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)

D3.5 Protocols for linking cross-folksonomy networks

Period covered: from 01/06/2007 to 31/05/2008	Date of preparation: 31/05/2008
Start date of project: June 1st, 2006	Duration: 36 months
Due date of deliverable: July 15th, 2008	Actual submission date: June 20th, 2008
Distribution: Public	Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: "La Sapienza" Università di Roma
Lead contractor for this deliverable: University of Southampton

2008 © Copyright lies with the respective authors and their institutions

Contents

List of Figures

Chapter 1

Introduction

While the future evolution of the Web is the subject of much speculation, recent trends suggest an increasing importance of social networking sites. Much like the dot-com surge in the late 1990s opened new opportunities to businesses through the proliferation of e-commerce, social networking is revolutionising the internet by empowering users with the means to share ideas, opinions and resources. Personal sharing and communication have reached unprecedented levels as users become more comfortable with the idea of sharing information with friends, both from the real and virtual world.

In recent years, increased connectivity and new collaborative software tools, such as Wikis and Folksonomies, have provided a suitable medium through which Web2.0 developers can engage a wide range of audiences using rich multimedia experiences. This, along with novel social networking features, has spurred an internet revolution in which users are increasing the amounts of information they expose on the web using a variety of online profiles or identities. A recent UK study by Ofcom (Ofcom, 2008) found that over one fifth of UK adults have at least one online community profile (54% for individuals aged 16-24). Silver (Silver, 2007) predicts that by 2010, each of us will have between 12 and 24 online identities.

In many cases, a number of different social networking sites have emerged to meet a particular use-case. For example, the social bookmarking sites del.icio.us[1], CiteULike[2], Connotea[3], and Bibsonomy[4] have been developed to support different communities in the tagging and sharing of resources spanning multiple domains of discourse. For photo sharing, Flickr[5], and Photobucket[6] have proved to be very popular, attracting around 26 million and 50 million users respectively from all over the world. In recent months (May 2008), many of the popular vendors, such as Google, MySpace and Facebook, have proclaimed an interest in linking this vast array of information spread over the web, striving to provide users with a sense of freedom of information that will allow them to share and manage data between their different online profiles. Against this background, we consider two important questions i) what are the benefits of linking such data?, and ii) how can one go about doing it?

If the information published by individuals is linked between different folksonomies, it would facilitate both search and retrieval. Since tagging is a simple and easy to understand organisation mechanism that users relate to readily, linking of resources across different domains by their common tags will enable users to locate resources of different media types, such as web pages, videos, and pictures. Such approaches could also be applied to communities of users, allowing recommender systems to suggest new friends or contacts in one site based on the commonalities of

[1] http://del.icio.us/
[2] http://www.citeulike.org/
[3] http://connotea.org/
[4] http://bibsonomy.org/
[5] http://www.flickr.com
[6] http://photobucket.com/

interests found in another. Furthermore, the nature of these tagging pursuits naturally leads users to expose various aspects of the their personality. For example, the tagging of pictures in Flickr discloses events and locations the user has attended, their bookmarking activities in del.icio.us provides an indication of their topics of interest, and the music they listened to may be recorded in Last.fm. If such data is combined, a complex picture may be constructed of the individuals activities across space and time that could be exploited for recommendation purposes.

The engineering of a solution that would enable this kind of cross-folksonomy integration is a complex task. One major obstacle is the amount of data: del.icio.us has more than 3 million users who have bookmarked in excess of 100 million urls[7]. It is therefore infeasible to consider a monolithic solution where all the data from different sites is stored in a central location and linked up. Instead, one observes the emergence of a set of protocols for integrating different elements of the folksonomies, such as the users or tags, to enable the consolidation of information to meet specific purposes, such as the sharing of friends and contacts between different online profiles.

In this document, we present a review of the current state-of-the-art, outlining the protocols that have emerged to satisfy different integration problems. We also refer to specific work within TAGora project that has contributed to the area of cross-folksonomy integration, as well as the applications that make use of it. This Document is organised as follows: Chapter 2 provides a detailed explanation of how folksonomies can be integrated, the challenges involved in exchanging data between different sites, and how RDF can be used as a tool for the construction of cross-folksonomy networks. Chapter 3 presents two TAGora applications that utilise cross-folksonomy integration techniques, before we conclude and give future work in Chapter 4.

[7] Recorded September 2007

Chapter 2

Cross-Folksonomy Integration

A folksonomy is conventionally described using three finite sets U, T, and R whose elements are called *users*, *tags*, and *resources*. In (Szomszor et al., 2008b), we extended this model to cater for two distinct folksonomy data sources: del.icio.us and Flickr. Under the assumption that a resource is defined by a url (and can therefore exists in either the del.icio.us or Flickr folksonomy as a resource), and that a tag is defined as a string (and can also exist in both folksonomies), we distinguish between the two folksonomies using two *tag assignment* sets: $Y^d \subseteq U \times T \times R$ a ternary relation for del.icio.us tag assignments, and $Y^f \subseteq U \times T \times R$ a ternary relation for Flickr tag assignments. This model assumes that individuals with an account in both del.icio.us and Flickr are represented as a single user in U. Therefore, we define the del.icio.us folksonomy as a tuple $\mathbb{F}^d := (U, T, R, Y^d)$ and the Flick folksonomy as a tuple $\mathbb{F}^f := (U, T, R, Y^f)$. Through this view, we can consider cross-folksonomy integration in terms of their three constituent elements:

1. **User Correlation** Since the same individual may hold accounts in multiple social networking sites, two separate folksonomies may be joined at the user level. Such a joining enables one to analyse and study the tagging practices of individuals across different tagging platforms.

2. **Tag Correlation** It is likely that many tags will appear in more than one folksonomy. As a result, cross-folksonomy networks can be generated by joining the common tags.

3. **Resource Consolidation** Finally, the resources themselves may appear across multiple folksonomies. By understanding how resources in different folksonomies relate to each other, e.g. through shared tags or users, it would be possible to provide a consolidated view of resources distributed over multiple folksonomies.

