Social Networks and Semantics

Ioan Toma
STI Innsbruck, University of Innsbruck, Innsbruck, Austria
ioan.toma@sti2.at

James Caverlee
Department of Computer Science, Texas A&M University, College Station, TX, USA
caverlee@cs.tamu.edu

Ying Ding
School of Library and Information Science, Indiana University, Bloomington, IN, USA
dingying@indiana.edu

Elin K. Jacob
School of Library and Information Science, Indiana University, Bloomington, IN, USA
ejacob@indiana.edu

Erjia Yan
School of Library and Information Science, Indiana University, Bloomington, IN, USA
eyan@indiana.edu

Staša Milojević
School of Library and Information Science, Indiana University, Bloomington, IN, USA
smilojev@indiana.edu

ABSTRACT

This chapter discusses the relation between Social Networks and Semantics – two areas that have recently gathered a lot of attention from both academia and industry. We show how synergies between these two areas can be used to solve concrete problems, and we describe three approaches that demonstrate the potential for interconnecting these technologies. The first approach focuses on the semantic profiling of social networks. More precisely, we study the characteristics of large online social networks through an extensive analysis of over 1.9 million MySpace profiles in an effort to understand who is using these networks and how they are being used. We study MySpace through a comparative study over three distinct but related facets: the *sociability of users* in MySpace; the *demographic characteristics* of MySpace users; and the *text artifacts* of MySpace users. The second approach to interconnecting social networks and semantics focuses on a solution for mediating between social tagging systems. The Upper Tag Ontology (UTO) is proposed to integrate social tagging data by mediating between related social metadata schemes. We discuss how UTO data can be linked with other social metadata (e.g., FOAF, DC, SIOC, SKOS), how to crawl and cluster tag data from major social tagging systems, and how to integrate data using UTO. The third approach discusses the use of social semantics to qualitatively improve the task of service ranking. We explore the idea of using social annotation technologies for ranking web services; and we show how such an approach can be implemented using information provided by Delicious, one of the largest social networks.

1. INTRODUCTION

Online communities are the fastest growing phenomenon on the Web, enabling millions of users to discover and explore community-based knowledge spaces and to engage in new modes of social interaction. Web 2.0 sites such as Facebook[1], MySpace[2], Delicious[3], YouTube[4], Yahoo! Answers[5], and LinkedIn[6] have grown tremendously in the past few years, garnering increased media and popular attention. The result of this increased awareness is that the Web is socially linked more strongly now than ever before.

Generally speaking, Web 2.0 technologies are transforming the Web environment from a simple repository for documents into the Social Web, a communal platform for connecting people and sharing information. The phrase *Social Web* was introduced in 1998 by Peter Hoschka (1998), who wanted to stress the social functions and capabilities of the Web medium. According to Wikipedia, the Social Web is a global and open distributed data sharing network that links people, organizations and concepts. The Web 2.0 environment is the venue for the Social Web and provides platforms and technologies (e.g., wikis, blogs, tags, RSS feeds, etc.) that facilitate online collaboration and communication. Online social networking is part of the Web2.0 being defined according to Wikipedia as having a core focus on building and reflecting of social networks and social relations among who share interests and/or activities.

Online publishing in the Web 2.0 environment has become so easy that anyone who can write or type can publish on the Web. This revolution has stimulated an ever-growing number of ordinary users, many of them teenagers or seniors, to become involved in Web communication. One of the newest and easiest ways for these users to contribute to the Social Web is through the process of tagging. Tagging is a means for users to add keywords to resources as typed hyperlinks and, cumulatively, reflects community efforts to organize and share information resources. The growing popularity of tagging is furthering the evolution of the Web from a simple repository for hyperlinked documents to a typed hyperlinked Web of data.

As online social networks continue to grow and evolve, an important challenge we face is how to maintain the incredible success of Web 2.0. There is a growing need to understand this new social phenomenon; to understand the processes by which communities come together, attract new members, and develop over time; and to understand what it takes to empower online communities with the ability to gather and retain a core of actively participating members (Backstrom et al., 2006; Coleman, 1990).

Another challenge that we are facing is the increasing heterogeneity and growing amount of data, numbers of resources and users on the Web. Data mediation and data integration have been central concerns of IT for decades (Batini, Lenzerini, & Navathe, 1986; Rahm & Bernstein, 2001). With the advent of the Web, interest in these issues has exploded. Currently, there is a focus on providing machine supported meditation on the Web (Antoniou & Harmelen, 2004; Berners-Lee, Hendler, & Lassila, 2001) through the medium of machine-processable metadata that has been added to resources. In this context semantics, more particular Semantic Web (Berners-Lee, Hendler, & Lassila, 2001), could enable machine supported mediation on a large scale. Using semantics, information becomes machine processable making possible for agents to understand and fulfill users requests.

While Web 2.0 technologies enhance the socially oriented aspects of the Internet, Service Oriented Architectures (SOA) contribute an alternative approach to the current Internet. The service-oriented perspective promotes the notion of service as central to system development, abstracting from implementations and the underlying hardware. While this abstraction provides little more than a common philosophy for the design of distributed applications, the paradigm shift introduces a new set of

[1] http://www.facebook.com/
[2] http://www.myspace.com/
[3] http://delicious.com/
[4] http://www.youtube.com/
[5] http://answers.yahoo.com/
[6] http://www.linkedin.com/

challenges, including how to organize, search, rank, and select services. Thus, for example, ranking of Web services is a core challenge that any SOA-based system must address. Existing solutions for service ranking are tightly integrated with service discovery and selection solutions and often use service data such as non-functional properties or Quality of Service (QoS) mechanisms to compute rank values for services. In most cases, multiple non-functional properties (e.g., price, availability, etc.) as well as dynamic values are considered (Hwang & Yoon, 1981; Liu, Ngu, & Zeng, 2004; Zeng et al., 2004; Zhou, Chia, & Lee, 2005); but social information from Web 2.0 is not generally considered in this approach.

In this chapter, we provide a set of solutions to the challenges mentioned above. First, we perform a large-scale study over MySpace, the largest and most active online social network. By studying the characteristics of MySpace, we hope to provide insight into the types of users of these online social networks, how the network itself is organized, and the important text artifacts that may distinguish users. In particular, we study over 1.9 million actual social network profiles with an emphasis on:

- The sociability of users in MySpace, based on relationships, messaging, and group participation;
- The demographic characteristics of MySpace users in terms of age, gender, and location and how these factors correlate with privacy preferences;
- The text artifacts of MySpace users that can be used to construct emergent language models that can distinguish between MySpace users not only by who they say they are but also by the language model they employ.

By studying how MySpace users participate in the social network (sociability), how they describe themselves (demographics), and how they communicate their personal interests and feelings (language models), we hope to encourage the development of new models, algorithms, and approaches for the further enhancement and continued success of online social networks.

The second contribution of this chapter is an analysis of social phenomena on the Social Web in order to identify ways for mediating and linking social data. This analysis is carried out with three major social tagging systems as examples -- Delicious, Flickr and YouTube -- and focuses on:

- Modeling social tagging data according to the proposed Upper Tag Ontology (UTO);
- Linking data from related social metadata schemes (e.g., FOAF, DC, SIOC, SKOS, etc.) using UTO;
- Crawling data from major social tagging systems and integrating them through UTO; and
- Clustering crawled tagging data.

Last, but not least, this chapter proposes an approach to social ranking for Web service selection and ranking that is based on an analysis of Delicious, one of the largest social networks. We discuss the use of social semantics to qualitatively improve the service ranking task, and we explore the idea of using social annotation technologies for ranking Web services. Annotation data from Delicious is used to discover and rank Web services associated with Delicious bookmarks. Given a set of Web services, the system checks to determine if there are Web pages in the Web services domain that are bookmarked in Delicious; when this is the case, the services are ranked based on the number of users in Delicious who have tagged the associated Web pages. As part of this third approach, an algorithm is proposed that considers interdependencies between services, Web pages, annotations and users to compute the global rank of each service.

The chapter is organized as follows. In Section 2, related work is surveyed and analyzed. In Section 3, we present an extended study of large online social networks that focuses on semantic profiling in social networks. In Section 4, we offer a solution for mediating between social tagging systems and linking social data that is based on an analysis of tagging phenomena on the Social Web. In Section 5, we explore the idea of using social annotation data to improve the quality of the service-ranking task. Section 6 concludes the chapter.

2. RELATED WORK

In this section we survey some of the existing approaches related to our work. We investigate related studies for each of the three approaches proposed in this chapter.

2.1 SOCIAL NETWORKS

The study of social networks has a rich history (Milgram, 1967), and the recent rise of online social networks has seen increasing interest in this area. For example, a number of studies have examined the nature and structure of online social networks, including social networks derived from blogspaces (Backstrom et al., 2006; Liben-Nowell et al., 2005), email networks (Adamic & Adar, 2005), online forums (Zhang, Ackerman, & Adamic, 2007), photo sharing sites (Kumar, Novak, & Tomkins, 2006), and many others.

With respect to online social networks such as MySpace and Facebook, there has been some research interest, but most studies have been carried out on a much smaller scale. In one study, researchers analyzed the relationship between a user's profile and friendships over 31,000 Facebook profiles (Lampe, Ellison, & Steinfeld, 2007). Social capital has been studied over several hundred Facebook users by Ellison, Steinfield, and Lampe (2006), and the privacy attitudes of 7,000 Facebook users was studied by Acquisti and Gross (2006). Dwyer, Hiltz, and Passerini (2007) surveyed trust-related issues for over 100 MySpace and Facebook users; and Hinduja and Patchin (2008) studied the revelation of personal information among 10,000 young people on MySpace. Other studies have investigated membership formation for 200,000 Orkut members (Spertus, Sahami, & Buyukkokten, 2005) or looked at the messaging characteristics of four million Facebook users (Golder, Wilkinson, & Huberman, 2007).

