Project no. 34721

TAGora

Semiotic Dynamics in Online Social Communities

http://www.tagora-project.eu

Sixth Framework Programme (FP6)

Future and Emerging Technologies of the Information Society Technologies (IST-FET Priority)

D4.5: Deployment of a semantic recommender (OTHER)

Period covered: from 01/06/2008 to 31/08/2009	Date of preparation: 15/09/2009
Start date of project: June 1st, 2006	Duration: 39 months
Due date of deliverable: October 15th, 2009	Actual submission date: September 15th, 2009
Distribution: Public	Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: Sapienza Università di Roma
Lead contractor for this deliverable: UNI KO-LD

Contents

Chapter 1

Semantic recommenders

Semantic recommendation includes different aspects which help to improve the user experience. Hence, this deliverable is not a single semantic recommender but three implementations tailored for the different recommendation strategies.

1.1 Bibsonomy

Recommender type: Tag Recommender

Deployed as: Recommender Framework in BibSonomy

Data set: Content of BibSonomy

The recommender framework allows the integration and evaluation of different (semantic or 'non-semantic') recommender systems into BibSonomy. These recommender systems can be either installed locally or remotely (connected and queried via http), thus allowing other research teams to integrate their recommender systems and giving a broad base for evaluation. All incoming events and informations are logged for evaluation in a SQL database. The framework is described in more detail in Deliverable 2.5.

The design of the framework has been performed within Tagora – its implementation was not possible with the Tagora budget any more. However, the framework could be implemented still within the life time of Tagora (financed by a national follow-up project). The recommendation framework is deployed in the ECML PKDD Discovery Challenge. See `http://www.kde.cs.uni-kassel.de/ws/dc09/online` for details.

1.2 MyTag

Recommender type: Tag Meaning Recommender

Deployed as: Recommender in MyTag

Data set: Wikipedia and Tag assignments

MyTag[1] is cross-folksonomy search portal that enables tag based searching across 5 popular tagging systems: Delicious[2], Flickr[3], YouTube[4], Connotea[5], and Bibsonomy[6]. This places MyTag