Figure 2.1 is a graphic representation of this concept, describing how del.icio.us and Flickr can be integrated in terms of users, resources, and tags. We continue in Section 2.1 with a more detailed discussion on each of the integration axis, citing relevant work undertaken as part of the TAGora project. Section 2.2 discusses the issues of data portability and describes current efforts to link folksonomy data by large web-site vendors, such as Google and Facebook. We also introduce the ideas emerging from the research community regarding folksonomy integration and how we intend to utilise them as part of the TAGora project.

2.1 Aligning Folksonomy Elements

As the introduction to this Chapter states, we can consider integrating folksonomies in terms of their three constituent elements: users, tags, and resources. Each of the following Subsections is devoted to one of these parts:

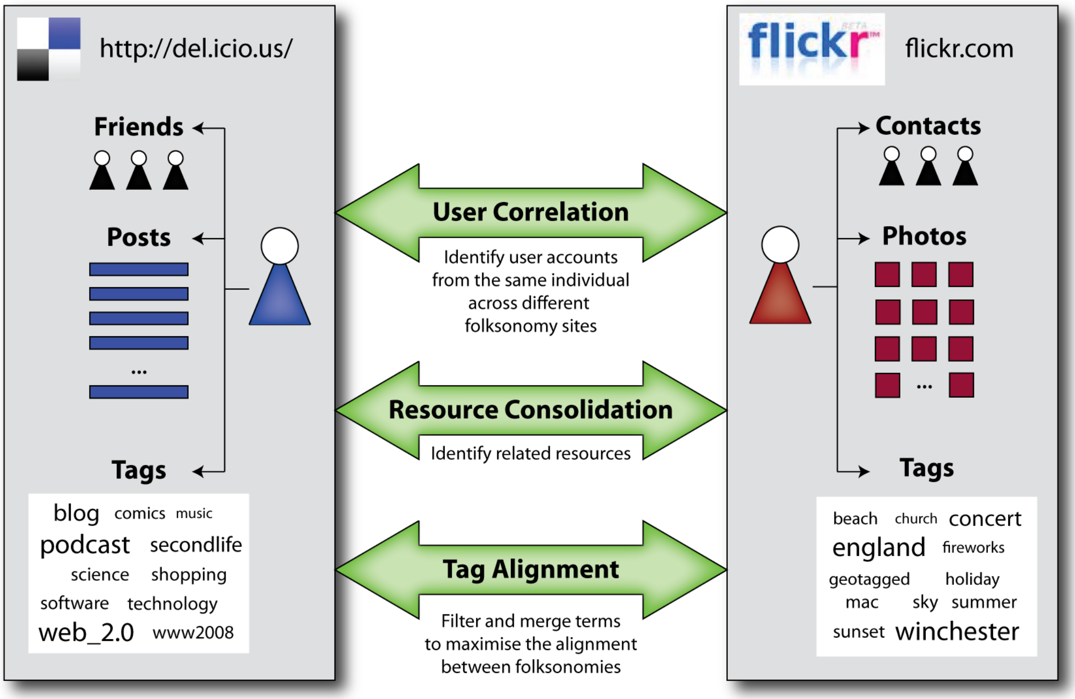


Figure 2.1: Integration of Folksonomies

2.1.1 User Profile Correlation

Many users create multiple profiles across a range of folksonomy sites to meet different social and information requirements. Since many of these sites are provided by different vendors, there are no provisions made to explicitly link accounts that belong to the same individual. In previous work (Szomszor et al., 2008b), we matched 502 user accounts between del.icio.us and Flickr by examining the usernames chosen by individuals. If the same username was found in both systems, and the string given as their real name was identical in both profiles, the accounts were matched. While such an approach is not particularly robust, the accuracy can be increased by matching other profile information such as age, sex, and location.

Through closer examination, it was apparent that many social networking sites supplied users with a field in their profile page to link to another resource that described them, such as a homepage url or blog url: When we examined a number of Last.fm profiles, we found that many individuals linked to their del.icio.us or Flickr profile. This kind of approach is more robust than matching on strings alone since it is unlikely that two accounts that point to the same url are *not* owned by the same individual. Fortunately, Google recently released an implementation of this matching technique as part of their Social Graph API [1] providing a powerful account correlation facility. We adopted this API in (Szomszor et al., 2008a) and created a substantially larger data-set containing 1998 individuals all holding an account in del.icio.us, Flickr, and Last.fm.

2.1.2 Tag Alignment

Initially, one might assume the integration of tags to be a trivial process: It could be assumed the tags that match each other at the symbol level (i.e. when the characters are equal) are referring to the same concept. However, tags are free text, and users can tag resources with any terms they wish to use. On the one hand, this total freedom simplifies the process and thus attracts

[1] http://code.google.com/apis/socialgraph/

users to contribute. It also avoids the problem of forcing users into using terms they do not feel apply, as opposed to enforcing the use of a set of terms (as well as conceiving of such a list). For these reasons, the lack of constraints seems essential. On the other hand, the free-form nature of tagging generates various vocabulary problems: tags may be too personalised to align with others, they may be formed using compound words, they may mix plural and singular terms indiscriminately , etc. (Golder and Huberman, 2006; Guy and Tonkin, 2006; Mathes, 2004). There is also a substantial set of tags that could be used to refer to one of many ambiguous meanings, such as `apple` (referring to the fruit or the technology company), or `sf` (referring to the genre science fiction or the location San Francisco). This total lack of control is resulting in some sort of tagging chaos, thus obstructing search (Guy and Tonkin, 2006) and analysis (Li et al., 2008).

Guy and Tonkin (Guy and Tonkin, 2006) suggest that users should be educated about how to author better tags, and that systems should implement procedures to check for problematic tags and suggest alternatives. Such steps could be useful for improving tag quality. In our work, we follow the approach of *cleaning* existing tags using a number of term filtering processes. In the same spirit of our tag filtering, Hayes and colleagues (Hayes et al., 2007) in their work on tag clustering have performed a number of filtering operations, such as stemming, stop word removal, tokenisation, and removal of highly frequent tags. Clustering of tags has been used by Begelman and colleagues for tag disambiguation (Begelman et al., 2006), where similar tags were grouped together to facilitate distinguishing between their different meaning when searching.