2.2 TAGGING ONTOLOGIES

In 2005, Tom Gruber first proposed the idea of using an ontology to model tagging data. His original idea was further formalized and subsequently published in 2007 (Gruber, 2007). His tag ontology contains the concepts Object, Tag, Tagger, and Source. He then introduced the + or - vote tag to the ontology to be used for collaborative filtering. UTO is based on Gruber's formulation, but it provides enhanced support for ontology alignment and data integration. When compared to Gruber's tag ontology, UTO contains more concepts and relations (e.g., Date, Comment, and has_relatedTaqg) and focuses on mediation between social metadata schemes in order to achieve data integration.

The Social Semantic Cloud of Tags Ontology[7] (SCOT) was developed to represent both the structure and semantics of a collection of tags and social networks of users based on tag sets. The core concepts of SCOT include Tagcloud and Tag. SCOT uses URIs as a mechanism for identifying a unique tag namespace to link a tag and a resource. The SCOT ontology is based on and linked to SIOC, FOAF and SKOS. It uses SIOC concepts to describe site information and relationships among site-resources, FOAF concepts to represent a human or machine agent, and SKOS to characterize relationships between tags. Although it does not include the concept Tagcloud, UTO is defined in such a way that it can be aligned with many other social metadata schemes in addition to SIOC, FOAF and SKOS (e.g., DC, microformats, etc.)

The Holygoat Tag Ontology[8] models the relationship between an agent, an arbitrary resource, and one or more tags. Using Holygoat, taggers are linked to foaf:agents and taggings reify the n-array relationship among tagger, tag, resource and data. This ontology is also linked to relationships in RSS and DC (e.g., rss:item, rss:category, rss:pubDate, rss:link and dc:subject) through use of rdfs:subClassOf or rdfs:subPropertyOf. Based on these links, it is possible to perform simple subsumption inferences using Holygoat metadata. Thus the Holygoat Tag Ontology provides some support for the emerging Semantic Web by utilizing ontology reasoning and inference. In contrast, because the primary objective of UTO is

[7] http://scot-project.org/

[8] http://www.holygoat.co.uk/projects/tags/

to support mediation and ease of alignment across metadata schemes, ontology reasoning and inference has not been considered at this stage in its development.

The MOAT Ontology[9] is a lightweight ontology intended to represent how different meanings can be related to a tag. It focuses on providing a unique identifier for each tag that serves to link an associated semantic meaning to the tag. MOAT is based on Holygoat Tag Ontology in its definition of a tag object and assumes that there exists a unique relationship between a tag and a label such that a tag can have a unique MOAT identifier in the Semantic Web. As noted previously, UTO is focused on the structure of the tagging behavior rather than the meanings of individual tags; but the provision of a unique identifier for a tag is always a helpful and an important contribution both to social tagging and to the Web in general.

2.3 WEB SERVICE RANKING

We now turn our attention to related investigations of the general ranking problem. Ranking various types of entities has been a challenging research problem over the years, but ranking has gained increasing attention in the context of improving Web search experience of user.

A variety of different approaches has been developed to improve ranking algorithms and ranking results. Classical approaches use statistical information such as term frequency, document length, etc. to compute the similarity degree of a document and a query. An interesting approach that uses social annotations to improve Web search was proposed in Bao et al. (2007). This work was motivated by two basic assumptions: (1) annotations are usually good summaries of the corresponding pages; and (2) the number of annotations indicates the popularity of a Web page. Two algorithms based on these assumptions have been proposed to define how the similarity between queries and annotations can be computed (i.e., the popularity of a Web page based on its relations to annotations and users).

A classification of various types of Web service ranking approaches is proposed in Gekas (2006). The author distinguishes two distinct categorization axes: one measures the localness of the ranking approach depending on whether local or global knowledge is used in computing ranking values; the other measures the absoluteness of the ranking approach depending on whether the measurement is of absolute scope or refers to a particular request.

Two of the more prominent categories of Web service ranking are based either on hubs and authorities, or HITS algorithms, which examine the relations between the number of services that link to a specific service (in-degree) and the number of services to which the service links (out-degree), or on non-functional rankings, which use nonfunctional aspects to compute the rank of services. The first category of methods exploits the given in/out relations between entities, computing rank values using global knowledge. The most prominent representative of this category is PageRank (Page et al., 1998). In the second category, the most prominent approaches have been offered by Hwang & Yoon (1981), Liu, Ngu & Zeng (2004), Zeng et al. (2004), and Zhou, Chia & Lee (2005). For example, in Zeng et al. (2004), a multi-criteria Quality of Service (QoS) model is used to determine the importance or rank of a service. Quality vectors are built for each service and a correspondence matrix is constructed between services and QoS. A simple additive weighting method is then applied to select the optimal Web service(s).

3. SEMANTIC PROFILING OF SOCIAL NETWORKS

3.1 APPROACH

In our first study of social networks and semantics, we provide the first large-scale demographic study of more than a million actual social network profiles. We analyze these profiles with respect to age, gender, and location, and we study how these factors correlate with privacy preferences. We compare two sampling approaches for extracting social network data, and we provide a unique analysis of text artifacts that can be used to distinguish between types of users.

[9] http://moat-project.org/ontology

Data and Setup

To study the characteristics of large online social networks, we selected MySpace as our target social network. MySpace is the largest social networking site, the sixth most visited Web destination according to Compete.com, and a website that has received a tremendous amount of media coverage. In addition to these characteristics, MySpace is one of the few online social networks that provide open access to user profiles. Many other sites require a user account and may restrict access to the entire social network even when a user has a valid account on the site.

On MySpace, as on most online social networks, the most basic element is a profile, a user-controlled Web page that includes descriptive information about the person it represents and that can be connected to other profiles through explicitly declared friend relationships and numerous messaging mechanisms. MySpace allows users to choose between making their profiles publicly viewable (the default option) or private. If a user's profile is designated as private, only the user's friends are allowed to view the profile's detailed personal information (e.g., the user's interests, blog entries, etc.). However, a private profile still provides information such as the user's name, gender, age, location, and last login date.

Because extracting and analyzing all 250 million MySpace profiles would place undue burdens on the resources and network of both MySpace and our local infrastructure, we adopted a sampling-based approach to extract representative profiles from MySpace. We considered two approaches – random-sampling and relationship-based (or snowball) sampling:

- *Random Sampling:* MySpace profiles are sequentially numbered and made publicly Web accessible by constructing a URL containing the profile's unique profile ID. Hence, we can randomly sample from the space of all MySpace profiles by randomly generating profile IDs. By construction, we expect a random sample of MySpace profiles to provide perspectives on the global characteristics of the entire MySpace social network.
- *Relationship-Based Sampling:* Unlike random sampling, the second approach leverages the relationship structure of the social network to select profiles. We begin by generating a set of randomly selected seed profiles. We then extract the IDs of their friends, add these friend IDs to the queue of profiles to sample, and continue in a breadth-first traversal of the social network. When the queue is empty, we generate a new random profile ID and continue the process. In contrast to the random sampling approach, we expect the profiles extracted through the relationship-based sampling approach to provide a more focused perspective on an active portion of the social network.

Data Collection

In practice, we collected two representative datasets from MySpace: the Random dataset was constructed using random sampling and the Connected dataset was constructed using the relationship-based sampling. We wrote two MySpace-specific crawlers (based on Perl's LWP::UserAgent and HTML::Parser modules). Both crawlers disregarded invalid profile IDs (i.e., profiles that were either deleted or undergoing maintenance at the time of the crawl) and entertainment profile IDs (i.e., profiles that were associated with bands, comedians, etc.) so as to focus our collections on active profiles that belong to regular individuals. In June 2006, we deployed ten instances of the random sampling crawler in parallel across ten different servers, collecting profiles for about a week. We repeated this setup in September 2006 with the relationship-based sampling crawler. Summary data for both the Random and Connected datasets are provided in Table 1.

We then wrote a custom MySpace parser to extract the name, age, and other pertinent information from each collected profile. Because some of these features are self-described by the owner of the profile (e.g., age and gender), they may or may not be truthful. Other features (e.g., number of friends) are maintained by MySpace and are therefore expected to be correct. However, MySpace has a limited validation process; thus, we have no assurances that a self-described 20-year old male from Texas is really who he says he is. Having said that, we do believe there is significant value in studying demographics in the aggregate, and, as we will demonstrate in the following section, certain text artifacts specific to certain groups on the social network could be used to detect deceptive profiles.

Table 1: Summary Data for the Two MySpace Datasets

	Public Profiles	Private Profiles	Total Profiles	Size
Random	859,347	101,158	960,505	52 GB
Connected	717,337	173,830	891,167	98 GB

3.2 RESULTS AND OBSERVATIONS

In this section we present the main findings of our study across a series of characterizations: sociability, demographics, language models, and privacy preferences.

Sociability Characterization

We begin our characterization of MySpace by examining the social aspects of users in the network. Since online social networks derive their value from users actively participating in relationships with other users, we are interested to observe to what degree users actually take advantage of these social aspects. To examine sociability over both datasets, we measured the number of friends, the number of comments, and the number of groups a user participates in, values that are only available for public profiles.

In Figure 1, we present the distribution of the number of friends for both the Random and the Connected datasets on a log-log scale. For the Random dataset, there is a heavy-tailed distribution: That is, most users have very few friends, but a few users have many friends. Such a heavy-tailed distribution has been observed in a number of related domains, and observing it here is no surprise. What was surprising was the number of users with zero or only one friend, which accounted for 426,926 or 50% of the public profiles in the Random dataset. Since MySpace provides each new user with a single default friend, we surmise that more than half of MySpace users had created an account and subsequently abandoned it.

