Semantic Modelling of User Interests based on Cross-Folksonomy Analysis

Martin Szomszor[1], Harith Alani[1], Ivan Cantador[2], Kieron O'Hara[1], Nigel Shadbolt[1]

[1] Intelligence, Agents, Multimedia
School of Electronics and Computer Science
University of Southampton, Southampton, UK
{mns03r, h.alani, kmo, nrs}@ecs.soton.ac.uk
[2] Escuela Politcnica Superior
Universidad Autnoma de Madrid
28049 Madrid, Spain
ivan.cantador@uam.es

Abstract. The continued increase in Web usage, in particular participation in folksonomies, reveals a trend towards a more dynamic and interactive Web where individuals can organise and share resources. Tagging has emerged as the de-facto standard for the organisation of such resources, providing a versatile and reactive knowledge management mechanism that users find easy to use and understand. It is common nowadays for users to have multiple profiles in various folksonomies, thus distributing their tagging activities. In this paper, we present a method for the automatic consolidation of user profiles across two popular social networking sites, and subsequent semantic modelling of their interests utilising Wikipedia as a multi-domain model. We evaluate how much can be learned from such sites, and in which domains the knowledge acquired is focussed. Results show that far richer interest profiles can be generated for users when multiple tag-clouds are combined.

1 Introduction

With the growth of Web2.0, it is becoming increasingly common for users to maintain a presence in more than one site. For example, one could be bookmarking pages in del.icio.us, uploading images in Flickr, listening to music in Last.fm, blogging in Technorati, etc. The nature of these pursuits naturally leads users to express the relevant aspects of their interests, which are likely to be different across the sites. If such multiple identities and distributed activities could be brought together independently of the Web 2.0 sites, far richer user profiles could be generated.

There would be a number of potential gains for recommender systems from the greater profile depth. Usually, such systems monitor in-house user activities over a certain period of time to build up profiles that support recommendations. As a result, they will be limited to the activities of users within those systems, and thus may fail to capture other user interests, resulting in potential recommendations, and eventually transactions, being lost. Furthermore, a fuller set of user activities can be captured when expanding data gathering to multiple sites, thus ensuring dynamic updates. For example, if someone reduces their use of Last.fm for a few months, their opinions of the latest music may not be properly captured, leading to a tailing-off of recommendation quality.

There is a strong push towards opening up social networking to support portability of data across various sites. Many popular sites are racing to develop tools to allow their users to port their personal profiles to other sites. Within days from each other in May 2008, Google, MySpace, and Facebook announced new initiatives for increasing social profile portability called Friend Connect[3], Data Availability [16], and Connect[18] respectively. Efficient cross-linking of user profiles should reduce tag-cloud maintenance, and facilitate search and retrieval of tagged resources from multiple sites. Tagging, a fast spreading activity where users assign terms to online resources, is an important discourse within the Web 2.0 phenomenon. Tags serve various purposes, such as for resource organisation, promotions, sharing with friends, with the public, etc. [14, 1]. However, studies have shown that tags are generally chosen to reflect their user's interests. Golder and Huberman [8] analysed tags on del.icio.us, and found that (a) the overwhelming majority of tags identify the topic of the tagged resource, and (b) almost all tags are added for personal use, rather than for the benefit of the public. These findings lend support to the idea of using tags to derive user profiles. But tags are free text, and thus suffer from various vocabulary problems [15, 8, 10]. If it were possible to *clean* such tags and render them somewhat more standardised, this could be helpful to improve tag-cloud compatibility.

The issue of modelling user interests based on cross-folksonomy activity is likely to become increasingly significant. In a recent survey, Ofcom found that 39% of UK adults with at least one folksonomy profile have indeed two or more profiles [19]. It has even been predicted that by 2010, each of us will have between 12 and 24 online identities [21]. Users are often forced to create separate accounts to participate in different activities. There are signs that many of these users are keen to link up their separate accounts. For example, many Last.fm users provided their Flickr or del.icio.us account URL as their homepages.

In this paper we will explore an approach for unifying distributed user profiles, and building semantic profiles of interests using FOAF and Wikipedia ontologies. The next section will review related work. Section 3 provides a full description of the approach, followed by an experiment in section 4 and an evaluation of results in section 5. Discussion and future works are covered in sections 6 and 7 respectively.

2 Related Work

2.1 Analytic Studies

The spread of tagging and the derivation of folksonomies is providing valuable data sources and environments for studying various user-related issues, such as online behaviour, tagging patterns, incentives for sharing, social networking, and opinion formation. A number of studies have focused on analysing user incentives and motivations behind tagging: Marlow and colleagues studied the effect of system design on tagging style, and the various incentives behind tagging [14]. Similarly, Ames and Naaman [1] studied the reasons why people tag images in Flickr, and articulated a taxonomy

[3] `http://www.google.com/friendconnect/`

of social and functional motivations. They found that users tag for various reasons, such as for organising their resources, sharing them with others, or simply to promote their work. As noted above, the motivations behind tagging tend to be almost always personally-focused [8], and the connection between tagging practice, user preferences and maximally effective profiling.