The filtering process we have developed (Cantador et al., 2008; Szomszor et al., 2008a,b) is a sequential execution of different morphologic filtering modules: the output from one filtering step is used as input to the next. The output of the entire filtering process is a set of new tags and their frequencies. Figure 2.2 provides a visual representation of the filtering process where a set of raw tags are transformed into a set of filtered tags and a set of discarded tags. Each of the numbers in the diagram corresponds to a step outlined below:

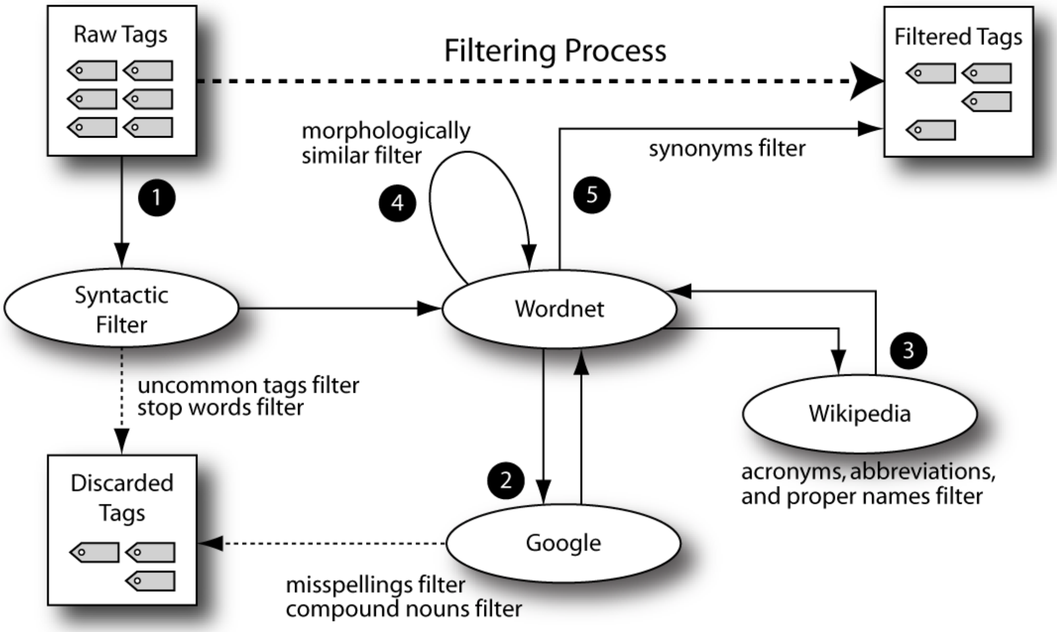


Figure 2.2: The Tag Filtering Process

Step 1: Syntactic Filtering

After the raw tags have been loaded, they are passed to the *Syntactic Filter*. First, tags that are too small (with length $= 1$) or too large (length > 25) are removed. Due to discrepancies regarding the use of *special* characters (such as accents, dieresis and the caret symbol), special characters are all converted to their base form. For example, the tag *Zürich* is converted to *Zurich*.

Tags containing numbers are also filtered according to a set of custom heuristics: To maintain salient numbers, such as dates (`2006`, `2007`, etc), common references (`911`, `360`, `666`, etc), or combinations of alphanumeric characters (`7up`, `4x4`, `35mm`), we consider the global tag frequency and discard any unpopular tags. Finally, common stop-words, such as pronouns, articles, prepositions, and conjunctions are discarded. After syntactic filtering, tags are verified against Wordnet (Fellbaum, 1998. p.423). If the tag has an exact match in Wordnet, we pass it directly to the set of filtered tags to avoid unnecessary processing.

Step 2: Compound Nouns and Misspellings

If the tags were not found in Wordnet, we consider possible misspellings and compound nouns. It is common for users to misspell tags, for example, the use of `barclona` instead of `barcelona`. To solve this problem, we make use the Google *did you mean* mechanism. When a search term is entered, Google will check to see if more relevant search results would be found using an alternative spelling. Because Google's spell check is based on occurrences of all words on the Internet, it is able to suggest common spellings for proper nouns (e.g. names and places) that would not appear in a standard dictionary.

The Google "did you mean" mechanism also provides an excellent way to resolve compound nouns. Since most tagging systems prevent users from entering white spaces into the tag name, users create compound nouns by concatenating two nouns together or delimiting them with a non-alphanumeric character such as a _ or `-`. This is an obvious source of complication when aligning folksonomy activity: users do not consistently use the same compound noun creation schema. By entering a compound terms into Google, we can resolve the tag into its constituent parts. For example, the tag `sanfrancisco` is corrected to `san francisco`. After using Google to check for compound nouns and misspellings, the results are validated against Wordnet. Any unmatched or unprocessed tags are passed to Step 3.

Step 3: Wikipedia Correlation

Many of the popular tags appearing in communal tagging systems do not appear in grammatical dictionaries, such as Wordnet, because they correspond to nouns (such as famous people, places, or companies), contemporary terminology (such as `web2.0` and `podcast`), or are widely used acronyms (such as `tv` and `diy`). In order to provide an agreed representation for such tags, we correlate them to their appropriate Wikipedia page. For example, when searching Wikipedia using the tag `nyc`, the entry for New York City is returned. If the search term `ny` is used, the entry for New York state is returned. The advantage of using Wikipedia to agree on tags from folksonomies is that Wikipedia is a community-driven knowledge base, much like folksonomies are, so it will rapidly adapt to accommodate new terminology. For example, Wikipedia contains extensive entries for terms such as `web2.0`, `ajax`, and `blog`.

Step 4: Morphologically Similar

An additional issue to be considered during the tag filtering process is that users often use morphologically similar terms to refer to the same concept. One very common example of this is the discrepancy between singular and plural terms, such as `blog` and `blogs`. Using a custom singularisation algorithm, and the stemming functions provided by the *snowball* library[2], we reduce morphologically similar tags to a single tag. The shortest term in Wordnet is used as the representative term.