[1] http://mytag.uni-koblenz.de

[2] http://delicious.com

[3] http://www.flickr.com

[4] http://www.youtube.com

[5] http://www.connotea.org

[6] http://bibsonomy.org

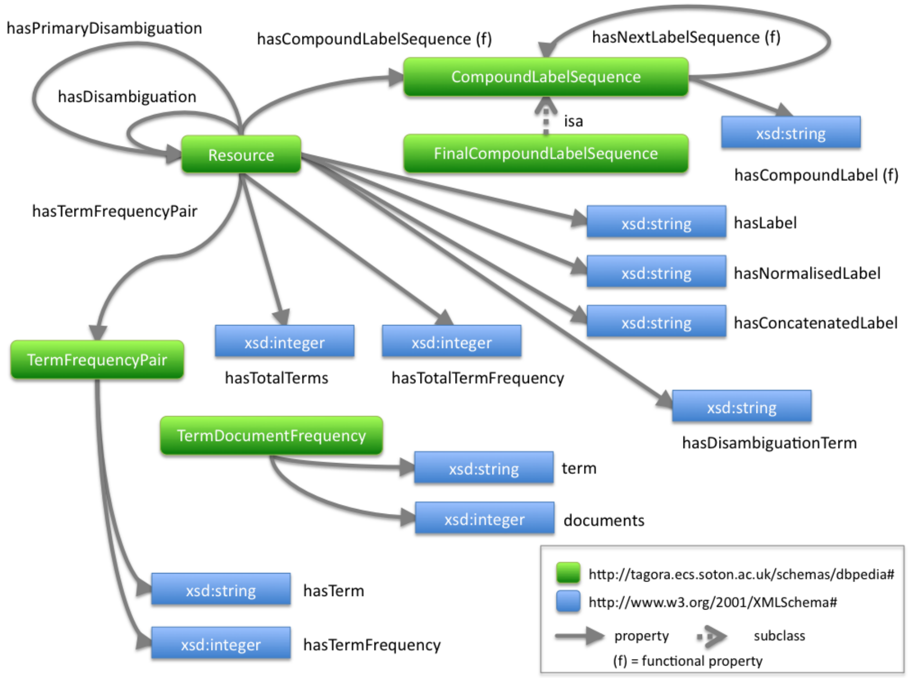


Figure 1.1: DBPedia ontology used to expose metadata about wikipedia resources

in a unique position to provide searching over various multimedia media types, such as videos, pictures, web pages, and scientific articles (Braun et al., 2008). In the process of developing this application, and other research on the semantics of tags in collaborative tagging systems (Szomszor et al., 2008a,b), we have found that many popular tags have multiple, ambiguous meanings. For example, the tag `apple` is often used in the Delicious bookmarking system to refer to the computer company, but in Flickr, pictures of the fruit are often tagged with apple. Similarly so for terms such as `windows` (the operating system and the building feature) and `leopard` (the Mac OS X operating system and the animal).

In this Section, we present the results of a collaboration between UNI-SOTON and UNI-KOBLENZ to enrich the MyTag portal by suggesting possible *senses* for a search tag and subsequently re-ranking the results according to the specified meaning. The results of this collaboration will be presented during the demo/poster session of ISWC 2009 (Dellschaft et al., 2009).

1.2.1 Recommending Senses

The TAGora Sense Repository[7] (TSR) is a linked data enabled service endpoint that provides extensive metadata about tags and their possible senses. When the TSR is queried with a particular tag string, by forming a URI that contains the tag in a REST style (e.g. `http://tagora.ecs.soton.ac.uk/tag/apple/rdf`), the tag is processed, grounded to a set of DBPedia.org resources, and an RDF document is returned containing the results.

Creating The Sense Index The first stage in building the TSR was to process the XML dump of all Wikipedia pages to index all titles, mine redirection and disambiguation links, and extract term frequencies for each of the pages. For the current version we use a dump available from `http://download.wikimedia.org`, created on the 08/10/2008. For each Wikipedia page in the dump, we extract and index the page title, a lower case version of the title, and a concatenated version of

