Correlating User Profiles from Multiple Folksonomies

Martin Szomszor
School of Electronics and
Computer Science
University of Southampton,
SO17 1BJ, UK
mns2@ecs.soton.ac.uk

Iván Cantador
Escuela Politécnica Superior,
Universidad Autónoma de
Madrid
Campus de Contoblanco,
28049, Madrid, Spain
ivan.cantador@uam.es

Harith Alani
School of Electronics and
Computer Science
University of Southampton,
SO17 1BJ, UK
ha@ecs.soton.ac.uk

ABSTRACT

As the popularity of the web increases, particularly the use of social networking sites and WEB2.0 style sharing platforms, users are becoming increasingly connected, sharing more and more information, resources, and opinions. This vast array of information presents unique opportunities to harvest knowledge about user activities and interests through the exploitation of large-scale, complex systems. Communal tagging sites, and their respective folksonomies, are one example of such a complex system, providing huge amounts of information about users, spanning multiple domains of interest. However, the current Web infrastructure provides no mechanism for users to consolidate and exploit this information since it is spread over many desperate and unconnected resources. In this paper we compare user tag-clouds from multiple folksonomies to: (a) show how they tend to overlap, regardless of the focus of the folksonomy (b) demonstrate how this comparison helps finding and aligning the user's separate identities, and (c) show that cross-linking distributed user tag-clouds enriches users profiles. During this process, we find that significant user interests are often reflected in multiple WEB2.0 profiles, even though they may operate over different domains. However, due to the free-form nature of tagging, some correlations are lost, a problem we address through the implementation and evaluation of a user tag filtering architecture.

Categories and Subject Descriptors

H.1.1 [**Systems and Information Theory**]: Information Theory; H.3.1 [**Content Analysis and Indexing**]: Linguistic Processing; H.3.5 [**Online information Services**]: Data sharing

General Terms

Design, Theory, Human Factors

HT'08, June 19-21, 2008, Pittsburgh, Pennsylvania, USA.

Keywords

folksonomy-alignment, tag-filtering, Web2.0, user profiling

1. INTRODUCTION

While the future evolution of the Web is 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. A recent UK study by Ofcom [14] found that over one fifth of UK adults have at least one online community profile (54% for individuals aged 16-24). Silver [16] predicts that by 2010, each of us will have between 12 and 24 online identities.

One significant catalyst underpinning the growth of the social networking phenomenon is increased connectivity, both in terms of the number of connected users and how pervasive such connections are: users can now view and publish rich multimedia content on a range of static and mobile devices. The future promises an even more connected world through the role out of municipal wifi networks, mobile broadband services, residential fibre optic backbones, and so on.

In addition to the increased connectivity, a number of technological advancements have made this new social web possible. WEB2.0 is a term often used to encapsulate this plethora of tools including wikis, blogs and folksonomies. Such tools are designed to promote creativity, collaboration, and sharing between users, and are responsible for some incredible feats of collaborative knowledge engineering, such as *Wikipedia* [1] (the de facto online encyclopedia), and *imdb* [2] (the biggest collection of information about movies and television shows in the world).

Against the background of increased connectivity and novel collaborative software tools, users of WEB2.0 are discovering new and exciting sites to meet emerging social demands: both music and video oriented sites, such as *last.fm* [3] and *youtube* [4], have engaged audiences through multimedia ex-

[1] http://www.wikipedia.org
[2] http://www.imdb.com
[3] http://www.last.fm
[4] http://www.youtube.com

periences, and socially focused sites, such as *Facebook* [5], have created new communication trends that go beyond conventional email and instance messaging. As a result, many users create and maintain profiles across different WEB2.0 sites, leaving a virtual trail of information strewn across the web revealing their tastes, interests, and activities. In order to exploit this data, for the purposes of personalised, multi-domain search and recommendation, alignment and consolidation is necessary. In fact, many aspects behind the future visions of the Web [2] rely heavily on connecting, understanding, and exploiting this vast array of data.

Tagging has proved to be a popular choice for WEB2.0 developers, supplying an intuitive and flexible mechanism to facilitate users in the organisation and retrieval of resources. Considering the wide adoption of tagging systems, and their ability to support many domains and resource types, we believe they play an important role in linking web data in meaningful ways. However, while tagging provides an excellent basis for users to organise and search content, the free-form nature of tagging leads to problems when large amounts of information are collated: syntactic and semantic differences in tagging habits mean closely related items are not always connected through a shared symbol.

In this paper, we describe our efforts to link-up user profiles across two popular, community driven tagging sites: *del.icio.us* [6], a site for bookmarking and sharing of Web resources, and *flickr* [7], a site for publishing and sharing photos. After harvesting and comparing the tagging history for many user accounts that have been correlated between both systems, we find that salient interests and activities are often prominent in profiles from both systems. However, due to the open and uncontrolled nature of community tagging, many correlations between user accounts are lost because tags do not match exactly. By using a number of term filtering processes, we are able to improve the alignment between an individual's tag cloud's, and improve the ability to distinguish them from others in the community for the purposes of account identification and verification.

Current social networking users are forced to create separate accounts to participate in multiple folksonomies. 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 homepage. Cross-linking these user profiles would have several advantages: From the user's perspective, it could reduce tag-cloud maintenance, and facilitate search and retrieval of tagged resources from multiple sites; From the system's perspective, bringing these profiles together enriches the knowledge about the individual users, which helps to improve personalisation and recommendation services. While in the future, we believe that users will control the ways in which their profiles are linked, current social networking sites do not share a policy to express this, and are unlikely to do so until some benefits have been demonstrated. Therefore, some this work is focussed on automatically identifying and linking these profiles in order to bootstrap an investigation into account correlation.

This paper is organised as follows: Section 2 provides a background on folksonomies and the study of community tagging behaviour and provides a motivating example. Section 3 explains our data gathering process and presents some initial analysis on the overlap that exists between user tagging folksonomies. Section 4 presents our filtering architecture before an evaluation is given in Section 5. Related work is discussed in Section 6 before conclusions and future work are presented in Section 7.

