzing different status scales used by different IM platforms we concluded that the complexity created by all the differences between them could be best resolved by introducing different Online Status Components whose combination would permit all existing IM scales to be mapped into one single model – the one used in OPO. We have defined the following components of the Online Status:

- Activity – denotes whether a user is present or away from the service;
- Disturbability – denotes whether a user wants to be contacted or declares himself/herself as busy;
- Visibility – denotes the possibility of others to view a user's actual state of presence;
- Contactability – denotes whether the possibility to contact a user is restricted.

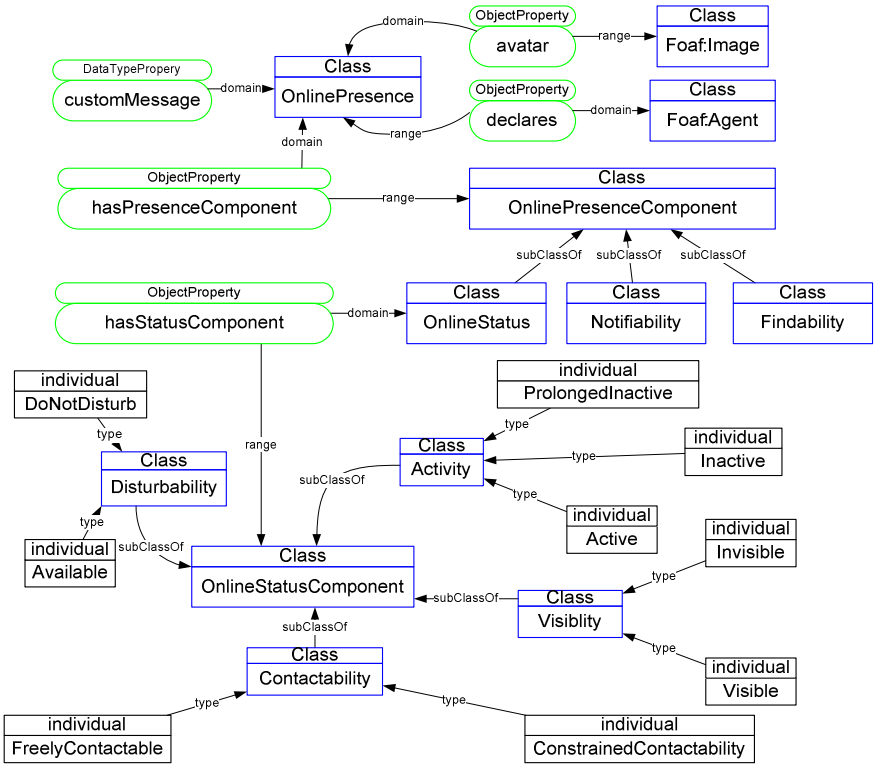

Fig. 1. The partial view of concepts and properties of OPO

These are modeled as subclasses of the OnlineStatusComponent class. By combining different predefined instances of these OnlineStatusComponents, every IM status scale that we took into consideration can be unambiguously described (see Section 3.3).

As for the custom message, and avatar, we modeled them as properties of the OnlinePresence class since their lack of complexity does not demand the creation of new classes for them.

The concept of Online Presence itself is connected to the class Agent from the FOAF vocabulary [1] using the property *declares* (see Figure 1).

The OPO is available at http://ggg.milanstankovic.org/opo/ns/.

3.3 Mappings of Online Status scales to the OPO

Representing different Online Status scales in OPO is one of the most complex and most important issues in the OPO design. The complexity arises from differences in meaning and usage of particular Online Statuses on different platforms. OPO delivers a flexible model to represent the semantics of Online Statuses thus making their descriptions precise and understandable for the IM platform importing them.

In order to demonstrate the actual benefit of the OPO in this domain, we will take the example of SkypeMe status used on Skype IM platform and show its OPO representation. The description of this status provided by Skype is the following:

"SkypeMe! mode allows everyone else on Skype know that you are available and interested in talking or chatting. This includes people who you do not know or you have not authorized but who can find you by searching the Skype directory. SkypeMe! mode ***disables*** *your privacy settings and allows anyone to contact you, whether you've authorized them or not."*[2]

The corresponding representation in OPO (Turtle syntax [3]) would be the following:

```
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix opo: <http://ggg.milanstankovic.org/ontologies/
OnlinePresence.owl#>.
:SkypeMe rdf:type opo:OnlineStatus ;
       opo:onlineStatusName "SkypeMe";
       opo:hasStatusComponent opo:Active, opo:Available,
       opo:FreelyContactable, opo:Visible.
```

This representation in OPO describes the SkypeMe status trough several statements that define it in terms of Online Presence aspects. After defining SkypeMe to be of type OnlineStatus, we assign it the adequate onlineStatusName (a string used by IM platforms to identify different statuses). Then we describe this Online Status in terms of Online Status Components. The following components are assigned to SkypeMe: Active – the user is active on the service (not away); Available – the user is available for contact (as opposed to busy); FreelyContactable – everyone can contact the user; and Visible – the user's onlinePresence is visible. With all these characteristics declared in OPO, our description of SkypeMe status is fully compliant to the textual description provided by Skype.

Let us recall our example user Sally, who wanted to propagate the Online Status from one IM platform to another, and suppose that she wanted to transfer her SkypeMe status to GoogleTalk. Without using OPO, relying on individual interpretations by IM platforms, GoogleTalk would just recognize her SkypeMe status as equivalent to Available on the GoogleTalk scale. In such an exchange there would be a significant loss of semantics, since the two statuses are not actually equal. Representing the status being exchanged in OPO preserves its semantics, allowing it to be correctly transferred. This way, GoogleTalk could import SkypeMe status and comprehend it as its Available status, since it does not support the variations of the Contactability dimension. However, in further transfers to other IM platforms, GoogleTalk could export the original OPO description allowing the application of all OnlineStatus dimensions on some other platform that may support them.

Thus the OPO serves, in this domain, as a mediator preserving the semantics of online status scales in their exchanges, enabling more precise transfers of data between heterogeneous services.

4 Related Work

One of the rare examples of related work in the field is the XMPP protocol [5]. However its main aim is to enable inter-platform IM, while in the field of interoperability of various IM status scales it does not provide much functionality. A large number of inter-platform IM tools built on top of XMPP have to conform to a very poor XMPP IM status scale. Creating mappings from that scale to others is left to individual implementations. In this area OPO can be of essential value for achieving a richer exchange of disparate online status scales by enabling an unambiguous description and understanding of semantic categories that determine them.

Another interesting example of related work is the MeNow Schema[1], aimed at enabling representations of various statuses that a user can assume online. The Schema's exceptional value is in the possibility to represent many different aspects of the context of user's presence online (e.g., current book, current music, company of others, etc.). On the other hand we find that the Schema underestimates the importance of the possibility to semantically represent different qualitative aspects of online status, a feature strongly supported by OPO.

4 Conclusions and Future Work

Building of the OPO represents, at its core, a task of bringing the Social and the Semantic Web closer together. It is inspired by the idea that the future of the Web lies in the merging of those two approaches [4].

The benefits of OPO and its flexible and extensible design are numerous. First of all, it enables interoperability between applications that collect online presence data. This interoperability could result in users being able to correctly transfer their online presence data from one service to another regardless of the type of the service, and possibly unify their appearance online over multiple services. The ontology itself is just a prerequisite for this goal, and applications would have to adopt the practice of exchange in accordance with the Data Portability[2] initiative in order for the goal to be achieved.

The favorable properties of Semantic Web technologies, allow for assembling partial semantic descriptions of Online Presence, published by various services, into one coherent description.

The future work will primarily focus on building plug-ins enabling applications and social websites to publish Online Presence metadata. Scenarios of metadata exchange will also be developed, resulting possibly in building a centralized server for resolving privacy issues concerning the exchange. We will also consider the possibilities to use semantic rules and policies to allow for defining restrictions of some aspects of Online Presence to some categories of users and other more sophisticated statements. Last, but not the least, we will consider the possibilities to integrate with the XMPP protocol, widely used in cross-platform chat, in order to enrich the OnlineStatus data being exchanged and build a ground for more meaningful mappings.

References

1. Brickley, D., Miller, L., 2006. Foaf vocabulary specification. Available at http://xmlns.com/foaf/spec/
2. Skype help record. Available at http://support.skype.com/index.php?_a=knowledgebase&_j=questiondetails&_i=442
3. Becket, D., 2007. Turtle - Terse RDF Triple Language. Available at http://www.dajobe.org/2004/01/turtle/
4. Bojars, U., Breslin, J., Peristeras, V., Tummarello, G., and Decker, S., Interlinking the Social Web with Semantics, IEEE Intelligent Systems, 23(3): 29-40, 2008
5. Extensible Messaging and Presence Protocol (XMPP): Core – RFC Document, 2004. available at http://www.xmpp.org/rfcs/rfc3920.html

[1] MeNow Schema specification can be found at http://crschmidt.net/foaf/menow/menow.rdf

[2] For details about the Data Portability initiative please see http://www.dataportability.org/

RDFohloh, a RDF wrapper of Ohloh

Sergio Fernández

Fundación CTIC
Gijón, Asturias, Spain
sergio.fernandez@fundacionctic.org
http://www.fundacionctic.org/

Abstract. Data on the Semantic Web is modeled and represented in RDF. In the Social Web people usually do not give a further thought about this kind of formalism, whereas they do take care about the content. That is why it could be useful to have tools capable to export that amount of content in machine-readable formats, such as RDF. In this demo paper we present RDFohloh, a RDF wrapper of Ohloh, a Web 2.0 open source directory.

1 Introduction

The original vision of the Semantic Web [3], as a layer on top of the current Web, requires that data is published on the Web, and ideally linked with other useful resources. During the last years the Semantic Web community has made a big effort to make available more and more RDF datasets. Data can come from legacy sources, relational databases, or just making Web scraping; but also from social sources. Web 2.0 applications commonly provide some of its content via public APIs; so it is another big opportunity where data can be extracted and exposed as Linked Open Data [2].

2 Ohloh

Ohloh[1] is an open source directory. Its main goal is to aggregate projects and developers from any Web site. By retrieving data from revision control repositories (such as CVS, SVN, or Git), Ohloh provides statistics about the longevity of projects, their licenses (including license conflict information) and software metrics such as lines of source code and commit statistics. At this moment[2] Ohloh lists 15,532 projects and 23,430 developers. Another goal of Ohloh is providing a public RESTful API[3] with the most important information and many of that metrics.

[1] http://www.ohloh.net/
[2] Data retrieved on September 18th, 2008
[3] http://www.ohloh.net/api

3 RDFohloh, wrapping the wrapper

RDFohloh[4] comes to fulfill the requirement previously mentioned with Ohloh: consuming the data provided by its API and publishing it in RDF [11] as Linked Data. So since Ohloh could be considered a Web 2.0 wrapper for open source projects and developers, RDFohloh could be deemed a wrapper of the wrapper; the n-layers architecture in pure state.

3.1 Related work

Obviously the idea behind RDFohloh is nothing really new. There are several applications that export FOAF or SIOC[5] from other Web 2.0 applications. But for DOAP it could be summarized mainly in two:

- doap:store [12] is an online DOAP directory of computing projects, collaboratively built, where people do not need to register to the service, because it constantly retrieves decentralized projects description to build its database thanks to Ping The Semantic Web service.
- DOAPspace[6] is a registry/repository that contains DOAP scrapped data from several sources including SourceForge, Freshmeat and the Python Package Index.

3.2 Data and links

RDFohloh mainly uses three popular ontologies: SIOC [6] and FOAF [7] for users and DOAP [9] for projects. At the moment of this writing, RDFohloh is publishing 23,430 instances of `sioc:User`/`foaf:Person` and 15,532 of `doap:Project`. The result dataset has `skos:subject` links with DBpedia [1] concepts (for the moment only with the programming language of projects) and `owl:sameAs` links with DOAPspace projects.

3.3 Publication

One of the details specially attended in RDFohloh was how to publish the data. Using cool URIs [13] and content negotiation [10], it provides three views (RDF/XML, N3 and XHTML+RDFa) of each resource. All the data is published attending the best practice recipes [5], and the final result was successfully tested with Vapour [4].

[4] `http://rdfohloh.wikier.org/`
[5] `http://sioc-project.org/applications`
[6] `http://doapspace.org/`

4 Conclusions and Future Work

RDFohloh comes to expand the actual horizon of the Linked Data planet, providing social data from a rich source of information. However it is necessary to improve the project including some new features:

- Providing dumps of the all the data, properly described with Semantic Sitemaps [8], but first it is necessary to find how to cope with the limitations of the number of requests per day of Ohloh's API. With that dumps, it would be easier to provide also a SPARQL endpoint to query the dataset.
- Including source code metrics from Ohloh that now are missing in the RDF export, allowing possible semantic analysis of it.
- Improving actual links and add new ones to other open datasets.

All that features are in the roadmap of the project from its beginning, so hopefully it will be soon available.

References

1. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *Aberer et al. (Eds.): The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735, Busa, Korea, November 2007. Springer 2007.
2. T. Berners-Lee. Linked Data Design Issues. Available at `http://www.w3.org/DesignIssues/LinkedData.html`, 2006.
3. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 2001.
4. D. Berrueta, S. Fernández, and I. Frade. Cooking HTTP content negotiation with Vapour. In *Proceedings of 4th workshop on Scripting for the Semantic Web 2008 (SFSW2008). co-located with ESWC2008*, Tenerife, Spain, June 2008.
5. D. Berrueta and J. Phipps. Best Practice Recipes for Publishing RDF Vocabularies. Working Draft, W3C, 2008.
6. U. Bojars and J. G. Breslin. SIOC Core Ontology Specification. Member submission, W3C, 2007.
7. D. Brickley and L. Miller. FOAF Vocabulary Specification. Technical report, 2005.
8. R. Cyganiak, R. Delbru, and G. Tummarello. Semantic Web Crawling: A Sitemap Extension. Technical Report, DERI, 2007.
9. E. Dumbill. DOAP: Description of a Project. `http://usefulinc.com/doap/`.
10. K. Holtman and A. Mutz. Transparent Content negotiation in HTTP. RFC, IETF, 1998.
11. G. Klyne and J. J. Carroll. Resource Description Framework (RDF): Concepts and abstract syntax. Technical report, W3C Recommendation, 2004.
12. A. Passant. A user-friendly interface to browse and find DOAP project with doap:store. In *Proceedings of the 3rd workshop on Scripting for the Semantic Web (SFSW2007), co-located with ESWC2007*, Innsbruck, Austria, May 2007.
13. L. Sauermann and R. Cyganiak. Cool URIs for the Semantic Web. Interest Group Note, W3C, March 2007.

Semantify del.icio.us: automatically turn your tags into senses

Maurizio Tesconi, Francesco Ronzano,
Andrea Marchetti, and Salvatore Minutoli

Institute for Informatics and Telematics
National Research Council - C.N.R.
Via G. Moruzzi, 1 - 56100 Pisa - Italy
{maurizio.tesconi, francesco.ronzano, andrea.marchetti,
salvatore.minutoli}@iit.cnr.it

Abstract. At present tagging is experimenting a great diffusion as the most adopted way to collaboratively classify resources over the Web. In this paper, after a detailed analysis of the attempts made to improve the organization and structure of tagging systems as well as the usefulness of this kind of social data, we propose and evaluate the Tag Disambiguation Algorithm, mining del.icio.us data. It allows to easily semantify the tags of the users of a tagging service: it automatically finds out for each tag the related concept of Wikipedia in order to describe Web resources through senses. On the basis of a set of evaluation tests, we analyze all the advantages of our sense-based way of tagging, proposing new methods to keep the set of users tags more consistent or to classify the tagged resources on the basis of Wikipedia categories, YAGO classes or Wordnet synsets. We discuss also how our semanitified social tagging data are strongly linked to DBPedia and the datasets of the Linked Data community.

1 Introduction

Tagging is currently one of the most widespread patterns to create, collect and share huge amounts of social data over the Web, represented by the set of tags adopted by the community of users of a tagging service to describe resources of interest. The number of Web tagging services and, in particular, the amount of social bookmarking sites, that are tagging services devoted to tag URLs, is rapidly growing [4]: among them Del.icio.us[1], mainly, but not only, used by Web experts, Bibsonomy[2], exploited by researchers to share links to papers and other relevant works and Technorati[3], widely adopted by communities of bloggers, represent relevant examples.

