A Prototype to Explore Content and Context on Social Community Sites

Uldis Bojārs
Digital Enterprise Research Institute
National University of Ireland, Galway
Galway, Ireland

Benjamin Heitmann
Digital Enterprise Research Institute
National University of Ireland, Galway
Galway, Ireland

Eyal Oren
Digital Enterprise Research Institute
National University of Ireland, Galway
Galway, Ireland

Abstract: The SIOC Ontology can be used to express information from the online community sites in a machine-readable form using RDF. This rich data structure allows us to easily analyse and extract social relations from these community sites. We use SIOC information to analyse the social relations between users through the content that they create. We introduce metrics for social neighbourhood and social reputation, formally expressed as SPARQL queries over SIOC data. Finally, we demonstrate these algorithms in our Social SIOC Explorer prototype.

1 Introduction

Online community sites (such as blogs, wikis and bulletin boards) are playing an important role in keeping people informed and facilitating communication on the Web. Some of these site are more centralised, others are more decentralised, but from an abstract perspective all such communities play a similar role: they allow users to gather together online, create content and enter into discussions about their topics of interest.

Often, discussions range over several of these communication channels. People try to keep up with these discussions by following web feeds, however, current feed formats only allow to see a single stream of content (posts or comments) and does not provide enough information on the social aspect of these online discussions, e.g., who replies to posts by this author and how often or how to find more information about the author of these posts.

SIOC (Semantically Interlinked Online Communities) [BHBD05] is an RDF vocabulary which allows online community sites to export their data in a semantically rich and interlinked manner. Until now, SIOC browsing interfaces have mainly focused on exploring the content itself or on providing a graph view of this information [BBP06, HO07]. The work presented in this paper looks further: towards the social relations manifested by the content created by users across the online community sites.

1.1 Social context in online communities

Feed readers and search engines for online communities in general focus on the content created in these communities. However, the social context of the message and its author are equally important. Therefore, we would expect that adding the social context to the browsing process would significantly enrich the user experience in selecting information.

Some of these relations may be expressed explicitly, through personal FOAF[1] profiles in RDF or through lists of friends on the social networking sites. However, these explicit relations require constant maintenance and tend to become outdated over time.

Therefore we turn towards object-centered sociality [KC97] and evidence of the connections between people gleaned from the content they create, co-annotate, and reply to. These collaborations uncover the implicit relations between people, but are typically ignored by metadata exporters and feed reader applications.

The SIOC Ontology gives us access to exactly this structure of the content created in the community, and can therefore be used for such object-centered social analysis. By combining all three aspects of the online community: the user-created content, the explicitly defined relations and the social relations derived from the content, we can provide a unified user interface that facilitates both the exploration of the content and the social browsing.

1.2 Outline

To realise these goals, we have started at the first chronological step, by enriching the social community sites with SIOC RDF exporters that automatically create high-quality data; having that data relieves us from screen-scraping and reconstructing the relations between online information. We have then proceeded to consolidate and extract the implicit social relations from that data, to finally build an exploration interface that uses both the community content and the social relations in that community:

1. Produce high-quality data SIOC RDF data from these sites
2. Consolidate distributed SIOC data and extract implicit social relations
3. Build a prototype SIOC explorer for exploring social communities

The paper is structured according to these three steps: we provide a short overview of the SIOC standard in Section 2 and describe the data sources and our integration methodology in Section 3. Section 4 describes the extraction of social context from the integrated community information, and Section 5 introduces the Social SIOC explorer prototype to browse and explore community content. We summarise and conclude in Section 7.

[1] `http://foaf-project.org/`

2 Semantically interlinked online communities (SIOC)

The SIOC initiative aims to ease the integration of online social community information [BHBD05]. SIOC provides on the one hand a unified ontology to describe such online communities, and secondly, several exporters from that translate community information from weblogs, forums, and mailing lists into RDF using the SIOC vocabulary.