In a study from Yahoo! on the del.icio.us data, Li and colleagues found that tags are better representatives of users' interests than the keywords of the tagged Web pages, because (a) they offer a higher level of content abstraction, and (b) they are better representations of the user's perception of that content [12]. The authors investigated matching users based on the similarity of tag clusters in del.icio.us. In our work however, we are interested in identifying the specific interest of the users as an independent attribute, and not only the similarity of his/her interests with others.

Investigations in related fields have shown that there are interesting correlations between social networking environments and the domains to which they relate. For instance, De Choudhury and colleagues found a correlation between certain blogs and the movement of the stock market [5], while Singla and Richardson analysed MSN Messenger chat logs and the search queries of the chatters, and found that those who exchanged short messages frequently were more likely to issue similar search queries [22]. In our work, we are investigating the correlation between user tagging activities across multiple folksonomies.

2.2 Normative Accounts of Tagging Practice

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 users to contribute. It also avoids the problem of forcing users into using terms they do not feel apply, a situation that arrises when vocabularies are enforced. For these reasons, the lack of constraints seems essential. On the other hand, it generates various vocabulary problems: tags can be too personalised, made of compound words, mix plural and singular terms, meaningless, synonymous, etc. [15, 8, 10]. This total lack of control is resulting in some sort of tagging chaos, thus obstructing search [10] and analysis [12].

Guy and Tonkin [10] 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. While 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 [11] 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 [2], where similar tags were grouped together to facilitate distinguishing between their different meaning when searching.

2.3 Collection and Semantic Representation of User Interests

This paper is mainly concerned with learning about user interests, and there is a strong tradition of work in this area. Mori and colleagues investigated extracting information from web pages using term co-occurrence analysis to build FOAF files [17]. Diederich and Iofciu [7] tried to identify user interests based on tag clustering. However, the tags used were in fact DBLP keywords, resulting in a system serving a different purpose to free tagging. Demartini suggests using the history of users' edits in Wikipedia to find out about their expertise [6]. Such an approach will obviously only work for users who actively edit Wikipedia pages. In contrast, the work reported in this paper exploits the resources of Wikipedia, but our primary interest is in identifying and semantically representing the general interests of users, based on what they tag and how they tag across several folksonomies.

Semantic representation should ideally involve associating user interests with appropriate URIs, thus moving folksonomy user profiles closer to the Semantic Web and moving the agenda of using Semantic Web technology to organise collectively assembled information characteristic of Web 2.0 [9]. Semantically-Interlinked Online Communities (SIOC) is an ontology that provides a foundation for semantically representing user activities in blogs and forums [3]. To facilitate representing tags with URIs, Meaning Of A Tag (MOAT) was developed as a framework to help users manually select appropriate URIs for their tags from existing ontologies [20], in contrast to the work reported in this paper, which explores the strategy of automating the selection of URIs to maintain the essential simplicity of tagging. Specia and Motta [23] investigated reusing existing ontologies to link tags automatically with pre-crafted concepts and relations. Here we are concerned with selecting URIs that represent *topics*, and not just any concept in any ontology.

The novelty of the work reported here is in the amalgamation of multiple Web 2.0 user-tagging histories to build up personal semantically-enriched models of interest. Correlating different folksonomies is a very new art, and has received surprisingly very little attention so far. In a previous study, we compared the tag clouds of users in both Flickr and del,icio.us and found that they tend to be more similar to each other than to other users tag clouds [24]. That indicated that users often carry some of their tagging selections and patterns across different folksonomies, and this insight is an important motivation for this work.

3 An Architecture for Building Semantic Profiles of Interest

The objective of this work is to supply an architecture that constructs a model of user interests by examining their interaction with various folksonomy sites. We are 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 its semantics.

While previous semantic profiling work has concentrated on using well defined ontologies for this purpose, it is not practical for a general solution since information within folksonomy sites such as del.icio.us and Flickr 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”.

Broadly, the architecture is split into four sections, as depicted in Figure 1:

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 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 a Tag Filtering architecture, developed in previous work [24], to filter and merge tags into a canonical representation. This stage allows us to resolve compound nouns (for example, the tags `second_life` 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.

The following Subsections give a more detailed account of each of these processes.