The possibility of freely choosing tags is probably one of the main reasons for the popularity of social tagging but it also makes difficult to produce a clean

[1] http://del.icio.us/
[2] http://www.bibsonomy.org/
[3] http://www.technorati.com/

and consistent organization and classification of the tagged resources. During the last few years, starting from a lot of different statistical analyses of tagging data collections, many distinct approaches to better create, structure and search for information querying tagging services have been proposed; the most relevant are briefly reported in Section 2. These studies mainly point out problems like synonymy, polysemy and in general all the different lexical forms or tags that can be used by each user to describe a concept; these issues represent the main causes of loss of consistency in tagging data collections as well as of decrease of precision and recall of tag based searches. The general and globally agreed way to face these issues consists of using some sort of semantic classification process to give an explicit meaning to each freely chosen tag; some of these techniques rely only upon the data retrievable from a tagging service, others also exploit external semantic resources.

In this context, we propose the *sense-based tagging*, a new approach to automatically structure the set of tags collected by each user of a tagging system. We automatically disambiguate the meaning of tags mining the information contained in Wikipedia: for this purpose we present and evaluate the Tag Disambiguation Algorithm, described in details in Section 3. In Section 4, we expose and evaluate all the advantages of the *sense-based tagging* in terms of classification of resources and creation of semantic metadata.

In conclusion, in Section 5, we expose our future works as well as the ongoing effort to implement all the approaches described creating our Web sense-based tagging service.

2 Related work: how to improve tags organization

Many studies have been carried out to describe and make new proposals to better organize and expose the information collected by tagging services. We can find a lot of statistical investigations about the structure of tags collections and the dynamics of their usage together with all the subsequent considerations about the most frequent tagging patterns: among them two interesting works based on del.icio.us bookmarks are [14] and [11]. There are also relevant analysis concerning other quantitative studies of social bookmarking systems aiming at finding how tagging data can support Web search and improve its results, providing additional information like in [6] and [16].

Summarizing the results of many relevant works regarding the possibility to improve the effectiveness of tagging, we have pointed out that all of them have identified two main causes for the poor structure and organization of tag-based classifications: the complete freedom users have when they choose tags and the lack of any semantic information to support tagging activity. In order to face these issues, different strategies have been set up. Tag recommendations systems have been proposed to keep users tagging data more consistent [19]; some of them exploit also external semantic resources like Wordnet[4] to perform

[4] http://wordnet.princeton.edu/

this task [15]. Moreover we can also find services that analyze the tags of a specific user to detect tag usage inconsistencies like slightly different keywords: relevant examples are Bookmark Cleaner[5] and Del.icio.us tag cleaner[6], all related to del.icio.us bookmarks.

The systems and the procedures that use some sort of semantic information to better organize tags and understand their meaning can be divided into *two groups*. The first one comprises *all the methods that introduce some sort of structure to the sets of tags taking into account only the information retrievable from tagging services*, that is the collections of users, tags and tagged resources. They mainly try to group together similar tags on the basis of their relations with users and resources. In this way they identify sets of strictly related tags or understand the sense of ambiguous ones: some of these procedures are described in [5] and [10]. The second group of semantic based approaches *exploits external semantic resources to structure sets of tags*. Some of them try to define the right meaning of each tag retrieving the semantic relations that occur between related tags so as to visualize tags on the basis of their sense and relevance. In order to achieve that, data extracted from different ontologies available over the Web are collected and merged; examples of this kind of methods are [13], [17] and [3]. They often suffer the poor terms coverage of Web ontologies, providing encouraging results only in particular domains. To point out the right meaning of tags other techniques exploit Wikipedia as well as its "semantic version", the DBPedia ontology, connected with a growing number of external datasets. Among them we can point out our last project, SemKey [1] that help users to disambiguate tags relying upon Wikipedia and Wordnet. Also the MOAT Project [2] provides an infrastructure to collect the concepts associated to tags, identified by means of DBPedia URIs. Similarly, the Faviki [7] Web system allows users to describe Web resources through Wikipedia meanings. All these projects don't support users enough in the management of the semantic data needed to describe their tags; often the choice of a particular tag or better a specific meaning, is not easy and requires a lot of additional user interactions. Moreover it is limited by the set of available word-meaning associations. In this context, also different ontologies have been defined to provide a reference model to describe the tagging activity in terms of the relations between tags, resources, users and tagging actions: a relevant example is the SCOT Ontology [7], that reuses and extends the previous Tag Ontology [8].

We propose and evaluate a methodology to overcome these limitations using an algorithm that automatically points out the right meaning of each tag, considering those available in Wikipedia, thus semantifying the tag collection of each user of a tagging service. We allow users to provide new tags to refer to a specific meaning thanks to the exploitation of Tagpedia, a semantic resource built ad hoc from Wikipedia to disambiguate tags. We rearrange user resources

[5] http://www.bookmarkcleaner.com/
[6] http://delicious.isnotworking.com/
[7] http://www.faviki.com/

on the basis of external classifications like Wikipedia categories or Yago and Wordnet classes.

3 From tags to Wikipedia senses: Tagpedia and the Tag Disambiguation Algorithm

Our idea of *sense-based tagging* is grounded on the possibility to define the meaning of each tag chosen by a user to describe one or more resources over the Web. In order to find out the right meaning of a tag we need a sort of **global semantic reference containing a rich collection of senses to search into**. Currently Wikipedia[8] represents the richest and constantly updated encyclopedic reference over the Web with a huge set of semantic contents, even if not explicitly exposed and easily accessible. Because of its features, Wikipedia is an ideal starting point to retrieve the information needed to define the meaning of a tag. Based on this assumption, by mining the contents of Wikipedia, we have built *Tagpedia*: this is a semantic reference for organization and classification of tags, intended as words or more in general as short textual expressions used to refer to a specific topic. Tagpedia is based on the model of term-concept networks [12]; *for each meaning of Wikipedia, Tagpedia groups together all the different words used to refer to it.* Currently Tapgedia includes more than *1,92* millions of distinct concepts and more than *4,23* millions of words used to refer to these meanings: these data have been extracted from Wikipedia pages. Tagpedia, built ad hoc to support the semantic characterization of Web contents through sense-based tagging and thus to easily disambiguate the meaning of a tag, is accessible over the web at the URL *http://www.tagpedia.org/*. It can be also queried by means of a dedicated Web API and it can be collaboratively edited. To get more information about the structure and the usage of Tagpedia see [9].

In this section we propose and evaluate a *Tag Disambiguation Algorithm* (TDA), implemented relying upon Tagpedia. Our implementation collects the tags of a user from a tagging service (del.icio.us, in our case) and for each of them finds out the relative sense by linking it to the corresponding page of Wikipedia.

In particular the TDA identifies for each tag t a list of candidate senses, referred to also as concepts or meanings in the rest of this paper and assigns them a number, called **sense-rank** SR: **the higher the rank of a meaning, the better that meaning defines the sense intended by the user for that tag**. In the remaining part of this Section we explain in more details the TDA and provide some evaluation of its disambiguation effectiveness.

In order to calculate the sense-rank of a meaning of the tag t of the user U we consider the text of Wikipedia describing that particular meaning and we base our algorithm on the following assumptions:

- the more the meaning described by that text is similar to the one intended by the user U, the higher is, in that text, the number of occurrences of tags that are in some way connected to the tag t.

[8] http://en.wikipedia.org/wiki/

– the meaning intended by the user U for the specific tag t doesn't change while tagging [5]. This is particularly true if we think that normally the interests of a user are generally focused on defined domains, so the meaning of a tag is usually unique.

We want to apply the TDA in order to disambiguate the tag t of the user U, ranking all its meanings to choose the one with the highest sense-rank. In particular, **starting from the texts of Wikipedia, we base our analysis on the calculation of the number of occurrences of the following two groups of relevant tags**: those used by the user U along with the tag t to describe Web resources of interest and the del.icio.us popular tags used to characterize the same set of Web resources, along with the popularity value of each of the selected popular tags.

Generalizing we can state that the user U has tagged m Web resources, R_1, R_2,..., R_m with the tag t . There are n meanings M_1, M_2,..., M_n of the tag t in Wikipedia; we want to calculate the sense-rank $SR_t(M_x)$ of each of them.

In order to do that we exploit **two different groups of parameters**: some of them are retrieved from del.icio.us and other ones are retrieved from Tagpedia and Wikipedia. In particular we consider:

- from **del.icio.us**:

• all the y tags CT_1, CT_2, ..., CT_y of the user U that co-occur with the tag t, considering the Web resources R_1, R_2,..., R_m, along with their respective frequency of co-occurrence $CF(CT_1)$, $CF(CT_2)$,..., $CF(CT_y)$;

• all the z popular tags PT_1, PT_2, ..., PT_z used in del.icio.us to describe the Web resources R_1, R_2,..., R_m along with their total popularity frequency $PF(PT_1)$, $PF(PT_2)$, ..., $PF(PT_z)$; the popularity frequency of a tag for a particular resource is the number of times that word has been used to tag the same resource.

- from **Tagpedia**:

• the n meanings M_1, M_2, ..., M_n of the tag t in Wikipedia; for instance the tag *owl* can be used to refer to the Web Ontology Language but also to a nocturnal bird or to an Australian rugby union club and so on;

• the n texts $T(M_1)$, $T(M_2)$, ..., $T(M_n)$ of the articles of Wikipedia describing each of the meanings of the tag t and for each of those texts, the number of occurrences of a particular tag W, referred to as $OCC(W, T(M_n))$.

All these values are used to calculate $SR_t(M_x)$, that is the sense-rank of the meaning M_x of the tag t for the user U. In particular, the value of the sense-rank is the weighted sum of two contributions: the first one, $SRU_t(M_x)$, is related to the tags of the user U that co-occur with the tag t; the second one, $SRP_t(M_x)$, deals with the popular tags used in del.icio.us to refer to the resources tagged by the user U with t. The sense-rank of the meaning M_x of the tag t for the user U is equal to:

$SR_t(M_x) =$

$WU \times [SRU_t(M_x) \,/\, \max(SRU_t(M_v))] + WP \times [SRP_t(M_x) / \max(SRP_t(M_u))]$

where v and u range from 1 to n and $WU + WP = 1$.

We divide the values of $SRU_t(M_x)$ and $SRP_t(M_x)$ respectively by the max value of SRU_t and SRP_t for all the n meanings of the tag t: in this way we normalize all the sense-ranks of the tag t to the interval $[0, 1]$, making their values comparable in order to choose the highest one. We have that:

$SRU_t(M_x) = [CF(CT_1) \times OCC(CT_1, T(M_x)) + CF(CT_2) \times$
$\times OCC(CT_2, T(M_x)) + .. + CF(CT_y) \times OCC(CT_y, T(M_x))] * UNN/UTOT$

$SRP_t(M_x) = [PF(PT_1) \times OCC(PT_1, T(M_x)) + PF(PT_2) \times$
$\times OCC(PT_2, T(M_x)) + .. + PF(PT_y) \times OCC(PT_y, T(M_x))] * PNN/PTOT$

where:

- UNN is the total number of values of $OCC(CT_o, T(M_x))$ that are not equal to 0

- $UTOT$ is the total number of tags of the user U co-occurring with the tag to disambiguate t in at least a description of a Web resource.

- PNN is the total number of values of $OCC(PT_o, T(M_x))$ that are not equal to 0

- $PTOT$ is the total number of del.icio.us popular tags related to the Web resources tagged by the user U with the tag t.

The values of the weights WU and WP in our TDA evaluation phase are both set to 0.5, but we are experimenting the possibility to adapt them to the quality and the origin of the set of disambiguation data, that is the set of tags that are related in some way to the one to disambiguate. For instance, if a user has chosen a tag to characterize a small number of resources or if the tag to disambiguate has a small number of co-occurring tags chosen by the user, when we calculate the sense-rank of the meanings of that tag, we can decrease the importance of the user tags contribution (decreasing the value of WU). When the sense-rank of each meaning of a polysemous tag has been calculated, the TDA chooses the sense with the highest rank as the correct one.

3.1 TDA evaluation

Now we provide some evaluation of our Tag Disambiguation Algorithm. As the starting point, querying del.icio.us, we have chosen to consider the tagging profile of 9 del.icio.us users (U_1, U_2,..., U_9) as it was on July 2008, ranging from very active taggers to people that only occasionally save some bookmark: globally we have collected 3520 tags used to bookmark 3926 URLs with a total average number of tags per bookmark equal to $3,38$. In the first columns of Table 1 are synthesized the distinctive features of the tagging profile of each user.

We have queried Tagpedia and in case of polysemous tags we have applied the TDA to choose the best meaning to be associated to each of them. We aim at evaluating respectively the coverage of the tags collection of Tagpedia in terms of the number of tags that can be associated to at least one sense and the effectiveness of the Tag Disambiguation Algorithm in terms of the correctness of the sense associated to each of the tags.

From the the last two columns of Table 1 we can notice that thanks to the support of Tagpedia and the execution of the TDA we can point out the

User	**Bookmarked URL**	**Tags**	**Average number of tags per bookmark**	**Disambiguated tags**	**Distinct concepts**
U_1	136	166	2,72	148 (89,16%)	141
U_2	275	355	3,48	321 (90,42%)	302
U_3	279	396	3,89	345 (87,12%)	323
U_4	541	511	4,02	449 (87,87%)	403
U_5	754	1149	6,54	1026 (89,30%)	885
U_6	428	175	1,38	165 (94,29%)	160
U_7	69	142	3,62	133 (94,66%)	129
U_8	453	76	2,08	71 (93,42%)	71
U_9	991	550	2,67	501 (91,09%)	470
Average	—	—	**3,38**	**3159 (89,74%)**	**2884**

Table 1. Tagging profile of del.icio.us test users and tags disambiguation results

meaning of 89, 74% of the tags of the 9 del.icio.us users. Considering the 3159 disambiguated tags, 2884 different senses have been identified.

The 11,71% of the considered tags have not been associated to a concept, because in Tagpedia there are no words for them.

In conclusion, we have evaluated the TDA effectiveness through a process of human review of results: among 2589 polysemous tags the TDA has chosen the right meaning of the 89, 15% of them. Globally, the 2891 (91, 52%) of the 3159 disambiguated tags has been correctly associated to the right meaning.

4 Advantages of sense-based tagging

In this section we explore the advantages of the sense-based tagging considering new ways to group similar tags, to classify resources and to produce semantic tagging metadata.

4.1 Sense-based grouping of tags

A first advantage of the adoption of sense-based tagging is represented by the possibility to group together user tags that refer to the same concept. In fact, considering the set of 9 del.icio.us users analyzed in the previous tables, we can notice that only 2884 distinct concepts, out of 3159 disambiguated tags, have been pointed out. This means than *on average the 9% of the tags chosen by a user to describe a resource has a sense referred to by other tags already present in the tagging profile of the user*. In the following table is shown some example of groups of tags having the same meaning, defined as the outcome of our disambiguation procedure based on the TDA; all the tags grouped together are words that refer to the same sense in Tagpedia and thus the same associated page of Wikipedia.

Wikipedia concept name (page title)	Brief Wikipedia concept description	Group of tags referring that concept
Carpool	Carpooling is the shared use of a car by...	rideshare, ridesharing, carpooling, carpool
Humour	Humour or humor is the tendency to provoke laughter and provide amusement..	funny, humor, humour
Film	Film is a term that encompasses individual motion pictures, the field of film as an art form...	film, films, movies, movie

Table 2. Grouping of tags having the same meaning

4.2 Exploiting Wikipedia categories, YAGO classes and Wordnet Synsets to classify resources

Once characterized all the bookmarked URLs of a user through the concepts of Wikipedia associated exploiting our Tag Disambiguation Algorithm, we have tried to create different alternative views of user resources exploiting three classification systems: the Wikipedia Categories, the YAGO classes and the Wordnet synsets. The mapping of Wikipedia concepts and thus of the corresponding Wikipedia pages to Wordnet synsets and YAGO classes has been derived from those available in the DBpedia datasets[9].

The Wikipedia categories is a collaboratively built categorization of Wikipedia articles: Wikipedia users can place one article in one or more categories or also create new categories and connect them to the other categories through subsumption relations. Almost all the articles of Wikipedia have been placed in at least one of the more than 312 thousands categories of Wikipedia. Because of its collaborative definition, the Wikipedia categorization system is untidy and includes many subsumption cycles or other kinds of inconsistencies.