The SIOC Core Ontology[2] 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 Figure 1. Users create content Items (e.g., Posts) that reside in Containers (e.g., Posts in Forums) on data Spaces (e.g. Sites).

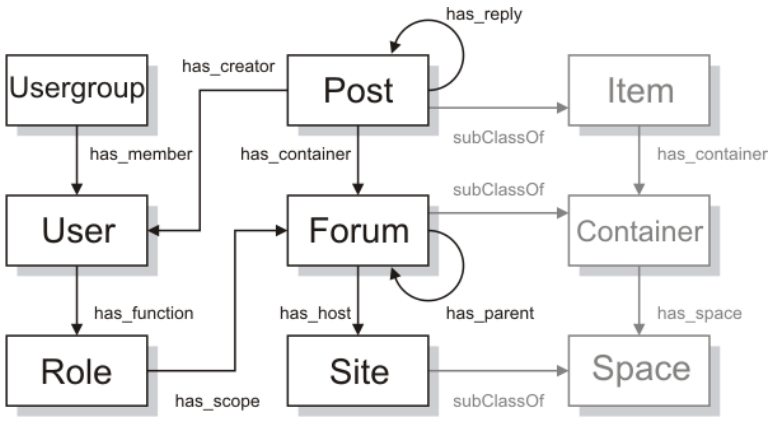

Figure 1: Main SIOC classes and properties

These classes allow to structure the information on online community sites and distinguish between different kinds of objects. Some of the main SIOC properties also play an important role in the context of this paper. `sioc:has_reply` property links replies (e.g., comments) to the original posts while `sioc:has_creator` and `foaf:maker` properties links all the user-created content to more information about its authors. Together these properties form a core set of relations for extracting the social neighbourhood information described in Section 4.

One of the problems with combining social media data is in knowing which accounts users hold on different social media sites. SIOC attempts to solve this by re-using the FOAF (Friend of a Friend) vocabulary which can describe links between a person and accounts it holds in a distributed manner. By combining SIOC with FOAF data we can also re-use the information from personal FOAF profiles, e.g., the `foaf:knows` relationships.

SIOC RDF data export tools[3] have been developed for different types of online community sites and content including *blog engines* such as WordPress, DotClear, b2evolution; *forum*, *bulletin board* and *CMS engines* such as Drupal or PhpBB, *microblogging* tools such as

[2] `http://www.w3.org/Submission/sioc-spec/`
[3] `http://sioc-project.org/exporters`

Twitter and Jaiku, and *mailing lists and IRC conversations*.

These tools make available an RDF representation of all the essential information about the content and users creating it on a Social Media site. This RDF representation contain semantically rich and inter-linked information which can be crawled using a generic RDF crawler by following rdfs:seeAlso links between different data pages. Typically such a crawler would be able to start at site's main SIOC profile page and retrieve RDF about all the publicly available content that a site contains. More advanced tools can be used to incrementally crawl the new information as it appears on a site.

3 Data and methodology

SIOC data were crawled[4] from ten online community sites (namely weblogs) from the list of SIOC-enabled sites[5]. FOAF profiles about the persons, authors of `sioc:Posts`, were retrieved. There was a difficulty retrieving FOAF profiles for people because homepages and blogs often do not have a machine-readable indication of where FOAF profiles can be found. As a result, finding the FOAF profiles was mainly a manual task.

The combined dataset contained $\pm$ 118.000 triples, including 62 `sioc:Users` (users registered on these community sites), 5815 posts and comments (an average of nearly 600 posts per weblog which is reasonable since SIOC exports the full history of a weblog), and in total 3310 `foaf:Persons` (more than the 62 users, since this number also includes all comment authors). The data furthermore contained 1421 unique homepages (because those are usually required for each comment), only 6 unique email addresses (since these are for privacy reasons by default not exported), and 91 hashcodes of email addresses (optionally exported for weblog owners or comment authors that supply email addresses).

