SIOC: an approach to connect web-based communities

John G. Breslin*, Stefan Decker, Andreas Harth and Uldis Bojars

Digital Enterprise Research Institute
National University of Ireland, Galway
University Road, Galway, Ireland
Fax: +353 91 512541
E-mail: john.breslin@deri.org
E-mail: stefan.decker@deri.org
E-mail: andreas.harth@deri.org
E-mail: uldis.bojars @deri.org
*Corresponding author

Abstract: Online communities are islands of people and topics that are not interlinked. Complementary discussions exist on disparate systems but it is currently difficult to exploit the available distributed information. A Semantically Interlinked Online Community (SIOC) can enable efficient information dissemination across communities by creating an ontology that will model concepts identified in discussion methods. Data instances can be accessed from community sites using this ontology, enabling connections between local and remote concept instances, and allowing queries on, or transfer of, the data. By searching on one forum, the ontology and interface will allow users to find information on other forums that use a SIOC-based system architecture. Other uses include cross-site querying, topic-related searches, and the importing of SIOC data into other systems. Fusing information and inferring links among various applications and types of information with SIOC provide relevant insights that make the community information available on the internet more valuable.

Keywords: weblogs; semantic web; ontologies; forums.

Reference to this paper should be made as follows: Breslin, J.G., Decker, S., Harth, A. and Bojars, U. (2006) 'SIOC: an approach to connect web-based communities', *Int. J. Web Based Communities*, Vol. 2, No. 2, pp.133–142.

Biographical notes: John G. Breslin received his PhD from the National University of Ireland, Galway (NUI Galway). He is a Postdoctoral Researcher at the Digital Enterprise Research Institute (DERI), NUI Galway. His research interests include social networks and online communities.

Stefan Decker received his PhD from the University of Karlsruhe, Germany. He is a Research Fellow, Adjunct Lecturer at DERI, NUI Galway. His research interests include the semantic web and P2P technologies.

Andreas Harth is currently studying for his PhD at DERI, NUI Galway. His research interests include RDF storage and querying.

Copyright © 2006 Inderscience Enterprises Ltd.

Uldis Bojars is also studying at DERI, NUI Galway. His research interests include semantic matching of skills and community discussions.

1 Introduction

Computer-supported collaboration and discussion systems for closed and open domains are in widespread use on intranets and the internet. The closed community collaboration model usually has a limited and controlled audience where restricted information access and workflow management are the main requirements (for example, commercial groupware products, such as Lotus Notes for businesses, or open source CSCW (Grudin, 1994) products, such as NetOfRce for researchers). The open community collaboration model facilitates information exchange with emphasis on open involvement, participation, circulation of information and feedback (examples include public bulletin boards or archived mailing lists, Usenet newsgroups, social networks, weblogs and wikis).

Online communities (McArthur and Bruza, 2001) using open collaboration systems have the potential to replace the traditional means of keeping a community informed via libraries and publishing (Millen, 2000). These sites allow improved communication and interactive contact within a community, by providing an online collaboration space for members to find and contribute certain interest-related or regional information (Wellmann and Gulia, 1999).

However, it is difficult to exploit the available information in such community sites on the internet, especially when most online communities are hosted on stand-alone sites that cannot be interconnected due to application and interface differences. Also, each community site will normally have a unique entry point to its own discussions. Parallel discussions on interrelated topics may exist on a number of sites that are not linked. There is a huge amount of related information that could be harnessed across such online communities, from similar member profile details to common-topic discussion forums.

This paper addresses how to maximise the usage of this potentially valuable information and how to enable the location of relevant information in online communities. The research question is to identify and model the concepts found in online discussion methods, and to create a data infrastructure among different community sites. This will aid in a reduction of the information overload from existing search engines, and will use semantic web technologies to make the information useable by applications.

SIOC faces the following interesting challenges:

- The grand challenge is adoption by community sites, *i.e.*, how a critical mass can be reached by enticing users to make use of the SIOC ontology. By using concepts that can be easily understood by site administrators, and by providing properties that are automatically created by an end user, the SIOC ontology can be adopted in a useful way.
- A second challenge is how best to use SIOC with existing ontologies and collaboration technologies. This challenge can be partially solved by mappings and interfaces to commonly used ontologies, such as DC[1] FOAF[2] and RSS,[3] and by wrappers to technologies such as NNTP, SMTP and SQL databases.