YAGO (Yet Another General Ontology)[10] is a large semantic knowledge base, that is automatically extracted from Wikipedia and uses Wordnet to organize information. It has been developed by the Max Plank Institute for Computer science. We have considered the mapping of Wikipedia pages to the corresponding YAGO classes: 1,412 millions of pages of Wikipedia has been mapped to at least one class in YAGO. Thus YAGO covers about 74% of the pages of Wikipedia. YAGO classes are arranged through the YAGO Class Hierarchy. The number of Wikipedia pages that have been mapped to one or more synsets of Wordnet is considerably lower.

Our first analysis aims at defining, considering the previously defined group of 9 del.icio.us users, how many senses of their profiles are covered by each classification system and, as a consequence, we can have an idea of the adequacy of the particular classification system to provide a new structure to organize and present the tagged resources to the users. The results are presented in Table 3.

[9] http://wiki.dbpedia.org/Downloads31/

[10] http://www.mpi-inf.mpg.de/∼suchanek/downloads/yago/

Distinct concepts	Concepts coverage of:		
	Wikipedia categories	**YAGO classes**	**Wordnet synsets**
2884	**2749 (95%)**	**1667 (58%)**	**507 (18%)**

Table 3. Total concepts coverage of Wikipedia categories, YAGO classes and Wordnet synsets

We can notice that the synsets of Wordnet cover an irrelevant portion of the senses of the 9 test users, while the YAGO classes includes the 58% of the concepts. The Wikipedia category system manages to classify almost all the senses of each user.

Each sense exploited by a single user describes more than one Web resource. Table 4 is similar to Table 3; it is intended to test the coverage of the three considered classification systems in terms no more of senses, but of bookmarked Web resources. We define how many tagged Web resources can be associated to at least a Wikipedia Category, a YAGO class or a Wordnet synset.

URLs	URLs described by disambiguated tags	URLs coverage of:		
		Wikipedia categories	**YAGO classes**	**Wordnet synsets**
3926	**3864**	**3852 (100%)**	**3284 (85%)**	**1749 (45%)**

Table 4. URL coverage of Wikipedia categories, YAGO classes and Wordnet synsets

Our results in this case are encouraging. Because each Web resource is on average tagged through two or more senses, we can see that, considering Wikipedia Categories, practically all of them are placed at least in a category; thus, exploiting Wikipedia Categories and the sense-based tagging, we can provide a new classification of all the bookmarks of our users. Also considering YAGO classes, we can map on average 85% of users resources on one or more of them. The coverage of the resources considering Wordnet classes is still too low to provide a valuable new classification of user resources through Wordnet hierarchy of classes.

In order to evaluate category-based classifications of user resources we have made some initial test related to YAGO classes. Considering the YAGO classes hierarchy, we have defined how many classes are necessary to classify the resources of each user, considering that 85% of these resources are present at least in a class. Moreover, we have calculated how the number of classes decreases and thus how the class-based classification of user resources gets more coarse-grained when we move to higher levels of the hierarchy. We have considered seven

different levels of ancestors of the direct YAGO classes that are those containing at least one tagged resource and we have defined, considering the classes of each level, how many classes are involved in the classification of user resources. The result of our analysis, considering the average number of classes needed to classify user resources for each level of ancestors, are shown in Table 5.

User concepts belonging to YAGO classes	**Direct YAGO classes**	**Levels of ancestors**						
		I	II	III	IV	V	VI	VII
185	**162**	**132**	**102**	**69**	**42**	**29**	**19**	**12**

Table 5. Average number of classes needed to represent user resources considering seven levels of ancestors

We can notice that, going back up to the seventh level of ancestors of the YAGO classes containing at least one tagged resource, the average number of different classes needed to classify the resources in at least one of them decreases from 162 to 12 different ones, thus allowing us to organize the users resources in a number of sets that can vary from 162 to 12.

In conclusion, in this Section we have shown, thanks to some initial analysis, that once classified users resources through senses and thus through the concepts of Wikipedia, the Wikipedia Categories and the YAGO classes can provide new classifications of these resources based on global shared classificatory schemas.

4.3 Linking del.icio.us to Linked Data datasets

The *sense-based tagging* is also a way to connect the social data collaboratively created through tagging and the Semantic Web. The Linked Data community actually represents one of the most relevant attempts to collect, interlink and semantically expose over the Web the information contained in many different datasets, through the adoption of the RDF and RDF triples, putting in practice the vision of the Semantic Web. It tries to define a common set of rules and best practices to publish and browse semantic-aware information. DBPedia[11][18] is a sort of alignment ontology created by mining Wikipedia, representing in some way the glue of the Linked Data community: by exploiting DBPedia each concept of Wikipedia can be univocally referenced through a specific URI that represents also the way to retrieve over the Web the RDF triples describing that concept. The Tag Disambiguation Algorithm manages to automatically convert each tag into the intended concept of Wikipedia and thus into the related URI of DBPedia. In this way we are able to generate for each user of a *sense-based tagging service*, a set of RDF triples describing his tagging profile; they include, for instance, a triple for each sense associated to a tagged resource. Identifying each user through the URL of his FOAF profile[12] and pointing out each referred sense

[11] http://dbpedia.org/About
[12] http://www.foaf-project.org/

through a DBPedia URI, each sense-based tagging service, can make accessible over the Web throug user-dedicated URLs the RDF triple-based descriptions of its users, following the publishing rules of the Linked Data Community. In this way we can automatically link the social data of a tagging service to the datasets of the Linked Data Community, thus providing a huge amount of continuously growing collaboratively created descriptions of Web resources.

5 Conclusions and future works

In this paper, we have described and evaluated a new way to automatically semantify the tags of the users of a Web tagging service, thanks to the Tag Disambiguation Algorithm. Considering del.icio.us and the tags of 9 users, we have managed to correctly find out the meanings of the $91,52\%$ of these tags, linking them to the right concepts of Wikipedia. Consequently we have described an evaluated the possibility to clean users tags grouping them by sense, but also to classify the tagged resources on the basis of Wikipedia categories, YAGO classes and Wordnet synsets. Wikipedia categories and YAGO classes, because of their wide coverage of the concepts of Wikipedia, can support the definition of new way to classify users resources. Characterizing Web resources through Wikipedia concepts we can also connect the social data produced by tagging systems to the datasets of the Linked Data community.

In our future works we would like to improve and better tune our Tag Disambiguation Algorithm. We want also better investigate and test the adequacy of Wikipedia categories and Yago class hierarchy to provide new views of the user resources of a tagging system. Moreover, we are developing a Web system that enable users to automatically semantify their tagging profile, retrieving it from other tagging services, thus making them easily shift to the adoption of the sense-based tagging.

In conclusion, we believe that the sense-base tagging thanks to the automated semantification of tags supported by the Tag Disambiguation Algorithm and because of its strict connection with the Linked Data community, can represent a valuable way to improve the quality and the organization of social tagging allowing to produce a considerable set of interlinked semantic data over the Web.

This work is funded by the European Community, in the 7th Framework Project KYOTO[13].

References

1. F. Ronzano, M. Rosella, S. Minutoli, A. Marchetti, M. Testoni. Semkey: A semantic collaborative tagging system. In Proc. of The workshop Tagging and Metadata for Social Information Organization of the World Wide Web Conference 07, May 8-12, 2007, Banff, Canada.

[13] http://www.kyoto-project.org/

2. A. Passant, P. Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In Proc. of The Linked Data on the Web Workshop of the World Wide Web Conference 08, April 19-25, 2008, Beijing, Cina.
3. T. Roth-Berghofer, B. Adrian, L. Sauermann. Contag: A semantic tag recommendation system, 2007.
4. L. Baker. 125 social bookmarking sites : Importance of user generated tags, votes and links. Blog article, December 2007.
5. N. Gibbins, C. M. Au Yeung, N. Shadbolt. Understanding the semantics of ambiguous tags in folksonomies. At The International Workshop on Emergent Semantics and Ontology Evolution at ISWC/ASWC 2007, 12 November 2007, Busan, South Korea.
6. Shuyi Zheng-Hongyuan Zha, C. Lee Giles Ding Zhou, Jiang Bian. Exploring social annotations for information retrieval. In Proc. of The World Wide Web Conference 08, April 19-25, 2008, Beijing, Cina.
7. Hak Lae Kim, J. G. Breslin, S. Scerri, S. Decker, Hong Gee Kim, Sung Kwon Yang SCOT Ontology Specification. DERI Galway at the National University of Ireland, Galway, Ireland, August 2008.
8. R. Newman Tag ontology design. Blog Article, 2005.
9. M. Tesconi F. Ronzano, A. Marchetti, S. Minutoli. Tagpedia: a semantic reference to describe and search for web resources. In Proc. of The workshop Social Web and Knowledge Management of the World Wide Web Conference 08, April 19-25, Beijing, Cina.
10. P. Keller G. Begelman, F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In Proc. of The World Wide Web conference 06, April 23-26, 2006, Edinburgh, Scotland.
11. Scott A. Golder, Bernardo A. Huberman. The structure of collaborative tagging systems. Technical report, Information Dynamics Lab, HP Labs, 2005.
12. A. Gregorowicz, M. A. Kramer. Mining a large-scale term-concept network from wikipedia. Mitre Technical Report, October 2006.
13. M. Espinoza J. Gracia, R. Trillo, E. Mena. Querying the web: a multiontology disambiguation method. In Proc. of The 6th international conference on Web engineering, July 10-14, 2006, Menlo Park, California.
14. M. E. I. Kipp, D. Grant Campbell. Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. In Proc. of The Annual General Meeting of the American Society for Information Science and Technology, 2006, Austin, Texas (USA).
15. F. Lanubile, G. Semeraro, P. Basile, D. Gendarmi. Recommending smart tags in a social bookmarking system. In Proc. of The European Semantic Web Conference 07, June 3-7, 2007, Innsbruck, Austria.
16. G. Koutrika, P. Heymann, H. Garcia-Molina. Can social bookmarking improve web search? In Proc. of The WSDM08, February 11-12, 2008, Palo Alto, California.
17. L. Specia, S. Angeletou, M. Sabou, E. Motta. Bridging the gap between folksonomies and the semantic web: An experience report. In Proc. of The European Semantic Web Conference ESWC 2007, June 7, 2007, Innsbruck, Austria.
18. G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, S. Auer, C.Bizer. Dbpedia: a nucleus for a web of open data. In Proc. of The International Semantic Web Conference 07, November 11-15, 2007, Busan, Korea.
19. Jianchang Mao Zhichen Xu, Yun Fu, Difu Su. Towards the semantic web: Collaborative tag suggestions. In Proc. of The World Wide Web conference 06, April 23-26, 2006, Edinburgh, Scotland.

Towards Opinion Mining Through Tracing Discussions on the Web

Selver Softic and Michael Hausenblas

Institute of Information Systems & Information Management,
JOANNEUM RESEARCH, Steyrergasse 17, 8010 Graz, Austria
firstname.lastname@joanneum.at

Abstract. This paper reports on our ongoing work regarding opinion mining from Web-based discussion forums in the realm of the Understanding Advertising (UAd) project. Our approach to opinion mining is to first RDFise discussion forums in SIOC, and in a second phase to interlink the so created data with linked datasets such as DBpedia. We are confident that this should allow a market researcher to formulate queries using domain semantics and hence understand what people think about a certain product or service. The system's architecture, preliminary results, and the current available demonstrator are discussed in this work.

1 Introduction

Products or services are often discussed by customers on the Web. Whilst (official) company sites usually tell a certain side of the story, having users discussing advantages or issues with certain products offers a source for a deeper understanding of a market. Equally found in common social life, the communities on the Web can have a strong impact on trend setting. This observation is eligible, if we behold a trend as a spot where subjective views of different people lead by their affections may cross and merge with each others.

We found that valuable data relevant for market research on the Web is neither easy accessible nor processable. Time expenses to collect and evaluate data needed for a better market understanding are still tremendous. As recently pointed out by Peter Mika (Yahoo! Research) [1]:

> Current search technology is unable to satisfy any complex queries requiring information integration such as analysis, prediction, scheduling, etc. An example of such integration-based tasks is opinion mining regarding products or services. While there have been some successes in opinion mining with pure sentiment analysis, it is often the case that users like to know what specific aspects of a product or service are being described in positive or negative terms and to have the search results appear aggregated and organized.

In the Understanding Advertising (UAd) project[1] we aim at developing a methodology allowing a market researcher to understand a certain market.

[1] http://www.sembase.at/index.php/UAd

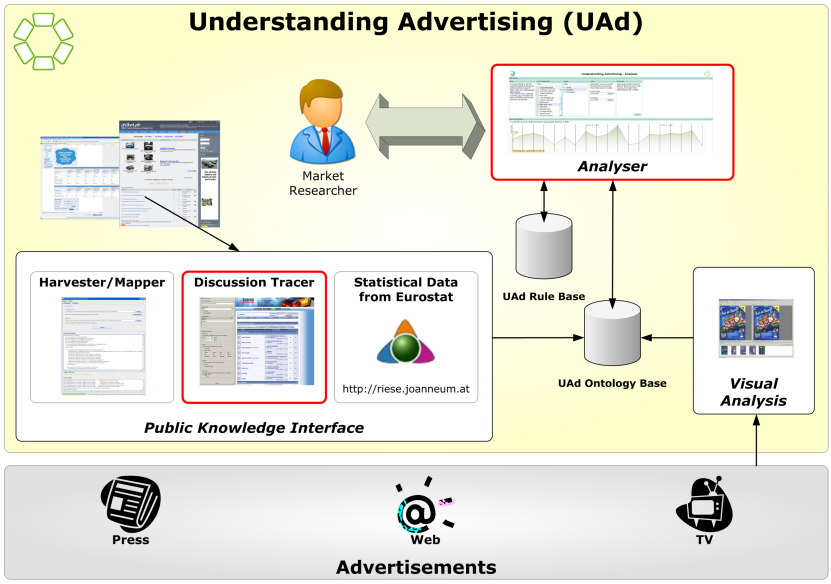

Fig. 1. The UAd system architecture.

The analysis performed in UAd is twofold, (i) by visual interpretation of advertisements (from print media, Web and TV), and (ii) by using information available on the Web. Fig. 1 depicts the overall UAd system architecture, consisting of (i) the UAd Analyser (the front-end for the end-user), (ii) the "Public Knowledge Interface", and (iii) the Visual Analysis module. Information about products and services are gathered from the Web through the so called UAd "Public Knowledge Interface" (PKI). We have developed three methods converting plain (HTML) Web content into structured data represented in RDF allowing us to be both flexible and comprehensive:

1. Plain old screen scraping (in the so called UAd Harvester/Mapper module);
2. Pattern-based RDFising and Interlinking for online discussions (the UAd Discussion Tracer);
3. Schema-based a-priori RDFising and Interlinking (for statistical data from Eurostat; described elsewhere [2, 3]);

In this paper we focus on tracing discussions on the Web, hence the two components involved in this task (Discussion Tracer and Analyser) are highlighted in Fig. 1.

This paper is structured as follows: First, we review related work in section 2. Then, in section 3 we discuss our approach representing discussions and opinions. In section 4 we present the system's architecture, discuss the data acquisition and the market researcher's interface. We present preliminary results in section 5. Finally, we discuss our findings and highlight future work in section 6.

2 Related Work

Recent research on opinion mining has focused on sentiment analysis, simple "pro" and "cons" classification [4] and determination of semantic orientation in opinion models using feature-based opinion summarisation on word, sentence or document level. Typically, Natural Language Processing (NLP) [5–7] and machine learning techniques [4, 8] have been utilised in supervised or unsupervised modes [9, 10] allowing the extraction and classification of sentiment and opinions polarisation. The workflow usually comprises three major phases: extraction, structuring and summarisation of results. In general we subscribe to this pattern, however differ in a number of details mostly regarding the explicit representation of the information.

Motivated by earlier experiences [11, 12] our approach is based on Semantic Web technologies (RDF, SPARQL, etc.). Further, in contrast to existing work, we use widely deployed vocabularies—e.g. Semantically-Interlinked Online Communities (SIOC)—along with existing APIs [13] for the extraction and structuring phase. Regarding the formal representation of products and their characteristics it is worth noting that the W3C has recently launched the "Product Modelling Incubator Group"[2] aiming at creating a product modelling ontology.