2. BACKGROUND AND MOTIVATION

In this Section, we provide a summary of collaborative tagging literature, revealing the current state of the art. We continue with a discussion of why it might be useful to connect user profiles distributed across multiple, folksonomy driven, sites. Example tagging history gathered from *del.icio.us* and *flickr* is used as a motivating example, highlighting some of the problems that arise when such folksonomies are connected.

2.1 Folksonomies

The term *folksonomy* was first coined by T. Vander Wal [18] to describe the taxonomy-like structures that emerge when large communities of users collectively tag resources. These *folk taxonomies* reflect a communal view of the attributes associated to items, essentially supplying a bottom-up categorisation of resources [9, 13].

Since individuals from different communities utilise different tags, often reflecting their degree of knowledge in the domain, folksonomies can support highly personalised searching and navigation. For example, an article in the social bookmarking site *del.icio.us* concerning web programming may have the tags `programming`, `ajax`, `javascript`, `tutorial`, and `web2.0`. With tags describing resources at varying levels of granularity, users may seek out their desired resources using terms they are familiar with.

As much as the popularity of WEB2.0 applications has grown, so to have research efforts to investigate, analyse, and understand the complex dynamics of community tagging. Research [8] has shown that tagging distributions tend to stabilise into power law distributions - providing enough users tag the resource. Over time, the most popular tags provide an emergent categorisation of resources, with many idiosyncratic tags appearing in the long-tail. As well as categorising resources, tag use can also be used to identify emergent communities of resources [4] that correspond to distinct tagging patterns with a specific meaning.

As well as investigating the large-scale emergent behaviour of collaborative tagging, other research has centered on understanding the human process of tagging, how tags are conceived, and how they are perceived. Marlow *et al* [12] examined usage patterns in the popular photo sharing site *flickr* to determine the user incentives for tagging. On top of the desire to tag for personal benefits (such as organisation and future retrieval of resources), the social networking element plays an important role: users can share pictures with others through the creation and subscription to groups, create networks of friends, and comment on other's photos.

In terms of the tags themselves, a number of classification schemes have been proposed. Through analysis of the social bookmarking site *del.icio.us*, Golder and Huberman [7] propose the following scheme:

[5] http://www.facebook.com

[6] http://del.icio.us

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

- Tags may be used to identify the topic of a resource using nouns and proper nouns such as `news`, `microsoft`, `vista`.

- To classify the type of resource (e.g. `book`, `blog`, `article`, `review`, `event`).

- To denote the qualities and characteristics of the item (e.g. `funny`, `useful`, and `cool`).

- A subset of tags, such as `mystuff`, `myphotos`, and `myfavourites`, reflect a notion of self reference, often used by individuals to organise their own resources.

- Much like self referencing tags, some tags are used by individuals for task organisation (e.g. `to read`, `job search`, and `to print`).

More abstractly, Coates [5] highlights the use of tags to place resources into categories, as opposed to classifying the resource directly. For example, the terms `blog` and `blogs` may not seem that different, but it is suggested that they show two subtle differences in the way a user perceives tags. Use of the `blog` tag suggests a direct classification (i.e. item x *is a* `blog`), and the tag `blogs` indicates a categorisation (i.e. item x is in the category of `blogs`). However, while this does highlight a difference in the a way a user percieves the tagging process, the resources reffered to are still closely related. Linguistic approaches have also been used in an attempt to understand the origin of a tag and what it represents in the human thought process. Veres [19] argues that in addition to taxonimc classification, tags can be used to describe functional properties (e.g. `shopping`) or resource attributes (e.g. `english`).

Even though tagging has proved extremely useful to users who expose large amounts of content, many problems have also been associated with its free-form nature. Principal among these are *polysemy*, *synonymy*, and morphologic variety [7]. Polysemy is common because many popular tags often have multiple meanings. For example, the tag `apple` is used frequently on *flickr* and *del.icio.us*, but could refer to the fruit or the computer company. Synonymy, or multiple words that have the same (or a very closely related) meaning, is also common because different users are likely to associate one particular word over another based on their own experiences and knowledge. Finally, even if a particular word has been agreed on, morphologic variety means that some discrepancies occur: users often use plural and singular forms interchangeably.

2.2 Multiple User Profiles

Web users are presented with a plethora of sites designed to satisfy many different usage scenarios. Users can organise and share bookmarks using *del.icio.us*, notify others of interesting articles using *digg*, communicate and share resources with friends using *Facebook*, organise and share photos with *flickr*- the list is almost endless. In many cases, different aspects of a user's personality are exposed through different sites, such as music tastes through *last.fm* and movie preferences through *imdb*. It is also possible to tailor the information users expose to fit particular domains. A user page on *Facebook* (used to communicate with friends) will often contain different information than the same user's page on *linkedin* [8] (a business oriented social networking site).

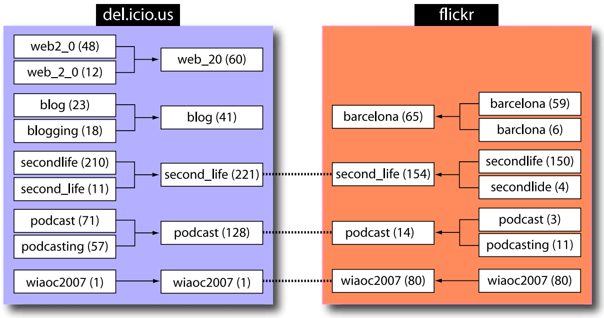


Figure 1: A set of sample tags (and their frequencies) used by an individual in *flickr* and *del.icio.us*.

This vast array of data published about individuals provides a unique opportunity to construct complex profiles that represent many aspects of a user's personality including their topics of interest, places of importance, significant events, and the individuals with whom they are virtually connected. Collecting and understanding this information will yield the following benefits:

- *To help understanding the dynamics underlying the tagging process from a global (multi-domain) perspective:* For example, it would be useful to understand how particular tags arise and propagate across different online communities.