The crawled data from disparate online community sites had to be integrated and consolidated. Using RDF and the SIOC vocabulary, the integration part was straightforward: using for example the NTriples RDF serialisation, multiple datasets can simply be concatenated and the duplicate lines can be removed. This was done using the Redland rapper[6] tool for converting the SIOC data from RDF/XML into NTriples, and the Unix tools `cat` and `uniq` for concatenation.

Next, the data needed to be consolidated. We used the inverse functional properties in the dataset to consolidate similar resources with different URIs. However, given the size of the integrated dataset direct OWL reasoning over this dataset was not practically possible. Instead, we separated schema-level reasoning from instance-level reasoning. We first recursively extracted and fetched a list of all used schemas in the dataset (including SIOC, FOAF, WordNet, XML Schema, and many others) and loaded all schemas (totaling $\pm$83.000 triples) into the OWL reasoner, namely the OWLIM extension [KOM05] to Sesame [BKvH02]. We then, after reasoning, extracted a list of all inverse functional properties in these schemas (16 in total), and wrote a manual algorithm to process the

[4] `http://rdfs.org/sioc/applications/#crawling`
[5] `http://esw.w3.org/topic/SIOC/EnabledSites`
[6] `http://librdf.org`

Listing 1: Query for user's direct neighbourhood (L1)

```
SELECT DISTINCT ?relatedPerson
WHERE {
  ?relatedPerson rdf:type foaf:Person .
  ?profileA foaf:knows ?relatedPerson .
  ?profileA foaf:holdsAccount personA . }
```

instance-level triples. The algorithm grouped synonym resources, known to be the same through their inverse functional properties, into a common bucket, and, for each bucket, rewrote all statements to use one canonical URI for all the synonym URIs.

4 Extracting social context

We extract two types of social contextual information from the online community sites. On the one hand, we extract the social neighbourhood of each site member, formed by the set of people that he knows directly or indirectly via online interaction. On the other hand, we extract the social reputation of each member, based on their community involvement, on their activity level and on their connectedness to their peers.

4.1 Extracting social neighbourhood

We define three levels of neighbourhood that can be extracted from the data, either explicitly stated, or implicitly derived from having a small social distance or from co-authoring or co-producing community content. Each neighbourhood is defined as a SPARQL query on our structured SIOC data, leading to a clean and formal definition of these relations and enabling straightforward analysis on actual community data, namely by executing these queries on the instance data. We consider three neighbourhood levels from `personA` to others.

Level 1 consists of the explicitly stated `foaf:knows` relations from `personA` to others. Listing 1 shows this neighbourhood defined as a query, using the direct `foaf:knows` links between people. Since SIOC data uses a `foaf:holdsAccount` property to link a `foaf:Person` and a `sioc:User` account together, the query has to traverse those links. By default, `foaf:knows` relations are not defined to be symmetric; if we want to treat them as such we can union the results with the same query but in the opposite direction.

Level 2 is an implicit neighbourhood relation, meaning that the neighbourhood is not explicitly stated but rather constituted by people who have replied to content created by `personA`. Listing 2 shows this neighbourhood expressed as a query. Note that the query uses a "select distinct"; to rank the neighbourhood by the amount of replies, a count of results returned by an ordinary "select" can be used.

Listing 2: Query for user's indirect neighbourhood (L2)

```
SELECT DISTINCT ?relatedPerson
WHERE {
  ?p rdf:type sioc:Post .
  ?p sioc:has_creator personA .
  ?p sioc:has_reply ?reply .
  ?reply foaf:maker ?relatedPerson . }
```

Listing 3: Query for user's indirect neighbourhood (L3)

```
SELECT DISTINCT ?relatedPerson
WHERE {
  ?p rdf:type sioc:Post .
  ?p sioc:has_reply ?reply1 .
  ?reply1 foaf:maker personA .
  ?p sioc:has_reply ?reply2 .
  ?reply2 foaf:maker ?relatedPerson . }
```

Expanding the neighbourhood further through the usage of shared objects, we define the Level 3 neighbourhood as all people who participated in the same conversations as `personA` (having all replied to the same post), as shown in Listing 3.