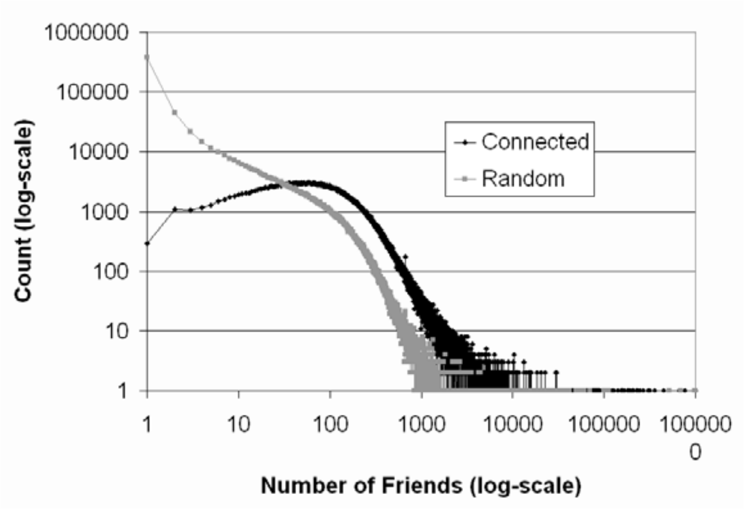


Figure 1: Distribution of Friends: The x-axis is the number of friends a user may have; the y-axis is a count of the number of users with a particular number of friends.

In contrast, we see that, for the Connected dataset, most users had many friends and were actively participating in the MySpace social network. Only 2.5% of the public profiles in the Connected dataset had no or only one friend. Obviously, the relationship-based sampling method used to construct the Connected dataset favored users with many friends.

To further validate the sociability divide, Figure 2 shows the distribution of the number of comments posted to a user's profile for both datasets. The commenting feature of MySpace is one of several avenues for users to communicate with other users. Because comments written to a particular user are posted on that user's profile, we would anticipate that users with many comments are well known and active in the social network. Again, there is a heavy-tailed distribution for the Random dataset, whereas the Connected dataset shows more skew, since it is, by construction, more connected.

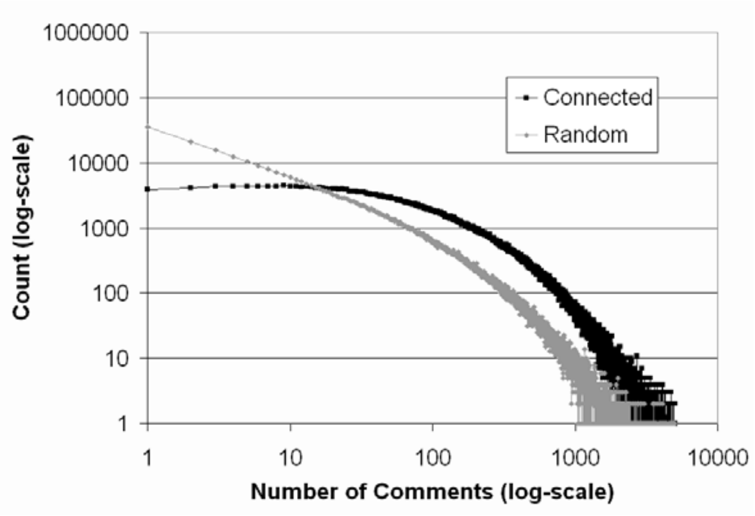


Figure 2: Distribution of Comments: The x-axis is the number of comments posted on a user's profile; the y-axis is a count of the number of users with a particular number of comments posted on their profiles.

Group participation is another metric of the sociability of a social network. While over 80% of the users in the Random dataset (and, by extrapolation, of the users of MySpace as a whole) participate in no groups, we found that slightly less than half of the users in the Connected dataset belong to at least one group and that nearly 20% of users in the Connected dataset belong to at least eight groups. This evidence further confirms what we observed for the friend and comment measures of sociability: Most MySpace users appear to have effectively abandoned their online profiles, but there remains a large core of active users within MySpace who account for the vast majority of friends, comments, and group activity.

Who, then, are these active users? In an effort to understand if some users are more likely to be active than others, we considered a number of features, including age, location, gender, and the length of time a profile had existed on MySpace. We found that California and other western U.S. states dominated the total number of profiles on MySpace, but that these differences were minor across the Connected and Random datasets, which indicates that location is not a strong indicator of sociability. Likewise, we found little evidence that a profile's self-declared age or gender impacted its relative sociability. In contrast, we did find that the length of time a user has participated in MySpace was a strong indicator of sociability. In order to measure the length of participation on MySpace, we had to augment our original sampling process. Because the profile creation date is provided for each profile on a separate blog page linked off the public profile Web page, retrieving the date of a profile's creation required accessing an additional page for every profile sampled. However, because profile IDs are assigned sequentially in MySpace, we can interpolate the date of creation for each profile sampled. Therefore, in an attempt to avoid overburdening MySpace with a doubling of page requests, we sampled a handful of profiles (e.g., profile 10,000, profile 100,000, etc.) to retrieve creation dates and thus construct a time series.

Thus, in Figure 3, each point serves as a bucket that represents all profiles created before that date, extending back in time to the previous point. The y-axis measures the rate of sampling for each bucket. As expected, the random sampling approach provided a nearly uniform sample for each bucket, although we do see a hiccup at the beginning because the bucket is so small and again at the end because the sampling periods are slightly different. In contrast, the relationship-based sampling approach used to construct the Connected dataset indicates that users overwhelmingly joined at an early date. These long-time users are presumably more plugged-in and are thus more active participants in the social network.

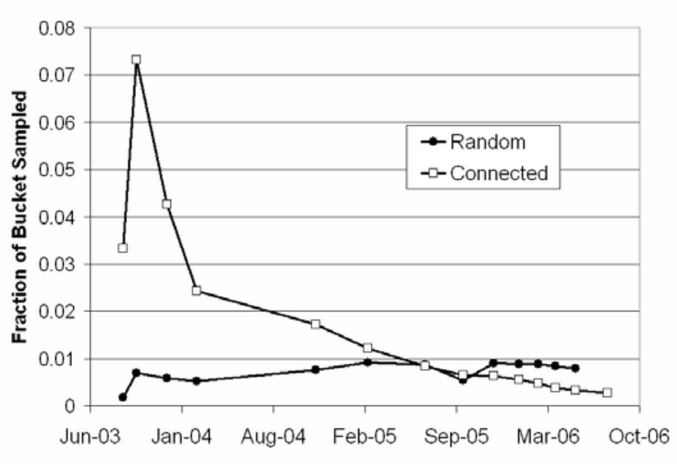


Figure 3: Sampling by Date: The x-axis shows buckets of profiles organized by their dates of creation; the y-axis shows the fraction of all sampled profiles that were created within a bucket's date range..

Demographic Characterization

In the previous section, we studied the sociability of MySpace users – how active they are and to what degree are they connecting to other users. In this section, we expand our analysis of MySpace to consider the demographics of participants. How old are they? Are they predominately male or female? Where are they located? The answers to these questions can provide us with added insight into how a social network grows and what features are attractive to certain participants as well as indicating other interesting avenues. Both the public and the private profiles on MySpace provide basic demographic information; and we found that nearly all MySpace users (> 99.9%) provide some information about age, gender, or location. Only 1,311 profiles in the Random dataset and 1,203 profiles in the Connected dataset declared no age, gender, or location.

Figure 4 shows the distribution of ages in both datasets. As expected, MySpace is dominated by younger users: Nearly 85% of the users on MySpace are 30 or younger. Interestingly, we observed that the Random dataset, which peaks at 17 years of age, skews slightly younger than the Connected dataset, indicating that the most active users on MySpace may, in fact, be in their 20s. We also observed a peak at the age of 69, presumably either a joke age or an age intentionally selected by users interested in sex in order to find one another through the age-based search facility of MySpace (Gradijan, 2007). We also observed another peak around 100, but we can presume that many of these self-reported ages are also false.

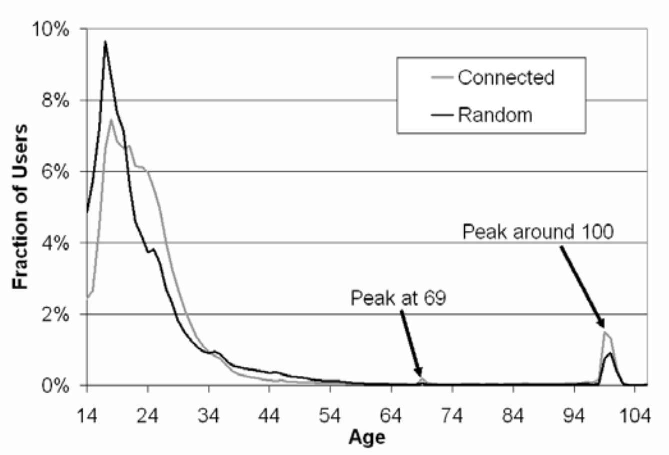
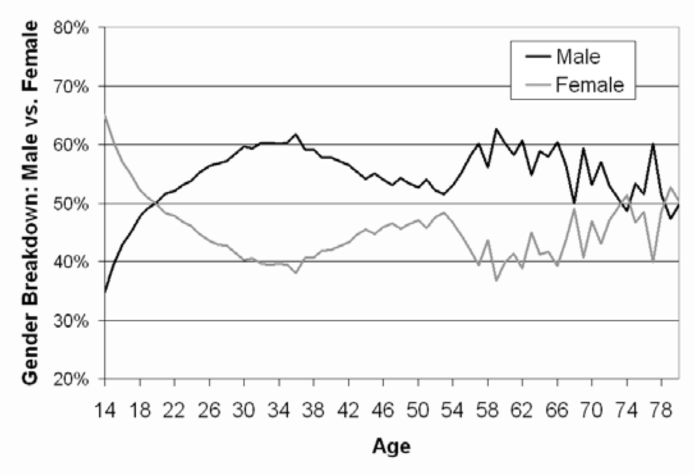


Figure 4: Distribution by Age: The x-axis is the self-reported age on a user's profile; the y-axis is the fraction of all profiles declaring a particular age.