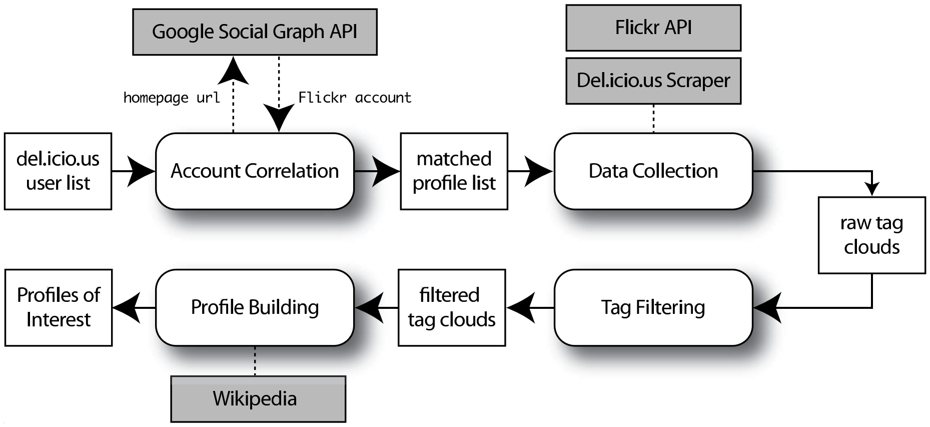


Fig. 1. An Architecture for building Semantic Profiles of Interest

3.1 Account 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 [24], we matched 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 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 only 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 [4] providing our profile building architecture with a powerful account tracing facility.

3.2 Data Collection

The Data Collection module is responsible for harvesting information from a range of social networking sites. In the case of sites such as Flickr and Last.fm, public APIs are provided that allow us to download a complete history of user tagging activity. For other sites, such as del.icio.us, public APIs are very limited so custom screen-scraping scripts were developed.

3.3 Tag Filtering

When users choose to tag a resource, be it a web page, photo, or video, they are free to choose any tag(s) they please. While it has been shown that this uncontrolled behaviour does result in meaningful structures at the global level, the tag-clouds of particular individuals often contain misspellings, synonyms and morphologic variety. As a result, important correlations between resources and users are sometimes lost simply because of the syntactic mismatches in the tags they have used. To cater for this problem, we developed a tag filtering architecture that cleans and reduces user generated tag-clouds [24].

The filtering process 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 provides a visual representation of the filtering process where a set of raw tags is 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:

[4] `http://code.google.com/apis/socialgraph/`

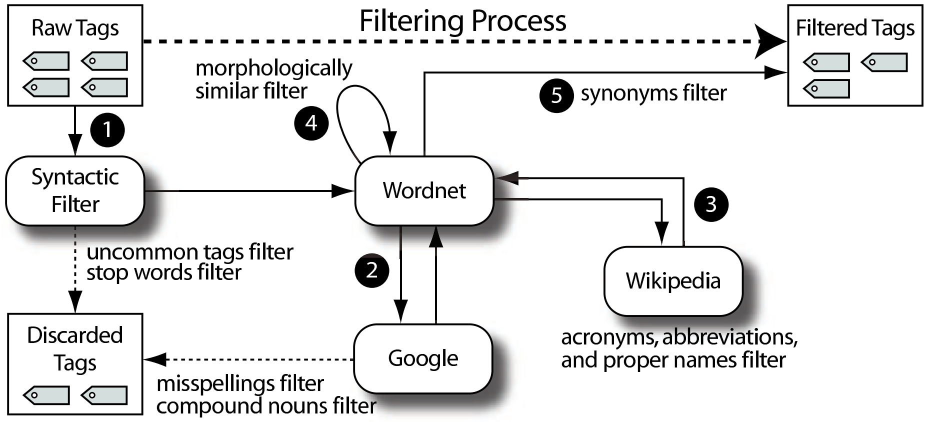


Fig. 2. The tag filtering process

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 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`, `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. 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[5], 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 [24] and explained in full with pseudocode.

3.4 Building Profiles of User Interests

The final stage of our profile building architecture turns a set of filtered tag-clouds to a single FOAF file representing as many of the user's interests as possible. To accomplish this, a three-stage process is followed: **(Stage 1)** Each filtered tag-cloud is transformed to a weighted list of Wikipedia categories. For example, if a del.icio.us and Flickr account are discovered for a particular individual, a separate category list is generated for each. **(Stage 2)** then combines these category lists and filters out the uncommon terms to produce the final interest list. **(Stage 3)** turns this list into an RDF representation using the FOAF and Wikipedia ontologies.

The process of transforming a filtered tag-cloud to a Wikipedia category list (Stage 1) is explained below:

1. **Identify Wikipedia Page:** For every tag, we attempt to identify its corresponding Wikipedia page. For example the tag `perl` is matched to the Wikipedia page `http://en.wikipedia.org/wiki/Perl`.
2. **Extract Category List:** For some terms, such as `directory`, the page returned by Wikipedia is a disambiguation page - one that does not define the term itself, but simply references a list of other pages associated with the title. In these cases, no Wikipedia category is found and we move on to the next tag. In the cases where a page is found, the list of categories (found at the bottom of the page) is extracted.