The extraction queries presented here should be considered preliminary results. Of course, there are many more types of queries that could be run on the dataset, connecting users through the tags and topics of posts, through a shared geographical region, and more. It should be clear though, that by using SIOC data we are able to very easily extract a particularly defined social relation, without the need for manual data collection, laborous cleansing and integration.

4.2 Extracting social reputation

After extracting the social neighbourhood of a site member, we extract his social reputation. The way in which a user of multiple sites produces and publishes content, and in which his involvement is received by other users in his community are indicators of the social reputation of a site member. Since we cannot measure the reputation directly (social community sites typically do not include "reputation" feedback, as implemented on peer-to-peer trading sites such as eBay.com or Amazon.com), we approximate a user's reputation through his activity level and community involvement.

We associate the activity level (R1) of a user with the number of posts on his own site and with the number of replies that he has written on other sites. A lower number of posts and replies generally indicates a lower activity level. Listing 4 shows the queries used to extract the number of published posts and replies. In the SIOC data, original posts have a `sioc:has_creator` and a `sioc:has_container` property, whereas replies have a `foaf:maker` and are connected to a post through a `sioc:has_reply` property.

Listing 4: Query for user's activity level (R1)

```
SELECT COUNT DISTINCT ( ?publishedPost )
WHERE {
  ?publishedPost rdf:type sioc:Post
  ?publishedPost sioc:has_creator personA .
  ?publishedPost sioc:has_container ?postsContainer . }

SELECT COUNT DISTINCT ( ?publishedReply )
WHERE {
  ?publishedReply rdf:type sioc:Post
  ?publishedReply foaf:maker personA .
  ?anyOtherPost sioc:has_reply ?publishedReply . }
```

Listing 5: Query for user's community involvement (R2)

```
SELECT COUNT DISTINCT ( ?replyingPerson )
WHERE {
  ?originalPost sioc:has_creator personA .
  ?originalPost sioc:has_reply ?reply .
  ?reply foaf:maker ?replyingPerson . }
```

The community involvement of a user (R2) is associated with the number of people from the community that have replied to his posts, where a higher number indicates relevance and visitors' involvement, as shown in Listing 5. This measure corresponds to the *indegree prestige* method in social network analysis [WF94, Sec. 5.3.2] applied to a graph formed between users via `sioc:has_reply` relationships.

SPARQL currently does not provide aggregate functions, therefore Listings 4 and 5 for illustration purposes use a pseudo-operator COUNT which we evaluate by counting a number of rows returned.

5 The Social SIOC Explorer prototype

Our prototype Social SIOC Explorer operates on the aggregated data described in section 3, extracts the social context as described in section 4 and allows users to browse and explore all disparate information in an integrated manner. The prototype can be used as a "social reader" to explore and subscribe to SIOC-enabled community sites such as weblogs, mailing lists, forums and IRC chats and includes the extracted social context of community information in the user interface.

Overview All SIOC content is integrated into a local RDF store and then displayed in various ways. The start page shows an overview of community sites and people. Users can decide to browse a particular forum, see the posts aggregated from all sites, or explore the profile of a particular user. Note that the posts here are not just weblogs posts but include posts from forums, IRC chats, mailing lists, etc. which are all described using the same SIOC RDF vocabulary.

Figure 2: An example post and aggregated replies

After selecting a particular forum, the user is presented with a list of posts in this forum in a reverse chronological order. As usual in feed readers, each post is summarised and can be opened to read the full content. An example of detailed view is shown in Figure 2, which includes metadata about the post, its complete text and all replies (which might have been posted outside of this particular weblog). Also, "lateral" social browsing is supported: clicking on an author or commenter jumps to this person's profile which includes all posts and replies written by this person across all forums; clicking on a topic shows all posts tagged with this topic, again across all forums. In contrast to ordinary feed readers, our lateral browsing works across all types of community forums: clicking on the user "Cloud" will not only show all his weblog posts, but also his contributions to mailing lists, IRC discussions, bulletin boards, etc.