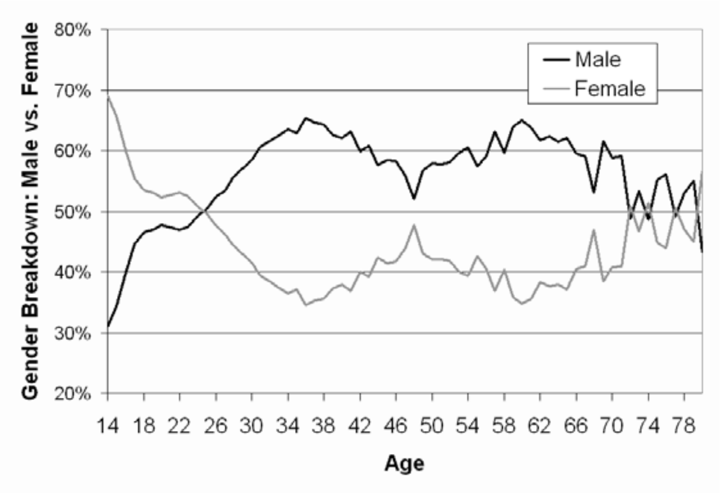
In Table 2, we show the gender breakdown for each dataset. The split between male and female is nearly even: 52% male and 48% female in the Random dataset versus 50% male and 50% female in the Connected dataset. The Other category is a placeholder for profiles that list either no gender information or non-standard gender information. In Figure 5, we consider the gender distribution across both datasets. The results are intriguing: Women are more prevalent at the youngest ages, whereas men are more prevalent at all other ages (barring a few hiccups at the older end, where the data is sparser).

Table 2: Gender Breakdown for MySpace Datasets

	Random	Connected
Male	505,357	440,330
Female	452,240	448,920
Other	2,908	1,917



(a) Random Dataset



(b) Connected Dataset

Figure 5: Gender Breakdown by Age: The x-axis is the self-reported gender on a user's profile; the y-axis is the fraction of all profiles of a particular age declaring a particular gender.

Why are women more active participants at younger ages? It may be either that women intentionally self-report a younger age or that men intentionally self-report an older age; or it may be that there are clear gender differences in how users participate in a social network, with younger women more attracted to certain social aspects than their male counterparts. These are interesting and open questions that deserve further exploration.

Finally, we studied the self-reported location information for each profile. MySpace users hail from all fifty U.S. states, and a significant percentage come from other countries. Not all profiles list an intelligible location (e.g. "Somewhere Over the Rainbow"), and some list multiple locations (e.g., "Honolulu and Metro DC"), so we built a best-effort parser. In our initial analysis, we found that the U.S. was the most prevalent location by far, followed by the United Kingdom and Canada. Accordingly, we focused on states in the U.S. for further analysis of location information. For the Random dataset, we found that 77% listed a U.S. state in the location; for the Connected dataset, we found that 87% listed a U.S. state.

In Table 3, we report the top five states that are overrepresented on MySpace, relative to their actual population, as well as the top five under-represented states. We measured the relative presence of a state on MySpace versus its relative share of the U.S. population as:

$$rel_i = 1 - \frac{\frac{pop_{i,MySpace}}{\sum_j pop_{j,MySpace}}}{\frac{pop_{i,us}}{\sum_j pop_{j,us}}}$$

where $pop_{i,us}$ is the population of state i based on the latest U.S. Census data and $pop_{i,MySpace}$ is the number of profiles in our dataset that listed state i as their location. For the Random dataset, Table 3 demonstrates that California and other western U.S. states were the most over-represented on MySpace relative to their actual population, while southern and mid-western states tended to lag behind relative to their actual population.

Table 3: U.S. States Most Over-Represented and Most Under-Represented in MySpace Random Dataset Relative to U.S. Census Population.

Most Over-Represented	**Most Under-Represented**
Hawaii [+115 %]	Mississippi [-58 %]
California [+61 %]	West Virginia [-53%]
Washington [+41 %]	Arkansas [-52 %]
Alaska [+40 %]	Missouri [-49 %]
Nevada [+39 %]	South Dakota [-48 %]

We attribute much of this geographic discrepancy to My-Space's initial launch by a California-based company and its success with Los Angeles area bands (Rosenbush, 2005). Although California accounts for only 12% of the U.S. population, users from California dominated the early adopters of MySpace.

Characterizing Language Models

Thus far in our study, we have characterized how users participate in the MySpace social network (e.g., friendships, comments, etc.) and how users describe themselves. In this section, we examine what users are saying on their profiles through an analysis of the language models of social network users. Our goal is to understand how language use varies across classes of users. For example, do women express themselves differently than men? Do older MySpace users describe themselves differently than younger MySpace users?

We begin with some definitions. We treated each profile as a sequence of terms drawn from a vocabulary set $V=\{t_1, t_2, \ldots, t_{|V|}\}$. We considered all terms in a profile that had been generated by the user (e.g., in About Me, Interests, etc.), and we excluded all terms that were most likely generated by other users (e.g., comments). Following the standard information retrieval approach, we can describe the language model of all profiles as a probability distribution over the terms in the profiles according to a unigram language model defined as:

$$\{p(t)\}_{t\in V} \ s.t. \sum_{t\in V} p(t) = 1$$

Terms with high probability are more likely to be observed on a profile than are low probability terms. We can compute *p(t)* as a function of the count *count(t)* of profiles containing term *t* relative to the total number of profiles *n*: *p(t)* = *count(t)/n*. For example, the top five most probable terms in the Connected dataset are *the*, *and*, *straight*, *friends*, and *with*. Because these common terms provide little insight, we augmented the basic language model by identifying class-specific distinguishing terms for classes based on age, gender, and location. Our goal was to identify terms that were more likely to be generated by a certain class of users (e.g., by women).

To identify these class-specific distinguishing terms, we relied on the information theoretic measure Mutual Information for assessing the importance of a term to a particular class. Mutual Information (MI) between a term and a class is defined as:

$$MI(t,c) = p(t|c)p(c)\log\frac{p(t|c)}{p(t)}$$

where *p(t|c)* is the probability that a profile contains term *t* given that it belongs to class *c*, *p(c)* is the probability that a profile belongs to class *c*, and *p(t)* is the unigram language model described above for the probability of term *t* across all profiles. Letting *count(c)* denote the count of profiles belonging to class *c* and letting *count(c,t)* denote the count of profiles containing term *t* that belong to class *c*, we have:

$$p(t|c) = \frac{count(c,t)}{count(c)} \ and \ p(c) = \frac{count(c)}{n}$$

MI measures how much information a particular term *t* tells us about class *c*. Higher MI values indicate stronger associations. In this raw form, however, rare terms that by chance happen to occur only in profiles belonging to a particular class will score highly using MI. Hence, a natural correction is to replace *p(tjc)* with a "smoothed" version that gives every term a non-zero probability of occurrence across all classes:

$$p^*(t|c) = \alpha p(t|c) + (1 \quad \alpha)p(t)$$

where $0 \le \alpha \le 1$. In practice, we selected a smoothing factor of $\alpha = 0.9$. Thus we can interpret

$$\{p^*(t|c)\}_{t \in V} \ s.t. \sum_{t \in V} p^*(t|c) = 1$$

as a class-specific language model.

Class-Specific Distinguishing Terms

Using the MI measure for identifying distinguishing terms, we explored language models of MySpace users according to three characteristics: gender, location, and age. Since we are primarily interested in users who were actively using the social network, we report results from the Connected dataset. Superficially, however, we saw many similarities with the Random dataset relevant to the presence of distinguishing terms. Furthermore, only public profiles were included in this analysis since the contents of private profiles are hidden.

First, we considered class distinction by gender (i.e., male and female). In Table 4, we report the top 16 class-specific distinguishing terms for profiles declared to be male and for profiles declared to be female. The differences are stark.

Table 4: Distinguishing Terms by Gender (Ranked by MI)

Male		**Female**	
dating	sport	love	people
networking	metal	dancing	life
serious	football	shopping	can
relationships	s***	girl	family
single	wars	hearts	being
straight	band	have	notebook
video	f***	are	dance
guitar	gay	favorite	things

Second, we considered class distinction by location for all fifty U.S. states. In Table 5, we report the results from three states that represent distinct geographic regions of the U.S.: the south, the Pacific northwest, and the northeast. We see an interesting mix of geography-specific identifiers (e.g., *protestant* in Alabama versus *catholic* in Connecticut), interests (e.g., *football* in Alabama versus *rafting* and *snowboarding* in Oregon), and word constructions (e.g., *yall*, *pdx*, *rad*, and *nas*).

Table 5: Distinguishing Terms for Three Representative Locations (Ranked by MI): Popular location names (e.g., Birmingham, Portland) within each state are excluded.

Alabama	**Oregon**	**Connecticut**
christian	camping	catholic
African-descent	pdx	Yankees
tide	hiking	nyc
jesus	northwest	uconn
football	pixies	Hispanic
bama	snowboarding	Bronx

church	coast	boston
chrish	rafting	sox
protestant	floater	nas
gospel	rad	Italian
yall	wine	goodfellas
nascar	vegan	sneakers

Finally, we consider how the language models of MySpace users vary by age. In Table 6, we report distinguishing terms for ages ranging from 16 to 100. We see how the language model shifts in focus with age and educational level (e.g., high school, college, college graduate, etc.). MySpace users 30 and older used terms like *married*, *parent*, and *proud*, whereas members under 30 used terms like *lol*, *single*, and *love*. We can also make a few comments about the older (and perhaps less truthful) age groups. The 69-year-olds have a clearly expressed interest in sex. The odd language model of 80-year-olds is skewed by the presence of many tribute profiles to Marilyn Monroe, who would have been 80 at the time the datasets were constructed: All of the terms here are relevant to Monroe's movie career and relationships. In contrast, the 100-year-olds display a less coherent language model, perhaps due to the diversity of users declaring this advanced age.