We aim at orienting the opinion holder context on domain semantics [8] along with exploiting linked datasets (such as DBpedia [14]) and domain delimited query expansion [15]. Furthermore, the creation of opinion ranking primary for sentiment classification [16, 17] will be considered in greater detail in our future work.

To the best of our knowledge there exists no other work in the area of opinion mining that deals with explicitly modelled opinions along with linked data sets for its domain knowledge.

The basic idea of linked data was outlined by Sir Tim Berners-Lee [18] in 2006. The Linking Open Data (LOD) community project[3] is an open, collaborative effort applying the linked data principles. It aims at bootstrapping the Web of Data by publishing datasets in RDF on the Web and creating large numbers of links between these datasets [19]. The datasets included in the project are diverse in both nature and size. Currently, the project includes some 30 different datasets, ranging from rather centralized ones (such as DBpedia [14]) to those that are very distributed (for example the FOAF-o-sphere). While some of the datasets focus on certain domains (for example the Eurostat data [3]), others are more of a generic type, such as Revyu.com [20].

3 Representing Discussions and Opinions

To support a market researcher in analysing a certain market, one of the sources used in the UAd PKI are Web-based discussion forums. For enabling structured

[2] `http://www.w3.org/2005/Incubator/w3pm/`
[3] `http://linkeddata.org/`

queries and browsing it is necessary to represent the discussions in a machine-interpretable way and enhance it with domain semantics. Web-based discussion forums offer a well-structured source for this purpose, hence the idea to exploit them along with linked datasets.

Our goal is it to explicitly model the opinions in a discussion being compliant to the Web of Data. We decided to reuse an existing vocabulary to represent the discussions rather than reinventing the wheel. Due to its popularity and wide-spread use, the Semantically-Interlinked Online Communities (SIOC) vocabulary[4] has been selected to represent discussion threads and posts.

However, in case of explicitly representing opinions we did not manage to find an appropriate vocabulary. Although one could for example use a review vocabulary[5] as a base and extend it, we found it better suited to define a dedicated vocabulary for this task.

Our "Opinion Mining Core Ontology"[6] (cf. Fig. 2) basically defines the following classes and properties:

- `opm:DiscussionOpinion`, the central hub that connects discussion threads with opinions about a certain entity;
- `opm:Opinion`, an abstract representation of an opinion;
- `opm:Topic`, a proxy concept to trigger aspects of a certain topic.

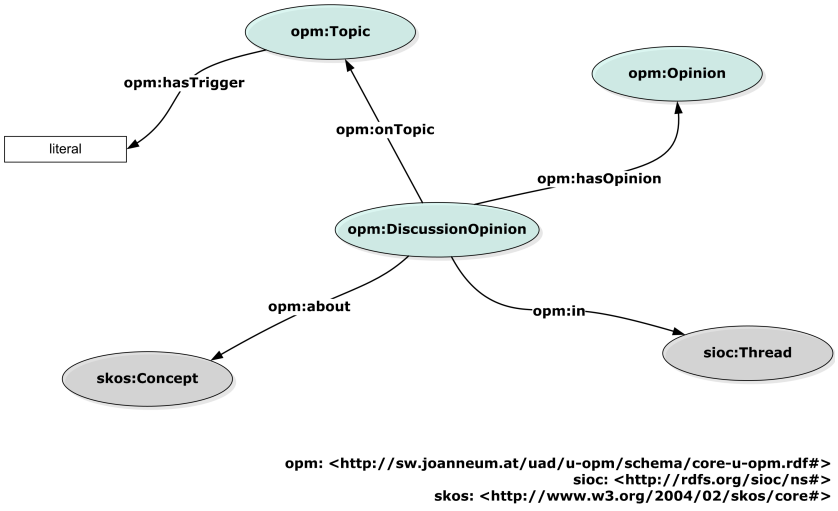

Fig. 2. UAd's Opinion Mining Core Ontology.

[4] `http://www.sioc-project.org/ontology`
[5] Such as `http://danja.talis.com/xmlns/rev_2007-11-09/index.html`
[6] `http://sw.joanneum.at/uad/u-opm/schema/core-u-opm.rdf`

We use `skos:Concept` of SKOS [21] to represent what a discussion *is about*, for example, a certain car such as the Alfa Romeo 156; we note that this design decision also supports the straight-forward utilisation of data from DBpedia. Further, we use the `sioc:Thread` from the SIOC vocabulary to indicate *where* the discussion has been taken place.

It has to be noted that `opm:Opinion` is currently deliberately underspecified. We intend to extend and refine this part to the ontology based on our experiences with the system and regarding earlier work from [9, 10]. Further, we want to point out that the `opm:Topic` concept is used to represent a certain aspect of a discussion, that is, it might indicate that users discuss about the pricing, about problems with a certain product or simply express their satisfaction. The semantics of this concept are such that if one of the assigned trigger words has been found in a discussion, the topic is believed to match (hence the labelling of the datatype property `opm:hasTrigger`).

The introduced lightweight ontology above plays a decisive role in our opinion mining process. In order to achieve better scalability and reusability, it acts as a nexus between the domain of concern and the RDFised data. This is why it makes no difference for our opinion mining model if there is the DBpedia categorisation behind or some other domain specific ontology. Therefore, our approach offers flexibility by choice of domain and yields a generic opinion creation.

4 Discussion Tracing

In the process of discussion tracing in UAd, two major components are involved (Fig. 3), namely (i) the UAd Data Acquisition (highlighted), where Web-based discussions are harvested, RDFised and interlinked, and (ii) the UAd Analyser, allowing to query and access the data.

4.1 Data Acquisition

The data acquisition in UAd is performed in three phases; in a first step the most common data in a Web-based discussion forum, such as title, author, creation date, etc. is RDFised using SIOC. In a second phase the entities occurring in the discussion posts are identified and interpreted regarding a certain domain (in our demonstrator this domain is "cars"). This second step involves the interlinking to linked datasets such as DBpedia or instances of some other domain specific ontology. For our purposes DBpedia offers enough adequate instances and well formed domain model respectively area of interest. However, as mentioned in Section 3 this is not mandatory and DBpedia can be easily replaced by any other domain ontology. Currently, interlinking with DBpedia is done manually, however in the final version we are aiming to automate this task. In a third phase the (subjective) statements of participants are analysed, and further added to the knowledge base. This is mainly achieved by the creation of `opm:DiscussionOpinion` instances and their respective properties. To this

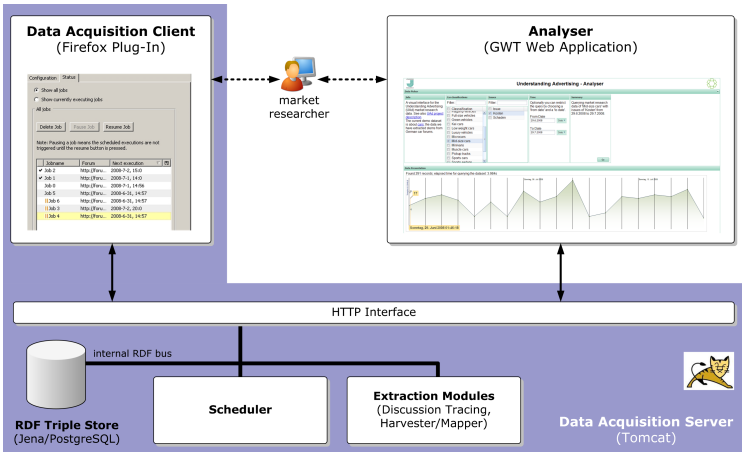

Fig. 3. Discussion Tracing in UAd.

end, we use a manually pre-configured list of possible topics, that is instances of `opm:Topic` to trigger the creation of opinions.

We have implemented a client/server system (Fig. 3, left and bottom) to perform the data acquisition in UAd. Within the scope of our research we support RDFising popular discussion forum types[7] such as vBulletin or phpBB. Data extraction occurs automatically using extracion profiles, manually defined for several forum types; a single acquisition task represents a single job on the server. The server has been implemented using a Java application server (Tomcat) along with a Jena 2/PostgreSQL RDF store taking care of the scheduling and execution of the acquisition tasks.

At the client side, a Firefox plug-in (Fig.4) allows a user to define, control and monitor the tracing tasks. The plug-in has been developed in JavaScript and XUL[8]. A user typically adds the link of a discussion forum and selects the forum type. Currently, only entire forums can be extracted. We plan to support the selection of sub-forums independently from each other for the extraction task. The user can also specify time parameters for the acquisition tasks, for example how often per week a job should be triggered to update the store.

4.2 Analyser

The UAd Analyser is a Web Application allowing a market researcher to examine the data gathered by the UAd Acquistion Server. In Fig. 5 the current state of the implementation (implemented with the Google Web Toolkit[9]) is depicted. The user can limit the data by selecting certain car classifications and issues

[7] `http://www.big-boards.com/statistics`
[8] `http://developer.mozilla.org/en/docs/XUL`
[9] `http://code.google.com/webtoolkit/`

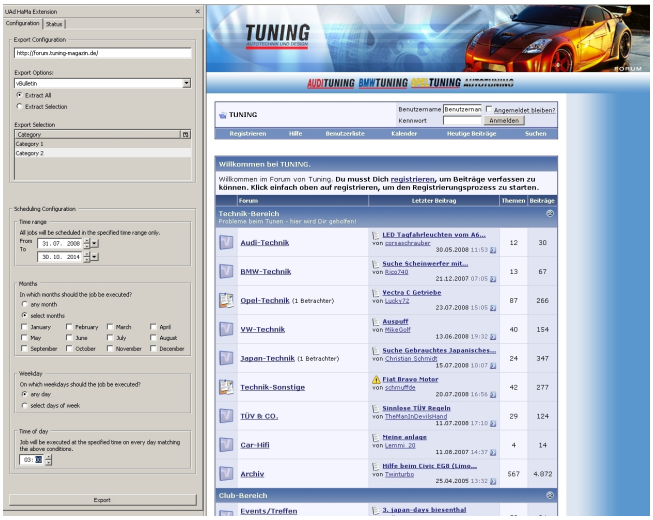

Fig. 4. The DT Plugin.

as well as by restricting the time period. The queried data is visualised with a Simile Timeplot[10] module, displaying the time on the X-axis and the number of discussion posts in the Y-axis. Discussion threads are illustrated as red vertical lines; the users may retrieve detailed information by clicking on the line and browse to the discussion thread where the matching post is located.

The post count, respectively a single time unit, in the X-axis reflects the occurrence frequency of topic. Additionally information about, diversity of authors who posted that day can be explored. The knowledge about authors diversity can be used to underline for instance how reliable or unreliable is the sentiment in chosen posts. The most important contribution of this visualisation is to offer an overview on diverse discussion forums regard a topic of interest.

5 Preliminary Results

In order to assess our opinion mining system, a baseline-evaluation using two standard information retrieval measures (precision and recall) has been performed. We have compared our approach to a full-text index (Lucene[11]). The domain is currently limited to "cars" (as we have mostly advertisements for the visual analysis available) although we note that the methodology is expected to yield similar results for other domains. The flexibility of our approach is mainly determined by the availability of appropriate instances from DBpedia.

[10] `http://simile.mit.edu/timeplot/`
[11] `http://lucene.apache.org/java/docs/`

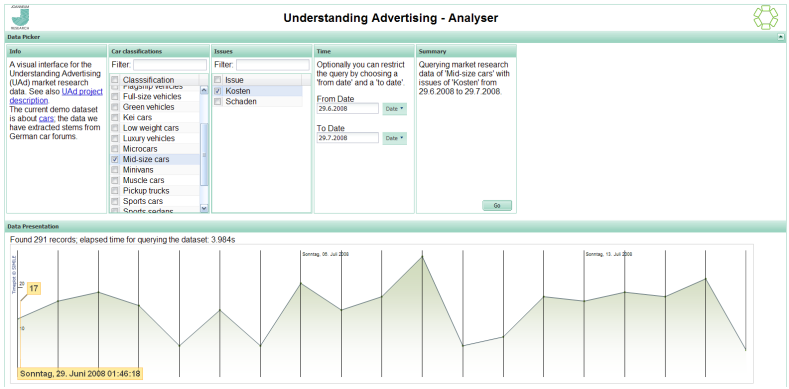

Fig. 5. The UAd Analyser.

5.1 Reference Data Set

Our reference data set contains approximately 1000 posts that have been extracted from a single discussion forum[12], focusing on the content of three sub-forums including all threads and posts about certain car types. Two of the extracted sub-forums contain discussions about cars belonging to the mid-sized car class according to categorisation from DBpedia[13]. The working data set includes 60 representative posts (20 per car type). We have manually selected posts containing discussion on topics such as "performance and problems" and "popularity".

The extracted posts were firstly used to generate opinions on the discussion topics, and secondly for the initialisation of the index over the reference test data (for Lucene). We have converted each of them into a single file containing information on the posting date, author, post URI, the content and the title of the thread the post belongs to, allowing to create an index searchable by Lucence. The Lucene index contains the fields author, title, summary, content and link to post corresponding to the properties in RDFised data and with the intention to provide as similar as possible initial point to RDFised data, for comparison and measurement of results.

Prior to the manual creation of the triggers for discussion topics we have analysed the initialised fields of the Lucene index for occurrence frequency of specific keywords and the "Zipf" distributions [22]. As depicted in listing 1.1 topic triggers contain words or word stems that serve as annotation events. Opinion generation is initiated by accordance of trigger words with words from the content or title of posts. An example discussion opinion generated in this way is shown in listing 1.2.

[12] `http://www.automotiveforums.com`
[13] `http://dbpedia.org/resource/Category:Mid-size_cars`

```
@prefix : <http://sw.joanneum.at/uad/cars/topics#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix opm: <http://sw.joanneum.at/uad/u-opm/schema/core-u-opm.rdf#> .

:performance_and_problems a opm:Topic;
 dc:subject "performance and problems";
 opm:hasTrigger "damage",
                "performance",
                ...
                "problem" .
```

Listing 1.1. Sample discussion topic snippet.

```
@prefix : <http://sw.joanneum.at/uad/cars/opinions#> .
@prefix utop: <http://sw.joanneum.at/uad/cars/topics#> .
@prefix opm: <http://sw.joanneum.at/uad/u-opm/schema/core-u-opm.rdf#> .

:do11 a opm:DiscussionOpinion;
 opm:about <http://dbpedia.org/resource/Alfa_Romeo_156>;
 opm:in <http://www.automotiveforums.com/vbul...php?t=173469>;
 opm:onTopic utop:performance_and_problems .
```

Listing 1.2. Sample generated discussion opinion.

5.2 Results

For the evaluation we have compared our method with the standard Lucene retrieval results of simple queries. Additionally we had a look at extended Lucene-queries; these extended queries have been used to decrease the influence of a single trigger. Listing 1.3 shows a sample SPARQL query we have used for our approach.

```
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix utop: <http://sw.joanneum.at/uad/cars/topics#> .
prefix opm: <http://sw.joanneum.at/uad/u-opm/schema/core-u-opm.rdf#>

SELECT * FROM <http://sw.joanneum.at/uad>
WHERE {
 ?do a opm:DiscussionOpinion ;
     opm:about ?about;
     opm:in ?in ;
     opm:onTopic utop:performance_and_problems .
 ?about owl:sameAs <http://dbpedia.org/resource/Alfa_Romeo_156> .
}
```

Listing 1.3. Sample opinion mining SPARQL query.

From table 5.2 we learn that regarding recall our method unsurprisingly seems to outperform simple full-text indexing. Even in the extended mode Lucene's precision and recall values are below our approach.

		Lucene		UAd Analyser	
		"performance and problems"	*"popularity"*	*"performance and problems"*	*"popularity"*
Precision	simple	0.4	1	0.76	0.86
	extended	0.2–0.62	0.56–0.86		
Recall	simple	0.1	0.05	0.95	0.6
	extended	0.05–0.8	0.3–0.7		

Table 1. Results from the Evaluation (Lucence vs. UAd).

Although we have used a rather limited working set in this evaluation we are optimistic that the results scale well both regarding size and other domains; further evaluations are in the scope of our current research.