Step 5: Wordnet Synonyms

The final step in the filtering process is to identify tags that are non-ambiguous synonyms, and merge them. This process must be carefully executed because many terms have ambiguous meaning. The algorithm for this process is present in (Szomszor et al., 2008b) and explained in full with pseudocode.

[2] `http://snowball.tartarus.org/`

2.1.3 Resource Consolidation

Since most resources in a folksonomy are defined by a uri, the matching of resource between multiple folksonomies is a simple task. The likelihood of two folksonomies referring to the same resource depends largely on the focus of the folksonomies. In cases where resources do appear in multiple folksonomies, different information about the resource can be extracted. For example, a photograph in a user's Flickr profile may also appear in their Facebook profile. Since the tagging focus of Facebook is oriented around identifying other Facebook users who feature in the photo, it would be possible to combine this information with that from Flickr to build an integrated view of the resource, proving details of the location, attributes of the content, and who is in the photo and how they relate to the author.

In a more general sense, we can consider the integration of information about resources that doesn't originate from a folksonomy. For example, the popular social bookmarking sites Digg[3], and reddit[4] do not allow users to tag resources (as del.icio.us does). Instead, they employ a voting system where users *digg* an article and other users vote it up or down. The result is an ordered list of resources reflecting the popularity of the resource by that community at a particular point in time. Since many of the resources posted in digg and reddit also appear in del.icio.us, it would be possible to combine popularity information with the tags people have used to describe it. This kind of information would be valuable to recommender systems attempting to notify users of the most important and relevant resources.

2.2 Creating Cross-Folksonomy Networks

With the increasing amount of distributed Web2.0 data, users are faced with the challenge of managing information spread over a number of sites in a range of heterogeneous formats. In many cases, there are overlaps between the functionality offered by such sites. For example, Flickr and Photobucket are popular sites for sharing photos, but Facebook also provides a photo sharing mechanism. Initially, users began duplicating data between profiles, uploading the same information to different sites to maximise their profile content. As demand increased, developers started creating widgets that allow users to include third-party data in their profile. For example, the Flickr-Facebook application[5] allows Facebook users to add a widget to their profile page that links to their Flickr account and displays photo streams. While these developments fostered new interest from users as they began to realise the benefits of linking their online identities, such solutions did not provide integration between data sources. For example, searching for photos in Facebook for a particular person would not return results from their Flickr account even if they had linked the profiles together using a third-party widget.

In May 2008, Google, MySpace, and Facebook all announced initiatives in the area of data portability: Google released their Friend Connect service[6], Facebook announced their Facebook Connect Application[7], and MySpace declared their interest in supporting the Data Portability Initiative[8] along with other partners such as Yahoo, eBay, and Twitter. In these cases, the notion of data portability is fairly limited: current specifications are mainly concerned with the interchange of social networking information only. For example, allowing different social networking sites to exchange friend lists enabling users to connect to existing friends when joining new social networks.

[3] http://digg.com/
[4] http://reddit.com
[5] http://www.facebook.com/apps/application.php?id=2498985378
[6] http://www.google.com/friendconnect
[7] http://developers.facebook.com/news.php?blog=1&story=108
[8] http://www.dataportability.org/

2.2.1 An RDF Based Solution

Within the research community, the Resource Description Framework (RDF)(Klyne and Carroll, 2004) is attracting much attention as a possible solution to these data portability issues (Berners-Lee, 2007). Large information providers, such as Wikipedia and Last.fm, have counterpart sites that expose their information in RDF, such as DBpedia[9] and DBTune[10]. We have developed a suite of software tools to automatically generate an RDF representation of an individuals del.icio.us, Flickr, Last.fm, and Facebook activity. By crawling the public data posted by the 1998 individuals with whom we discovered a number of online accounts (described previously in Section 2.1.1), and following one level of redirection (e.g. by following links to their friends and contacts), we have create a dataset describing the activity of 6,861 del.icious, 7,978 Flickr, and 105,557 Last.fm users. To model tagging activity, we use the SCOT ontologies[11], summarised in Figure 2.3, since they provide the ability to model tagging events (through the SKOS ontology (Miles et al., 2005)), resource attributes (through the SIOC ontology (Bojars et al., 2008)), and social relationships (through the FOAF ontology(Brickley and Miller, 2007)). The complete dataset currently contains 177,822,400 triples, but is continuing to increase as we harvest more data. In a sense, an RDF representation of this tagging activity only provides the raw data model required to link folksonomies: The user's raw tagging behaviour would need to be refined, for example through tag filtering, to support a meaningful integration of tags across folksonomies. Frameworks, such as MOAT(Passant and Laublet, 2008), have been developed to allow the semantic of tags to be expressed explicitly, through a link to an ontology concept, and would facilitate in the understanding of ambiguous terms.