Table 6: Distinguishing Terms by Age (Ranked by MI)

16	18	20	25	30	40	60	69	80	100
high	high	college	graduate	networking	parent	parent	networking	scudda	swinger
school	school	someday	college	graduate	proud	proud	swinger	mortenson	our
hearts	someday	student	networking	parent	married	president	sex	gable	kids
junior	love	love	grad	proud	networking	swinger	a**	jeane	capricon
single	best	straight	professional	married	kids	his	f***	showgirl	networking
best	boy	Caucasian	relationship	grad	great	married	rock	asphalt	virgo
hair	ever	white	traveling	professional	our	kids	islander	dimaggio	artists
friend	hair	like	some	art	divorced	united	real	Dougherty	their
lol	lol	girl	reading	cure	daughter	began	our	harlow	please
play	single	know	working	travel	years	retired	night	actress	official

Identifying Language Model Clusters

In the previous section, we saw how certain classes of MySpace users could be described by distinguishing terms that are relatively strong indicators of class membership. In this section, we continue this analysis by considering clusters of related classes. For example, assuming that most self-declared 100-year-old members of MySpace are not actually 100, what is their true age? MySpace has made some efforts to remove self-declared older members (Gradijan, 2007) through manual inspection. Can the language models we have identified provide a scalable solution?

We begin with the class-specific language models of interest (e.g., by age: $\{p^*(t|c = 16)\}_{t\in V}$, $\{p^*(t|c = 17)\}_{t\in V}$ and so on). Are there clusters of language models by age or by location? In this initial study, we considered a similarity measure for determining the "closeness" of two language models based on the Kullbeck-Leibler divergence (KL-divergence). KL-divergence, or relative entropy, measures the difference between two probability distributions *p* and *q* over an event space *X*:

$$= \sum_{x\in X} p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

Intuitively, the KL-divergence measure indicates the inefficiency (in terms of wasted bits) of using the *q* distribution to encode the *p* distribution. In this case, we can measure the divergence of two class-specific language models (i.e. $p = \{p^*(t|c = 16)\}_{t \in V}$ and $q = \{p^*(t|c = 17)\}_{t \in V}$). Note that KL-divergence is not symmetric, so we will typically find $KL(p, q) \neq KL(q, p)$.

In Figure 6, we report the KL-divergence for 16-year-olds versus other ages, for 20-year-olds versus other ages, and for 30-year-olds versus other ages. Since there are very few profiles listing an older age, we have omitted these from the graph.

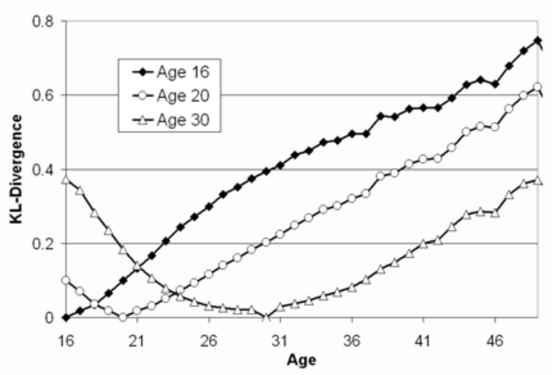


Figure 6: KL-Divergence by Age: Comparison of class-specific language models using KL-divergence (where a lower value is better).

The KL-divergence for 16-year-olds is lowest for profiles closest in age, which indicates that the language model of a 16-year-old is closest to that of a 17-year-old, then an 18-year-old, and so on. A similar pattern holds for the language models of both 20-year-olds and 30-year-olds, pointing to clear clusters based on age.

What do we observe when we consider profiles that are more likely to be deceptive about their true age? Table 7 shows the closest language models for profiles listing an age of 69 and for profiles listing an age of 100.

Table 7: Identifying Outliers: Language model(s) that most closely match the language model of self-described 69-year-olds and self-described 100-year-olds using KL-divergence.

Rank	Age 69	**Age 100**
1.	100 [0.017]	99 [0.047]
2.	99 [0.021]	101 [0.103]
3.	101 [0.047]	30 [0.105]
4.	33 [0.068]	31 [0.105]
5.	31 [0.072]	29 [0.106]

For the 69-year olds, the closest matches are other outlier ages: 100, 99, and 101. This provides some evidence that the type of user who lies about his age is bound by some common language model cues. The next two closest matches are with users in their 30s. This is a bit surprising, since we might have expected teenagers to be more likely to engage in such behavior. For the 100-year olds, we see a similar pattern: close matches with other outlier ages (99 and 101) and then close matches with younger profiles that could be presumed more likely to declare true ages. We believe this line of inquiry could be extended in a number of fruitful directions.

Privacy Preferences

Finally, we turned our attention to the important issue of privacy in social networks. A number of researchers have examined some of the aspects impacting privacy on social networks (Barnes, 2006; Boyd, 2007; Nussbaum, 2007) in an effort to comprehend user understanding of privacy and the limits of privacy controls. We examined the privacy choices of MySpace members through the lens of our demographic study. As noted previously, MySpace users can elect to declare their profile as public or private, although younger members of MySpace (i.e., 14- and 15-year-olds) are required to have a private profile that displays only limited information, such as name, age, gender, and location.

In Table 8, we report on the privacy preferences of the randomly selected MySpace users who comprised the Random dataset, which was constructed to reflect the general MySpace population; and we contrast these preferences with the privacy preferences of the more sociable members of the Connected dataset. Members of the Connected dataset selected private profiles by nearly 2-to-1 over the average MySpace user. These findings are especially surprising since the relationship-based sampling technique used to extract the Connected dataset relied on the friendships declared on public profiles to identify profiles to sample: Private profiles reveal no friendships and so the sampling terminates when it arrives at a private profile. We further examined the private profiles in each dataset and found that nearly all (99.9%) of the private profiles in the Random dataset belonged to 14- and 15-year-olds (see Table 9). In contrast, we found that over 73.7% of the private profiles in the Connected dataset are those of MySpace members who were 16 or older.

Table 8: Privacy Preferences for MySpace Datasets.

	Random	**Connected**
Private	101,158 (10.5 %)	173, 830 (19.5 %)
Public	859, 357 (89.5 %)	717,337 (80.5 %)
Total	960,505	891,167

Table 9: Private Profiles by Age for MySpace Datasets.

	Random	**Connected**
14/15 Years Old	101,017 (99.9 %)	45,633 (26.3 %)
All Other Ages (16+)	141 (00.1 %)	128,197 (73.7 %)
Total	101, 158	173,337

Overall, very few users elected private profiles when given the opportunity (0.1%); but, for users who actively participated in the MySpace social network, a larger percentage preferred privacy. These results lend credence to the hypothesis that more sociable members tend to be more likely to choose private profiles.

To further explore the impact of demographics on privacy preferences, Figure 7 presents the percentage of private profiles in the Connected dataset by age and gender. We have truncated the graph over the age of 60 since there are very few profiles for those ages and hence more noise. We find that women favor private profiles 2-to-1 over men and, perhaps counterintuitively, that younger users are more likely to adopt a private profile than older users. Why is this? It may be that older users are less technically savvy and have more difficulty understanding how to configure the privacy setting; or it may be that younger users are more attuned to privacy and security concerns in social networks. We believe this is an area deserving more attention.

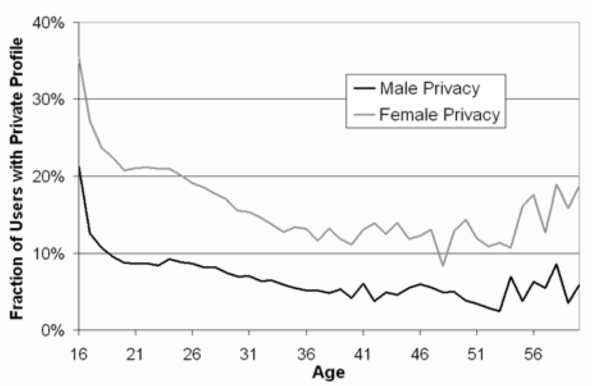


Figure 7: Privacy Breakdown by Age for MySpace Connected Dataset: The x-axis is the self-reported age on a profile; the y-axis is the percentage of all profiles declaring a particular age that are private.

Finally, we considered how privacy preferences have changed over time. In Figure 8, each point represents a bucket of all profiles created before that date and extending back to the previous point. The y-axis measures the percentage of profiles created within that bucket that are private (again, relying on MySpace's use of sequential IDs to interpolate profile creation dates). After an initial drop in the privacy rate, we saw a fairly steady growth in the adoption of privacy settings by new members. Overall, the percentage of private profiles increased over time, indicating that new adopters of social networks tended to be more attuned to the privacy risks inherent in the adoption of a public Web presence. We also investigated privacy preferences by location, but found no dramatic swings from state to state.

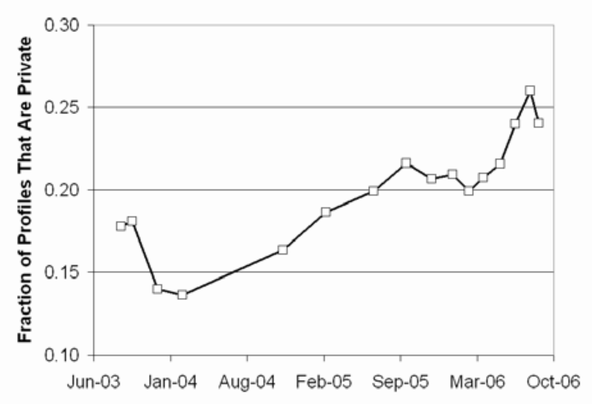


Figure 8: Privacy Over Time for MySpace Connected Dataset: The x-axis shows buckets of profiles organized by date of creation; the y-axis shows the percentage of private profiles created within a bucket's range.