6 Conclusion

In this paper we have proposed a novel approach to opinion mining on the Web by using Web of Data technologies and linked datasets. Our goal is to explicitly model opinions found in discussions on the Web; we have developed an according vocabulary to represent these opinions formally (in RDF) and have reported on an implementation of this approach.

We contemplate on using GoodRelations[14]—an ontology for linking product descriptions and business entities on the Web—in order to more accurately describe the target of an discussion in our realm.

To increase the precision we ponder about extending our opinion mining core mechanism with Natural Language Processing techniques and/or use neural networks to categorise topics automatically. As a part of the sentiment classification we aim to use SentiWordnet [23] or other similar approaches for the creation of opinion ranking based on trigger occurrences and the so called PN-polarity [15] of the content.

Currently, we summarise results visually respectively topics, identities, time and occurrence frequency to mirror the sentiment intention in opinions environment. However, currently we do not dive into sentiment interpretation of opinions. Considering the visual analysis, it is important to mention that sentiment interpretation underlies the judgement of end user and his observation standpoint. Anyway, objective parameters such as time period, identities, number of posts etc. can be evaluated independent of matter of particular interest. For further evaluations, user annotated content like reviews or similar will be used additionally.

[14] `http://www.heppnetz.de/projects/goodrelations/`
[15] P stands for "Positive" and N for "Negative" in this context.

Acknowledgements

The research reported in this paper has been carried out in the "Understanding Advertising" (UAd) project, funded by the Austrian FIT-IT Programme. The authors would like to thank their colleagues Magdalena Lauber, Wolfgang Weiss, and Werner Bailer for their support and valuable comments.

References

1. P. Mika. Microsearch: An Interface for Semantic Search. In *Proc. of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008) , Tenerife, Spain*, volume 334 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.
2. W. Halb, Y. Raimond, and M. Hausenblas. Building Linked Data For Both Humans and Machines. In *WWW 2008 Workshop: Linked Data on the Web (LDOW2008)*, Beijing, China, 2008.
3. M. Hausenblas, W. Halb, and Y. Raimond. Scripting User Contributed Interlinking. In *4th Workshop on Scripting for the Semantic Web (SFSW08)*, Tenerife, Spain, 2008.
4. S. Kim and E. Hovy. Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 483–490, Morristown, NJ, USA, 2006. Association for Computational Linguistics.
5. M. Hu and B. Liu. Mining opinion features in customer reviews. In *American Association for Artificial Intelligence at AAAI-04*, 2004.
6. M. Hu and B. Liu. Mining and summarizing customer reviews. In *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining at KDD-2004*, pages 168–177, 2004.
7. K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *WWW2003 - The Twelfth International World Wide Web Conference, Budapest, HUNGARY*, 2003.
8. N. Kobayashi, K. Inui, and Y. Matsumoto. Opinion Mining from Web Documents: Extraction and Structurization. *Informational and Media Technologies 2(1)*, 12(1):326–337, 2007.
9. A. Ghose, P. Ipeirotis, and A. Sundararajan. Opinion Mining using Econometrics: A Case Study on Reputation Systems. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2007.
10. M. Gamon and A. Aue. Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms. In *Proceedings of the ACL-05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, 2005.
11. M. Hausenblas and H. Rehatschek. mle: Enhancing the Exploration of Mailing List Archives Through Making Semantics Explicit. In *Semantic Web Challenge 2007 at the 6^{th} International Semantic Web Conference (ISWC07)*, Busan, South Korea, 2007.
12. S. Fernandez, D. Berrueta, and J.E. Labra. Mailing Lists Meet The Semantic Web. In *Proc. of the BIS 2007 Workshop on Social Aspects of the Web*, Poznan, Poland, 2007.

13. S. Fernandez, F. Giasson, and K. Idehen. SIOC Ontology: Applications and Implementation Status. `http://www.sioc-project.org/applications#creating-api`, 2007.
14. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, pages 722–735, 2007.
15. J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Inf. Process. Manage.*, 43(4):866–886, 2007.
16. A. Esuli. Opinion Mining. Presentation slides, Language and Intelligence Reading Group, June 14, 2006, Pisa, Italy, Istituto di Scienza e Tecnologie dell' Informazione Consiglio Nazionale delle Ricerche, 2006.
17. B. Liu. Opinion Mining and Summarization, Sentiment Analysis. Presentation slides, Tutorial given at WWW-2008, April 21, 2008 in Beijing, China, Department of Computer Science University of Illinois at Chicago, 2008.
18. T. Berners-Lee. Linked Data. `http://www.w3.org/DesignIssues/LinkedData.html`, 2007.
19. C. Bizer, T. Heath, D. Ayers, and Y. Raimond. Interlinking Open Data on the Web (Poster). In *4th European Semantic Web Conference (ESWC2007)*, pages 802–815, 2007.
20. T. Heath and E. Motta. Revyu.com: a Reviewing and Rating Site for the Web of Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, pages 895–902, 2007.
21. Semantic Web Deployment Working Group. SKOS Simple Knowledge Organization System Reference. W3C Working Draft, Semantic Web Deployment Working Group, 2008.
22. G. K. Zipf. *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology.* Addison-Wesley, 1949.
23. A. Esuli and F. Sebastiani. SentiWordnet: A Publicly Available Lexical Resource for Opinion Mining. In *5th Conference on Language Resources and Evaluation (May 22–28, 2006)*, Genova, Italy, 2006.

Towards Socially Aware Mobile Phones

Alessandra Toninelli[1], Deepali Khushraj[2], Ora Lassila[2], and Rebecca Montanari[1]

[1] DEIS – Università di Bologna, Viale Risorgimento 2, 40136 Bologna, Italy
{alessandra.toninelli, rebecca.montanari}@unibo.it,
[2] Nokia Research Center, 3 Cambridge Center, Cambridge, MA 02142, USA,
{deepali.khushraj, ora.lassila}@nokia.com

Abstract. Mobile phones currently represent the most pervasive technology for enabling social networking. They collect a wide range of socially meaningful data, such as address book contacts, pictures, call logs and exchanged messages. This data, however, is highly underutilized as it is managed by separate applications which are typically unaware of the user's social setting. This "unawareness" makes the user a slave of his device instead of helping him achieve his goals or manage his everyday activities. Enhancing next generation mobile phone with socially-aware features will provide significant benefits. In this paper, we present real data collected from a user study about mobile phone usage, with the aim of providing evidence for the need of socially-aware phone applications. As a relevant example of socially-aware capability, we analyze the case of interruption management, i.e., how and why users respond to an incoming call or message by interrupting their current activity. To automate interruption management on mobile phones, we suggest the adoption of a policy-based approach to express socially-aware policies based on Semantic Web technologies.

1 Introduction

Social bindings characterize each individual's life. From friendship to professional activities, to family bindings, every person is connected to a number of other individuals within the framework of a so-called social network. Technology advances in portable devices, such as mobile phones, offer a unique chance to support and improve social networking activities. In particular, mobile phones represent the most pervasive social networking tools that users currently exploit to build, maintain and manage the social networks they participate in [5]. Within their social networks users tend to coordinate mobile phone use according to group needs, expectations and social context [1].
Despite their role in social networking, mobile phones are currently equipped with software applications that are largely unaware of the users' social setting. For example, we witness how often people are interrupted by incoming phone calls and messages, which not only disturb the user but also his surroundings. Many of these interruptions could be avoided or adequately managed, only if the phone could adapt its behavior to the specific circumstances when the call is received. On the contrary, to utilize mobile phone applications the user has to adapt himself to the device and application logics- which happens every time we are forced to learn how to use another application running on our phone. Meanwhile, the

burden of managing social norms and patterns implied by mobile phone usage is left to the user. For instance, when one is in a meeting, he must remember to switch the phone to silent and decide whether to pick up any incoming call. If the caller could be informed that the called person is in a meeting, she might decide to call at a better time, thus avoiding useless interruptions and possible socially embarrassing situations.

The mobile phone has great potential as enabling means of social networking, but it is currently equipped with inadequate software applications that are unable to exploit the huge amount of socially-related data they collect from the user. Users' mobile phones carry in fact a considerable amount of socially meaningful data, such as contacts in the address book, pictures, call logs recording communication activity between users, and exchanged text messages or emails. This data, however, is highly underutilized as they are managed by separate applications and accessed by phone owners only when a specific information is needed.

We claim that next generation mobile phone applications should be enhanced with socially-aware features. To validate our arguments, we present real data collected from a user study we have conducted about mobile phone usage. As a relevant example of socially aware capability, we analyze the case of interruption management, i.e. how and why users decide to respond to an incoming call or message by interrupting their current activity. Users' activity of managing interruptions has been shown to have strong social underpinnings [5]. Novel mobile phone applications should therefore support users in regulating interruptions by (partially) automating features that, at present, are totally managed by hand.

Semantic technologies and policies seem to represent a promising solution to address this issue. Semantic languages permit to build an interconnected graph of socially meaningful data modeled according to an unambiguous semantics. This allows the exchange of social data without loss of meaning between both different applications and different users. In particular, in this paper we suggest the adoption of a policy-based approach to express socially-aware policies regulating mobile phone interruptions. At a high level, policies can be defined as directives ruling the behavior of (entities within) a system. Policies can therefore be used to establish under which conditions the user can be interrupted by a call or a message.

This paper will present some results from our study to provide useful insights about users' social needs and preferences with respect to mobile phone interruptions. The paper is organized as follows. Section 2 provides an overview of motivations and goals for our study, which is presented in Section 3. Section 4 presents our approach to a semantic policy-based framework to handle interruptions on mobile phones. Ongoing work and future research directions are outlined in Section 5, before conclusions in Section 6.

2 Analyzing Interruption Management to Understand Users' Social Needs

We have conducted a user study to determine and analyze which elements play a relevant role in users' decisions about mobile phone interruption management.

People are often interrupted by incoming calls and messages having both a personal and a social impact [4]. By observing how interruptions are currently handled by users, we might derive useful design insights to (at least partially) automate interruption management. Understanding when the user can be interrupted is, however, a challenging issue. Users themselves often find it difficult to think of all situations in which they will be interruptible a priori. There are different strategies for the user to limit and/or avoid such interruptions, including:

- filtering calls from certain callers, or at certain times
- being provided with information about the call (e.g. urgency, topic) and the caller to decide whether to respond to it
- provide information about the user situation to the caller, and/or a better time to call, and let him decide whether to call
- adapt the phone settings to the social protocol of the situation/location, e.g. switching the phone to silent mode

Currently, all the above mentioned strategies are manually performed by the user, who decides when to respond to incoming calls; might be able to figure out when a call is urgent or expected; might provide information, e.g. via SMS, about his current status; and is in charge of switching the ringing tone to silent when required by the situation, e.g. at the movie theatre. These strategies for interruption management can be grouped under two main categories of activities, namely:

- call filtering, e.g. based on caller identity, situation and time
- status information sharing, such as current location, activity, etc.

Our study is focused on the two above outlined activities. In particular, for each of them, we try to identify relevant *factors* and *actions*. With the term factors we denote decisional elements that influence the user in making choices about call filtering and status sharing, such as the identity of the contact or the particular situation (e.g. work, at home). By actions we mean those actions that are currently performed by users to filter calls and/or messages and to share or disclose information about their status, like for instance send a text message to say that one is in meeting, therefore he cannot answer.
Based on our factor and actions analysis, in the next section we provide insights about the results of our study.

3 Study Results And Evaluation

3.1 Methodology

The study was conducted in two phases. In phase 1 we conducted 14 face-to-face interviews. The results from phase 1 inspired us to do a larger study using which we could substantiate our findings. In the larger study, conducted in phase 2, there were about 50 participants split across Italy and USA. The key idea behind splitting participants across 2 countries was to analyze (and account for) culture specific social behaviors and norms.
In phase 1, we conducted 1 hour interviews with participants from 3 different

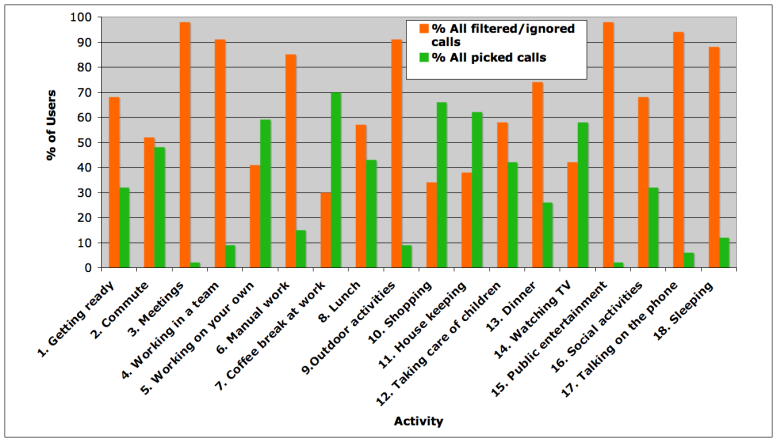

Fig. 1. Activity Based Response to Incoming Calls

organizations. During the interview we asked several questions related to the subject's daily routine. The questions were targeted to determine the subject's phone usage and interruptibility patterns, and also to determine willingness to share information (like presence, status, settings etc.). We also analyzed some objective data by looking at the user's call logs. In phase 2, we designed an online questionnaire. There were 26 participants from USA and 24 from Italy. The study included 24 males and 26 females in the age range of 20 to 61, with mean 35. We did not have any participants with a software or computer science related background, to avoid any bias due to personal expertise while answering questions, which we experienced in phase 1 of the study.
The important findings from the survey are discussed in the next section.

3.2 Study Results

Activity based call filtering: Despite users' reluctance or possible inability to think in advance of which factors influence their decision to pick up a call, the analysis of everyday activities provides evidence to support the idea of activity-based filtering criteria. As seen in Figure 1, most users tend to ignore or filter calls during activities, such as: meetings, while working in a team, during outdoor activities, while sleeping etc. Most of the calls are ignored or filtered by switching the phone to silent mode and by manually screening calls. In many cases, the current activity and the relationship to the caller both contributed to decision making.

Call filtering: In our study, we asked users if they would be willing to filter incoming calls from some people based on the current activity or event. 67% of the participants did not like the idea of filtering calls automatically, instead they preferred to decide when the call arrived. Of the participants who said "no" to call filtering, 82% did not want filtering because they thought it was too difficult

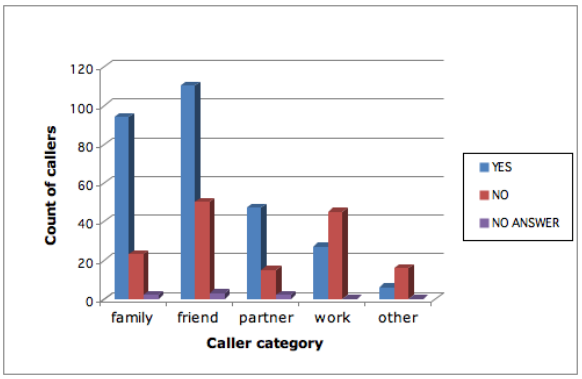

Fig. 2. Willingness to Share Status with Caller

to decide in advance what calls to filter. 21% were not willing to filter because it could lead to a lack of control.

Sharing status with address book contacts: Sharing the user's status using messages (like busy, free, at work, in meeting etc.) with incoming callers or address book contacts can help avoid interruptions. In our study, 78% of the participants indicated that they would be willing to share status messages with contacts in their address book. 8% were willing to share with all contacts, 62% indicated that they would share with some contacts only and 8% were willing to share based on the situation rather than just the caller. It is interesting to notice that all participants who already share status by using instant messenger applications (IM) were also willing to share status messages with some or all contacts in their address book. 84% of the overall participants were also willing to share with some or all callers the reason for not picking up the call, such as poor network coverage, low battery life or cost (ordered by importance).
We also collected objective information by analyzing the users' recent (last 5) received and missed calls. We observed that most of the callers fell under well defined categories (suggested by users) such as family, work etc. Figure 2 shows willingness to share status based on category. We can see that willingness to share status is quite high for categories such as family, friend and partner.
Willingness of users to use someone else's status message is an important factor in determining the popularity of status sharing. Overall 84% of the participants said they would use status information before calling another person, for example to avoid disturbing him. Of this, 40% were willing to use this information only for people who frequently updated their status; 18% said they would use this information, but still send a text message to cross-check and 26% said they would use this information anyway.

Sharing location with address book contacts: Sharing location information (also called presence) could provide some context about the user, which could help avoid interruptions.