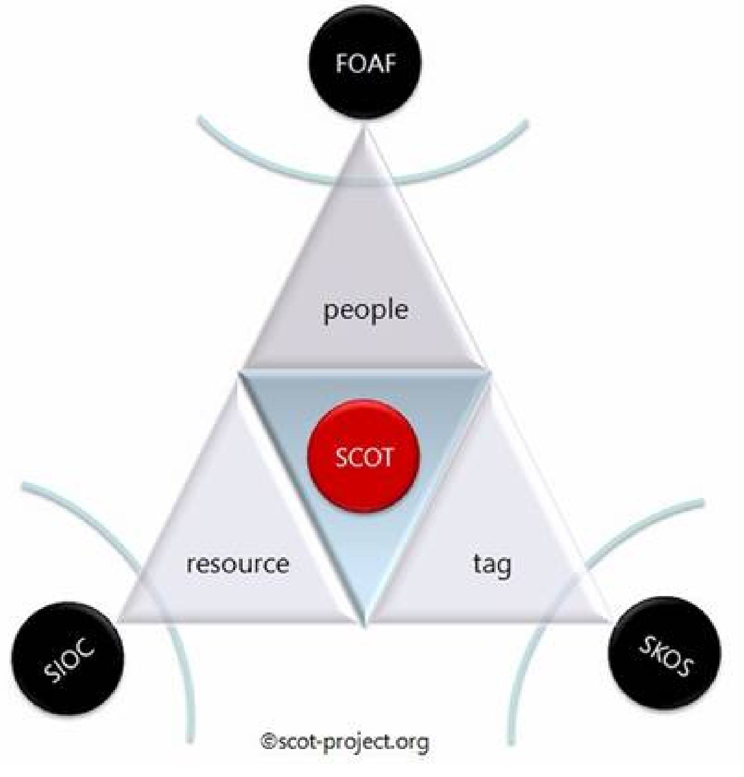


Figure 2.3: The SCOT ontologies

[9] http://dbpedia.org/
[10] http://dbtune.org
[11] http://scot-project.org/

Chapter 3

Applications

Two applications developed within the TAGora project cover the topic of cross-folksonomy integration: The MyTag application, described in Section 3.1, and an application for the semantic modelling of user interests, presented in Section 3.2.

3.1 MyTag

There are many Web 2.0 platforms, such as YouTube, Flickr, and del.icio.us, that provide large amounts of information in a variety of multimedia formats, including videos, photographs, and web pages. However, there are many different sites catering for a particular media type. For example, Flickr, Photobucket, and Facebook, are all popular places for users to publish photos. Hence, searching for a single media type that could exists in multiple sites, or for information that could be presented in a variety of formats, has become and increasingly convoluted process requiring much intervention on the part of the user who must search multiple platforms manually (Anadiotis et al., 2007). Further more, the raking mechanism employed by such sites is fairly crude: the most highly rated or recently viewed items are usually pushed to the top of the list. Current platforms lack the ability to rank results according to preferences expressed by the user, either explicitly or implicitly.

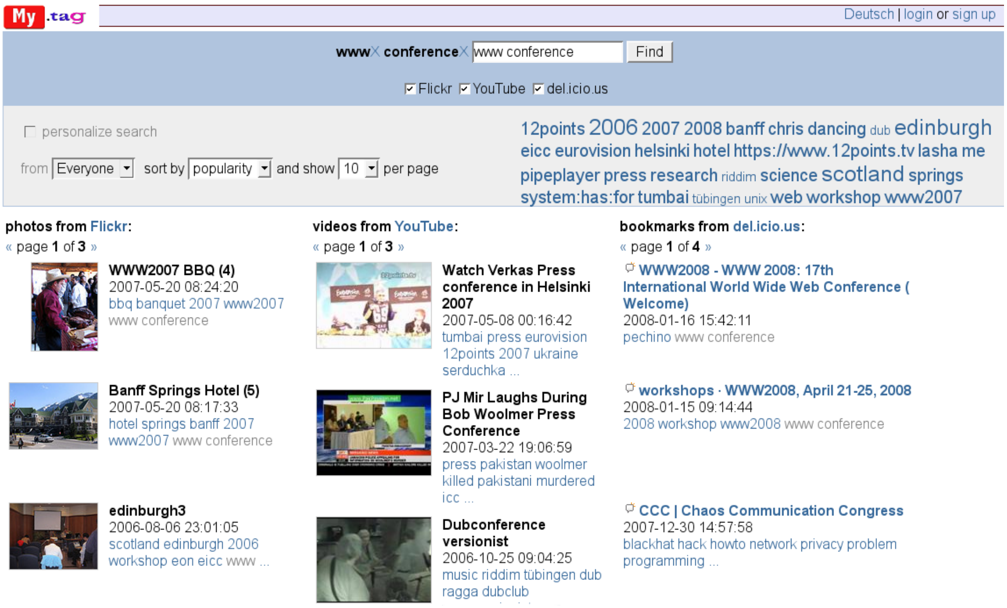
The MyTag[1] application has been developed to address these limitations, enabling cross-media searching over images, videos, and social bookmarks (Braun et al., 2008). MyTag provides transparent search across multiple tagging platforms, each providing different media content. The current implementation covers Flickr, YouTube, and del.icio.us. Furthermore, it incorporates the notion of personalisation, allowing users to grow a profile that, in turn, ranks search results in more suitable and refined manner. A sample screen-shot is provided in Figure 3.1 showing the result set obtained when searching for `www conference`.

3.1.1 Personalisation

Two personalisation features are provided for search: First, a search can be restricted to resources uploaded by the user. This feature requires that a user enters their account names for Flickr, del.icio.us, and/or YouTube into their profile. Searching over user resources is implemented simply by using the corresponding search feature from the source tagging platform.

The second personalisation feature allows for ranking search results based on the user's *personomy*. The personomy is automatically built based on the resources the user picks from the search results. It is modeled by a vector $\mathbf{p}$ of tag frequencies representing the previous search interests of the user. As it is based on the implicit feedback given by selecting from the search results, no additional user effort is required to gain personalisation. Using implicit user feedback is a very

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


My.tag
Deutsch | login or sign up
www X conference X
www conference
Find
☑ Flickr ☑ YouTube ☑ del.icio.us
□ personalize search
from Everyone ▾ sort by popularity ▾ and show 10 ▾ per page
12points 2006 2007 2008 banff chris dancing dub edinburgh eicc eurovision helsinki hotel https://www.12points.tv lasha me pipeplayer press research riddim science scotland springs system:has:for tumbai tübingen unix web workshop www2007

photos from Flickr:

« page **1** of **3** »

WWW2007 BBQ (4)
2007-05-20 08:24:20
bbq banquet 2007 www2007 www conference

Banff Springs Hotel (5)
2007-05-20 08:17:33
hotel springs banff 2007 www2007 www conference

edinburgh3
2006-08-06 23:01:05
scotland edinburgh 2006 workshop eon eicc www ...

videos from YouTube:

« page **1** of **3** »

Watch Verkas Press conference in Helsinki 2007
2007-05-08 00:16:42
tumbai press eurovision 12points 2007 ukraine serduchka ...

PJ Mir Laughs During Bob Woolmer Press Conference
2007-03-22 19:06:59
press pakistan woolmer killed pakistani murdered icc ...