[5] http://snowball.tartarus.org/

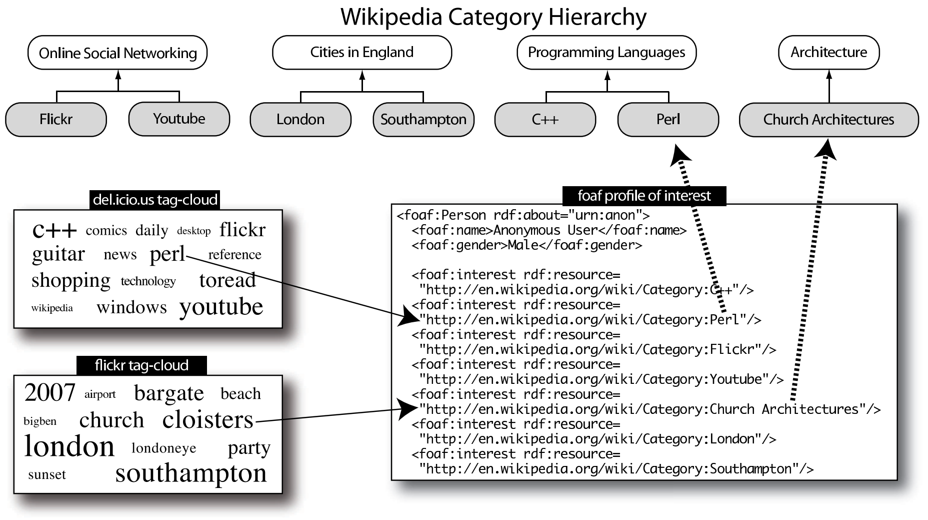
3. **Selection of Representative Categories:** Initially, we believed it would be useful to record *all* the categories associated with a particular page. For example, the Wikipedia page for Blogs is associated with the categories Blogs, Blogging, Digital Revolution, Internet terminology, Politics and technology, and Technology in society. However, due to the diversity of categories used in Wikipedia, the final category list was often be dominated by spurious terms such as "host cities of the summer olympic games" (from the entry for London), "Christmas nomenclature and language" (from the entry Christmas), as well as Wikipedia specific meta-categories such as "needs more sources". To compensate for this, we decided to only include a category if: a) there is only one category associated with the page, b) the category matches the page name exactly to maximise accuracy, or c) the category is a pluralisation of the page name (e.g. `http://en.wikipedia.org/wiki/Blog` has the category `http://en.wikipedia.org/wiki/Category:Blogs`). If an appropriate category is found, a weight is associated to it corresponding to the frequency of the tag. If more than one tag is matched to the same category, the category weight is the sum of all the tag frequencies.

For Stage 2, a global category list is generated by combining all the category lists generated from stage 1. If a category appears in more than one list (e.g. the user has used the same tag in del.icio.us and Flickr), its final weight is the sum of all weights. These final lists often have a characteristic long-tail: For most users, there are many categories that appear with a weight of only 1 or 2. Since these categories are the product of a tag with a low frequency, and therefore do not necessarily correspond to something which the user is particularly interested in, they are filtered out - the final category list contains only categories with a weight above the average for that user.

Finallly, Stage 3 constructs the RDF profile of interest using the FOAF `interest` property to link the person to each of the Wikipedia categories. Figure 3 presents a partial 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 del.icio.us tag-cloud and correlated with the appropriate Wikipedia categories. Such terms are often related by a common super category such as "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 correlated to the category "Church Architectures", a match that would not be possible without semantic techniques.

4 Experiment

To build a suitable test-set to evaluate our semantic profiling architecture, we bootstrapped our system with a list of 667,141 del.icio.us users obtained in previous work [4]. For each user that specified an account url in their del.icio.us profile, we queried the Social Graph API to find all other accounts held. By filtering out those who also held a Flickr account, as well as those with low activity (i.e. less than 50 distinct tags in del.icio.us and Flickr), we obtained a final list of 1,392 users. For each individual, a complete history of their tagging activity was harvested using the Data Collection



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

Fig. 3. An example FOAF file. Its contents relates to the users tag clouds, and the categories matched are represented in the Wikipedia hierarchy

Module. The table below provides a summary of the data collected. In general, users had posted more information in Flickr, using a wider variety of tags.