- *To facilitate cross-folksonomy, multi-domain searching:* This would allow users to automatically search across a range of different sites for diverse media types such as videos, photos, articles, and reviews.

- *To provide personalised recommendation based on popular user tags:* By examining the tagging habits of a particular user, we can build complex profiles that represent the topics, people, and places a user is most interested in.

- *Provide cross-domain recommendations based on tags used in a different sites:* For example, it would be useful to recommend articles tagged in *del.icio.us* based on a user's *last.fm* or *imdb* profile.

To illustrate this concept, we examine the *del.icio.us* and *flickr* profiles for a particular individual who describes themselves as a Second Life resident, blogger, and podcaster from Barcelona, Spain. These profiles were correlated by examining the username, realname, and homepage referenced in both profiles. Many of the tags used reflect their areas of interest: `blog`, `secondlife`, and `podcast` being the most obvious. However, through closer examination, it is apparent that this user is not consistent when tagging resources in *del.icio.us* and *flickr*. Figure 1 illustrates this argument by showing a set of tags used by our example user in *del.icio.us* and *flickr*. Many of the popular tags are used in both sites, increasing our confidence that these two profiles represent the same person, but often a number of different variations

[8] http://www.linkedin.com

appear. For example, `blog` and `blogging` are used to describe similar sets of similar resources in *del.icio.us*. In this example, the tag `wiaoc2007` is used to denote an event (Webheads in Online Convergence) that took place in 2007. The photos tagged on *flickr* include slides from presentations and the *del.icio.us* bookmark points to the conference page. Through further investigation, we find the raw tagging information can be noisy and inconsistent - a property that does not lend itself to integration. Grammatical mistakes are often made; people tag concepts indistinctly in singular and plural form; they may add pronouns, adjectives, adverbs or prepositions to the main concept of the tag; and use non-alphanumeric characters.

In the next 4 Sections of this paper, we explain the process used to create a test-set of correlated *del.icio.us* and *flickr* accounts, present the overlaps that emerge between these folksonomies, describe a filtering architecture to improve the alignment between them, and evaluate it by comparing user overlap in user tag-clouds pre and post filtering.

3. DATA GATHERING

We begin this Section with a brief summary of the notation throughout this paper. This is followed by an explanation of the process we used to correlate different user account between *del.icio.us* and *flickr*, a demonstration and explanation of the folksonomy overlap, and a presentation of much user tag-clouds align.

3.1 Terminology and Notation

We adopt conventional notation to describe folksonomies. U, T, R are finite sets, whose elements are called *users*, *tags*, and *resources*. Since we are working with two separate tagging datasets (*del.icio.us* and *flickr*), we distinguish between them 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. Thus, we define the *del.icio.us* folksonomy as a tuple $\mathbb{F}^d := (U, T, R, Y^d)$ and the *flickr* folksonomy as a tuple $\mathbb{F}^f := (U, T, R, Y^f)$. With this view, no distinction is made between the users, tags and resources of *flickr* and *del.icio.us*.

3.2 User Correlation

To correlate a set of user accounts between *del.icio.us* and *flickr*, we bootstrap using a list of 667,141 *del.icio.us* account names obtained in previous research [3]. These account names provide a unique identifier to the individual's profile within *del.icio.us* site (e.g. `http://del.icio.us/username`). A similar identification procedure is used in *flickr*: users may select a username to distinguish themselves from other users (e.g. `http://www.flickr.com/people/username`). The first stage in the correlation process was to create a list of potential user matches simply by searching for *del.icio.us* usernames in the *flickr* site.

After discovering a candidate list of 232,391 usernames, we refined the matching further by comparing real name descriptions - In both *del.icio.us* and *flickr*, users have the option of filling a form with their real name. By examining all the users in our candidate list and keeping only those whose real name description matched *exactly*, as well as discarding those with low activity, a final list of 502 matching users was produced. While this approach was quite restrictive, e.g. the usernames had to match exactly, we wanted to maximise the probability that the two profiles refer to the same individual. A more accurate and scalable solution is discussed later in Section 7. A complete history of each user's tagging activity in both *del.icio.us* and *flickr* was harvested for analysis. The table below summaries the data collected:

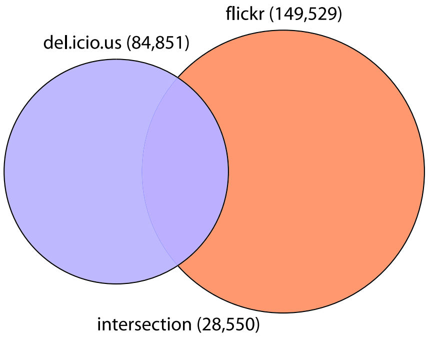


Figure 2: A Venn Diagram showing the intersection of tags between *del.icio.us* and *flickr*.

Posts	
del.icio.us	1,639,639
flickr	4,694,161
Distinct Tags	
del.icio.us	83,851
flickr	149,529
Users	
del.icio.us	502
flickr	502

3.3 Tag-Cloud Intersection

Using the data collected from *del.icio.us* and *flickr*, we are able to determine the tags used in both systems. Out of the $84,851$ distinct *del.icio.us* tags, and $149,529$ distinct *flickr* tags, $28,550$ are used in both systems, as depicted with the Venn diagram in Figure 2. Formally, we define the intersection I_t of a tag t as the set of users who have tagged resources in both *del.icio.us* and *flickr* with tag t:

$$I_t := \{u \in U \mid [(u,t,r_1) \in Y^d] \wedge [(u,t,r_2) \in Y^f] \wedge r_1 \in R \wedge r_2 \in R\}$$

Hence, we define the intersection weight i_t of a tag t as the sum of all users $i_t := |I_t|$ allowing us to build an *intersection tag-cloud* between *del.icio.us* and *flickr*. Figure 3 is an intersection tag-cloud where higher tag intersection weights depicted using a larger font. Due to space constraints, only the most popular tags are shown. From this tag-cloud, we see that tags the intersection of *del.icio.us* and *flickr* represents are range of tags: high-level classifications (e.g. `architecture`, `design`, `food`), dates (`2006` and `2007`), and functional descriptions (`shopping` and `cooking`) are good descriptions of user interests. Locations (`nyc` and `sanfrancisco`) and events (`christmas` and `conference`) provide good indications of prominent activities and places of importance.