54% of the participants were willing to share location with some or all contacts at all times or based on the situation. Only 29% of Males were willing to share location, in contrast 77% of the Females were willing to share location. Willingness to share location was equally distributed across people from Italy and USA.
Participants with over 200 contacts (20%) were the least willing to share location– 82% said no to location sharing. Participants who share status on IM were the least averse to sharing location– only 9% said no to location sharing.

Traveling scenario: While traveling, people are often concerned about additional costs and timezone differences, their phone usage can often change due to this.
Approximately 65% of the participants said they are more selective with picking up calls, either because of cost (main reason) or because of other reasons (e.g. timezone, don't like getting disturbed etc.). Most people who were more selective, picked up calls based on relationship to caller or based on mental priorities. Calls from unknown numbers were often not picked.
If the phone could inform contacts about travel status (e.g. with messages like "in Italy", "not in town", "reachable after July 6th"), 74% said they would use this feature with some or all contacts or based on situation.

Phone behavior in public situations: People are often embarrassed in a public place (e.g. library, movie theatre etc.) by having their phone ring because they forgot to put it on silent. This data was generally confirmed by our study. In a public place, the phone could be programmed to turn silent automatically e.g. when in a library, if 40 other phones are on silent, your phone could automatically turn silent. 68% of the participants were willing to use this feature. Of the overall participants, 42% said they would like the phone to prompt first before going silent.
When asked if they would be willing to share their ringer settings with others around (assuming no identifying information is revealed), only 36% said yes. Most participants were not willing to share this information with people around.

Calendar based automation: The phone could use the user's calendar (from PC or phone) to automatically turn silent, when in a meeting. 73% of the users who use a calendar, were willing to use this feature. In some cases people wanted to explicitly select the events for which the phone should turn silent, while some others wanted the phone to prompt them before going silent. 70% of the participants who said no, did not want to use this feature because their calendar was not up-to-date.

Tolerance to failure: The survey also collected information on tolerance to failure while automating the phone's behavior or for information sharing.
Both in the case of sharing location and filtering calls, on average 50% of the participants willing to use these automated features said they would use it only if it worked always. Of the rest, participants seemed to be more tolerant to failure for location sharing: for example, 13% would use it even in case of wrong behavior, but expect others to not rely on this information. Between 30% and 35% of the

people said that they would use the location sharing or call filtering feature only if it was free of charge.
Of the people who were willing to use the auto silent feature in public situations, 38% said they would not use this feature unless it worked always and 22% said they would use this feature but not rely on it.

3.3 Evaluation of Results

As far as call filtering is concerned, we observed a strong need to filter calls based on activity, yet we see that most users do not like the option of call filtering. We believe that users would be more willing to use call filtering if the phone made it easy to do so. For instance, the phone could recommend settings based on observed phone usage patterns (that varied by activity and/or relationship to caller). Also, we observed that the same factors, i.e. activities and relationship to the caller, tend to be recurring in users' call filtering decisions. If there were predefined templates available for popular activities (such as sleeping, in meeting etc.), or for popular relationships (like spouse, close friends etc.,) then people would be more willing to filter calls.
We also noticed that in some situations most users generally adopt similar behaviors about call filtering, such as meetings and sleeping time, while in others their decisions depend on several factors, such as expectations and current mental priority order. Another interesting scenario is represented by traveling, when most people tend to adopt regular strategies to pick up calls.
Regarding status sharing, users generally feel confident about sharing some information with selected contacts. This is supported by both answers to direct questions and call log analysis. For some categories, location seems to be a critical piece of information, probably due to privacy concerns that are generally dependent on social settings and habits. Quite surprisingly, users did not like the idea of sharing their ringer settings, which we considered a not particularly sensitive information. One explanation might be that they don't like to share any kind of information with people they don't know.
Finally, the tolerance to failure is not very high as users expect no mistakes or at least 99% accuracy. It must be said, however, that it might be very difficult for users to assess their tolerance to failure for applications they have never seen nor used.

4 Semantic Policies: a Viable Approach Towards Socially-Aware Mobile Phone Applications

The results of our study show that users currently manage interruptions on their mobile phone both by filtering calls/messages and sharing status information, e.g. via text messages. In particular, the study helped us to outline a number of factors that play a role in the user's decision about answering to calls and showing status information to other people.
Regardless of the specific factors and actions identified in the study, it is worth underlining that users actually discriminate situations (in terms of caller, time, activity, etc.) when they have to make decisions about calls and status information.

In other words, users have *strategies* in mind, albeit not always explicit, which they put in action to manage specific situations. As stated above, while some strategies are generally constant with respect to a certain situation (e.g. meetings), others are more difficult to predict since they tend to dynamically change. Therefore, only the former represent a possible choice for supporting automated features.
In particular, in order to build an automated interruption management system, we need to be able to express and enforce user defined "rules of conduct" about call filtering and status sharing. For instance, we might wish to explicitly specify that, during meetings, only urgent calls from family members are allowed. It is interesting to note that both cases, call filtering and status sharing, lend themselves to be modeled as involving *access control* decisions. In the latter case, the accessed resource is status information, while in the former it is the "user's attention" itself, which we can think of as a particular type of resource owned by the user.
We claim that such user defined strategies can be expressed as access control policies. Policies represent an emerging research direction in the area of access control and security in general. At a high level, policies are defined as directives regulating the behavior of (entities within) a system. Policies have been extensively studied over the last decade and applied to several application fields, from network management, to multi-agent systems regulation, to security [6]. In particular, access control policies define which subject is allowed to access which resources under which circumstances.
Access control policies provide a powerful and expressive model to represent and enforce user preferences and constraints with respect to interruptions. Previous research on policies provides a well-established foundational model for representing directives about how entities operating in a system are allowed/not allowed to access resources, as well as reference architectural models for evaluating and enforcing policies [6]. Existing work also includes tools for policy specification, management and enforcement that have already been utilized in different applications domain [2]. In particular, relevant research efforts have been spent in recent years to integrate semantic technologies within policy definition, thus enabling automated reasoning over expressive policy definitions [3, 9]. We believe that the adoption of semantically rich policies provides a suitable solution to the issue of representing and enforcing user defined strategies for interruption management.

4.1 Socially Aware Policies

The adoption of a policy approach to control interruptions on mobile phones requires the definition of a policy model that can precisely enough identify the basic types of policies required to control access to the user's attention (call filtering) and to information about his status (status sharing), can specify how to express and represent policies in a semantically expressive form, and how to enforce them. In particular, based on the analysis of the results from our study, we consider the following to be requirements that should be considered in the design of a socially aware policy model to regulate interruptions:

– Support for *intensional* rather than extensional definitions of policies. For example, a user would like to define a policy applying to all his "friends" or to "all people depicted in this picture", rather than explicitly naming each person to whom the policy applies.
– Support for *social-aware* modeling of access conditions that reflect the actual relationships between each user and the social environment in which he operates. Social relationships and activities play a crucial role in the user's mental model of sharing and controlling access to his resources, as shown by previous literature in the field [7, 5], and by the results of our study as well.

We suggest that semantic technologies are well suited to model socially aware policies. In particular, we adopt the Semantic Web language RDF to model access control policies. Semantic languages support the intensional definition of policies by allowing the expressive representation, at a high level of abstraction, of the conditions under which a resource can be accessed. Semantic languages also represent a promising solution to the issue of properly representing user's social environment, as efforts like the "Friend-of-a-Friend" (FoaF) initiative demonstrate [3]. In addition, thanks to the ability of performing automated reasoning over social information, semantic techniques might increase the expressivity of user defined socially-aware policies. For instance, a policy applying to a "call" would also apply to a "video call" if we define the latter as an RDF subclass of the former.
Furthermore, a crucial issue in current mobile phone applications lies in the extreme fragmentation of social data. Disseminated within different applications, social data cannot be connected due to the boundaries of the applications that collect and manage them. RDF graphs not only provide a uniform and semantically defined representation for social data, but they also offer the great potential of interconnecting them via semantic links, thus creating a global graph of social information about the user's world. Let us note that interconnection might be needed at the application level, i.e. between different applications running on the device(s) of the same user, and at the social level, i.e. between applications running on behalf of different users or organizations. For example, information about a contact in the user's personal address book could be connected to information extracted from a corporate address book, or a social networking application such as Facebook [4].
Finally, we make the note that social data are constantly changing according to user's situation, activities and social relations. Thanks to their extensibility features, semantic techniques are well suited to accommodate (possibly unforeseen) modifications/additions to existing social data, thus allowing for the greater flexibility in policy specification.

4.2 An Example of Socially-Aware Policy

In this section we provide a brief insight on the socially-aware policy model we have developed. A detailed description of the model is out of the scope of this paper. We illustrate an example to show how semantic technologies can be used

[3] http://xmlns.com/foaf/0.1/
[4] http://www.facebook.com

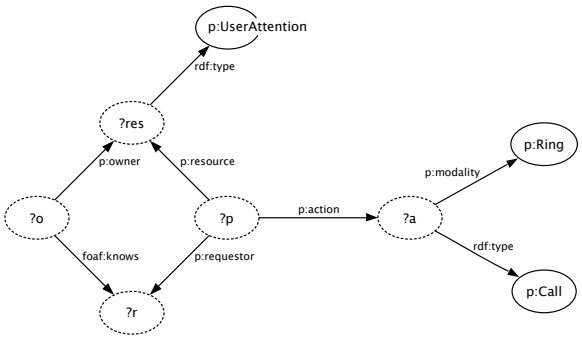

Fig. 3. Query conditions represented graphically as a graph

to model socially aware policies.
In our model, a policy defines all the characterizing information that is considered relevant for access control. This includes information about the entities that might operate on the resource, about the resource itself or other properties, e.g., conditions of the surrounding environment, such as time.
To represent policies we adopt RDF graphs: each policy context is represented as a set of RDF statements about the characterizing elements of a policy, or as SPARQL triple patterns (in the case when there are "unknowns" that have to be matched at policy enforcement time). The use of property paths allows us to represent relationships between the resource, requestor and policy context conditions.
Let us consider for example the following policy: "My phone will ring when people I know call me":

?p `p:requestor` ?r
?p `p:resource` ?res
?res `rdf:type p:UserAttention`
?p `p:action` ?a
?a `rdf:type p:Call`
?a `p:modality p:Ring`
?o `p:owner` ?res
?o `foaf:knows` ?r

This set of (conjunctive) conditions can be thought of as a "graph template", as illustrated in Figure 3.
Figure 4 illustrates a possible (partial) graph that could be matched by the above conditions – note that the nodes inside the dashed oval are data typically stored in the user's social network profile or (with the advent of socially-aware software) possibly in the user's address book.
Let us note that our model does not impose any limitation to the kind of attributes and values that can be defined for policy specification. In fact, the possibility to define customized and application-specific policy constraints allows for great flexibility in policy definition. In our example policies, we used the common

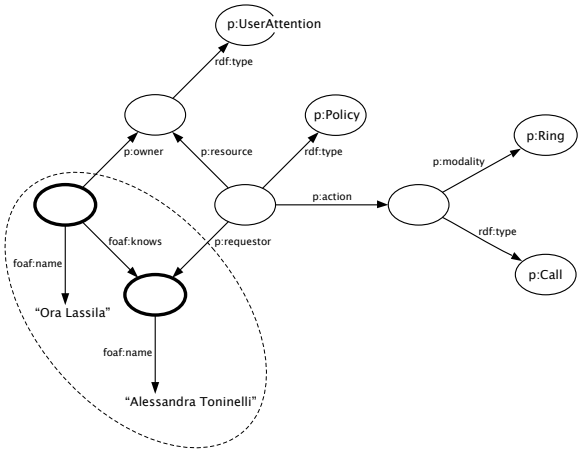

Fig. 4. Possible (partial) graph representing the results of the query

attributes to all policy definitions, such as "requestor" or "resource", while others are policy-specific, such as "knows".

5 Ongoing Work

Based on the policy model presented in Section 4 and on previous work [8], we are working on the design and development of a socially aware policy framework for interruption management on mobile phones.
The framework should provide support for the specification, retrieval and enforcement of semantic policies. In particular, when creating a new policy, a key issue is not only "how" to specify policies, for example via user-friendly graphical interfaces, but also "when" and "where" the user is allowed to specify a policy. As revealed by our study, users often find it difficult to make a priori decisions about how to respond to calls and to share status. Hence, we can expect them to define and/or manage access control policies not as a separate process, but within a certain application context, in which those policies make sense, e.g. when they actually receive a call or need to share status information.
In order to integrate policy specification directly into applications, we are focusing on a customizable policy interface that is designed not only to execute also as a stand-alone application, but also to be activated from the applications installed on mobile devices, such as the calendar or the address book. Similarly to policy definition, policy retrieval mechanisms should be integrated with user applications. It is worth noting that such integration feature strongly relies on the graph-based nature of the RDF-based policy model. A graph-based policy definition can be built and browsed in multiple directions from different starting nodes, thus allowing multiple definitions and interpretations of the same policy from different perspectives.

In addition, we are further analyzing the results provided by our study, particularly to identify possible correlations between various factors and actions that might provide us with useful insights for the design of our policy framework.

6 Conclusions

Despite being the most pervasive social networking tools, mobile phones are currently equipped with software applications that are largely unaware of the user's social setting. In this paper, we highlight the need for socially-aware mobile phones and the mechanisms needed to enable such social features.
Understanding the user and his needs are fundamental to building a socially-aware phone. Therefore, we conducted a cross-cultural user study with a specific focus on interruption management i.e., how can we better manage interruptions to mobile phone users and their surroundings. The study provided useful insights into scenarios and phone features where automation could help. It has also remarked the importance of keeping automated features easy to use.
We propose to adopt a semantic policy-based approach to express socially-aware policies that can regulate interruptions on mobile phones. The use of a semantic model enables us to utilize the user's fragmented social data and allows support for intensional policy descriptions. The design and implementation of the policy model (and framework), presented in this paper, are still under development.

References

1. W. Campbell and T. Russo. The social construction of mobile telephony. *Communications Monographs*, pages 317–334, 2003.
2. N. Damianou, N. Dulay, E. Lupu, and M. Sloman. The Ponder policy specification language. In *POLICY 2001*.
3. L. Kagal, T. W. Finin, and A. Joshi. A policy language for a pervasive computing environment. In *POLICY 2003*.
4. A. Kahlil. *Context-Aware Telephony and Its Users: Methods to Improve the Accuracy of Mobile Interruptions*. PhD thesis, Indiana University, 2006.
5. J. E. Katz and M. Aakhus. *Perpetual Contact: Mobile Communication, Private Talk, Public Performance*. Cambridge University Press, New York, NY, 2002.
6. M. Sloman. Policy driven management for distributed systems. *Journal of Network Systems Management*, 2(4), 1994.
7. D. S. Tara Whalen and E. F. Churchill. User experiences with sharing and access control. In *CHI '06 extended abstracts on Human factors in computing systems*. ACM Press, 2006.
8. A. Toninelli, et al. A semantic context-aware access control framework for secure collaborations in pervasive computing environments. In *ISWC*, volume 4273 of *LNCS*, pages 473–486. Springer, 2006.
9. A. Uszok, et al. Kaos policy and domain services: Toward a description-logic approach to policy representation, deconfliction, and enforcement. In *POLICY 2003*.

Wikipedia Mining for Triple Extraction Enhanced by Co-reference Resolution

Kotaro Nakayama

The Center for Knowledge Structuring
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
TEL: +81-3-5841-0462 FAX: +81-3-5841-0454
nakayama@cks.u-tokyo.ac.jp

Abstract. Since Wikipedia has become a huge scale database storing wide-range of human knowledge, it is a promising corpus for knowledge extraction. A considerable number of researches on Wikipedia mining have been conducted and the fact that Wikipedia is an invaluable corpus has been confirmed. Wikipedia's impressive characteristics are not limited to the scale, but also include the dense link structure, URI for word sense disambiguation, well structured Infoboxes, and the category tree. In previous researches on this area, the category tree has been widely used to extract semantic relations among concepts on Wikipedia. In this paper, we try to extract triples (Subject, Predicate, Object) from Wikipedia articles, another promising resource for knowledge extraction. We propose a practical method which integrates link structure mining and parsing to enhance the extraction accuracy. The proposed method consists of two technical novelties; two parsing strategies and a co-reference resolution method.