Del.icio.us		Flickr	
Total Posts	1,134,527	Total Posts	2,215,913
Distinct Tags	138,028	Distinct Tags	307,182

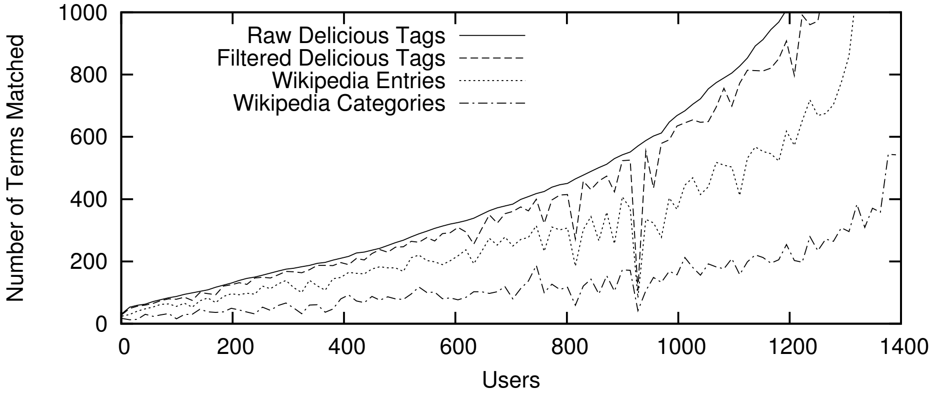
After collecting the user tag-clouds, they were filtered and then passed to the Profile Builder which constructed a list of Wikipedia categories describing their interests (see section 3).

5 Evaluation of Results

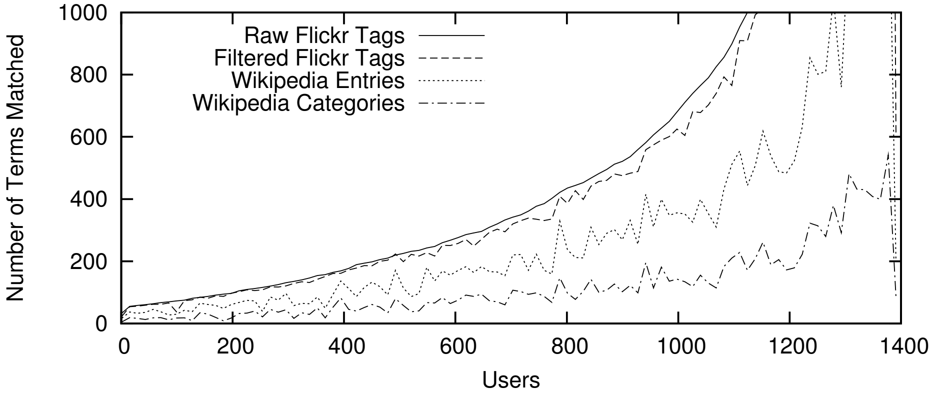
We present and evaluate our semantic profiling architecture in four ways: (1) the performance of the Tag Filtering and mapping to Wikipedia entries, (2) the difference between the most common categories (or interests) in del.icio.us and Flickr, (3) the amount learnt from merging profiles from the two folksonomies, and (4) the accuracy of the matching of tags to Wikipedia categories (concepts).

5.1 Tag Filtering and Category Matching

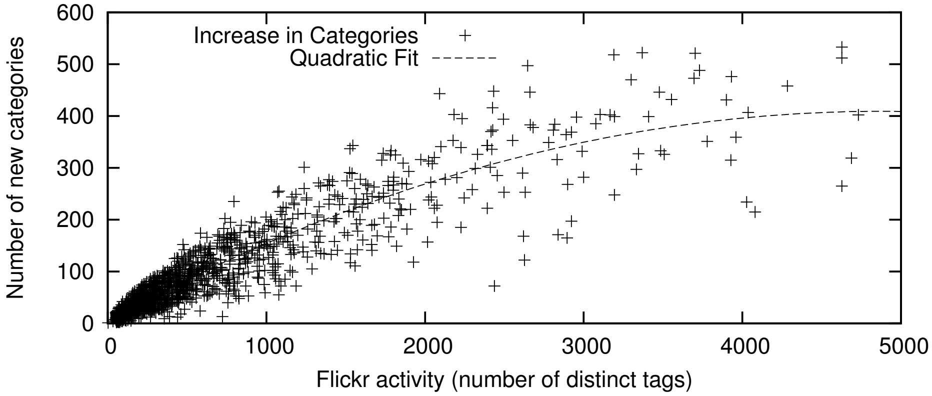
The process of correlating a raw user tag to a Wikipedia category follows three steps: i) first the tag is filtered ii) it is matched to a Wikipedia page, and iii) a suitable category is selected. The table below provides as summary of how many terms were matched during the filtering step, the page matching step, and the category selection step.



(a) Matching del.icio.us tags to Wikipedia ontology. Graph shows how many tags the user had in the raw tag cloud, how many tags were filtered, how many corresponded to a Wikipedia entry, and finally how many categories were selected to represent the given tag cloud.