4. MEDIATING AND ANALYZING SOCIAL DATA

4.1 MODELING SOCIAL TAGGING DATA

In our second study of social networks and semantics, we focus on the rise of social tagging and how tagging can be used for mediating and linking social data. A tag is essentially a typed hyperlink: It is a keyword contributed by a user that categorizes or describes an online object. The goal of tagging is to make a body of information increasingly easier to search, discover, share and navigate over time. But tagging by users is not simply the addition of keywords to resources; rather, these tags are social metadata generated through collective intelligence. Thus, social tagging is a bottom-up approach to indexing that reflects the collective agreement of and speaks the same language as the users, making online objects easier to find and leading to the creation of systems of social semantics called folksonomies.

There are many social networks providing tagging services. Users can tag a range of resources, including bookmarks (Delicious), photos (Flickr), videos (YouTube), books (LibraryThing), Music (Last.fm), citations (CiteULike), and blogs (Technorati). Here we have selected three major social tagging systems -- Delicious, Flickr, and YouTube -- and analyze their social tagging behavior. Based on this analysis, we propose the Upper Tag Ontology (UTO), which is an adaptation of the Tag Ontology proposed by Tom Gruber (2007). In his tag ontology, Gruber outlined five key concepts: *object*, *tag*, *tagger*, *source* and *vote*. In UTO, we have added an additional three concepts: *comment*, *date* and *tagging*. Most social networks contain information about user-contributed comments, whether they be comments on tags or on objects; and this information can contribute to an understanding of tags or objects. Date is another important concept because it allows us to track the evolution of tags and tagging behaviors. It can also help to reveal hidden social changes occurring within a social network. Although it does not have real meaning, the tagging concept functions to link the core concepts. We have also added the *has_relatedTag* relationship to the tag concept itself. Additional information about modeling social tagging data with UTO is discussed in Ding, Toma, Kang, Fried, and Yan (2008).

The Upper Tag Ontology (UTO) is defined as follows:

Let O be UTO ontology,

$$\mathrm{O} = (C, \Re)$$

Where $C = \{c_i, i \in N\}$ is a finite set of concepts

$\Re = \{(c_i, c_k), i, k \in N\}$ is a finite set of relations established among concepts in C.

In UTO,

$$C = \begin{Bmatrix} Tag, Tagging, Object, Tagger, Source, \\ Date, Comment, Vote \end{Bmatrix},$$

$$\Re = \begin{Bmatrix} has\_relatedTag, has\_tag, has\_object, \\ has\_source, has\_date, has\_creator, \\ has\_comment, has\_vote \end{Bmatrix}$$

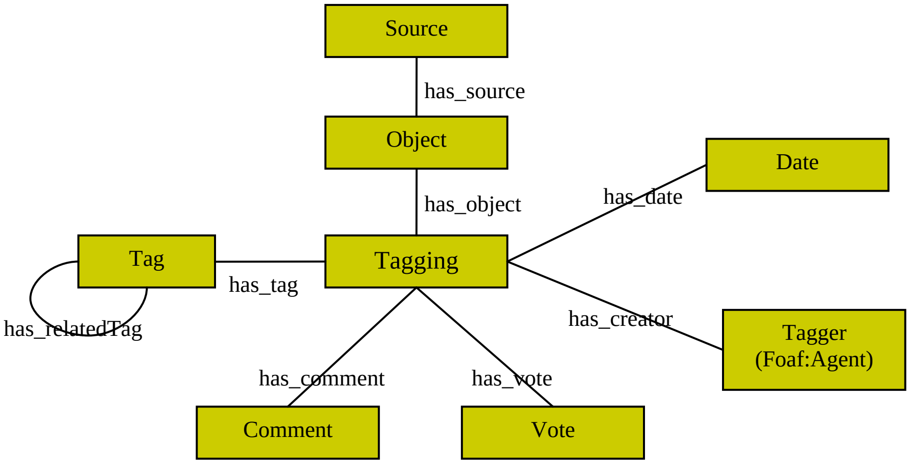


Figure. 9 Upper Tag Ontology (UTO)

Figure 9 represents the concepts and relations in UTO. Obviously, UTO is a small and simple ontology with only eight concepts and eight relationships. The tagging concept acts as a virtual connection for the other concepts of UTO, serving the function of linking core concepts. For example, it may be difficult to tell whether a date is related to the tag or to the tagging behavior or a comment could be viewed as added to a tag or directly to an object. For this reason, most of the relations in UTO are defined as transitive so that a comment can be connected to an object via tagging or to a tag via tagging.

UTO is very different from a folksonomy, which focuses on the meaning of tags. Following the basic design principle of "making it easy and simple to use", UTO is designed to capture the structure of social tagging behaviours rather than the meanings of the tags themselves. By focusing on the structure of social tagging behaviours rather than the actual tag semantics, UTO aims to model the structure of tagging data in order to integrate tagging data from one social tagging application with tagging data from other applications.

4.2 LINKING SOCIAL DATA

It is becoming increasingly important for data to be interlinked. Linking itself is moving from the traditional hyperlinks of Web 1.0 to the typed hyperlinks of Web 2.0 and, ultimately, to the semantic links of Web 3.0 in the Web 1.0 environment, we began by simply linking documents. Then, we added more metadata to these resources and turned unstructured information into structured information. Now, we are striving to provide semantic links between those structured resources so as to form Web 3.0 or the so-called Semantic Web. Social tagging plays an important role in this process of linking not only by structuring data but also by linking this structured information.

UTO can be aligned with other social metadata schemes such as Friend of A Friend (FOAF), Dublin Core (DC), Semantically-Interlinked Online Communities (SIOC) and Simple Knowledge Organization System (SKOS). With UTO, we try to make alignment between schemes as simple as possible since a more complicated alignment may actually generate more problems or double the complexity of application. In UTO, we have focused primarily on mapping of classes with the consideration of equivalent and sub-class mapping. For instance, the UTO *Tagger* concept is equivalent to foaf:Person and sioc:User and is a subclass of foaf:Agent. The *Tag* concept is equivalent to skos:Concept; the *Object* concept is defined as a superclass of foaf:Document, foaf:Image, sioc:Post, sioc:Item, dc:Text and dc:Image; and the *has_relatedTag* relationship is defined as a superordinate property of skos:narrower, skos:broader and skos:related.

Aligning UTO with other social semantics schemes enables easy data integration, mash-ups, and the interlinking of structured data. Using such integrated data, we can perform tag searches across multiple sites, applications, sources, and hosts and mine relationships (associations) across different platforms and applications. For example, it would be possible to find the friends of Stefan who have used the tag *spicy-Chinese-food* by aligning FOAF with UTO or to identify blogs, wikis or discussion groups where Stefan and his friends have discussed "spicy Chinese food" by aligning FOAF and SOIC through UTO. Associations among tags, taggers and objects can also be mined. For example, we can mine the social network relations of taggers through foaf:knows by aligning FOAF with UTO; we can mine relations between tags by aligning skos:broader, skos:narrower or skos:related with the UTO relation *has_relatedTag*; and we can use co-occurrence technologies to mine associations among tags, taggers and objects.

4.3 CRAWLING SOCIAL TAGGING DATA

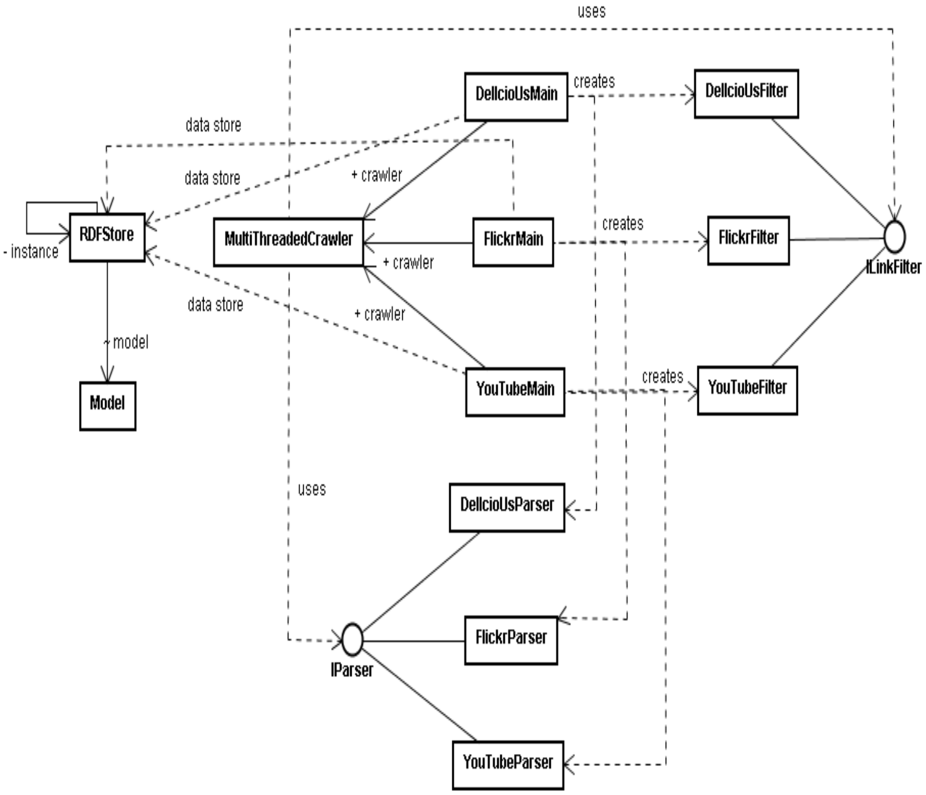
The Social Tagging crawler (or ST crawler) was developed for crawling major social tagging systems, including Delicious, Flickr, and YouTube (Fried, 2007). It is a multi-crawler based on the "Smart and Simple Webcrawler"[10] and UTO. Figure 10 is a detailed class diagram of the crawler.

The ST crawler is written in Java with Eclipse IDE 3.2 on Windows XP and Ubuntu 6.04, and data is cleaned using linux batch commands. ST crawler can start from either one link or a list of links and has two crawling models:

- Max Iterations: Crawling a website through a limited number of links, which needs a small memory footprint and CPU usage.
- Max Depth: A simple graph model parser without recording incoming and outgoing links, which uses a filter to limit the links to be crawled.