SIOC also has to deal with the addition and removal of topics. An evolving category hierarchy is essential for any living online community, but this presents challenges with respect to the matching of recently created or deprecated topics.

Another challenge is how well SIOC will scale. If there are more sites to query, then there are more potential relevant results, but also longer response times and higher load on the participating community sites. We must keep this scaling challenge in mind when creating the architecture of an interconnected system of the community sites.

2 Why is this problem significant?

Some of the main problems in relation to existing online community technologies are:

- Information on community sites cannot be harnessed correctly by search engines (Bruza *et al*., 2000) limited to syntactic matching, *e.g*., by keyword matching on bulletin boards.
- Many isolated communities that discuss complementary topics exist.
- Information is being repeatedly requested across separate sites, and people are wasting time searching for relevant community information by waiting for answers that have already been posted elsewhere.

Linking individual posts with others is possible on the HTML level, but a forum search will not represent this link. Using the example in Figure 1, a user is searching for information on installing broadband on a Linux-based PC in their house in Galway. There is a post A discussing local ISPs on site 1, a bulletin board dedicated to Galway, that references (on the HTML level) both a Usenet post B comparing broadband modems and a mailing list post C detailing how to install broadband on Linux. Previously, the user would have had to traverse three sites to find the relevant information. However, by making use of the SIOC ontology and remote RDF querying, a search for broadband on the Galway bulletin board will also yield the relevant text from the interlinked Usenet and mailing list posts B and C.

In reality, explicit references between posts do not often exist, and therefore similarities among community sites could be quantified (for example in Figure 2, an explicit linking of sites could be weighted as 0.1, forums as 0.01, sites as 0.001, and the weighted combination is then used to determine sites with a ranking greater than 0.1 for a cross-community search).

Once there exist enough sites that have richer query facilities to instances of SIOC data, then these different sites can be interlinked. If a user has an account at site 1, then site 2 could pull that user information from site 1 and that user would not need to maintain his/her own accounts database. Other benefits include: applying reasoning facilities (such as those provided by OWL-S time (Pan and Hobbs, 2004)) to make use of events descriptions; visualisation of scheduling information of individuals or groups of people; and representations of where people related to a certain topic are located geographically.

Figure 1 Retrieving related posts through SIOC

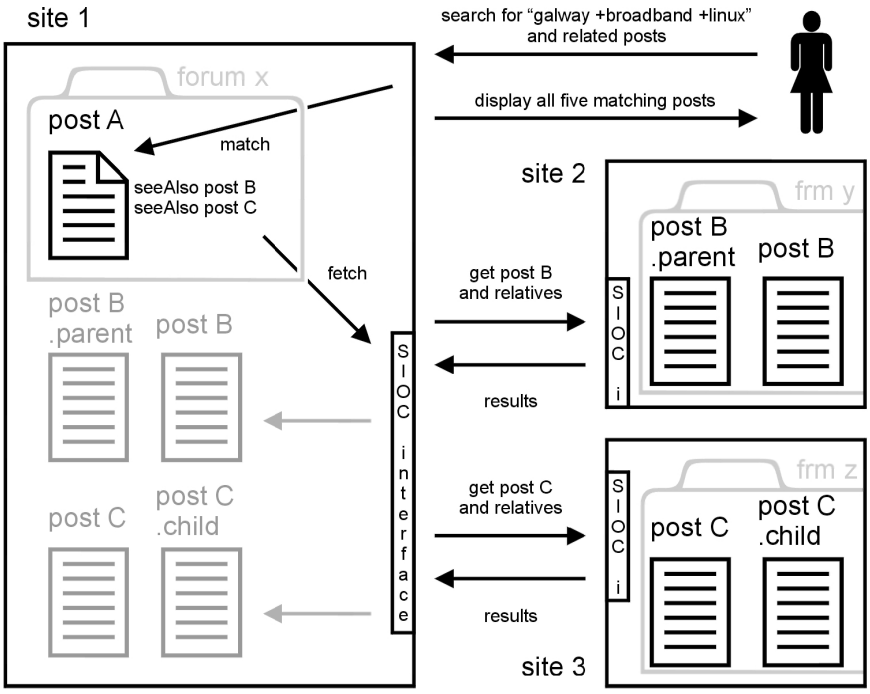
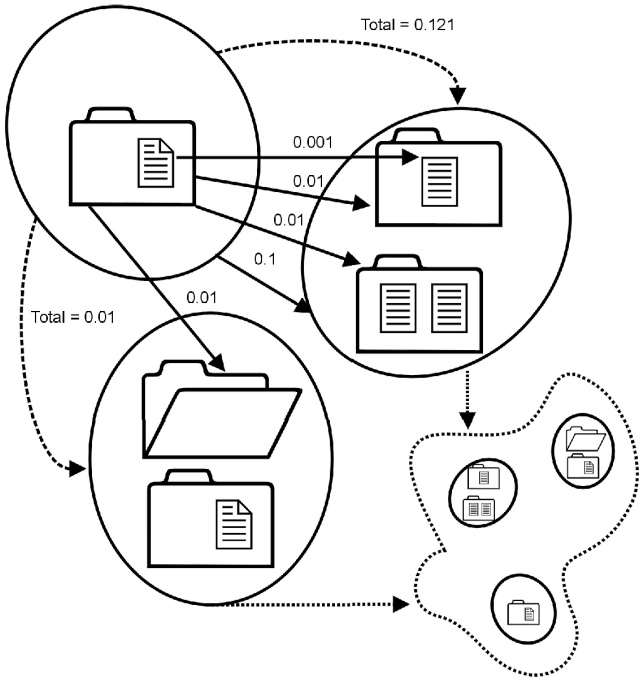