1 Introduction

Even though the importance of ontology construction is widely recognized and a considerable number of Semantic Web implementations based on standardized formats (such as RDF and OWL) are being built/published on the WWW, what seems lacking is the mapping of ontologies due to the nature of distributed environments. Since it is difficult to map local ontologies one by one, an approach based on the global ontology approach seems a solution having capability to intermediate local ontologies. However, previous methods for constructing huge scale ontologies faced technical difficulties, since it was impossible to manage a huge scale global ontology due to the lack of human resources.

Meanwhile, Wikipedia, a collaborative wiki-based encyclopedia, has become a phenomenon among Internet users. According to statistics of Nature, Wikipedia is about as accurate in covering scientific topics as the Encyclopedia Britannica. It covers concepts of various fields such as Arts, Geography, History, Science, Sports, Games. Wikipedia contains more than 2 million articles (Oct. 2007, English Wikipedia) and is becoming larger day by day while the largest paper-based encyclopedia Britannica contains only 65,000 articles. As a corpus for knowledge extraction, Wikipedia's impressive characteristics are not limited to the scale, but also include the dense link structure, sense disambiguation based on URL, brief link texts (a. k. a. anchor texts) and well structured sentences. The fact that these characteristics are valuable to extract accurate knowledge from Wikipedia is strongly confirmed by a number of previous researches on Wikipedia

Mining [1–8]. Besides, we proposed a scalable link structure mining method to extract a huge scale association thesaurus in a previous research [2]. In that research, we developed a huge scale association thesaurus dictionary extracting a list of related terms from any given term. Further, in a number of detailed experiments, we proved that the accuracy of our association thesaurus achieved notable results. However, association thesaurus construction is just the beginning of the next ambitious research on huge scale Web ontology construction from Wikipedia.

Semantic Wikipedia [9] is an impressive solution for developing a huge scale ontology on Wikipedia. Semantic Wikipedia is an extension of Wikipedia which allows editors to define semantic relations among concepts manually. Another major approach is to use Wikipedia's category tree as an ontology [7, 8]. These researchers proved that Wikipedia's categories are promising resources for ontology construction by showing significant results.

In contrast to these approaches, we propose a full-automated consistent approach for semantic relation extraction by mining Wikipedia articles. Since a Wikipedia article is a set of definitive sentences, the article text is yet another valuable resource for ontology construction. However, co-reference resolution will be one of the serious technical issues for this aim since a lot of abbreviations, pronouns and different expressions are used to point an entity in a Wikipedia article. Therefore, we propose a co-reference resolution method based on synonym information and an improvement method by using important sentence detection.

The rest of this paper is organized as follows. In section 2, we explain a number of researches on Wikipedia Mining for knowledge extraction in order to make our stance clear. In section 3, we describe our proposed integration method based on parsing and link structure mining. We describe the results of our experiments in section 4. Finally, we draw a conclusion in section 5.

2 Related Works

2.1 Relation Acquisition from Text Corpora

In the statistical NLP research area, a significant number of researches on relation acquisition from large scale text corpora have been conducted. For instance, Hearst [10] is one of the researchers who has pointed out that lexico-syntactic patterns (mainly for is-a relation) can be extracted from large scale corpora. Berland and Charniak [11] have proposed similar methods for part-whole relations. Kim and Baldwin [6] focused on nominal relations in compound nouns.

These researches are targeting ordinary text corpora or Web corpora, thus in order to apply these methods for Wikipedia, we need to consider about the characteristics and reconstruct the methods since Wikipedia has various unique characteristics compared with other corpora.

2.2 Wikipedia Mining

"Wikipedia mining" is a new research area that is recently addressed. Researches on semantic relatedness measurement are already well conducted [1–3]. WikiRelate [3] is one of the pioneers in this research area. The algorithm finds the shortest path between categories which the concepts belong to in a category tree. As a measurement method for two given concepts, it works well. However,

it is impossible to extract all related terms for all concepts because we have to search all combinations of category pairs of all concept pairs (2 million × 2 million). Therefore, in our previous research, we proposed $pfibf$ (Path Frequency - Inversed Backward Link Frequency)[1], a scalable association thesaurus construction method to measure relatedness among concepts in Wikipedia. The basic strategy of $pfibf$ is quite simple. The relativity between two articles v_i and v_j is assumed to be strongly affected by the following two factors:

- the number of paths from article v_i to v_j,
- the length of each path from article v_i to v_j.

The relativity is strong if there are many paths (sharing of many intermediate articles) between two articles. In addition, the relativity is affected by the path length. In other words, if the articles are placed closely together in the graph of the Web site, the relativity is estimated to be higher than that of farther ones. Therefore, by using all paths from v_i to v_j given as $T = \{t_1, t_2, ..., t_n\}$, the relativity pf (Path Frequency) between them is defined as follows:

$$pf(v_i, v_j) = \sum_{k=1}^{n} \frac{1}{d(|t_k|)}, \tag{1}$$

$$pfibf(v_i, v_j) = pf(v_i, v_j) \cdot \log \frac{N}{bf(v_j)}. \tag{2}$$

$d()$ denotes a function which increases the value according to the length of path t_k. N denotes the total number of articles and $bf(v_j)$ denotes the number of backward links of the page v_j. Wikipedia Thesaurus [2] [2] is an association thesaurus search engine that uses $pfibf$ in its behind. It provides over 243 million relations for 3.8 million concepts in Wikipedia.

2.3 Wikipedia and Web Ontology

Semantic Wikipedia [9] is one of the pioneers that remarked the effectiveness of Wikipedia style editing for making a huge ontology covering wide range topics. Semantic Wikipedia is an extension of Wikipedia which allows editors to define relations among concepts manually. The contribution of Semantic Wikipedia is that it showed a new direction to achieve the vision of the Semantic Web. While Semantic Wikipedia is a promising approach for a huge scale Web ontology construction, it needs human-effort. Therefore, we try to develop a completely-automated method without any additional human-effort since Wikipedia articles already include rich semantic relations.

Another interesting approach is to use Wikipedia's category tree as an ontology [7, 12]. In previous researches on Wikipedia mining, a large number of researches were based on category tree analysis since Wikipedia categories are a promising resource for ontology construction. For instance, DBPedia [5] uses several types of information on Wikipedia such as InfoBox, article texts, categories in order to extract structured knowledge and provide Web APIs.

[1] The method name was $lfibf$ in the past and was changed to $pfibf$

[2] http://wikipedia-lab.org:8080/WikipediaThesaurusV2

In this research, in contrast to these approaches, we developed a full-automated consistent approach for semantic relation extraction by mining Wikipedia article texts. Wikipedia article texts are promising resources to extract semantic relations but a small number of researches have been conducted in this area.

2.4 Characteristics of Wikipedia

As a Web corpus for knowledge extraction, URL for word sense disambiguation is one of the most notable characteristics of Wikipedia. In Wikipedia, almost every page (article) corresponds to exactly one concept and has an own URL respectively. For example, the concept apple as a fruit has a Web page and its own URL. Further, the computer company Apple also has its own URL and these concepts are semantically separated. This means that it is possible to analyze term relations avoiding ambiguous term problems or context problems.

Hyperlinks do not just provide a jump function between pages, but have more valuable information than we expect. There are two type of links; "forward links" and "backward links". A "forward link" is an outgoing hyperlink from a Web page, an incoming link to a Web page is called "backward link". Researches on Web structure mining, such as Google's PageRank [13] and Kleinberg's HITS [14], emphasize the importance of backward links in order to extract objective and trustful data. "Link texts" also contains valuable information.

Link texts in Wikipedia have a quite brief, clear and simple form compared with those of ordinary Web sites. Among the authors of Wikipedia, it is a common practice to use the title of an article for the link text but users also have the possibility to give other link texts to an article. This feature makes another important characteristic; the "variety of link texts," which can be used to extract valuable information. However, what seems interesting is that link texts do not contain any wordy information in most cases. Since no link text data is available on Wikipedia dump data, we customized the Wiki parser engine on Wikipedia to extract the link text data.

3 Proposed method

In order to extract semantic relations from Wikipedia, we propose a method that analyzes both the Wikipedia article texts and link structure. Basically, the proposed method extracts semantic relations by parsing texts and analyzing the structure tree generated by a parser. However, parsing all sentences in an article is not efficient since an article contains both valuable sentences and non-valuable sentences. We assume that it is possible to improve accuracy and scalability by analyzing only important sentences on the page. Furthermore, we use synonyms to enhance co-reference resolution. In a Wikipedia article, usually a number of abbreviations, pronouns and different expressions are used to point to an entity, thus co-reference resolution is one of the technical issues in order to make the parsing process accurate.

Figure 1 shows the whole flow of the proposed method. The method consists of three main phases; parsing, link (structure) analysis, and integration. First, for a given Wikipedia article, the method extracts a list of related terms for an article using $pfibf$ [2]. At the same time, it provides synonyms by analyzing the link texts of backward links of the article. Second, the method analyzes the article text to extract explicit semantic relations among concepts by parsing the

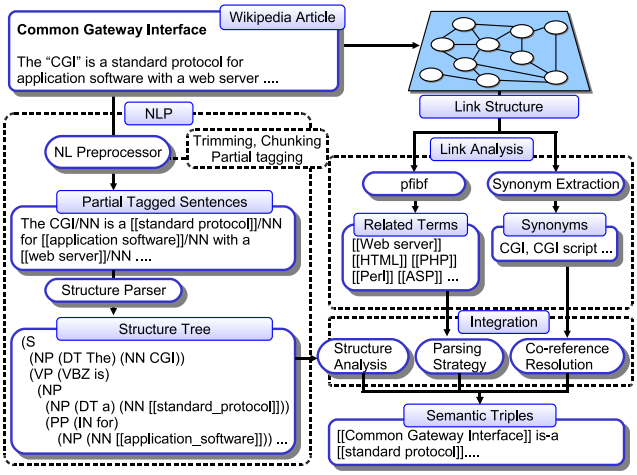

Fig. 1. Whole flow of the proposed method.

Table 1. Synonym extraction by link text analysis.

Concept	Synonyms
Apple Computer	'Apple' (736), 'Apple Computer, Inc.' (41), 'Apple Computers' (17)
Macintosh	'Apple Macintosh' (1,191), 'Mac' (301), 'Macs' (30),
Microsoft Windows	'Windows' (4,442), 'WIN' (121), 'MS Windows' (98)
Intl. Organization for Standardization	'ISO' (1026), 'international standard' (4), 'ISOs' (3)
Mobile phone	'mobile phones' (625), 'cell phone' (275), 'Mobile' (238)
United Kingdom	'United Kingdom' (50,195), 'British' (28,366), 'UK' (24,300)

(): Number of backward links
(Link texts corresponding to the title of an article are excluded).

sentences. Finally, in the integration phase, three steps for triple extraction are conducted; 1) analyzing the structure tree generated by the parser, 2) filtering important semantic information using parsing strategies, and 3) resolving co-references by using synonyms. The main steps of the proposed method are described as follows.

3.1 Synonym Extraction

We describe our co-reference resolution method by using synonyms extracted from anchor texts. A synonym word has one meaning but various expressions. Since backward links of a web page have a "variety of backward link texts," this variety can be used to extract synonyms of a concept (article). For instance, the computer company "Apple" is sometimes referred to as "Apple", but it is sometimes also written as "Apple Computer, Inc," "Apple Computers," etc. Table 1 shows a number of examples of randomly chosen synonym terms.

The article "Apple Computer" has 1,191 backward links with the link text "Apple Macintosh" and 301 backward links with the link text "Mac." This shows that both words are typical synonyms for the concept "Apple Computer." Statistical data unveiled that backward link texts analysis can extract high quality synonyms by specifying a threshold to filter noisy data such as 'international standard' and 'ISOs' for ISO.

Synonyms are helpful information to detect whether two sentences are describing the same subject. In other words, the information is needed for coreference resolution. For example, there is an article about "United Kingdom" in Wikipedia and it contains "UK" many times. However, if the machine does not know that "UK" is a synonym of "United Kingdom," it can not extract many relations on the topic. Therefore, we use the extracted synonyms in the following steps to improve the coverage.

For a given article a and the synonym candidate s, we define a simple scoring function $syn(a, s)$ as follows;

$$syn(a, s) = \frac{log\ num\_bk(a, s)}{log\ num\_bk(a, *)}. \tag{3}$$

$syn(a, s)$ basically measures the popularity of the label for the concept by calculating ratio of total backward links and the link texts. $num\_bk(a, s)$ is the number of backward links of a with link text s. $num\_bk(a, *)$ is the total number of backward links of a. We defined a threshold for $syn(a, s)$ to filter irrelevant synonyms by 200 training data evaluated by human effort.

3.2 Preprocessing

Since the structure and syntax of a Wiki is much different from natural languages, we need to modify and optimize the parser by considering special syntax composed of HTML tags to achieve better accuracy. Basically, special Wiki command tags such as triple quotation, brackets for hyperlinks and tables, prevent correct parsing. However, it is also true that this kind of information is helpful to analyze the content since it contains hyperlinks and helpful information to compound words into semantic chunks. Therefore, we constructed a preprocessor by ourselves to achieve better accuracy. The preprocessor trims the Wikipedia article to remove unnecessary information such as HTML tags and special Wiki commands first. It also removes table tags because contents in tables are usually not sentences. However, it does not remove link tags ("[[...]]") because links in Wikipedia are explicit relations to other pages and we use the link information in the following steps. Finally, phrases in quotations and link tags are tagged as nouns to help the following parsing step.

Parsing and Structure Tree Analysis After the preprocessing, it provides partially-tagged sentences. In this step, the method parses the sentences to get a structure tree and analyzes the structure tree to extract semantic relations. To parse sentences, we adopted a lexcicalized probabilistic parsing method based on the factored product model. We used the Stanford parser [15] for this purpose. It can parse a sentence accurately if the sentence is trimmed, chunked and tagged correctly by preprocessing. A list of main POS (Part Of Speech) tags used in this step is shown in Table 2 (Right).

Table 2. Wikipedia statistics and POS tags.

Statistics of Wikipedia articles.

# of concept pages (exc. redirect and category pages)	1,580,397
# of pages having more than 100 backward links: P_a	65,391
# of pages (in P_a) begin with is-a definition sentence: P_b	56,438
# of pages (in P_a) that the 1st sentence has links: P_c	62,642
# of $P_b \cap P_c$	56,411

POS Tags.

Tag	Description
NN	Singular or mass noun
NNS	Plural noun
NNP	Singular proper noun
NNPS	Plural proper noun
NP	Noun phrase
VB	Base form verb
VBD	Past tense
VBZ	3rd person singular
VBP	Non 3rd person singular present
VP	Verb phrase
JJ	Adjective
CC	Conjunction, coordinating
IN	Conjunction, subordinating

For example, for a sentence "Lutz_D._Schmadel is [[Germany]] [[astronomer]]." about the person with the name "Lutz_D._Schmadel," the parser generates a structure tree like this;

```
(S (NP (NN Lutz_D._Schmadel) (VP (VBZ is) (NP (NN [[Germany]]) (NN [[astronomer]])))))
```

In our proposed method, the parser takes a partially tagged sentence made by preprocessing and generates a structure tree from the sentence. After that, the structure tree is analyzed in order to extract triples (Subject, Predicate, Object) in the following steps:

1. Extract "(NP ...) (VP (VBZ/VBD/VBP ...) (NP ...))" pattern from the parsed sentence.
2. For both NP, replace the NP by the last NN/NNS in the NP if the NP parts consist of JJ and NN/NNS.
3. For both NP, split the NP into two NP parts if the NP contains CC. After that, perform step 2 again.
4. If the 1st NP is a synonym of the concept representing the article, replace the NP part by the title of the main subject.
5. Finally, extract the 1st NP part as a subject, VB part as a predicate, the 2nd NP part as an object.

In the first step, we extract "(NP ...) (VP (VBZ/VBD/VBP ...) (NP ...))" and assume that the 1st NP part is the subject, the VB part is the predicate, the 2nd NP part is the object respectively.

In the second step, for both NP parts, we replace NP by the last NN/NNS term (or hyperlink) because the last term is the mainstay of the phrase. For instance, the 2nd NP in the sentence about "Lutz_D._Schmadel" consists of two NN and both of them have a hyperlink to other pages and the 1st NN has a link to a country "Germany". So in this case, it obtains "[[astronomer]]" as the mainstay of the object part.