[7] `http://tagora.ecs.soton.ac.uk/`

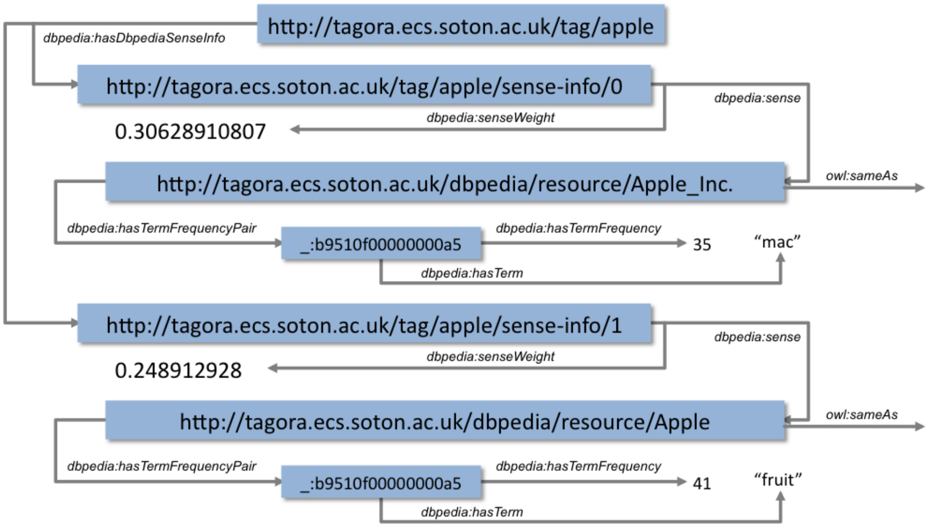
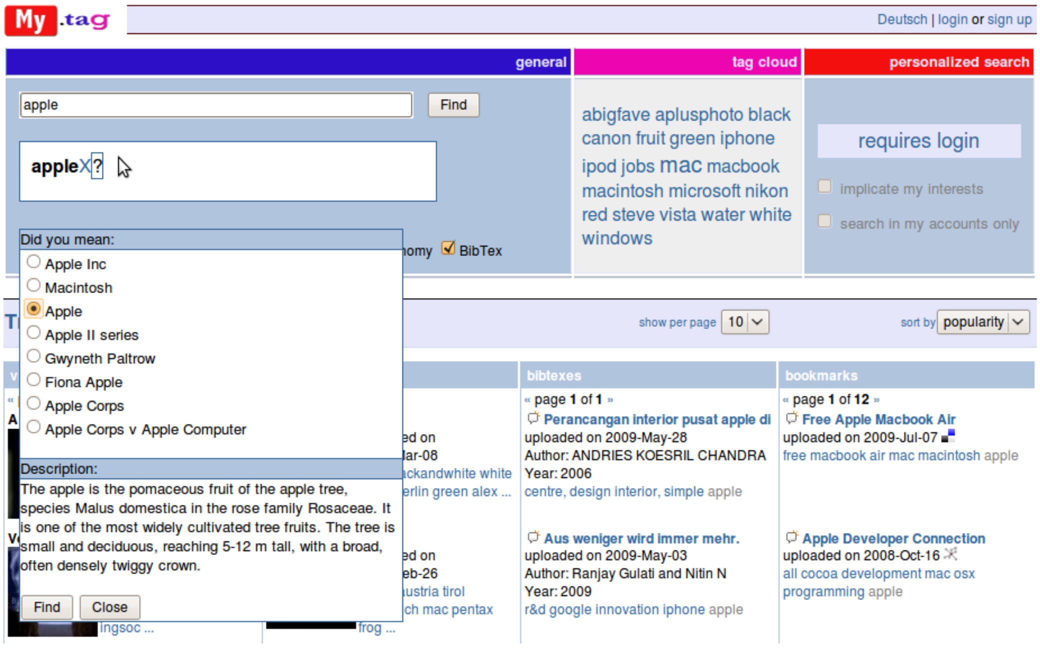

http://tagora.ecs.soton.ac.uk/tag/apple
dbpedia:hasDbpediaSenseInfo
http://tagora.ecs.soton.ac.uk/tag/apple/sense-info/0
dbpedia:sense
dbpedia:senseWeight
0.30628910807
http://tagora.ecs.soton.ac.uk/dbpedia/resource/Apple_Inc.
owl:sameAs
dbpedia:hasTermFrequencyPair
_:b9510f00000000a5
dbpedia:hasTermFrequency
35
"mac"
dbpedia:hasTerm
http://tagora.ecs.soton.ac.uk/tag/apple/sense-info/1
dbpedia:sense
dbpedia:senseWeight
0.248912928
http://tagora.ecs.soton.ac.uk/dbpedia/resource/Apple
owl:sameAs
dbpedia:hasTermFrequencyPair
_:b9510f00000000a5
dbpedia:hasTermFrequency
41
"fruit"
dbpedia:hasTerm

Figure 1.2: Linked data representation of tag sense information

the title (i.e. the title Second_life becomes secondlife). This style of multiple title indexing enables us to match more easily tags that are made up of compound terms. We also extract redirection links, disambiguation links, as well as the terms contained in the page and their frequencies. During this indexing process, we also store a list (and total) of all incoming links to each page as well as a term-document total for the purposes of TF-IDF analysis. Since the dump is large, we only store the top 20 most frequent terms in a document This data is stored in a Triple Store using our own extended DBPedia ontology since we are providing more detailed metadata about the entries than DBPedia.org, such as the term frequencies. Each Wikipedia page in the TSR is also linked to DBPedia via the owl:sameAs property. Figure 1.1 shows the ontology we use to expose term frequency metadata about DBPedia resources.