In the summer of 2007, we used the ST crawler to crawl social tagging data from Delicious, Flickr and YouTube and to model them according to UTO. The data are represented in RDF triples and stored in Jena. After one-week of crawling, the output was contained in a number of RDF files with a total file size of 2.10GB:

- 16 Delicious data files with a total size of 1.64GB
- 3 Flickr data files with a total size of 233MB
- 3 YouTube data files with a total size of 234MB



[10] https://crawler.dev.java.net/

Figure 10. Class Diagram of the ST Crawler

4.4 SOCIAL TAGGING DATA

Table 10 presents an overview of the data collected from the Delicious, Flickr and YouTube social networks. The total dataset contains approximately 1 million bookmarks, 2.8 million taggers and 9.3 million tags from Delicious; 300,000 photographs, 150,000 taggers and 1.4 million tags from Flickr; and 500,000 videos, 200,000 taggers and 1.35 million tags from YouTube. The average number of tags per object ranges from a low of 2.74 in YouTube to a high of 9.31 in Delicious. The average number of tags a single tagger assigns to a resource ranges from a low of 3.33 in Delicious to a high of 8.79 in Flickr. The average number of objects tagged by a single tagger ranges from a low of 0.36 in Delicious to a high of 2.84 in YouTube. The seeming disparity reflected in the low average for objects tagged by a Delicious tagger can be accounted for by the fact that, while users are required to provide a title when uploading bookmarks to Delicious, they are not required to include tags in the tag field: Thus, there may be many bookmarks in Delicious that have titles but no tags. Combined data from the three social networks totals approximately 1.8 million objects, 3.1 million taggers, and 12.1 million tags, of which 648,368 tags are unique.

Table10. Data from Delicious, Flickr and YouTube for the Years 2005-2007

Social Network	Objects	Taggers	Tags	Tag/Object	Tag/Tagger	Objects/Tagger
Delicious	996,748	2,787,860	9,282,058	9.31	3.33	0.36
Flickr	295,837	153,778	1,351,201	4.57	8.79	1.92
YouTube	527,924	185,975	1,443,924	2.74	7.76	2.84
Total	1,820,509	3,127,613	12,077,183	5.54	6.63	1.71

Note: Cells in the column labeled *Tags* represent the total number of tags assigned by taggers (e.g., when TagA is assigned by Tagger X and by Tagger Y, it is counted as two tags).

All tag data are represented in RDF and stored in Jena. We used the tag data as they were retrieved, although we did perform some data cleaning (e.g., stemming and checking with WordNet). By querying the data, we identified the 20 most highly ranked tags and the 20 most highly ranked bookmarks in Delicious for the period 2005-2007, as shown in Table 11.

Table 11. Top 20 Tags and Bookmarks in Delicious for the Period 2005-2007

Rank	Tag	Tag Frequency	Bookmark	Bookmark Frequency
1	blog	141,871	en.wikipedia.org	26,745
2	system	120,673	www.youtube.com	14,990
3	design	109,249	community.livejournal.com	6,594
4	software	87,719	www.google.com	6,376
5	programming	83,665	www.w3.org	6,193
6	tool	83,461	news.bbc.co.uk	5,718
7	reference	74,602	www.flickr.com	5,645
8	web	70,538	java.sun.com	5,538
9	video	65,226	www.nytimes.com	5,222
10	music	61,246	www.microsoft.com	5,219
11	art	57,970	lifehacker.com	5,207
12	linux	47,965	www-128.ibm.com	4,569
13	tutorial	41,844	www.codeproject.com	4,429
14	java	40,780	www.wired.com	4,269
15	news	40,652	video.gooogle.com	4,261
16	game	39,391	www.techcrunch.com	3,818
17	free	39,006	www.bbc.co.uk	3,318

18	development	37,914	www.readwriteweb.com	3159
19	business	35,272	blogs.msdn.com	3,121
20	internet	34,580	msdn2.microsoft.com	2,950

The tag *blog* dominates in Delicious in the period from 2005 to 2007. Because most taggers on Delicious are presumed to be IT gurus, it is not surprising that tags such as *system*, *design*, *software*, *programming*, and *tool* are also included in the top 20 tags. The tags *Web* and *Internet* represent evergreen topics of this community. Because people in general like to share music, video, news, and games, these are also popular topics in Delicious. And, because people like things that available without charge, the tag *free* is seventeenth on this list of most highly ranked tags. Highly ranked bookmarks in Delicious include other major social networks (YouTube, livejournal, wikipedia, and Flickr), major news (BBC and the New York Times), and major computer giants (Microsoft, Google, IBM, and Sun), which indicates the social impact of these websites.

Power law distribution

The tagging data from Delicious, Flickr and YouTube was merged to form a single, comprehensive dataset. Using this combined dataset, an analysis of tag frequency was conducted. Figure 11 and Table 12 demonstrate that the distribution of tag frequency follows a power law distribution that conforms to Zipf's Law. Table 12 shows the details of this distribution: Only 1,363 of the 648,368 unique tags (or approximately 0.2% of all tags assigned between 2005 and 2007) were assigned more than 1000 times each, while 357,028 (or approximately 55% of all tags) were assigned only once.

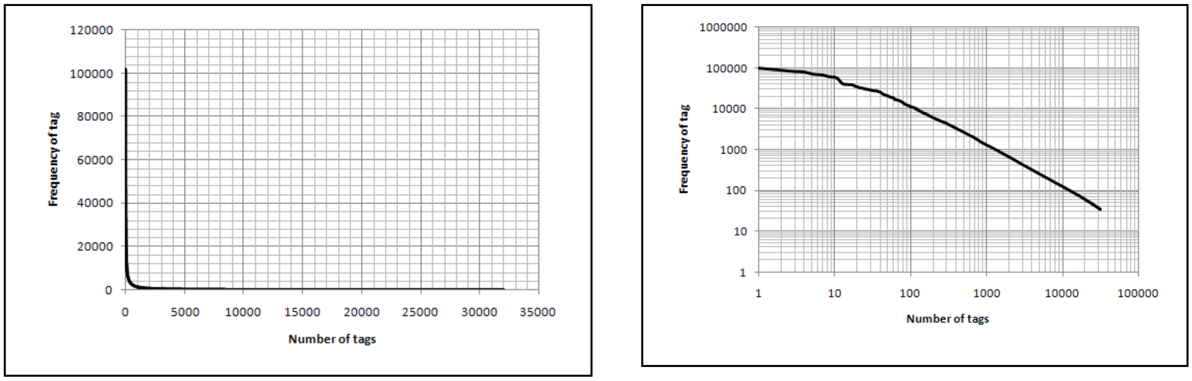


Figure 11. Distribution of Tag Frequency

In the combined dataset, the most frequently occurring tag is *design*, which accounts for 101,786 or nearly 1% of all tag occurrences. The second most frequently occurring tag is *blog* and accounts for 90,242 or 0.7% of the total tags assigned between 2005 and 2007. The 1,363 most frequently occurring tags account for a total of 6,210,163 tagging instances; these 1,363 tags comprise a core tagging vocabulary that represents more than 50% of the entire corpus of 12,077,183 tagging instances. (See the Appendix for a list of the 1,363 tags that make up the combined core tagging vocabulary of Delicious, Flickr and YouTube.) It is hoped that linguistic analysis of this core set of tags will reveal features of the evolving vocabulary of tags in each social tagging network.

Table 12: Tag Frequency Distribution

Tag Frequency Range	No. Unique Tags	Cumulative %
1	357028	55.07%
2-10	217746	88.65%
11-20	27404	92.88%

21-30	11524	94.65%
31-40	6656	95.68%
41-50	4454	96.37%
51-60	3387	96.89%
61-70	2461	97.27%
71-80	2066	97.59%
81-90	1597	97.83%
91-100	1348	98.04%
101-200	6193	99.00%
201-300	2151	99.33%
301-400	1044	99.49%
401-500	645	99.59%
501-1,000	1301	99.79%
1,001-120,000	1363	100.00%

Note: Cells in the column labeled *No. Unique Tags* represent the total of unique tags (e.g., when TagA is assigned by Tagger X and by Tagger Y, it is counted as one tag).

4.5 SOCIAL TAGGING ANALYSIS

In order to generate individual portraits of tag use and the composition of tag vocabularies in Delicious, Flickr and YouTube, data from each social network were analyzed independently using three time frames (2005, 2006, 2007).

Delicious

Table 13 shows the 20 most frequently assigned tags in Delicious for the years 2005, 2006, and 2007. These tag sets appear to be relatively stable across the three years. The tags *xml*, *science*, *search*, *games*, *technology*, and *security* appear among the top 20 tags for 2005 but are dropped from the lists of top 20 tags for 2006 and 2007; and the tags *imported*, *research*, and *internet* are dropped from the list of top 20 tags for 2007. The tags *development*, *howto*, *tutorial* and *Web 2.0* appear in the lists for both 2006 and 2007, and the tags *webdesign*, *free* and *opensource* are introduced in 2007, pointing to the emergence of new trends in user interests. Overall, 85% of the top 20 tags are stable across 2006 and 2007, indicating that a shared social vocabulary may be emerging in Delicious.

A profile of Delicious users can be generated through analysis of the lists of popular tags. The dominance of tags such as *blog*, *web*, *programming*, and *design* indicate key interests of Delicious users who are tagging bookmarks to store or share. While the tags *music*, *video*, *art* and *news* indicate a level of general interest that spans all three years, actual tagging evidence strongly supports the popular assumption that Delicious is a social network for IT gurus and other individuals interested in Web and programming skills. Furthermore, the tags introduced in 2006 and 2007 indicate a growing interest in free or open source resources as well as tutorials and how-to resources that support learning programming languages or applications and developing new computer skills.