Figure 2 Inferring implicit inter-site connections



Interlinking community concepts into a coherent representation enables more sophisticated applications and therefore will result in more efficient information dissemination in communities. SIOC effectively performs the functions of a broker between different communities, thus, adding to the informational diversity of these communities. Research has shown that much important knowledge is produced on the border areas between communities (Burt, 2003). Hence, important collaboration and knowledge generation may be enabled by interconnecting diverse communities.

3 How is this question being addressed?

The research approach for SIOC involves two phases, along with evaluation:

1 The creation of an ontology by identifying and modelling the concepts in community sites, *e.g.*, people, groups, posts, forums, locations, events, categories, resources, *etc*. Semantic web ontologies can be used to enrich community sites and to make the underlying information available to both humans and software agents. The SIOC ontology identifies the main community concepts and enables the linking of these concepts by modelling the relationships between them.

2 The building of a data infrastructure to query and exchange instances of the ontologised concepts, and to identify the topics related to these data instances. This infrastructure will allow sites to make messages, events, user profiles, *etc*. available in a common format for cross-community querying and storage. The data format is machine readable, allowing automatic detection of connections among people and related forums on separate sites.

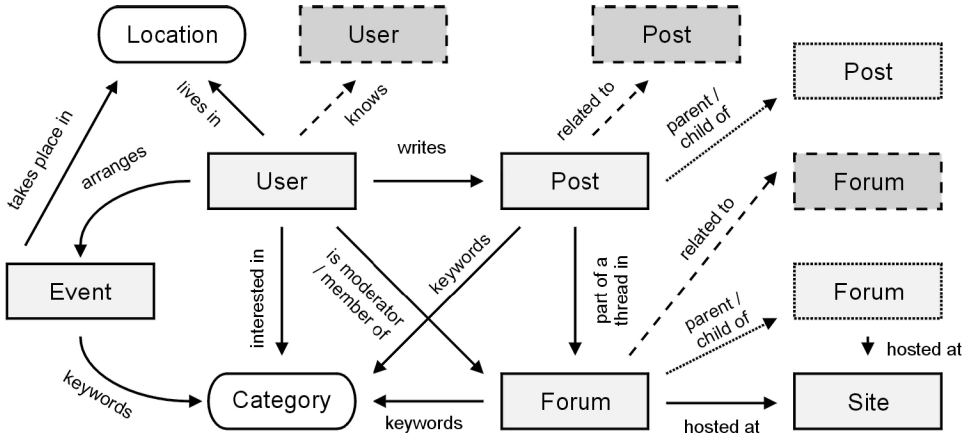
3.1 Ontology development

3.1.1 Construction and refinement

An analysis has been performed on existing discussion systems to identify the concepts and properties to be modelled in the SIOC ontology. Preliminary work on a first version of the ontology has been published (Harth *et al*., 2004), which details current community building primitives and an initial research approach. The ontology namespace is available online.[4]