Searching For Senses When the TSR is queried with a tag, the first step is to find a list of candidate DBPedia resources that represent possible senses of the tag. We begin by normalizing the tag string i.e. removing non-alphanumeric characters as described in (Szomszor et al., 2008a). The Triple Store is then queried for all entries with the same lowercase title or concatenated title as the tag. During this process, we are likely to encounter redirection links and /or disambiguation links, both of which are followed. When a set of candidate senses has been created, we calculate the total number of incoming links for each resource (including the sum of incoming links for any pages that redirect to it). Finally, a weight is associated with each possible sense as the fraction of incoming links associated with that sense / the total number of incoming links for all senses associated with the tag. This basic page rank inspired measure means senses that have very specific meanings receive much lower weights than general those associated with general concepts. Figure 1.2 provides a visual example of the linked data associated with the tag `apple` - a common tag that could refer the computer company (Apple_Inc.), or the of fruit (Apple). In this example, the URI for apple (center, top) is linked to a number of sense-info instances (only two of which are shown here) via the dbpedia:hasdbpediaSenseInfo. Each sense-info pair gives the weight (0.306 for Apple_Inc. and 0.249 for Apple) and corresponding DBPpedia resource. Each resource is linked to a set of blank nodes (of type termFrequencyPair) that states the frequencies of terms within the Wikipedia page of that resource.


My.tag
Deutsch | login or sign up
general
tag cloud
personalized search
Find
abigfave aplusphoto black canon fruit green iphone ipod jobs mac macbook macintosh microsoft nikon red steve vista water white windows
requires login
implicate my interests
search in my accounts only
BibTex
show per page
sort by
bibtexes
bookmarks
Find
Close

apple

appleX?

Did you mean:

○ Apple Inc
○ Macintosh
● Apple
○ Apple II series
○ Gwyneth Paltrow
○ Fiona Apple
○ Apple Corps
○ Apple Corps v Apple Computer

Description:

The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family Rosaceae. It is one of the most widely cultivated tree fruits. The tree is small and deciduous, reaching 5-12 m tall, with a broad, often densely twiggy crown.

show per page 10

sort by popularity

bibtexes

« page 1 of 1 »

Perancangan interior pusat apple di
uploaded on 2009-May-28
Author: ANDRIES KOESRIL CHANDRA
Year: 2006
centre, design interior, simple apple

Aus weniger wird immer mehr.
uploaded on 2009-May-03
Author: Ranjay Gulati and Nitin N
Year: 2009
r&d google innovation iphone apple

bookmarks

« page 1 of 12 »

Free Apple Macbook Air
uploaded on 2009-Jul-07
free macbook air mac macintosh apple

Apple Developer Connection
uploaded on 2008-Oct-16
all cocoa development mac osx programming apple

Figure 1.3: MyTag sense searching dialogue

1.2.2 MyTag Interface and Ranking

MyTag analyzes the current search terms of the users and sends it to the disambiguation web service which returns possible related terms. MyTag filters the meanings which are not represented in the search results retrieved from the tagging platforms. The remaining terms are then presented to the user along with a short description (see Fig. 1.3). The user can then select the intended meaning of the search term and re-rank the current list of results so that resources corresponding to the intended meaning are ranked higher. A more detailed description of the single steps are described below.

Query TAGora Sense Repository As a first step, the TAGora Sense Repository is queried for possible senses of the current search terms. Because the repository is implemented using a REST API, this corresponds to retrieving the content from a specific URL. As the exchange format between the Sense Repository and MyTag we use JSON[8]. For example, if we want to have the possible senses for the search term *apple* we would open the URL `http://tagora.ecs.soton.ac.uk/tag/apple/json`. For each resource URI, which is a DBPedia link and corresponds to a possible sense, the returned JSON document contains the following information:

1. The **weight** of the sense which is computed using the number of incoming links (see Subsection 1.2.1 for more details). It corresponds to the relevance of this sense.

2. The **abstract** consists of the first three sentences of the corresponding DBPedia article. This information is shown to the user in the *Description* box of the user interface if he selects the corresponding sense.

3. The list of **term frequencies** gives terms related to the current sense and their frequencies which represent their importance with regard to this sense. The list of term frequencies is used for ranking the documents that were retrieved from the tagging systems (see below).