3.4 User Tag-Cloud Alignment

To measure the alignment between two user tag-clouds, we measure the frequency of tags common to *del.icio.us* and

2006 2007 advertising airport animals apple architecture art baby beer bike blog book books building california camera canada car cars cat cats chicago christmas church coffee color comics computer conference cooking cute design desktop diy dog drawing email family fashion film fire firefox flash flickr food football france free friends fun funny gallery game games geek google graffiti green guitar halloween history home hotel house humor illustration internet iphone ipod itunes japan kids laptop library light linux logo london mac magazine map maps me media microsoft mobile money movie museum music nature news newyork nintendo nyc office osx painting paper party people phone photo photography photoshop podcast politics radio religion restaurant sanfrancisco school screenshot sculpture security shoes shop shopping software street subway sun technology television tools toys traffic train travel tree tv uk urban usa video wallpaper water weather web wedding windows wine work writing yahoo

Figure 3: Tag-cloud showing the intersection of tags for *del.icio.us* and *flickr*.

flickr. For a user u, we define the tag-frequency for a tag t for both *del.icio.us* $n_t^d(u)$ and *flickr* $n_t^f(u)$ as the number of times a resource r has been tagged with t by user u:

$$\begin{aligned} n_t^d(u) &:= \mid \{ r \in R \mid (u,t,r) \in Y^d \} \mid \\ n_t^f(u) &:= \mid \{ r \in R \mid (u,t,r) \in Y^f \} \mid \end{aligned}$$

We then define the set of tags used by a user u in both *del.icio.us* (T_u^d) and *flickr* T_u^f as:

$$\begin{aligned} T_u^d &:= \{ t \in T | (u,t,r) \in Y^d \wedge r \in R \} \\ T_u^f &:= \{ t \in T | (u,t,r) \in Y^f \wedge r \in R \} \end{aligned}$$

Hence, the total intersection frequency N_u for a user u is defined as the sum of the frequencies of all tags appearing in their *del.icio.us* and *flickr* tag clouds.

$$N_u := \sum_{t \in T_u^d \cap T_u^f} n_t^d(u) \; + n_t^f(u)$$

We compare this intersection frequency against the total number of tag assignments made. For *del.icio.us* and *flickr*, we define the set of couples (t,r) for a user u to represent the all tag assignments made:

$$\begin{aligned} A_u^d &:= \{ (t,r) \in T \times R \mid (u,t,r) \in Y_d \} \\ A_u^f &:= \{ (t,r) \in T \times R \mid (u,t,r) \in Y_f \} \end{aligned}$$

Hence, the total tag assignments for a user u is specified:

$$A_u := \mid A_u^d \mid \; + \; \mid A_u^f \mid$$

Figure 4 presents a plot of total tag assignments against total intersection frequency for each user tag-clouds in our sample dataset. The x axis shows the total tag assignments and the y axis shows the total intersection frequency. Essentially, this plot tells us as users tag more resource in *flickr* and *del.icio.us*, their intersection frequency will increase. Therefore, we have increasing confidence that two correlated profiles in *del.icio.us* and *flickr* refer to the same individual as their total intersection frequency increases. We will also use this as a baseline for an evaluation of our filtering algorithm, comparing the results after filtering to the raw tag-clouds.

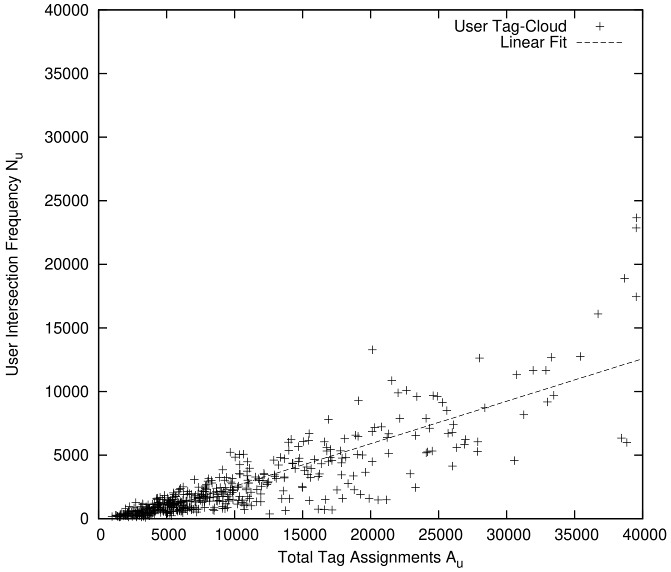


Figure 4: A plot showing the total intersection frequency between user tag-clouds in *del.icio.us* and *flickr*.

4. FILTERING ARCHITECTURE

In this Section, we present our filtering architecture. The aim is to transform a set of raw tags to set of filtered tags that are better aligned between folksonomies. As we highlighted in Section 2 through a motivating example, most users are inconsistent with respect to their tagging habits. This means that while it is possible to garner information from multiple folksonomy sites, such as *del.icio.us* or *flickr*, inconsistency will lead to confusion and loss of information when tagging data is compared. For example, if a user has tagged photos from a recent holiday in New York with `nyc`, but also bookmarked relevant pages in *del.icio.us* with `new_york`, the correlation will be lost.

4.1 Overview

Broadly, the filtering architecture can be divided into four sections, as depicted in Figure 5:

- **Tag Reader**
  This module reads different user tagging datasets (e.g. from *del.icio.us* or *flickr*) and converts them to a internal representation.
- **Tag Filtering Module**
  This module contains a number of software components responsible for different stages in the filtering process. They are split into two categories: *syntactic filters* (on the left) and *semantic filters* (on the right). Tags are maintained, merged, or discarded according to different morphological filters.