Dubconference versionist
2006-10-25 09:04:25
music riddim tübingen dub ragga dubclub

bookmarks from del.icio.us:

« page **1** of **4** »

WWW2008 - WWW 2008: 17th International World Wide Web Conference (Welcome)
2008-01-16 15:42:11
pechino www conference

workshops · WWW2008, April 21-25, 2008
2008-01-15 09:14:44
2008 workshop www2008 www conference

CCC | Chaos Communication Congress
2007-12-30 14:57:58
blackhat hack howto network privacy problem programming ...

Figure 3.1: A Screenshot of a MyTag Search Result

promising approach to personalising search results or web browsing in general (cf. (Sugiyama et al., 2004) and (Mladenic, 2002)). This feature adds an advantage compared to systems such as Flickr and del.icio.us, where personalization requires adding resources to the system, i. e. the explicit feedback of users.

The current MyTag platform implements a ranking algorithm that combines information from the personomy and the tags assigned to resources of a result set. The tags of a resource are represented as a vector $\mathbf{v}$ of binary values indicating the presence of a tag. The rank r of a resource is then computed by the scalar product of the two vectors: $\mathbf{r} = \mathbf{v} \cdot \mathbf{p}$. It is then used for ordering the resources based on their rank value.

3.1.2 Architecture

The MyTag architecture realises the model-view-controller paradigm (MVC). The three layers of the MyTag architecture are shown in Fig. 3.2. The view layer at the top is responsible for the interaction with the user while the control layer in the middle processes data from the model layer, e.g. by computing personalized rankings. The model layer consists of two core parts: First, the interface to the local database that contains the user profiles and personomies. Second, a generic interface that abstracts core functionality provided by the APIs of the external tagging platforms, ensuring MyTag's future extensibility.

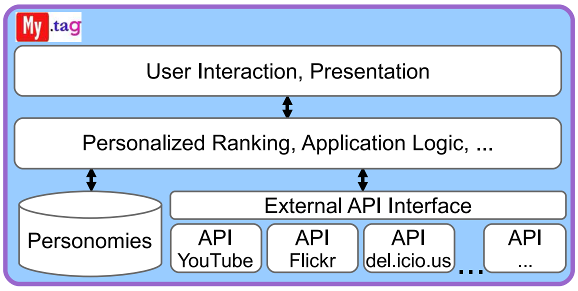


Figure 3.2: The MyTag Architecture

3.2 Building Semantic Models of User Interests

As part of the work carried out for Task 3.5, we explore an approach for unifying distributed user profiles and exploiting them to build semantic profiles of interest represented using FOAF and Wikipedia ontologies (Szomszor et al., 2008a). The objective is to supply an architecture that constructs a model of user interests by examining their interaction with various folksonomy sites working under the assumption that the tags used most often by an individual correspond to the topics, places, events, and people they are most interested in.

To maximise the utility of such profiles, semantic modeling is essential - tags themselves are only string literals and have no explicit semantics so there are no relationships between terms. For example, resources related to programming languages may be tagged in del.icio.us using the terms `perl`, `c++`, or `python`. While it is clear to the user that these tags are related, such a relationship is not modeled within the folksonomy. Hence, our approach relies not only on identifying the most important tags used, but also correlating them to a uri that has explicit references describing it's semantics.

While previous semantic profiling work has concentrated on using well defined ontologies for this pupose, it is not practical for a general solution since information within folksonomy sites such as del.icio.us, Flickr, and Last.fm is extremely diverse. Furthermore, folksonomies are dynamic systems that constantly evolve to accommodate new terminology and trends. Therefore, we decided to use Wikipedia categories to model user interests because Wikipedia covers a wide range of topics and is constantly updated by the community. Referring to the example above, the Wikipedia categories for perl and c++ are both subcategories of “Programming language families”.

3.2.1 Architecture

Broadly, the Semantic Profiling architecture is split into four sections, as depicted in Figure 3.3:

1. **Account Correlation** The first step is to identify the accounts held by a particular individual across a range of social networking sites. By using the Google Social Graph API, we are able to take a url denoting the user (such as their homepage) and discover the various online accounts they hold.

2. **Data Collection Module** Once the user accounts have been identified, the Data Collection Module harvests a complete history of their tagging activity within each site.

3. **Tag Filtering** After collecting an individual's raw tagging activity, we utilise the Tag Filtering process (presented earlier in 2.1.2) to filter and merge tags into a canonical representation. This stage allows us to resolve compound nouns (for example, the tags `second_life`, `secondlife`, and `second-life` are merged), cater for misspellings, identify acronyms, and identify synonyms.

4. **Profile Building** The final stage in the process consumes an individual's filtered tag-clouds and attempts to match each term to a Wikipedia category. Once the list of categories has been generated, a FOAF file is generated to express their interests using references to Wikipedia category urls.

3.2.2 Semantic Models of Interest

Figure 3.4 presents part of an example FOAF file (for an anonymous user), emphasising how tags extracted from del.icio.us and Flickr tag-clouds are associated with Wikipedia categories. In this example, the popular tags `Flickr`, `Youtube`, `C++`, and `Perl` have been extracted from their