In SIOC, new concepts are being added to the existing set (Harth *et al*., 2004) to allow integrated access to terms on various discussion systems, *e.g.*, users with roles such as moderator or subscriber; posts with associated content such as polls, events, announcements, attachments; user groups, which may be private or public, *etc*. Some unique properties of these concepts have also been defined, *e.g.*, on bulletin board systems, forums will have parent containers, and these may have various associated topic categories (see Figure 3). Other properties of the concepts (*e.g.*, topic, creator) can be used to create implicit links among concept instances.

Figure 3 Interlinking instances of classes in the SIOC ontology



3.1.2 Topic hierarchies

For the classes such as post, user, forum, *etc*., a topic definition will be used to match documents and people to each other. While it may be more difficult to require a user to assign a topic to a post at creation time, it is more likely that a forum will have an associated topic or set of topics that can be propagated to the posts it contains. Similarly, users or groups can define topics of interest when their profiles are created or modified. A proper use of topics can lead to many interesting scenarios in community sites. For example, a user has defined certain topics of interest on registering an account, after which forums matching those topics are suggested to the user.

In order to enable the location of related information among community sites, a common categorisation system has to be used. On a large scale, *e.g.*, in general interest community site, topics can be quite broad and a general categorisation, such as a DMOZ[5] category hierarchy may be used. On specialised sites, which may have a very specific category hierarchy, generic categorisation systems are not suitable because they are too broad and may not have the necessary level of detail. For these sites, we propose to define a category hierarchy in the SKOS[6] framework, and create mappings among these concepts and a common category system.

3.1.3 Mappings

There is the issue of whether one should link and reuse existing ontologies, or use mappings to an entirely new ontology and therefore require more intelligent applications. One of the main functions of SIOC is to provide a means for exchanging instances of data. We have therefore provided mappings in RDFS and OWL to allow the importing and exporting of SIOC instance data in different vocabularies. This provides more flexibility and we can also leverage any instance data that is already available.

We have provided different kinds of mappings[7] in RDFS for import and export using rdfs:subClassOf and rdfs:subPropertyOf, and also mappings in OWL using owhequivalentProperty and owhequivalentClass together with other OWL constructs. Vocabularies to be mapped to SIOC include FOAF, RSS 1.0, Atom[8] and various e-mail RDF schemas.[9–10]

3.2 Data infrastructure

3.2.1 Data storage and transfer

SIOC enables the linking of community sites in a machine-interpretable format. Automatic connections between community sites can be made in various ways: *e.g.*, we can infer a connection if the same person posts to different sites, or if one forum is explicitly linked to a forum at another site. By also allowing users to manually create links among different concept instances, the result is a useful network of community sites.

In the context of SIOC, it is appropriate to extend our notion of community sites. If we say sites, we mean all sorts of community tools from forums to weblogs. All sites should implement a common interface that allows access to the underlying data in a semi-structured format via the SIOC ontology. For that matter, it is not relevant whether the data comes from a site in a network of community sites, from a node in a P2P network, or from a web service.

For a site to access data from other sites, there are two options that have been investigated: warehousing of data one level or two levels away from a community into a local database (*e.g.*, using YARS[11]), or a query interface to a site's data by extending the HTTP GET access interface to allow it to perform queries on RDF data (Harth, 2004).

For use with SIOC, we propose a database architecture known as virtual integration. Data is fetched on demand when a query arrives. All necessary transformations (schema mapping, *etc*.) are carried out in this process. The query is translated, sent to all sources, and the resulting RDF is translated back into the caller's ontology.

3.2.2 Topic extraction and identification

In the ontology properties, we mentioned a topic hierarchy that can be associated with the discussion areas on a community site, and these topics can be propagated to the posts contained therein. Even if a common categorisation system is used, there may be ambiguities for different reasons: accidental (incorrect or no classification), deliberate (advertising by categorising data in many unrelated categories) or conflict-of-interest (differing opinions in a single discussion).