In the third step, NP will be separated if it contains CC such as "and" and "or". In the fourth step, if the 1st NP is a literal and it is a synonym of the concept representing the article, then the NP is replaced by the concept of the article. Finally, the first NP part is extracted as a subject, the VB part as a predicate, the 2nd NP part as an object.

The first step's POS tag pattern can be replaced by other alternatives. Currently, we prepared following three for the first step.

1. (NP ...) (VP (VBZ/VBD/VBP ...) (NP ...))
   Normal pattern. E. g. "is-a"
2. (NP ...) (VP (NP (NP ...) (PP (IN ...) ...))
   Subordinating pattern. E. g. "is-a-part-of"
3. (NP ...) (VP (VBZ ...) (VP (VPN ...) ...))
   Passive pattern. E. g. "was-born-in"

We can prepare further POS tag patterns to improve the coverage of triples. However, in this research, we applied these three basic patterns to confirm the capability of this direction of research.

In this research, we also extract a relation if the object part does not contain any hyperlinks to other pages. We call it "literal" object. For example, assume that there is a sentence "Brescia is a city" with the following structure tree;

```
(S (NP (NNP [[Brescia]])) (VP (VBZ is) (NP (DT a) (NN city))))
```

The subject part is "a city" but it is not a hyperlink to an article about "city" but it is just a literal. Literal objects are not machine understandable but the literal information is useful depending on the application even if the meaning of the term can not be specified. So we extract the literal information as well.

Co-reference Resolution In a Wikipedia article, usually a number of abbreviations, pronouns and different expressions are used to point an entity, thus co-reference resolution is one of the technical issues in order to make the parsing process accurate. In several previous researches on Wikipedia Mining, co-reference resolution methods optimized for Wikipedia article are proposed [4, 16]. Gang mentioned that emphasized words are likely

Let us assume that there is a Wikipedia article A_t which is describing the topic t (the main subject of the article). A_t is a set of sentences and each sentence a has triple; subject s_a, predicate p_a, and object o_a. Co-reference resolution is a procedure that judges whether s_a is describing about same topic as the main subject t or not. We use three co-reference resolution approaches (included one novel approach) considering following three factors; article title ($C1$), frequent pronouns ($C2$) and synonyms ($C3$).

$C1$ is an approach to detect co-references if the terms used in s_a are all contained in the title of A_t. $C2$ uses pronouns for the judgment. It judges s_a as a co-reference to t if s_a is the most frequently used pronoun in A_t. $C1$ and $C2$ were proposed in previous research [16], but $C3$ is a novel approach proposed by us. The main idea of the approach is to detect co-references if the s_a is a synonym of t. In addition, we investigated the effectiveness of combining these three approaches in detail.

3.3 Parsing Strategies

LSP: Lead Sentence Parsing LSP is a strategy that parses only the lead sentences (first n sentences). After a simple inspection, we realized that a considerable number of Wikipedia articles begin with definitive sentences containing relations (hyperlinks) to other articles (concepts). Especially, the first sentence often defines "is-a" relation to other article. We took detailed statistics (Table 2 Left) from the English Wikipedia (Sept. 2006) to confirm this phenomenon.

First, we removed all redirect pages and category pages from the target of the statistics because these pages are not concept pages but navigational pages. After that, we removed all pages having only few backward links (less than 100) because such pages often contain noisy information and are not structured well [1]. Then, we investigated how many articles begin with a definitive sentence (contain is/are/was/were). The result showed that over 86.3% (P_b/P_a) of all pages begin with a definitive sentence.

We also investigated whether the first sentences have hyperlinks to other pages. The results showed that over 95.7 % (P_c/P_a) of all pages begin with a sentence having hyperlinks to other pages. Further, over 85.5 % ($(P_b \cap P_c)/P_a$) of pages begin with a definitive sentence having hyperlinks.

To conclude this, the statistics unveiled that a large number of pages in Wikipedia has a high potential for extracting "is-a" relations to other concepts thus the first sentence analysis seems a promising approach.

ISP: Important Sentence Parsing ISP detects important sentences in a page if the sentence contains important words/phrases for the page. Our assumption is that the sentences containing important words/phrases are likely to define valuable relations to the main subject of the page, thus we can make the co-reference resolution accurate even if the subject of the sentence is a pronoun or another expression for the main subject. We use $pfibf$ to detect important sentences. By using $pfibf$, a set of important links for each article (concept) in Wikipedia can be extracted. ISP detects important sentences in a page from sentences containing important words/phrases for the page. It crawls all sentences in the article to extract sentences containing links to the associated concepts. The extracted sentences are then parsed as the important sentences in the article. For each links in a sentence, the parser calculates $pfibf$ and the max value denotes the importance of the sentence. The importance can be used for filtering unimportant sentences by specifying thresholds.

For example, when analyzing the article about "Google," associated concepts such as "Search engine", "PageRank" and "Google search" are extracted from the association thesaurus. Therefore, ISP crawls all sentences in the article to extract sentences containing links to the associated concepts.

4 Evaluation

To prove the effectiveness of our proposed method, we conducted two experiments. The first experiment was conducted to measure the co-reference resolution accuracy. The second experiment was conducted to measure the accuracy of the extracted triples. We describe these experiments in detail as follows.

4.1 Experiment 1: Co-reference resolution

In this experiment, we first filtered noisy pages by checking the number of backward links of the articles and extracted 65,391 pages as a test collection. After that, we parsed 2,508 sentences in 52 articles chosen randomly from the test collection. Then, totally 1,002 triples were extracted by parsing patterns described before. A list of term examples used in this experiment is shown as follows; Niagara Falls, Root beer, Deer, Arrow, Odonata, Marie Antoinette, Germany, Colorado, and Blizzard.

Table 3. Evaluation results.

Co-reference resolution.

Methods	Precision	Recall	F-Measure
$C1$	99.22%	59.26%	74.20%
$C2$	65.00%	18.06%	28.26%
$C3$	89.04%	60.19%	71.82%
$C1 \cup C2$	81.78%	81.02%	81.40%
$C1 \cup C3$	89.94%	70.37%	78.96%
$C2 \cup C3$	81.99%	80.09%	81.03%
$C1 \cup C2 \cup C3$	82.33%	81.94%	82.13%

Important sentence selection.

Method	Literal	Extracted Relations	Correct Relations	Precision
ASP	Includes	458	285	62.22 %
	Excludes	162	133	82.09 %
LSP	Includes	101	91	90.09 %
	Excludes	54	52	96.30 %
ISP	Includes	67	54	80.59 %
	Excludes	59	51	86.44 %
$LSP \cup ISP$	Includes	153	130	84.96 %
	Excludes	99	88	88.88 %

We manually checked whether the subject of each sentence is a co-reference of the main subject of the article. Totally 216 subjects of sentences were co-references of the article subject. We used the data set to calculate precision, recall and f-measure. The result is shown in table 3.

As we can see, not surprisingly, $C1$ (article title approach) achieved quite high precision. However, the precision of $C2$ (frequent pronouns approach) was rather low. We investigated the reason and realized that the approach to use frequent pronouns is an error prone strategy. In particular, the pronouns "it" and "he/she" are not used for representing the main subject of an article but for different meanings. We tried all combinations and realized that the combination of all methods achieved the highest f-measure. This means that the combination of these three methods compensates for the weak points of each method, and is therefore helpful to achieve a higher coverage.

4.2 Experiment 2: Triple extraction

In this experiment, we first randomly selected 110 articles and totally 1,016 sentences were extracted as a test set. After that, we applied the proposed method to extract triples. We used LSP and ISP to improve the accuracy of triples. As a baseline, we also parsed all sentences and call it "All Sentence Parsing (Hence ASP)" method. Table 3 shows the result of the experiment.

First of all, we would like to mention that the accuracy of the LSP method is quite high. It achieved high quality relation extraction for both literal objects and non-literal objects. This means that our conviction that the first sentence is useful information is strongly confirmed. We have no strong evidence but we think that this is because of the reliability of the sentences. Usually, the top part of a page attracts much more attention than the bottom part. Thus, the top part is edited by many authors and structured well in most cases. Several parsing misses happened when the sentence is too complicated which was the cause of accuracy loss.

Second, the ISP method also achieved better results than ASP. In particular for literal objects, the accuracy significantly improved. Furthermore, by using the ISP method, we can determine whether a sentence contains important concepts before parsing it, decreasing the analysis time significantly. We also believe that the combination of LSP and ISP is a balanced method because it achieves high coverage and high precision at the same time.

Table 4 shows some examples of explicit relations extracted by LSP. "Explicit relation" means a relation where the object part is a hyperlink to another article. As we can see, the extracted relations are very accurate. As we mentioned

Table 4. Examples of the results.

Extracted explicit relations by LSP samples.

Subject	Predicate	Object
Apple	is-a	Fruit
Bird	is-a	Homeothermic
Cat	is-a	Mammal
Computer	is-a	Machine
Isola d'Asti	is-a	Comune
Jimmy Snuka	is-a	Pro. wrestler
Karwasra	is-a	Gotra
Mineral County	is-a	County
Sharon Stone	is-a	Model
Sharon Stone	is-a	Film producer

Extracted explicit relations by ISP samples.

Subject	Predicate	Object
Odonata	is an order of	Insect
Clarence Thomas	was born in	Pin Point, GA
Dayton, Ohio	is situated	Miami Valley
Germany	is bordered on	Belgium
Germany	is bordered on	Netherlands
Mahatma Gandhi	founded	N. Indian Congress
Mahatma Gandhi	established	Ashram
Rice	has	Leaf
Rice	is cooked by	Boiling
Rice	is cooked by	Steaming

before, almost all articles of Wikipedia begin with a definitive sentence, so LSP extracted mainly "is-a" relations. While is-a relation is one of the most basic (and important) relations in Semantic Web, the result shows the capability of this approach for ontology construction and the possibility for making practical approach to achieve next generation WWW technologies.

Table 4 shows some examples of explicit relations extracted by ISP. Since ISP analyzes important sentences in the article, it extracts various relations such as "was born in," "founded" and "has". However, machines cannot understand the meaning "was born in" without any instruction from humans. So, in order to make the predicate part machine understandable, we have to define the relation between predicates. For example, "is" and "was" have the same meaning but the tense is different. By giving this kind of knowledge, machines can infer semantic relations between two concepts. We believe that the relations among verbs are quite limited compared with relations between nouns, thus do not cause enormous workload.

5 Conclusion

In this paper, we showed that Wikipedia article is yet another invaluable corpus for ontology extraction by showing both detailed statistics and the effectiveness of integrating parsing and link structure mining methods. The experimental results showed that the integration method and co-reference resolution significantly improves the accuracy of triple extraction. Especially, the conviction that lead sentences have rich semantic information is strongly confirmed. Furthermore, important sentence detection by using link structure analysis was helpful to filter inaccurate results.

More than anything else, what we are trying to show in this paper is the possibility and capability of semantic relation extraction using Wikipedia knowledge. We believe that this direction will be an influential approach for Semantic Web in near future since Wikipedia has great capability for constructing a global ontology. The extracted association thesaurus and semantic relations are available on our Web site.

Wikipedia Lab : http://wikipedia-lab.org
Wikipedia Thesaurus : http://wikipedia-lab.org:8080/WikipediaThesaurusV2
Wikipedia Ontology : http://wikipedia-lab.org:8080/WikipediaOntology

We hope the concrete results will be a helpful information to judge the capability of this approach. Our next step is to apply the extracted semantic relations to Semantic Web applications (esp. Semantic Web search). To do that, we need further coverage of relations by enhancing the POS tag analysis patterns and mappings among relations.
Acknowledgment: This research was supported in part of the Microsoft Research IJARC Core Project. We appreciate helpful comments and advices from Prof. Yutaka Matsuo at the University of Tokyo as well as from Prof. Takahiro Hara and Prof. Shojiro Nishio at Osaka University.

References

1. E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis.," in *Proc. of International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pp. 1606–1611, 2007.
2. K. Nakayama, T. Hara, and S. Nishio, "Wikipedia mining for an association web thesaurus construction," in *Proc. of IEEE International Conference on Web Information Systems Engineering (WISE 2007)*, pp. 322–334, 2007.
3. M. Strube and S. Ponzetto, "WikiRelate! Computing semantic relatedness using Wikipedia," in *Proc. of National Conference on Artificial Intelligence (AAAI-06)*, pp. 1419–1424, July 2006.
4. G. Wang, Y. Yu, and H. Zhu, "Pore: Positive-only relation extraction from wikipedia text," in *International Semantic Web Conference, Asian Semantic Web Conference (ISWC/ASWC)*, pp. 580–594, 2007.
5. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, "Dbpedia: A nucleus for a web of open data," in *International Semantic Web Conference, Asian Semantic Web Conference (ISWC/ASWC)*, pp. 722–735, 2007.
6. S. N. Kim and T. Baldwin, "Interpreting semantic relations in noun compounds via verb semantics," in *Proc. of Conference on Applied Computational Linguistics (ACL)*, 2006.
7. F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proc. of International Conference on World Wide Web*, pp. 697–706, 2007.
8. D. N. Milne, O. Medelyan, and I. H. Witten, "Mining domain-specific thesauri from wikipedia: A case study," in *Proc. of ACM International Conference on Web Intelligence (WI)*, pp. 442–448, 2006.
9. M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer, "Semantic wikipedia," in *Proc. of International Conference on World Wide Web (WWW 2006)*, pp. 585–594, 2006.
10. M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proc. of COLING*, pp. 539–545, 1992.
11. M. Berland and E. Charniak, "Finding parts in very large corpora," in *Proc. of Conference on Applied Computational Linguistics (ACL)*, 1999.
12. S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou, "Extracting semantics relationships between wikipedia categories," in *Proc. of Workshop on Semantic Wikis (SemWiki 2006)*, 2006.
13. P. Lawrence, B. Sergey, M. Rajeev, and W. Terry, "The pagerank citation ranking: Bringing order to the web," *Technical Report, Stanford Digital Library Technologies Project*, 1999.
14. J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, no. 5, pp. 604–632, 1999.
15. D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proc. of Meeting of the Association for Computational Linguistics (ACL 2003)*, pp. 423–430, 2003.
16. D. P. T. Nguyen, Y. Matsuo, and M. Ishizuka, "Relation extraction from wikipedia using subtree mining," in *Proc. of National Conference on Artificial Intelligence (AAAI-07)*, pp. 1414–1420, 2007.

Additional information

Workshop:

- E-Mail: sdow2008@easychair.org
- Web: http://sdow2008.semanticweb.org/

Chairs:

Dr. John Breslin

- Department of Electronic Engineering
- National University of Ireland, Galway
- Galway, Ireland
- Phone: +353 91 492622
- Fax: +353 91 494511
- E-Mail: john.breslin@nuigalway.ie
- Web: http://www.johnbreslin.org/

Uldis Bojārs

- Digital Enterprise Research Institute
- National University of Ireland
- Galway, Ireland
- Phone: +353 91 495079
- Fax: +353 91 495541
- E-Mail: uldis.bojars@deri.org
- Web: http://captsolo.net/

Alexandre Passant

- Digital Enterprise Research Institute
- National University of Ireland
- Galway, Ireland
- E-Mail: alexandre.passant@deri.org
- Web: http://apassant.net/

Sergio Fernández

- R&D Department
- Fundación CTIC
- Gijón, Asturias, Spain
- Phone: +34 984 29 12 12
- Fax: +34 984 39 06 12
- E-Mail: sergio.fernandez@fundacionctic.org
- Web: http://www.wikier.org/

Related communities:

FOAF Project

http://www.foaf-project.org/

http://lists.foaf-project.org/mailman/listinfo/foaf-dev

SIOC Project

http://sioc-project.org/

http://groups.google.com/group/sioc-dev

Social Media Research

http://socialmediaresearch.org/

http://groups.google.com/group/social-media-research

Data Portability

http://www.dataportability.org/

http://groups.google.com/group/dataportability-public/

Linked Data

http://linkeddata.org/

http://lists.w3.org/Archives/Public/public-lod/

Semantic Web

http://semanticweb.org/

http://lists.w3.org/Archives/Public/semantic-web/

ISWC 2008

The 7th International Semantic Web Conference

October 26 – 30, 2008

Congress Center, Karlsruhe, Germany