[8] http://www.json.org/

Removing Irrelevant Senses MyTag only retrieves a certain number of documents from each of the integrated tagging systems. The actual number for each tagging system is dependent on the restrictions of the corresponding API (e. g. for Flickr one can retrieve a maximum of 500 photos with a single call to the API). Usually, not all senses that were retrieved from the Tagora Sense Repository are contained in the set of documents from the tagging systems. Thus, in this step all senses are removed from the list which are not contained in the set of documents anyway. This helps to significantly reduce the number of possible senses that are shown by the user interface.

For removing irrelevant senses, we compare the term frequencies of each sense with the tag cloud of the current search results. The tag cloud contains all tags and how often they are assigned to the documents in the current search results. We remove all senses which do not have at least one tag and/or term in common between their *term frequencies* vector and the *tag frequencies* vector of the search results.

For example, the Tagora Sense Repository returns *Gwyneth Paltrow*[9] as a possible sense of apple because the first name of her daughter is *Apple*. Related terms in the Sense Repository are for example *daughter*, *actress* or *paltrow*. But because none of these related terms is contained in MyTag's tag cloud for this search, this sense is discarded and not shown to the user (see Fig. 1.3).

Ranking Search Results When the user selected one of the offered senses and clicked on the *Search* button in the search interface (see Fig. 1.3), MyTag calculates a new ranking for the documents in the result sets retrieved from the different tagging systems. For this purpose, we reuse the ranking algorithm that is also used for providing a personalized ranking of search results (see (Braun et al., 2008)). It rank *r* of a document is computed by the scalar product of the term frequencies vector retrieved from the Tagora Sense Repository and the vector that contains the tags assigned to the resource *r*. Before computing the scalar product, the term frequencies vector from the Sense Repository and the tag vector of the document are normalized so that both vectors are of length 1. The ranking value *r* is then used for ordering the documents.

1.2.3 Future Work

For the future it is planned to do an evaluation of the Tagora Sense Repository and the subsequent steps performed by MyTag. During the evaluation, we will cover the following questions:

1. Is disambiguation an urgent need of the user during searching in tagging systems?
   (a) For how many tags do the tagging systems return documents which are related to different senses of the tag?
   (b) Which strategies do the users apply for coping with ambiguous tags?
2. How often provides the Tagora Sense Repository an appropriate sense that can be used for disambiguation?
3. Is the algorithm able to rank relevant resources higher in the list of results?
4. Do the users understand how to use the disambiguation algorithm?
   (a) How is their subjective impression of the usefulness of the offered senses?
   (b) Are they able to quickly decide which of the senses they can use for their purpose?
   (c) How is their subjective impression of the quality of the re-ranked results?

[9] http://dbpedia.org/resource/Gwyneth_Paltrow

1.3 Live Social Semantics

Recommender type: Interest Recommender

Deployed as: Demo application at ESWC 2009

Data set: Web 2.0 data

Social interactions are one of the key factors to the success of conferences and similar community gatherings. In this Section, we describe Live Social Semantics (LSS), a novel application that integrates data from the semantic web, online social networks, and a real-world contact sensing platform provided by the SocioPatterns.org project. This application was successfully deployed at the European Semantic Web Conference in Crete (ESWC09), and the Hypertext conference in Turin, Italy (HT2009). Personal Profiles of Interests of the participants were automatically generated using several Web 2.0 sources, and integrated in real-time with face-to-face contact networks derived from wearable sensors. Integration of all these heterogeneous data layers made it possible to offer various services to conference attendees to enhance their social experience such as visualisation of contact data, and a site to explore and connect with other participants. This Section describes broadly the Profile Building and interest recommendation architecture of the application, the services we provided, and the results we achieved in this deployment. For further information on the experiment and a more detailed explanation of the architecture, please refer to (Alani et al., 2009).

1.3.1 Recommendation of User Interests

In previous work (Szomszor et al., 2008a), we devised an architecture to automatically generate a list of DBpedia URIs to represent interests a person might have by reasoning over their social tagging activity. Under the assumption that the tags used most often by an individual correspond to the topics, places, events and people they are interested in, we sought to provide a novel dimension to the social interaction at the conference by providing people with a basis to expose their interests, both professional and personal, and see those of others at the conference. Central to this idea is that these profiles can be built automatically, only requiring a short verification phase from the user.