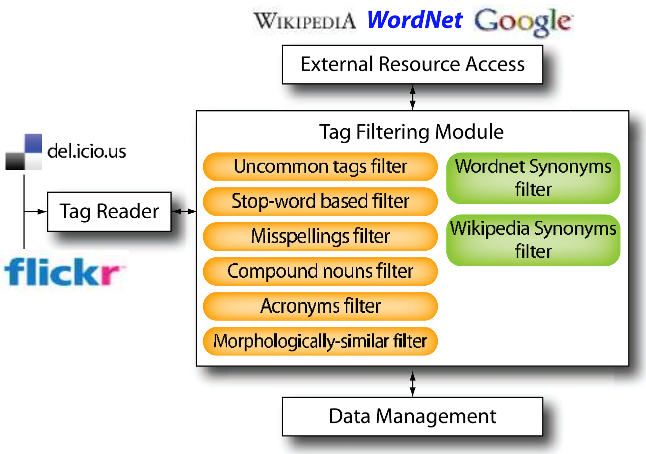


Figure 5: The tag filtering architecture

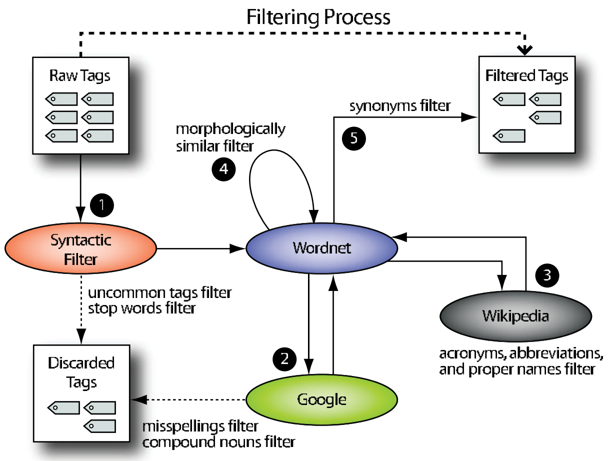


Figure 6: The tag filtering process

- **External Resource Access Module**
  This module provides a communication portal to external knowledge resources such as *Wikipedia*, *Wordnet* and *Google*.

- **Storage Module**
  The storage module supplies a database for tags and also manages the results from the various filtering steps.

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) that correspond to an agreed *intermediate* representation. This is achieved by correlating tags to entries in two large knowledge resources: *Wordnet* [6] and *Wikipedia*. Wordnet is a lexical database and thesaurus that groups English words into sets of cognitive synonyms called *synsets*, provides definitions of terms, and models various semantic relations between synsets. Hypernym relations for nouns and verbs (e.g. a dog *is a* carnivor, mamal, and animal) are also modeled, allowing lexical terms to be compared against a broad taxonomy.

Wikipedia is a multilingual, open-access, free-content encyclopedia on the Internet. Using a WIKI style of collaborative content writing, *Wikipedia* has grown to become one of the largest reference Web sites with over 75, 000 active contributors, maintaining approximately 9, 000, 000 articles in over 250 languages[9]. *Wikipedia* contains collaboratively generated categories that classify and relate entries, and also supports term disambiguation and dereferencing of acronyms.

4.2 Filtering Process

Figure 6 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:

4.2.1 Step 1: Syntactic Filtering

After the raw tags have been loaded by the *Tag Reader*, 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, as specified in the table below. For example, the tag *Zürich* is converted to *Zurich*.

Pre-filtering	Post-filtering
á, à, â, ã, ä, å	a
é, è, ê, ë	e
í, ì, î, ï	i
ó, ò, ô, õ, ö, ø	o
ú, ù, û, ü	u
ý, ÿ	y
ç	c

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 set a threshold manually discarding any unpopular tags. Finally, common stop-words, such as pronouns, articles, prepositions, and conjunctions are discarded. After syntactic filtering, tags are passed to the Wordnet module. If the tag has an exact match in Wordnet, we pass it directly to the set of filtered tags to avoid unnecessary processing.

4.2.2 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` by Alice featured in Section 2. 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

[9] As of February 2008

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 folksonomies: 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`. This mechanism works well for compound nouns with 2 terms, but is likely to fail if more than 2 terms are used. For example, the tag `unitedkingdomsouthampton` will not return any results from Google. However, by caching previous lookups, and matching the first shared characters of the tag string, we are able to split the tag into a prefix (previously resolved by Google) and a postfix. A second lookup will then be made using the postfix to see if further matches can be found. Compared to a bespoke string-splitting heuristic, this process has a low computational cost. Some examples of long compound nouns found in our sample dataset are `war of the worlds`, `lord of the rings`, and `martin luther king jr`.

After using Google to check for compound nouns and misspellings, the results are validated against Wordnet. Any unmatched or unprocessed tags are added to the discard pile.

4.2.3 Step 3: Wikipedia Correlation

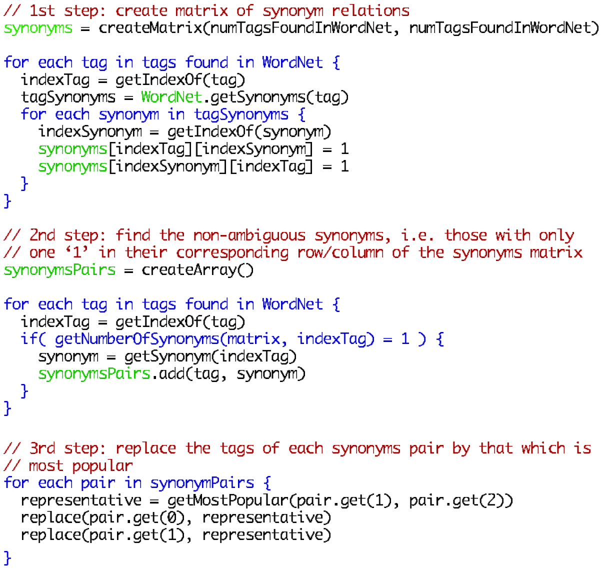
Many of the popular tags appearing in communal tagging systems do not appear in grammatical dictionaries, such as *Wordnet*, because they correspond to nouns (such as famous people, places, or companies), contemporary terminology (such as `web2.0` and `podcast`), or are widely used acronyms (such as `tv` and `diy`). In order to provide an agreed representation for such tags, we correlate tags to their appropriate *Wikipedia* entry. 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 accomodate new terminology. For example, *Wikipedia* contains extensive entries for terms such as `web2.0`, `ajax`, and `blog`.