Figure 3 shows an example of a person's description, including information from his FOAF profile such as his picture, homepage and interests, and also his extracted social context. The screenshot shows a summary of these relations, with more details and links to actual people in the social neighbourhood available by clicking the links. For this user, we see that he has written 338 posts and made 115 comments (R1), has received a total of 1454 replies, and that he knows 634 people through a shared discussion (R2) and knows 11 people directly (L1). All this information is extracted from multiple online community sites, e.g. replies and joint discussions that take place on another user's weblog are counted into this picture.

Development The prototype is built on the Ruby on Rails framework for Web application developmen and uses several components for consuming and processing Semantic Web data. One such component is ActiveRDF [ODG+07], which maps RDF data onto

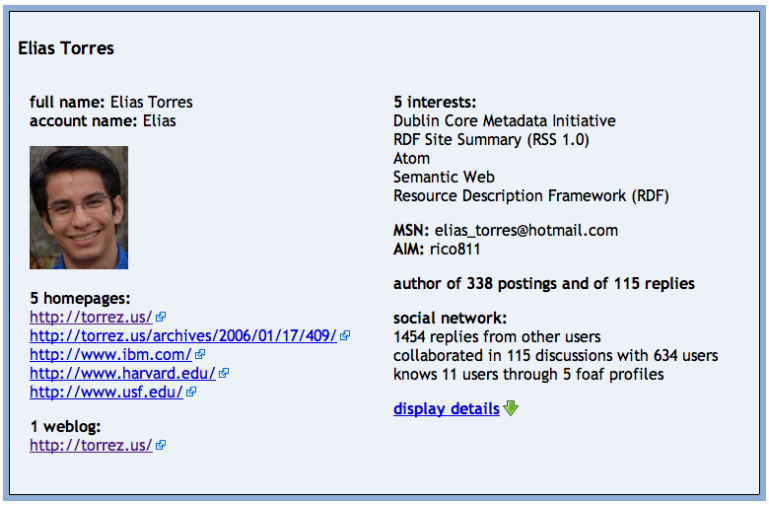

Figure 3: An example user profile with extracted social relations

programmatic objects. The second component is BrowseRDF [ODD06], a faceted browsing engine that enables navigation of large Semantic Web datasets without domain-specific navigation knowledge. The third component, written for a previous SIOC-based prototype, is a SIOC crawler that crawls, extracts, normalises, and integrates SIOC data from various community sites (which use different methods of exposing and linking to their SIOC data). The last component, added specifically for this use case, implements the social analysis algorithms described in section 4; it extracts social context information from the SIOC data and visualises this context in the user interface.

6 Related work

Flink [Mik05] is a web-based application for the extraction, analysis and visualisation of the social networks and research interests of Semantic Web researcher community. It uses electronic sources such as homepages, publications archives, and FOAF profiles as its data sources but does not consider online community sites such as weblogs. In comparison, SIOC provides us with rich and structured information from online community sites, including replies to posts, which enables us to perform this analysis.

Social network analysis methods [WF94] include calculation of metrics such as centrality, prestige, etc. The work presented here focuses on the prototype application for the exploration of online community sites and their networks. We therefore use rather simple and intiutive metrics such as indegree prestige. In future work more complex measures could be used to analyse the characteristics of these social networks, but that would require a larger dataset.

In terms of our navigation interface, related work can be found in other generic RDF

browsers that are also based on faceted navigation, such as Flamenco [YSLH03], mSpace [SWRS06]. Compared to these, our interface is more expressive in terms of navigation functionality [ODD06], but more importantly, these generic RDF browsers do not exploit the social aspect of the data.

Also, several graphical approaches exist for generic visual exploration of RDF graphs [FSvH04, FSvH02, Pie02] but these do not scale well for large graphs in terms of the user interface [FTH06].