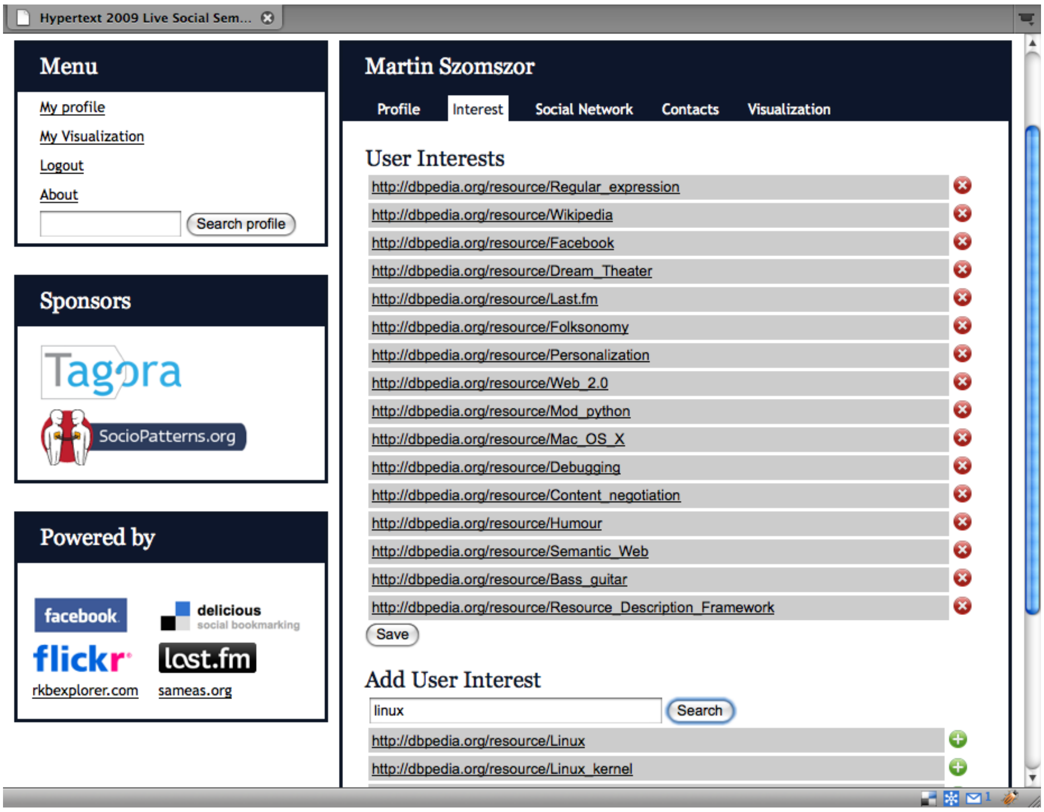
With the LSS website, users were able to associate their various social networking site (SNS) accounts with their conference profile. In the current version, we support Delicious, Flickr, Last.fm, and Facebook. Once a user has registered their SNS accounts, any social tagging information from Delicious and Flickr is collected and converted to an RDF representation according to the TAGora tagging ontology[10]). For each of the user's tags, we use the TAGora Sense Repository (TSR) to lookup possible meanings of the tag, providing a mapping between tags and DBPedia URIs. Using the profile building algorithm described below, we were able to suggest to users a list of possible interests that they may want to expose to other conference participants. Figure 1.4 show a screen shot from the website where a user profile has been suggested and edited. We also provided a simple search interface (using the TSR) to allow users to add other interests.