4.2.4 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[10], we reduce morphologically similar tags to a single tag. The shortest term in *Wordnet* is used as the representative term.

4.2.5 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 pseudocode listed in Figure 7 explains the merging process. In the first stage, a matrix of synonyms is created by using *Wordnet*. In the second stage, we find each non-ambiguous synonym, and finally, stage three replaces each of the synonym pairs with the term that is most popular.



```
// 1st step: create matrix of synonym relations
synonyms = createMatrix(numTagsFoundInWordNet, numTagsFoundInWordNet)

for each tag in tags found in WordNet {
  indexTag = getIndexOf(tag)
  tagSynonyms = WordNet.getSynonyms(tag)
  for each synonym in tagSynonyms {
    indexSynonym = getIndexOf(synonym)
    synonyms[indexTag][indexSynonym] = 1
    synonyms[indexSynonym][indexTag] = 1
  }
}

// 2nd step: find the non-ambiguous synonyms, i.e. those with only
// one ‘1’ in their corresponding row/column of the synonyms matrix
synonymsPairs = createArray()

for each tag in tags found in WordNet {
  indexTag = getIndexOf(tag)
  if( getNumberOfSynonyms(matrix, indexTag) = 1 ) {
    synonym = getSynonym(indexTag)
    synonymsPairs.add(tag, synonym)
  }
}

// 3rd step: replace the tags of each synonyms pair by that which is
// most popular
for each pair in synonymPairs {
  representative = getMostPopular(pair.get(1), pair.get(2))
  replace(pair.get(0), representative)
  replace(pair.get(1), representative)
}
```

Figure 7: Pseudocode for the tag merging by synonym algorithm

5. EVALUATION

We evaluate our work in two ways: (a) by measuring the improvement in tag-cloud alignment through each of our filtering processes, (b) by measuring the similarity between user tag-clouds in *del.icio.us* and *flickr* as means to correlate profiles.

5.1 Tag Filtering

The focus of our work is not to better align the *del.icio.us* and *flickr* folksonomies at a global level, rather, we aim to bring user tag-clouds that have been constructed in separate folksonomies closer together. The hypothesis we test is that better quality tag-clouds, i.e. those which terms have been filtered and modified to an intermediate, agreed representation, results in user tag-clouds that are more closely connected. Therefore, we measure the relative increase in alignment between user tag-clouds in *flickr* and *del.icio.us* at each step of the filtering process. For the purposes of comparison, we define two measures that reflect the alignment made between a user's *flickr* and *del.icio.us* tag-clouds: The *assignment intersection ratio* (α_u) for a user is their intersection frequency divided by the total tag assignments made:

$$\alpha_u := \frac{N_u}{A_u}$$

The *tag intersection ratio* (β_u) measures the number of distinct tags that feature in both the user's *flickr* and *del.icio.us* profiles divided by the total distinct tags:

$$\beta_u = \frac{|T_u^d \cap T_u^f|}{|T_u^d| + |T_u^f| - |T_u^d \cap T_u^f|}$$

We applied our filtering algorithm to each of our 502 test user's *del.icio.us* and *flickr* tag-clouds, calculating α_u and

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

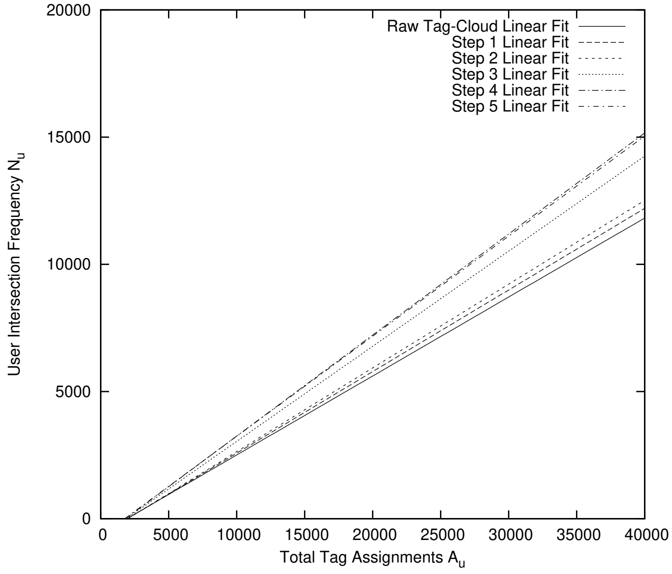


Figure 8: How each filtering step effects the user intersection frequency

β_u for each filtering step. A summary of the results follows, showing the mean and variance of α_u and β_u:

	Assignments		Tags	
	$\bar{\alpha_u}$	σ^2	β_u	σ^2
Raw	0.328	0.0524	0.0537	0.00056
Step 1	0.336	0.0558	0.0547	0.00058
Step 2	0.348	0.0635	0.0561	0.00062
Step 3	0.407	0.1107	0.0628	0.00076
Step 4	0.449	0.1265	0.0720	0.00091
Step 5	0.451	0.1273	0.0724	0.00091

In general, our filtering architecture increases both the assignment intersection ratio and the tag intersection ratio. On average, the number of overlapping tags (β_u) was increased by roughly 2% (from 5% to 7%) after the application of all filtering steps. The assignment intersection ratio has increased from 32% to 45%, on average. This means that small increase in the number of aligned tags can significantly increase the number of aligned tag assignments because many of the overlapping tags have a high frequency.

The plot given in Figure 8 provides a visual representation of how each filtering step effects the tag alignment of users. This plot uses the same axis as the one presented earlier in Section 3 so we can compare the relative increase given by each filtering step. For clarity, individual points are removed and a line of best fit indicates the approximate trend of each filtering step. Through inspection of this graph, it is apparent that the largest increase in alignment occurs between filtering steps 3 and 4. Step 4 corresponding to the alignment of tags to *Wikipedia* (including acronym resolution).