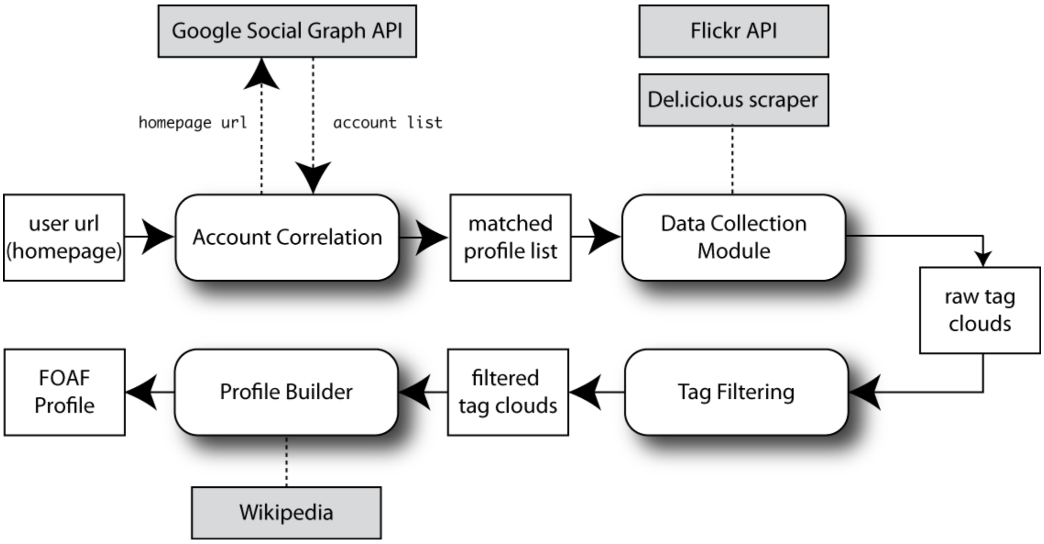
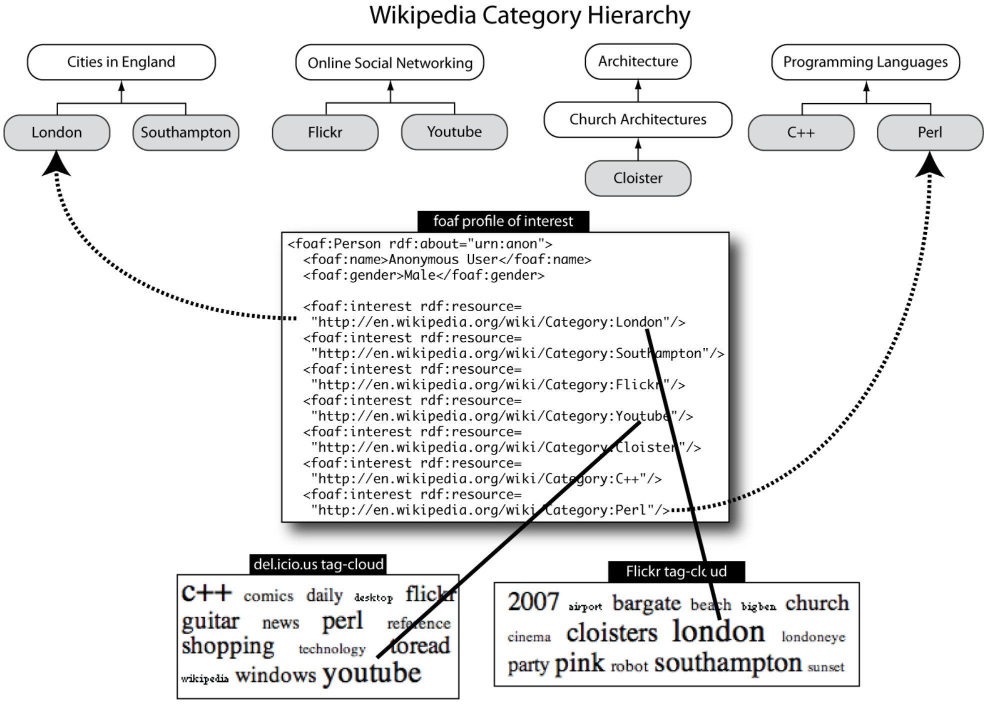


Figure 3.3: The Semantic Profiling Architecture

del.icio.us tag-cloud and correlated with the appropriate Wikipedia categories. Such terms are often related by a common super-category such “Online Social Networking” and ”Programming Languages”. From their Flickr tag-cloud, the terms `London`, and `Southampton` have been correlated. Furthermore, the tag `cloisters` has been matched to the the category “Church Architecture”, a correspondence that would not be possible without semantic techniques.


Wikipedia Category Hierarchy
Cities in England
London
Southampton
Online Social Networking
Flickr
Youtube
Architecture
Church Architectures
Cloister
Programming Languages
C++
Perl
foaf profile of interest
del.icio.us tag-cloud
C++ comics daily desktop flickr guitar news perl reference shopping technology toread wikipedia windows youtube
Flickr tag-cloud
2007 airport bargate beach bigben church cinema cloisters london londoneye party pink robot southampton sunset

```
<foaf:Person rdf:about="urn:anon">
  <foaf:name>Anonymous User</foaf:name>
  <foaf:gender>Male</foaf:gender>

  <foaf:interest rdf:resource=
   "http://en.wikipedia.org/wiki/Category:London"/>
  <foaf:interest rdf:resource=
   "http://en.wikipedia.org/wiki/Category:Southampton"/>
  <foaf:interest rdf:resource=
   "http://en.wikipedia.org/wiki/Category:Flickr"/>
  <foaf:interest rdf:resource=
   "http://en.wikipedia.org/wiki/Category:Youtube"/>
  <foaf:interest rdf:resource=
   "http://en.wikipedia.org/wiki/Category:Cloister"/>
  <foaf:interest rdf:resource=
   "http://en.wikipedia.org/wiki/Category:C++"/>
  <foaf:interest rdf:resource=
   "http://en.wikipedia.org/wiki/Category:Perl"/>
```

Figure 3.4: An Overview of the Semantic Profiles Generated

2008 © Copyright lies with the respective authors and their institutions

Chapter 4

Conclusions and Future Work

In this Document, we have introduced the idea of cross-folksonomy integration, providing some motivating examples as to why it might useful. One important benefit is the ability for users to search over multiple tagging platforms spanning different domains of interest, presenting different media types. The MyTag application has taken steps towards this goal, supporting the transparent searching of media across Flickr, YouTube, and del.icio.us.

Since many users hold profiles in multiple social networking sites, linking user account between different folksonomies is a beneficial activity. By examining an individual's tagging history across multiple sites, a complex picture of their activity can be built and used to construct semantic profiles representing their interests. The more profiles that are included, the more can be learnt about the user. In terms of the resources, linking folksonomies would enable one to combine information and enrich the amount of knowledge about the resource. For example, del.icio.us bookmarking tells us what tags people use to index the resource, and Digg will tell us how the popularity of the resource evolved over time. The same picture in Flickr a Facebook is likely to receive different annotations reflecting the focus of the site.