1.3.2 Profile Building Algorithm

To build a Profile of Interests (POI), we first check to see if the user has a LastFM account. Using DBTune, a linked data site providing metadata about music, we can map the MusicBrainz[11] ID associated to their top artists in LastFM to a resource in DBpedia. The top 5 artists with a DBpedia mapping are added to the user's POI. The second phase of the profile generation procedure is to

[10] `http://tagora.ecs.soton.ac.uk/schemas/tagging`

[11] `http://musicbrainz.org/`



User Interests

http://dbpedia.org/resource/Regular_expression
http://dbpedia.org/resource/Wikipedia
http://dbpedia.org/resource/Facebook
http://dbpedia.org/resource/Dream_Theater
http://dbpedia.org/resource/Last.fm
http://dbpedia.org/resource/Folksonomy
http://dbpedia.org/resource/Personalization
http://dbpedia.org/resource/Web_2.0
http://dbpedia.org/resource/Mod_python
http://dbpedia.org/resource/Mac_OS_X
http://dbpedia.org/resource/Debugging
http://dbpedia.org/resource/Content_negotiation
http://dbpedia.org/resource/Humour
http://dbpedia.org/resource/Semantic_Web
http://dbpedia.org/resource/Bass_guitar
http://dbpedia.org/resource/Resource_Description_Framework

Add User Interest

linux

http://dbpedia.org/resource/Linux
http://dbpedia.org/resource/Linux_kernel

Figure 1.4: A screen shot of the interests editing page from the the Live Social Semantics website running at Hypertext 2009

map the user's tags to DBpedia resources that represent their topics of interest. This is achieved with the following steps:

1. **Disambiguate Tags** When tags are associated to multiple senses (i.e. more than 1 DBpedia resource), we compare the similarity (using a cosine measure) of the user's cooccurrence vector for that tag (i.e. all other tags that occur in the same post, and their frequencies) against the term frequencies associated with the possible DBpedia senses. If one of the similarity scores is above a threshold value, (0.3 in this case), we conclude that this is the correct sense for that tag. If more than one (or zero) senses score above the threshold, we do not associate a meaning to the tag. By iterating through all tags associated to a user (i.e. through Delicious or Flickr), we are able to build a candidate resource list C.

2. **Calculate Interest Weights** For each DPpedia resource $r \in C$, we calculate a weight $w = f_r * u_r$, where f_r is the total frequency of all tags disambiguated to sense r, and u_r is a a time decay factor. This factor $u_r = \lceil days(r)/90 \rceil$. Hence tags used within the last 3 months are given their total frequency, tags used between 3 and 6 months ago are given $1/2$ their frequency, 6 - 9 months a third, etc.

3. **Create Interest List** If more than 50 candidate resources have been found, we rank them by weight and suggest the top 50. Since users are required to edit and verify this list, we believe it important to keep the number of suggestions to a reasonable amount.

1.3.3 Results

The LSS website allows users to declare their accounts on Delicious, Flickr, lastFM, and Facebook. Table 1.1 shows how many social networking accounts were entered into our system by the 139 registered participants at ESWC2009. Table 1.2 shows that about 35% of registered users did not declare any social networking accounts (49 users). It also shows that over 61% of the 139 users had more than one social networking account.

Account	Facebook	Delicious	lastFM	Flickr	**Total**
Quantity	78	59	57	52	**246**

Table 1.1: Number of social networking accounts entered by users into the ESWC2009

Number of Social Networking Accounts	0	1	2	3	4	**Total**
Number of Users	49	36	28	13	13	**139**

Table 1.2: Number of users who entered 0,1,2,3 or 4 social networking accounts.

Out of the 90 people who entered at least one social networking account (Table 1.2), 59 of them entered at least one account from Delicious, Flickr, or lastFM (remaining 31 only entered Facebook accounts, which we do not use when generating POIs). Although our profile building framework had the potential to utilise all three of these accounts, the linked data site DBTune was offline for the duration of the conference, and hence, we were unable to associate a user's favourite lastFM artists to a DBPedia concept. 41 individuals viewed and saved their POI, of which 31 had a non-empty profile generated. Empty profiles were generated for a number of users who registered that had a very small or empty tag-cloud. Table 1.3 summarises the results in terms of the number of concepts automatically generated, the number that were removed manually by users, the number that were added manually, and the size of the final profile they saved.