(b) Same as graph above but for Flickr tag clouds.



(c) Category increase from adding Flickr to del.icio.us. Graph shows that the higher the user activity is in del.icio.us or in Flickr, the more new Wikipedia categories are found.

	del.icio.us		Flickr	
	Average	Std. Dev.	Average	Std. Dev.
Tag Filtering	90.6%	6.8%	90.7%	8.0%
Wikipedia Page Matching	70.6%	9.6%	59.1%	12.2%
Wikipedia Category Matching	40.9%	7.2%	41.5%	7.5%

On average, 40.9% of del.icio.us tags and 41.5% of Flickr tags were correlated to a Wikipedia category (see 5.4). Figures 4(a) and 4(b) provide detailed plots representing each of these steps. For every user, four points are plotted corresponding to the number of terms matched at each stage (including the number of raw tags). For both del.icio.us and Flickr, the general trends are approximately the same, but there are some anomalies. For example, user 928 has 571 del.icio.us tags, yet only 7 were matched with a Wikipedia category. On closer examination of this user's activity, we discovered that many posts were made using a single tag that itself was a concatenation of many tags. For example, popular tags were `culture.humor`, `politics.party`, and `tech.computers`.

5.2 Global Category View

To understand the difference in what can be learnt from a user's del.icio.us and Flickr activity, we generated a global category frequency table. Each time a tag was matched to a category, we increment the global frequency by the number of times that tag was used. The following table shows the top 15 categories found in del.icio.us and Flickr.

del.icio.us				Flickr			
Wikipedia Categ.	Total Freq.	Wikipedia Categ.	Total Freq.	Wikipedia Categ.	Total Freq.	Wikipedia Categ.	Total Freq.
design	69,215	blogs	68,319	travel	51,674	australia	51,617
music	45,063	photography	41,356	london	46,623	festivals	42,504
tools	35,795	video	34,318	music	40,943	cats	38,230
arts	29,966	software	28,746	holidays	37,610	family	37,100
maps	26,912	teaching	22,120	japan	36,513	concerts	35,374
games	21,549	how-to	19,533	surnames	34,947	washington	33,924
technology	18,032	news	17,737	given names	32,843	dogs	32,206
humor	15,816			birthdays	22,290		

These results are a good indication of the types of interest one can learn from the two different domains. Del.icio.us tells us about the bookmarking habits of the user, and subsequently, the topics they are interested in reading about on the Web. For example, `design`, `software`, and `humor` account for many of the posts made. In Flickr, the tags tell us more about locations and events. This shows that it is very likely to learn about other user interests when such different folksonomies are correlated.

5.3 Learning more about users

One central argument behind our approach is that different online profiles for an individual tell us different things about what that person is interested in. In Section 5.2 (above), we summarised the different types of categories we learnt from del.icio.us and Flickr. To evaluate this at a user level, we consider the difference in profiles that would be generated if only their del.icio.us tag-cloud is used versus a profile generated from del.icio.us and Flickr. The underlying hypothesis is that one should increase the number of categories found by including their Flickr profile. On average, 94.8 new categories were found (15 were with above average frequency and thus added to the profile ontology) for the individuals in our test-set, with a standard deviation of 100. This high variation is accounted for by the fact that many users have a high activity in del.icio.us, but a low activity in Flickr (or vice versa). To account for this, the plot given in Figure 4(c) shows the number of new categories added for a user as a function of the their flickr activity.

5.4 Evaluating category matching

To evaluate the approach in terms of how well it identifies the relevant Wikipedia categories from tags, we generated a random sample of 100 users, and randomly selected from their tag clouds 1 tag from del.icio.us and 1 from Flickr. The following procedure was then followed:

1. Open the del.icio.us and Flickr pages for the user.
2. Open the list of resources that the user tagged with the randomly selected tag.
3. Establish from the content of these resources, as well as from the other tags in those postings, what the interest is likely to be.
4. Check if that interest is in the list of Wikipedia categories selected for that user.

The table below summarises the results of this evaluation.

	Represented with correct Wikipedia concepts	Unresolved	Ambiguous
del.icio.us	66	20	14
Flickr	63	25	12

As the table shows, about 13% of the tags lead to Wikipedia disambiguation pages, and thus were discarded (see Future Work section). On average, 64.5% of the tags were correctly represented with a Wikipedia Category. However, 22.5% of the tags were not mapped to any Wikipedia concepts, a situation that arrises when the tag is not well covered in Wikipedia (i.e. no Wikipedia entry was found), or when there is no unique category for representing it (i.e. multiple categories, non of which with the same label as the given tag - section 3). As noted earlier, our system currently ignores highly ambiguous terms in Wikipedia, hence the true accuracy of the results according to the results above is 76.7% for del.icio.us and 71.6% for Flickr (average is 74.2%).