If the category information defined in the topic property of a certain concept is incorrect (perhaps due to a non-cooperative site) or missing, we need to allow for both manual and automatic classification of that concept by other people and systems. For example, a moderator or post creator may manually change an incorrect topic for a post, or an automatic classifier may be built to detect and assign a missing topic to the post.

Alternatively, topic information may be completely absent and cannot be deduced from the post content or containers using automatic classification. In this case, if a search is being performed for related content, then sites, forums and user groups could be rated relevant to the posts host site, container forum or poster user group, and suggestions could be made based on these ratings or inferred connections among concept instances.

As new topics are created, these can be assigned to existing concept instances by mapping the new topic to similar existing topics. Similarly, as topics are deprecated, concept instances with no associated topic must be mapped to a replacement topic.

3.3 Evaluating results

Results are being evaluated throughout the SIOC project. A number of data collections are available for use-case testing: discussion forums from boards, *i.e.*, a community of webloggers at irishblogs.org, publicly available mailing lists and newsgroups. In these use cases, some information may be removed to evaluate the performance of SIOC. For example, two posts on different sites may be similar through related topics and connected users, and one of these properties is removed before testing. Or someone makes links between communities, and these are removed before testing to see how good or bad SIOC is at inferring these connections. Here are some typical use cases that will be used to evaluate the success of the SIOC project:

- A new post is created at a community site. The community site engine then creates links to the relevant and related information on this and other community sites. SIOC is being used to locate and identify the related information.
- A user enters keywords to search for in a query box. The query can be for a single community site (as already exists), for a number of sites (by querying sister-sites) or the overall data available to an aggregator or RDF data store.
- A user chooses a set of metadata to search for, for example, one or more concepts, people or other metadata that have been added to posts and, hence can be searched for.
- A user wishes to import information from other data sources, and can therefore obtain the required information in a medium of choice.

4 What is the value to online communities?

SIOC can connect people and discussions from a myriad of online communities: the Usenet community; bulletin board communities (*e.g.*, general discussion forums and specialised activism groups using software such as phpBB and vBul-letin); third-level students and researchers (through systems such as Drupal and uPortal respectively); *etc*.

Furthermore, the development of the SIOC infrastucture will create a hub of interconnected communities worldwide. Distributing SIOC under an open source licence will aid in its rapid dissemination and adoption by community sites.

For example, uPortal is a Java-based community portal system used by institutes of higher education that will serve as a target system for SIOC. Enabling SIOC functionality for uPortal will allow interconnections among uPortals used by colleges worldwide. Drupal, a content management system, and phpBB, a bulletin board system, are PHP-based software programmes that have been adopted by many student and non-profit groups due to their ease of installation and modification. SIOC functionality is being added to these[12–13] and other popular discussion systems (WordPress,[14] vBulletin, *etc*.).

SIOC will create many commercial opportunities to make use of the wealth of information available in online communities. For example, there is the potential for development of a search engine that will answer a user's questions rather than provide links to information that might potentially be correct based on syntactic matching. Targetted online advertising can also make use of the topic category information

associated with users and posts, as provided by SIOC. As a common format for storage and exchange of community information, SIOC will generate impact as it is adopted for both open source and commercial applications. An eventual aim is to develop SIOC into a recognised standard.

5 Conclusions

This paper addresses the issue of online communities being islands that are not interlinked, where complementary discussions exist on disparate systems and where it is difficult to exploit the available combined information. SIOC, or Semantically Interlinked Online Community, has been created to enable efficient information dissemination across communities. An ontology has been developed to model concepts identified in discussion methods. A data interface will be provided to community sites using this ontology, enabling connections between local and remote concept instances. Data is exported from open source discussion systems such as Drupal and WordPress to facilitate future use cases. SIOC is a prerequisite for a search engine that will answer questions rather than provide links to possibly relevant information.

Acknowledgement

The authors would like to acknowledge the support of Science Foundation Ireland which funded this project. This paper has evolved from work originally published at the 2nd IADIS International Conference on Web-Based Communities (Breslin *et al*., 2005).