7 Conclusion

We have presented an approach to extract social context from online social communities, and a prototype that exploits this information in the browsing process. By using the SIOC ontology we have access to high-quality data with rich structure, which we can directly analyse for implicit social relations. Relations between people can be derived from their online interactions, such as content that they create or reply to. We have introduced three levels of a user's social neighbourhood and two metrics of a user's social reputation, and have defined these as queries on the SIOC ontology. Finally, we have presented our prototype for browsing content from community sites based on these implicit social relations.

Acknowledgements This material is based upon works supported by the Science Foundation Ireland under Grants No. SFI/02/CE1/I131 and SFI/04/BR/CS0694. We gratefully acknowledge John Breslin's work on SIOC and feedback on this paper.

References

[BBP06] Uldis Bojārs, John G. Breslin, and Alexandre Passant. SIOC Browser – Towards a Richer Blog Browsing Experience. In Thomas N. Burg and Jan Schmidt, editors, *BlogTalks Reloaded: Social Software - Research & Cases*. Books on Demand GmbH, 2006.

[BHBD05] John G. Breslin, Andreas Harth, Uldis Bojārs, and Stefan Decker. Towards Semantically-Interlinked Online Communities. In *Proceedings of the European Semantic Web Conference (ESWC)*, 2005.

[BKvH02] Jeen Broekstra, Arjohn Kampman, and Frank van Harmelen. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 54–68, 2002.

[FSvH02] Christiaan Fluit, Marta Sabou, and Frank van Harmelen. Ontology-based Information Visualization. In *Visualizing the Semantic Web*, pages 36–48. Springer-Verlag, 2002.

[FSvH04] Christiaan Fluit, Marta Sabou, and Frank van Harmelen. Supporting User Tasks through Visualisation of Light-weight Ontologies. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, pages 415–434. Springer-Verlag, 2004.

[FTH06] Flavius Frasincar, Alexandru Telea, and Geert-Jan Houben. Adapting graph visualization techniques for the visualization of RDF data. In V. Geroimenko and C. Chen, editors, *Visualizing the Semantic Web*, chapter 9, pages 154–171. Springer-Verlag, second edition, 2006.

[HO07] Benjamin Heitmann and Eyal Oren. Leveraging existing Web frameworks for a SIOC explorer to browse online social communities. In *Proceedings of the ESWC Workshop on Scripting for the Semantic Web*, June 2007.

[KC97] Karin D. Knorr-Cetina. Sociality with Objects: Social Relations in Postsocial Knowledge Societies. *Theory, Culture and Society*, 14(4):1–30, 1997.

[KOM05] Atanas Kiryakov, Damyan Ognyanov, and Dimitar Manov. OWLIM – a Pragmatic Semantic Repository for OWL. In *Proceedings of the Conference on Web Information Systems Engineering (WISE) Workshops*, pages 182–192, 2005.

[Mik05] Peter Mika. Flink: Semantic Web technology for the extraction and analysis of social networks. *Journal of Web Semantics*, 3(2-3):211–223, 2005.

[ODD06] Eyal Oren, Renaud Delbru, and Stefan Decker. Extending faceted navigation for RDF data. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 559–572, November 2006.

[ODG+07] Eyal Oren, Renaud Delbru, Sebastian Gerke, Armin Haller, and Stefan Decker. ActiveRDF: Object-Oriented Semantic Web Programming. In *Proceedings of the International World-Wide Web Conference*, pages 817–823, May 2007.

[Pie02] Emmanuel Pietriga. *Environnements et Langages de Programmation Visuels pour le Traitement de Documents Structurés*. PhD thesis, Institut National Polytechnique de Grenoble, 2002.

[SWRS06] M. C. Schraefel, Max Wilson, Alistair Russell, and Daniel A. Smith. mSpace: Improving Information Access to Multimedia Domains with MultiModal Exploratory Search. *Communications of the ACM*, 49(4):47–49, 2006.

[WF94] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994.

[YSLH03] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. Faceted metadata for image search and browsing. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 401–408, 2003.