During this evaluation, two false positives were found (i.e. tag mapped to the wrong category). Those were the tags “oracle”, and “labrador”. The first was represented with the Wikipedia category “Divinity”, whereas the user was interested in Oracle the database. As for “labrador”, which was used to tag photos of a dog in Flickr, was incorrectly mapped to a city in Canada. To avoid such cases, a disambiguation process is required as explained in section 7.

6 Discussion

There were some interesting properties of the strategy reported here of using Wikipedia. The use of a rich source like Wikipedia means that tags that cannot be matched can still gain from matching related tags. For example, one user tagged a resource with “KL”.

This we were unable to match against Wikipedia, but other tags for similar resources by the user matched to “Kuala Lumpur” and “Malaysia”.

Some ambiguities are the result not of using ambiguous terms so much as the abstraction with which interests are represented, which may render the exact nature of the interest opaque. Again, other tags of the user may help resolve the abstraction down to a level which is meaningful. For instance, one person tagged a resource with “time”. This is over-general and not too helpful. However, when we looked at other tags for this user on comparable posts, it became clear that the particular interest of the user was astronomy, in which the concept of time plays a specific role. Both “time” and “astronomy” were selected for this user as *interests*, but it is clear that the real interest lies in the combination of the two.

At the other end of the abstraction spectrum, specific tags do not map easily onto concepts when considered as an expression of interest. Examples of specific tags include names of individual (non-celebrity) people, including friends and family of the tagger. These by their nature were more likely to turn up in Flickr than in, say, del.icio.us. Similarly, low-frequency tags were found to be less likely to indicate users interests, especially if the frequency did not grow over time. Again, exploiting the richness of the tag-cloud aided investigation of such anomalous issues, so for example one person tagged some contact “visa”, which is perhaps not best characterised as an ‘interest’, but a better expression of their interests – “travel” – emerged from their other tags.

In the current implementation, tags that lead to a Wikipedia disambiguation page (e.g. directory) are discarded and hence are not represented with an interest. As, noted, with the free text structure of tags, ambiguity and unclarity are endemic problems, and various methods for disambiguation have been proposed in the literature (see section 2). Some of these methods are based on clustering the whole collection of tags (e.g. [2]), or resources [13]), but such techniques are more suitable for static environments where the data do not change too often. In the world of tagging, this is hardly the case. Hence, we need less demanding methods to disambiguate tags, to cope with the highly dynamic nature of folksonomies, even if those methods could never be perfect (see next section).

An alternative approach to disambiguation is proposed in [23] where pairs of tags (e.g. “apple” and “computer”) are searched for in a collection of ontologies, and the ontology that contains both concepts will define the disambiguated domain. This approach has a number of limitations, such as when multi-domain ontologies exist (WordNet, Cyc), or when the tags are not found in any single ontology.

Nevertheless, representing user interests with an ontology enables us to benefit from the hierarchical structure when dealing with the user’s interests at different levels of granularity. For example, if someone is interested in “Visualbasic”, “Perl”, and “C++”, then one can infer that this person is into “Programming languages”. The hierarchy can show how general the user interest is, so one user may use the tag “music” very often, while another might tag with “jazz” or “Hip hop”, which are more specific concepts than “music”. People tag with different levels of specificity, and this usually reflects their level of expertise in the subject [8, 25]. More specific concepts (i.e. interests in our case) can be found at lower levels in the ontology hierarchy.

7 Future Work

According to Wikipedia, the majority of terms can be disambiguated (e.g. “iTunes” could be the device or the store, “furniture” is the object or the UK band). We will investigate using the *distance* between Wikipedia categories to disambiguate the tags. For example, if the user tagged a resource with “apple” and “computer”, then these will match to Wikipedia categories with the same labels, but there will be ambiguity to resolve. The path between these two concepts in the Wikipedia ontology is shorter if “apple” is matched to the technological concept, than if it was matched to the fruit. One approach is to select the category with the shortest path to represent this interest.

The use of an ontology will allow recommendation systems to find out how specific the user interest is, and use this information to fine tune recommendations. Inferring interests by analysing links or paths in the ontology can help uncover implicit interests. For example if the user is found to be interested in “Science fiction” and in “Books” then the system might assume that the user will be interested in science-fiction books. This can be refined further depending on, for example, places or authors of interest to the user. We plan to explore using the profiles of interest we produced for making cross-domain recommendations. In addition, we plan to expand this work to also include last.fm accounts, which we have already gathered for all the users in our current dataset.

Cross-domain interests could also be served, given the range of social networking sites and activities, by exploring the taggers social environment. Both Flickr and del.icio.us now allow users to form links with others (e.g. friends, groups). Such social links could be explored for further interest and recommendation analysis. We are currently collecting this information for the users in our dataset and will investigate whether there is a correlation between social links and user interests.