References

Breslin, J.G., Decker, S. and Harth, A. (2005) 'An approach to connect web-based communities', *The 2nd IADIS International Conference on Web Based Communities (WBC 2005)*, Carvoeiro, Portugal, pp.272–275.

Bruza, P.D., McArthur, R. and Dennis, S. (2000) 'Interactive internet search: keyword, directory and query reformulation mechanisms compared', *The 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* http://citeseer.ist.psu.edu/bruza00interactive.html.

Burt, R.S. (2003) 'Social origins of good ideas', *Workshop Manuscript Draft*, http://web.mit.edu/sorensen/www/SOGI.pdf.

Grudin, J. (1994) 'Computer supported cooperative work: its history and participation', *IEEE Computer,* Vol. 27, No. 5, pp.19–26.

Harth, A. (2004) 'SECO: mediation services for Semantic Web data', *IEEE Intelligent Systems, Special Issue on Semantic Web Challenge.*

Harth, A., Breslin, J.G., O'Murchu, I. and Decker, S. (2004) 'Linking semantically-enabled online community sites', *The 1st Workshop on Friend of a Friend, Social Networking and the (Semantic) Web (FOAF Galway)*, Galway, Ireland, pp.19–29.

McArthur, R. and Bruza, P. (2001) 'The ABCs of online community', *1st Asia Pacific Conference on Web Intelligence, Lecture Notes in AI*, Pittsburgh, PA: Springer-Verlag, http://citeseer.ist.psu.edu/mcarthur01abcs.html.

Millen, D.R. (2000) 'Community portals and collective goods: conversation archives as an information resource', *The 33rd Annual Hawaii International Conference on System, Sciences,* IEEE.

Pan, F. and Hobbs, J.R. (2004) 'Time in OWL-S', *1st Semantic Web Services Symposium.*

Wellman, B. and Gulia, M. (1999) 'Virtual communities as communities: net surfers dont ride alone', *Communities in Cyberspace,* pp.167–194.

Bibliography

de Pater, F. (2001) 'Enterprise information portals', *IMSE, Vrije Universiteit Amsterdam*, Amsterdam, The Netherlands: Deloitte and Touche.

Granovetter, M. (1973) 'The strength of weak ties', *American Journal of Sociology,* Vol. 78, pp.1360–1380.

Grinter, R.E. and Palen, L. (2002) 'Instant messaging in teen life', *CSCW '02,* New Orleans, LA, http://citeseer.ist.psu.edu/grinter02instant.html.

Lara, R., Han, S.H., Lausen, H., Stollberg, M., Ding, Y. and Fensel, D. (2004) 'An evaluation of Semantic Web portals', *International Conference in Applied Computing (IADIS04)*, Lisbon, Portugal.

Lee, L.Y., Liu, C.C., Hsu, C.C. and Chen, G.D. (2003) 'Using database technologies in building a learning community to improve knowledge exchange', *The 3rd IEEE International Conference on Advanced Learning Technologies,* IEEE.

Neumann, M., MacDonaill, C., Hogan, D., Reinsford, C. and Decker, S. (2004) 'A university online portal for enterprise learning communities', *UFHRD/AHRD 5th International Conference on HRD Research and Practice Across Europe*, University of Limerick.

van Dijk, T.A. and Kintsch, W. (1983) *Strategies of Discourse Comprehension*, New York: Academic Press.

Wilson, P. (1991) *Computer Supported Cooperative Work: An Introduction*, Kluwer Academic.

Notes

1 http://dublincore.org/schemas/

2 http://xmlns.com/foaf/0.1/

3 http://web.resource.org/rss/1.0/spec

4 http://rdfs.org/sioc/ns

5 http://www. dmoz.org/

6 http://www.w3.org/2004/02/skos/core/

7 http://rdfs.org/sioc/mappings

8 http://www.atomenabled.org/developers/syndication/atom-format-spec.php

9 http://www.w3.org/2000/04/maillog2rdf/email/

10 http://web.resource.org/rss/1.0/modules/email/

11 http://sw.deri.org/2004/06/yars/yars.html

12 http://rdfs.org/sioc/drupal

13 http://rdfs.org/sioc/phpbb

14 http://rdfs.org/sioc/wordpress