A total of 1210 DBPedia concepts were proposed (an average of 39 per person across the 31 non-empty profiles), out of which 247 were deleted. While it would be useful to know exactly why users deleted a concept, whether it be simply inaccurate (i.e. incorrect disambiguation), it didn't reflect an actual interest (i.e. a very general concept), or it was something they wished to keep private, we considered it too much of a burden to ask users this question when editing their profiles. The total number of concepts deleted was 20% of those suggested. Although a facility was included on the website for users to add their own interests, few did - only 19 new concepts were added. When comparing the results from Delicious and Flickr, we see that 17% of concepts proposed from Delicious Tags were deleted, and 32% respectively for Flickr tags. This suggests that the accuracy of topics harvested from Delicious tags was more accurate than those from Flickr. Inspection of the concepts removed shows that Flickr was likely to suggest concepts referring to years and names.
Table 1.4 shows the top 10 most common interests out of all participants that saved their Profiles of Interest.

1.3.4 Discussion and Future Work

The deployment of LSS at ESWC2009 was the first where all components were put together and a good number of participants got to use it. There are many social networking sites, but we

	Global	Delicious	Flickr
Concepts Generated	1210	922	288
Concepts Removed	247	156	91
Concepts Added	19		
Concepts Saved	982	766	197

Table 1.3: Statistics of the profile generation, editing, and saving.

Number of Participants	Interest
17	Semantics
16	Tutorial
16	Ontology
15	Research
14	Computer_software
13	Computer_programming
13	Application_programming_interface
12	Science
11	Travel
10	Reference

Table 1.4: The most common interests registered by users of the ESWC2009 LSS website.

only supported four currently popular ones. We are working on a open plug-in architecture that allows external parties to develop the functionality needed to connect to, and crawl data from, other networking systems. We also plan to let users submit their FOAF files.

The number of available social networking sites on the web is always on the increase, and the popularity of such sites is never constant. In our application, only four of such networking systems were taken into account. Although the ones we selected are currently amongst the most popular ones, several users wished to add other accounts, such as FOAF files, LinkedIn, and Twitter. One approach to increase extendibility and increase coverage is to use an open architecture to allow external parties to develop and plug applications and services to connect to, and crawl data from, other networking systems, or sources such as FOAF files.

Extractions of POIs has so far been limited to users' online tagging activities. However, many of the participants have authored papers which can be used to determine their research interests, and some of these interests are already available on `semanticweb.org` in the form of paper keywords. Acquiring such interests can be added to the system and used to improve recommendations on talks or sessions to attend, or people to meet. Also, information from social networking accounts can be used to avoid recommending existing friends. We furthermore believe it will be advantageous to organise the interests URIs into hierarchies, to support inference and fuzzy matching.

Bibliography

Harith Alani, Martin Szomszor, Gianluca Correndo, Ciro Cattuto, Alain Barrat, and Wouter Van den Broeck. Live Social Semantics. In *Proceedings of the International Semantic Web Conference (ISWC)*, Westfields Conference Center near Washington, DC, 2009.

Max Braun, Klaas Dellschaft, Thomas Franz, Dominik Hering, Peter Jungen, Hagen Metzler, Eugen Müller, Alexander Rostilov, and Carsten Saathoff. Personalized Search and Exploration with MyTag. In *Proceedings of the WWW 2008 Poster Session*, 2008.

Klaas Dellschaft, Olaf Görlitz, and Martin Szomszor. Sense Aware Searching and Exploration with MyTag. In *Proceedings of ISWC09 Poster and Demo Session*, 2009.

Martin Szomszor, Harith Alani, Ivan Cantador, Kieron O'Hara, and Nigel Shadbolt. Semantic Modelling of User Interests based on Cross-Folksonomy Analysis. In *submitted to Int. Semantic Web Conf., Karlsruhe, Germany*, 2008a.

Martin Szomszor, Ivan Cantador, and Harith Alani. Correlating User Profiles from Multiple Folksonomies. In *Proc. Int. Conf. Hypertext (HT08), Pittsburgh, PA, USA*, 2008b.