5.2 User Correlation

The second part of our evaluation investigates the similarity between user tag-clouds in *del.icio.us* and *flickr*, and how they might be used to correlate user accounts. For this experiment, we measure the similarity between a *del.icio.us* tag-cloud belonging to user i, and a *flickr* tag-cloud belonging to user j, using cosine similarity:

$$sim(i,j) = \frac{\sum_{t \in T} n_t^d(i) \ \cdot \ n_t^f(j)}{\sqrt{\sum_{t \in T} n_t^d(i)^2 \ \cdot \sum_{t \in T} n_t^f(j)^2}}$$

We performed this test for each of the 502 users in our test-set, comparing their *del.icio.us* tag-cloud to every other user's *flickr* tag-cloud, recording the most similar tag-cloud found, the average similarity to each of their neighbours, and the correlation to their own *flickr* tag-cloud. This test was performed using the raw tag-clouds harvested from their profiles, and aligned tag-clouds produced by our tag filtering process.

The plot in Figure 9 shows the results obtained using raw tag-clouds. Each point represents a single user, with a y-value calculated by subtracting the cosine similarity of their nearest *flickr* neighbour from the cosine similarity of their own *flickr* tag-cloud. With this representation, users with a point on the y-axis above zero correspond to user who's own *flickr* tag-cloud is more similar than any of the other 501 test users. The greater the y-value, the more distinct the user's own *flickr* profile is. The horizontal line just above 0 on the y-axis corresponds to the average similarity to all other users.

The plot in Figure 10 shows the same projection, this time with tag-clouds that have been filtered. The principal difference between the two is that the majority of points are below zero before any filtering, and above zero after filtering. These results indicate that while tag filtering only produces a small increase in the tag-cloud overlap between a user's *del.icio.us* and *flickr* tag-clouds (see Section 5.1), it is significant enough to make them stand-out from their neighbours. This technique that would be useful when trying to verify an individual's accounts when given a set of candidate profiles.

6. RELATED WORK

To aggregate users resources and content that are distributed across different WEB2.0 sites, Iturrioz *et al* [10] propose the *TAGMAS* (TAG Management System) architecture: a federation system that supplies a uniform view of tagged resources distributed across a range of WEB2.0 platforms. The TAGMAS system addresses the problem that users do not have consistent view of their resources or a single query end-point with which to search them. The TAGMAS architecture is based on a tagging ontology, proposed by Knerr [11], the provides a homogeneous representation of tags and tagging events. By aggregating user tagging events that span multiple sites, such as *flickr* and *del.icio.us*, it is possible to query TAGMAS using SPARQL [15], enabling users to find resources distributed across many sites by their tags, the date when tagged, which site they were tagged in. However, this work does not focus on the issues arising from free-form nature of tagging systems: the organisation and implementation of a consistent tagging schema is imputed on the user. We have shown, through the investigation of real-world tagging data, that users tend to change their tagging habits indiscriminately, using different schemas even within the same folksonomy.

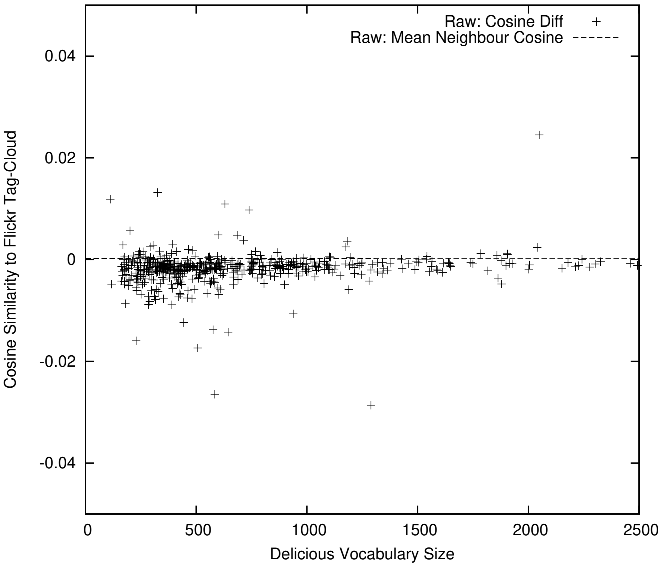


Figure 9: Raw cosine similarity measure between *del.icio.us* and *flickr* tag-clouds

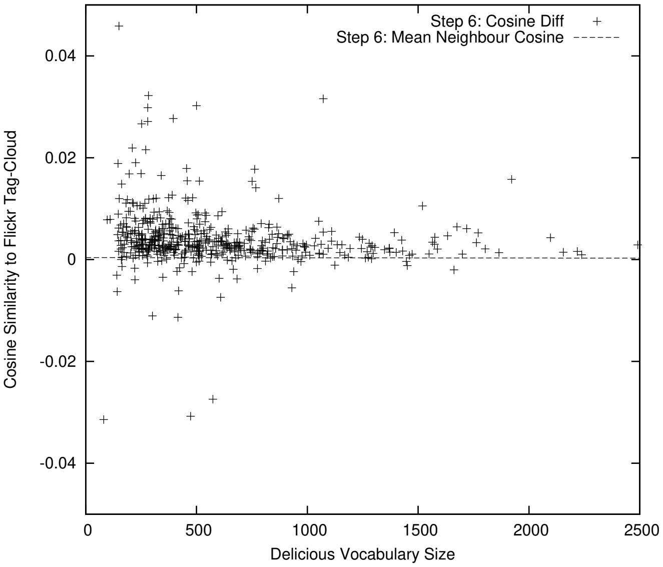


Figure 10: Post-filtering cosine similarity measure between *del.icio.us* and *flickr* tag-clouds

Specia and Motta [17] have also tackled the problem of integrating folksonomies. Using data collected from *del.icio.us* and *flickr*, they use co-occurrence analysis and clustering techniques to construct meaningful groups of tags that corresponds to concepts in an ontology. Their focus is primarily on understanding the relationships between tags, in particular synonyms and different levels of granularity. By exploiting external resources, such as *Wikipedia*, *Wordnet*, and semantic web ontologies, meaningful relationships can be established between such tag groups.