8 Conclusions

This paper investigated a novel idea of merging users’ distributed tag clouds to build richer profile ontologies of interests, using FOAF interest properties and Wikipedia categories. We experimented with over 1300 users with high activities in both del.icio.us and Flickr, and the result showed that on average 15 new concepts of interest were learnt for each users when expanding tag analysis to their tag cloud in the other folksonomy. We have also introduced a process to “clean” the data to maximise tag-cloud matching. Our initial evaluation showed that on average, 72% of the filtered tags have been correctly represented with a Wikipedia category (i.e identification of interest).

Acknowledgement

This research has been supported by the TAGora project funded by the Future and Emerging Technologies program (IST-FET) of the European Commission under the contract IST-34721. The information provided is the sole responsibility of the authors and does not reflect the Commission’s opinion. The Commission is not responsible for any use that may be made of data appearing in this publication.

References

1. M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *Proc. of Computer and Human Interaction (CHI), San Jose, CA*, 2007.
2. G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Proc. 17th Int. World Wide Web Conf., Edinburgh, UK*, 2006.

3. U. Bojars, J. G. Breslin, A. Finn, and S. 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.
4. C. Cattuto. Semiotic dynamics in online social communities. *The European Physical Journal C - Particles and Fields*, 46:33–37, 2006.
5. M. D. Choudhury, H. Sundaram, A. John, and D. Seligmann. Can blog communication dynamics be correlated with stock market activity? In *Proc. Int. Conf. Hypertext (HT08), Pittsburgh, PA, USA*, 2008.
6. G. Demartini. Finding experts using wikipedia. In *Proc. ExpertFinder Workshop, at ISWC, Busan, Korea*, 2007.
7. J. Diederich and T. Iofciu. Finding communities of practice from user profiles based on folksonomies. In *Proc. Workshop on Building Technology Enhanced Learning solutions for Communities of Practice, EC-TEL, Crete, Greece*, 2006.
8. S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32:198–208, 2006.
9. T. Gruber. Collective knowledge systems: Where the social web meets the semantic web. *Journal of Web Semantics*, 6(1), 2008.
10. M. Guy and E. Tonkin. Tidying up tags? *D-Lib Magazine*, 12(1), 2006.
11. C. Hayes, P. Avesani, and S. Veeramachaneni. An analysis of the use of tags in a log recommender system. In *Int. Joint Conf. Artificial Intelligence (IJCAI), Hyderabad, India*, 2007.
12. X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *Proc. 19th Int. World Wide Web Conf (WWW), Beijing, China*, 2008.
13. C. man Au Yeung, N. Gibbins, and N. Shadbolt. Tag meaning disambiguation through analysis of tripartite structure of folksonomies. In *Proc. Workshop on Collective Intelligence on Semantic Web (CISW), IEEE/WIC/ACM*, Los Alamitos, CA, USA, 2007.
14. C. Marlow, M. Naaman, danah boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proc. Int. Conf. Hypertext (HT06), Odense, Denmark*, 2006.
15. A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. *Computer Mediated Communication - LIS590CMC*, December 2004.
16. C. McCarthy. Myspace announces ’data availability’ project with yahoo, ebay, photobucket, twitter. CNET, http://www.news.com/8301-13577_3-9939286-36.html, 2008.
17. J. Mori, Y. Matsuo, M. Ishizuka, and B. Faltings. Keyword extraction from the web for FOAF metadata. In *Workshop on Friend of a Friend, Social Networking and the Semantic Web, Galway, Ireland*, 2004.
18. D. Morin. Announcing facebook connect. facebook developers, http://developers.facebook.com/news.php?blog=1&story=108, 2008.
19. Ofcom. Social networking: A quantitative and qualitative research report into attitudes, behaviours, and use. http://news.bbc.co.uk/1/shared/bsp/hi/pdfs/02_04_08_ofcom.pdf, 2008.
20. A. Passant and P. 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.
21. D. Silver. *Smart Start-ups: How to Make a Fortune from Starting Online Communities*. John Wily and Sons, Inc, 2007.
22. P. Singla and M. Richardson. Yes, there is a correlation - from social networks to personal behavior on the web. In *Proc. Int. World Wide Web Conf.*, Beijing, China, 2008.
23. L. Specia and E. Motta. Integrating folksonomies with the semantic web. In *Proc. 4th. European Semantic Web Conf. (ESWC), Innsbruck, Austria*, 2007.
24. M. Szomszor, I. Cantador, and H. Alani. Correlating user profiles from multiple folksonomies. In *Proc. Int. Conf. Hypertext (HT08), Pittsburgh, PA, USA*, 2008.
25. J. W. Tanaka and M. Taylor. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitve Psychology*, 23:457–482, 1991.