However, the construction of cross-folksonomy networks is a difficult task. On the one hand, the sheer amount of data causes many problems. When a social networking site becomes sufficiently popular, the amount of new data added will exceed the amount that can be crawled. Hence, it is infeasible to build a complete index spanning multiple tagging platforms. Instead, small cross-folksonomy networks can be built for specific tasks. As we demonstrated in (Szomszor et al., 2008b), correlating user accounts enabled us to examine the tagging activity of users in del.icio.us and Flickr and investigate the overlaps that exist in their tag-clouds as a possible mechanism for correlating accounts in different systems. Through further work, we found the most robust and successful solution to matching users is to examine the links between accounts (such as a common homepage or blog), as provided by the Google Social Graph API.

The nature of tagging results in many difficulties when aligning the terms used in different folksonomies. Morphologic variety, and ambiguous terms must be considered when trying to match tags. We have found it useful to relate tags to entries in Wordnet and Wikpiedia since it provides a canonical reference, as well as some useful semantic about the tags, such as the synonyms and closely related terms.

I terms of future work, the MyTag personalised ranking algorithm will be evaluated against other state-of-the-art approaches, such as FolkRank (Hotho et al., 2006), to compare performance and investigate possible improvements. Continuing from this, efforts will be made to consolidate the results from different domains and present a single results set to the user (rather than multiple result sets as is currently implemented). It is also our intention to incorporate more tagging platforms, such as Bibsonomy. Furthermore, we will investigate the possibility of integrating our semantic profile building technology with the MyTag application with an aim to enrich the personomies used to rank search results.

The framework we have developed for building semantic models of user interests will be extended in a number of ways: First, we will port the profile building technology to operate over the RDF models we described in Section 2.2.1. Then, by using a combination of our filtering algorithms, and the Wikipedia correlation mechanism developed in (Szomszor et al., 2008a), it will be possible to associate user tags to a Wikipedia uri, providing explicit semantics and therefore facilitate the integration of tags between folksonomies, especially with respect to ambiguous terms Finally, for Task 4.2 (Deployment of a semantic recommender), we will utilise the profiles of interest to drive a cross-domain recommendation system (Loizou, 2007).

Bibliography

George Anadiotis, Thomas Franz, and Susanne Boll. W3C Multimedia Semantics Incubator Group: Tagging Use Case. `http://www.w3.org/2005/Incubator/mmsem/wiki/Tagging_Use_Case`, 2007.

Grigory Begelman, Philipp Keller, and Frank Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Proc. 17th Int. World Wide Web Conference (WWW), Edinburgh, UK*, 2006.

Tim Berners-Lee. Giant global graph, November 2007. URL `http://dig.csail.mit.edu/breadcrumbs/node/215`.

Uldis Bojars, John G. Breslin, A. Finn, and Stefan Decker. Using the semantic web for linking and reusing data across web2.0 communities. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):21–28, 2008.

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. URL `http://www.uni-koblenz.de/~klaasd/Downloads/papers/Braun2008PSA.pdf`.

Dan Brickley and Libby Miller. FOAF vocabulary specification 0.91, 2007. URL `http://xmlns.com/foaf/spec/`.

Ivän Cantador, Martin Szomszor, Harith Alani, Miriam Fernändez, and Pablo Castells. Enriching ontological user profiles with tagging history for multi-domain recommendations. In *Proc. Workshop on Collective Semantics: Collective Intelligence and the Semantic Web (CISWeb 2008), in 5th ESWC, Tenerife, Spain*, 2008.

Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. Massachusetts: The MIT Press, 1998. p.423.

Scott A. Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32:198–208, 2006.

Marieke Guy and Emma Tonkin. Tidying up tags? *D-Lib Magazine*, 12(1), 2006.

Conor Hayes, Paolo Avesani, and Sriharsha Veeramachaneni. An analysis of the use of tags in a log recommender system. In *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI), Hyderabad, India*, 2007.

Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Trend detection in folksonomies. In *Proc. of SAMT Conference*, 2006.

Graham Klyne and Jeremy J Carroll. Resource description framework (RDF): Concepts and abstract syntax. Technical report, W3C, 2004. URL `http://www.w3.org/TR/rdf-concepts/`.

Xin Li, Lei Guo, and Yihong (Eric) Zhao. Tag-based social interest discovery. In *Proc. 19th Int. World Wide Web Conf (WWW), Beijing, China*, 2008.

Antonis Loizou. Unlocking the potential of recommender systems: A framework to acheive multiple domain recommednations. Technical report, Deparment of Electronics and Computer Science, University of Southampton, 2007. URL `http://eprints.ecs.soton.ac.uk/15462/1/al05r_mini-thesis.pdf`.

Adam Mathes. Folksonomies - cooperative classification and communication through shared metadata. *Computer Mediated Communication - LIS590CMC*, December 2004. URL `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`.

A. Miles, B. Matthews, D. Beckett, D. Brickley, and D. Wilson. SKOS core: Simple knowledge organisation for the web. In *Proc. International Coneference on Dublin Core and Metadata Applications (DC-2005), Madrid, Spain*, 2005. URL `http://epubs.cclrc.ac.uk/bitstream/675/`.

D. Mladenic. Web browsing using machine learning on text data. In P. Szczepaniak, J. Segovia, J. Kacprzyk, and L. Zadeh, editors, *Intelligent exploration of the web*. Physica-Verlag, 2002.

Ofcom. Social networking: A quantative and qualitative research report into attitudes, behaviours, and use., 2008. URL `http://news.bbc.co.uk/1/shared/bsp/hi/pdfs/02_04_08_ofcom.pdf`.

Alexandre Passant and Philippe Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Workshop on Linked Data on the Web (LDOW), Int. Word Wide Web Conference, Beijing, China*, 2008.

David Silver. *Smart Start-ups: How to Make a Fortune from Starting Online Communities*, page 5. John Wiley and Sons, Inc., 2007. ISBN 978 0 7499 2788 2.

K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proc. of 13th WWW*, 2004.

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.