In terms of social linking, Google's OpenSocial[11] API is the most promising step towards a vision of highly interconnected social networking sites. The aim of OpenSocial is not to provide a universal API coverage: many of the sites provide diverse functionality that would be difficult to abstract through a common interface. Instead, the focus of OpenSocial is on the common uses of social networking sites, namely friend connections (social graph), and activities (so users can notify others when they have posted a new blog or review). Already, the OpenSocial API is being exploited to build application that promote sharing of resources and recommendations between friends. By subscribing to the OpenSocial API, it is possible to connect users even though their connection may exist across a variety of social networking sites. As of February 2008, 76 social networking sites, including *mySpace*, *bebo*, and *orkut*, have subscribed to the OpenSocial API.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a method for correlating user accounts between *del.icio.us* and *flickr*, providing us with a large enough test-set to investigate the overlaps that occur between folksonomies. Through this investigation, we discovered that prominent user interests, important locations, and events, are often reflected in the intersection between tag-clouds, irrespective of the focus of the folksonomy. Such correlations could be exploited in the future to build a complex, inter-connected network of user interests and past activities. However, because the uncontrolled nature of tagging results in inconsistencies between folksonomies, many of these correlation will be lost unless some intelligent tag manipulation is performed.

Through evaluation, we have shown that our filtering process does increase the alignment between user tag-clouds, but often by only a small proportion. The most significant increases occurred when tags were grounded to an agreed representation, in this case, *Wordnet* and *Wikipedia*. However, such a small increase is sufficient to draw a user's tag-cloud away from those of their neighbours, supporting the claim that tag-cloud similarity can be used to measure the likelihood that two account have been correctly correlated.

While our filtering architecture does cater for synonymy, morphologic variety, and the use of compound nouns, it does not take polysemy into consideration. For that, we require more sophisticated techniques to disambiguate the use of a tag. Clustering techniques [20, 1] have proven to be a useful tool approach for solving this problem, and will be incorporated in future work.

Even though a naive user correlation approach was used, it was suitable for this work since it provided us with a large enough dataset with which we could be confident that most user account were accurately correlated. To create a larger, and more accurate test set, we intend to match user accounts based not on string matching of their usernames and real names, but by examining the user's link structure. When creating accounts in most sites, it is common for a user to register their homepage. By using a reverse lookup, provided by search engines such as altavista[12] and Google, it is possible to find all pages that link to the user's homepage. By filtering these hits, it is possible to find accounts they have registered on other systems, such as *last.fm* and *flickr*. This method of account correlation is especially important in the next stage of our research because we intend to incor-

[11] http://code.google.com/apis/opensocial/

[12] http://www.altavista.com

porate more user accounts than just *flickr* and *del.icio.us*: the future vision is one where many different profiles are linked together.

Building a larger, multi-folksonomy, test set will also facilitate a more in-depth investigation of social networking properties. By examining the tagging activity of users within explicitly defined groups, it might be possible to determine representative tags that describe that group of people, as well as suggest new groups the user was not aware of but it likely to be interested in.

8. ACKNOWLEDGEMENTS

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, and the Spanish Ministry of Science and Education (TIN2005–6885). 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.

9. REFERENCES

[1] G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, 2006.

[2] T. Berners-Lee. Giant global graph, November 2007. `http://dig.csail.mit.edu/breadcrumbs/node/215`

[3] C. Cattuto. Semiotic dynamics in online social communities. *The European Physical Journal C - Particles and Fields*, 46(Volume 46, Supplement 2 / August, 2006and Fields):33–37, August 2006.

[4] C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Emergent community structure in social tagging systems. In *Proceedings of the European Confeence on Complex Systems*, Dresden, Germany, October 2007.

[5] T. Coates. Two cultures of fauxonomies collide, Jun 2005. `http://www.plasticbag.org/archives/2005/06/two_cultures_of_fauxonomies_collide/`

[6] C. Fellbaum, editor. *WordNet: an electronic lexical database*. Massachusetts: The MIT Press, 1998. p.423.

[7] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[8] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 211–220, New York, NY, USA, 2007. ACM.

[9] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11, Apr 2005. 10.1045/april2005-hammond.

[10] J. Iturrioz, O. Diaz, and C. Arellano. Towards federated web2.0 sites: The tagmas approach. In *Tagging and Metadata for Social Information Organization Workshop, WWW07*, 2007.

[11] T. Knerr. Tagging ontology - towards a common ontology for folksonomies, 2007. `http://tagont.googlecode.com/files/TagOntPaper.pdf`

[12] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM.

[13] A. Mathes. Folksonomies – Cooperative Classification and Communication Through Shared Metadata, December 2004.

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

[15] E. Prud'hommeaux and A. Seaborne. SPARQL query language for RDF. W3c recommentation, W3C, January 2008.

[16] D. Silver. *Smart Start-ups: How to Make a Fortune from Starting Online Communities*, page 5. John Wiley and Sons, Inc., 2007.

[17] L. Specia and E. Motta. Integrating folksonomies with the semantic web. *The Semantic Web: Research and Applications*, pages 624–639, 2007.

[18] T. Vander Wal. Folksonomy definition and wikipedia, November 2005. `http://www.vanderwal.net/random/entrysel.php?blog=1750`

[19] C. Veres. The language of folksonomies: What tags reveal about user classification. *Natural Language Processing and Information Systems*, pages 58–69, 2006.

[20] C. M. A. Yeung, N. Gibbins, and N. Shadbolt. Understanding the semantics of ambiguous tags in folksonomies. In P. Haase, A. Hotho, L. Chen, E. Ong, and P. C. Mauroux, editors, *Proceedings of the International Workshop on Emergent Semantics and Ontology Evolution (ESOE2007) at ISWC/ASWC2007, Busan, South Korea*, November 2007.