Table 13: Top 20 Tags in Delicious for the Years 2005, 2006 and 2007

Rank	2005	2006	2007
1	blog/blogs	blog/blogs	blog/blogs
2	programming	programming	design
3	software	software	software
4	music	design	programming
5	design	reference	reference
6	web	music	tools
7	reference	web	Web 2.0
8	java	tools	web
9	art	art	video

10	tools	java	music
11	linux	video	art
12	news	Web 2.0	linux
13	xml	linux	webdesign
14	science	news	howto
15	search	tutorial	free
16	games	howto	tutorial
17	research	imported	news
18	technology	development	development
19	security	research	opensource
20	video	internet	java

Table 14: Top 20 Tags in Delicious for 2007 and Frequency of Assignment in the Years 2005, 2006 and 2007

Top 20 Tags in Delicious for 2007	2005	2006	2007	2006/2005	2007/2005	2007/2006
blog/blogs	6731 (1)	29485 (1)	90474 (1)	4	13	3.1
design	3045 (5)	19273 (4)	78115 (2)	6	26	4.1
software	3558 (3)	19533 (3)	60405 (3)	5	17	3.1
programming	4295 (2)	21789 (2)	55237 (4)	5	13	2.5
reference	2541 (7)	16643 (5)	53971 (5)	7	21	3.2
tools	1943 (10)	13340 (8)	53772 (6)	7	28	4.0
Web 2.0	658 (-)	10620 (12)	50270 (7)	16	76	4.7
web	2743 (6)	14115 (7)	44406 (8)	5	16	3.1
video	1114 (20)	11383 (11)	43847 (9)	10	39	3.9
music	3325 (4)	15523 (6)	39859 (10)	5	12	2.6
art	2344 (9)	12043 (9)	37518 (11)	5	16	3.1
linux	1799 (11)	10434 (13)	34241 (12)	6	19	3.3
webdesign	688 (-)	6542 (-)	33224 (13)	10	48	5.1
howto	962 (-)	8588 (16)	31701 (14)	9	33	3.7
free	643 (-)	5793 (-)	30750 (15)	9	48	5.3
tutorial	895 (-)	8683 (15)	30648 (16)	10	34	3.5
news	1712 (12)	8854 (14)	28086 (17)	5	16	3.2
development	1107 (-)	7588 (18)	27322 (18)	7	25	3.6
opensource	872 (-)	6468 (-)	25735 (19)	7	30	4.0
java	2449 (8)	11606 (10)	25732 (20)	5	11	2.2

Note: Numbers in parentheses in the columns labeled *2005*, *2006* and *2007* reflect the ranking of a term for that particular year. The column labeled *2006/2005* indicates that the value in each cell is the result of dividing the value for 2006 by the value for 2005. The result indicates the increase in raw numbers of frequency of tag assignment from 2005 to 2006. This also applies to the columns labeled *2007/2005* and *2007/2006*.

Figure 12 shows the evolution of dominant topical tags used in the Delicious social network for the period 2005-2007. The tag *Web2.0* shows the highest peak in both 2006 and 2007: The raw frequency with which *Web2.0* was used to tag bookmarks increased 16 times in 2006 and 76 times in 2007 when compared with its raw tagging frequency in 2005. The tags showing the most dramatic increase in raw tagging frequency from 2006 to 2007 were *webdesign*, *free* and *Web2.0*, indicating growing interest in these topics on the part of Delicious taggers. The three tags with the least impressive increase in raw tagging frequency from 2006 to 2007 were *java*, *programming*, and *music*. While this might seem to indicate waning interest in these topics, only the ranking for *java*, which dropped from eighth most popular tag in 2005 to twentieth most popular in 2007 (see Table 14), appears to support this conclusion. The tag *programming* drops from second position in 2005 and 2006 to fourth position in 2007; however, this is not a drop in popularity dramatic enough to justify any assumptions about waning interest on the part of Delicious taggers. The tag *music* does demonstrate a more dramatic drop in popularity -- from fourth position in 2005, to sixth position in 2006, and to tenth position in 2007 -- but the fact that Last.fm became one of the more popular social networks for sharing music during this period may help to explain why use of the tag *music* decreased from 2005 through September 2007.

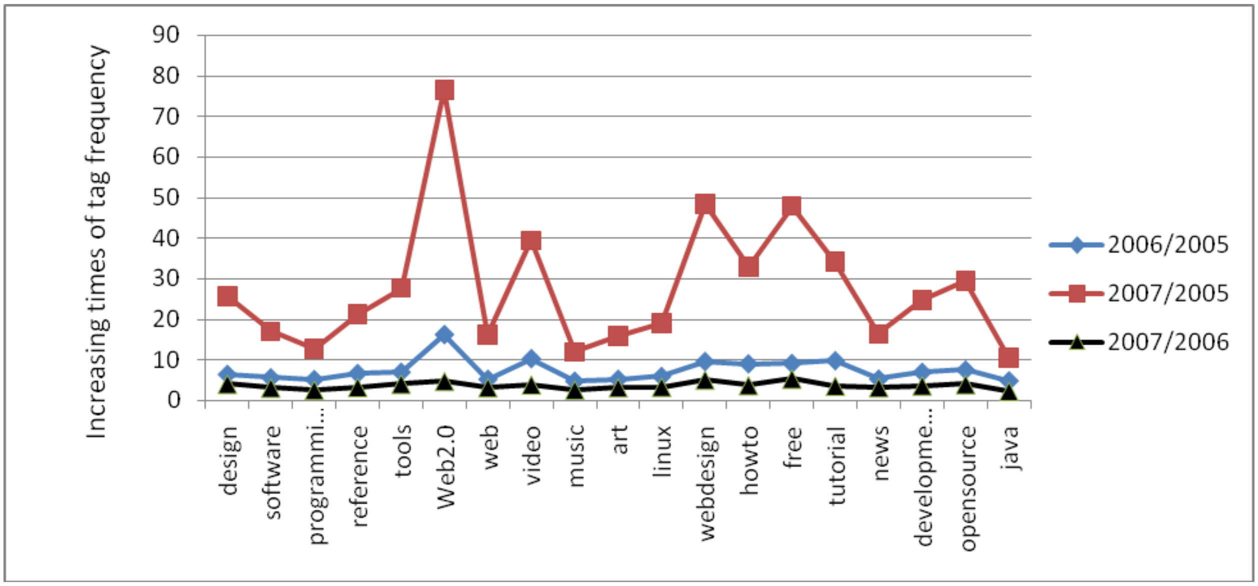


Figure 12. Evolution of the Top 20 Tags in Delicious for the Period 2005-2007.
The line with diamonds represents the increase in tag frequency from 2005 to 2006 (tag frequency for 2006/tag frequency for 2005). The line with squares represents the increase in tag frequency from 2005 to 2007 (tag frequency for 2007/tag frequency for 2005). The line with triangles represents the increase in tag frequency from 2006 to 2007 (tag frequency for 2007/tag frequency for 2006).

Flickr

Table 15 shows the 20 most frequently used tags in Flickr for the years 2005, 2006, and 2007. In sharp contrast to the more topical tagging culture of Delicious, Flickr taggers like to tag photographs with dates, locations, colors, and seasons. Favorite locations in Flickr include Hong Kong (2005), Germany (2005), USA (2006 and 2007), London (2005-2007), California (2006), and Japan (2007). Favorite color tags are *orange* (2005), *blue* (2006 and 2007), *red* (2006 and 2007), *green* (2006 and 2007), and black-and-white (i.e., *bw* in 2007). Most frequently used tags for seasons are *autumn* and *fall* (2007). In addition, users also favor tagging photographs with time of day or lighting conditions, especially when the photographs are night views. With the exception of tags in the categories year, color and location, topics of the top 20 tag sets differ widely across the three years.

Flickr taggers frequently assign informal tags to photographs (e.g., *me*), indicating that users may be tagging photographs for purposes of storing and retrieving them for their own use rather than with any intent to share them with others. When tagging photographs, users tend to emphasize the eye-catching features of an image such as color, subject (e.g., *sky*, *water*, *beach* and specific locations), and lighting conditions (e.g., *night* and *nightview*). Nonetheless, time (i.e., year, season, or month), locations and colors are the major features of images tagged by users. It could be useful to analyze the tagging culture of Flickr in greater detail given that the annotation of images is an important consideration in image retrieval.[11]

Table 15. Top 20 Tags in Flickr for the Years 2005, 2006 and 2007

Rank	2005	2006	2007

[11] An interesting example of ongoing research on social annotation of images and videos is GWAP, the "games with a purpose" project at Carnegie Mellon available at http://www.gwap.com/gwap/

1	2005	usa	2007
2	d70	california	canon
3	tsimshatsui	2006	nature
4	hongkong	cameraphone	autumn
5	nightview	celltagged	art
6	germany	zonetag	nikon
7	newkie	sanfrancisco	water
8	ragbrai	blue	bw
9	art	light	red
10	wonder	sky	blue
11	night	urban	sky
12	buttersweet	red	japan
13	15fav	sea	fall
14	central	me	beach
15	light	water	portrait
16	marco	nature	london
17	london	marco	night
18	apargioides	london	green
19	orange	green	usa
20	ads1	music	november

Figure 13 and Table 16 show the temporal history of tag popularity in Flickr for the period 2005-2007. In 2005 and 2006, tagging was not particularly popular in the Flickr community, with the total number of tags at 3,598 for 2005 and 23,066 for 2006. However, as tagging became more popular on the Web, tagging behavior in Flickr appears to have changed dramatically: 1,324,537 tags were assigned by Flickr users from January through September of 2007, approximately 50 times more tags than were assigned for all of 2006. Raw tagging frequency for *cannon*, the second most popular tag in 2007, increased 203.5 times over its total use in 2006; but *fall*, the thirteenth most popular tag in 2007, showed the greatest jump, increasing 672.5 times over its raw frequency of assignment in 2006.

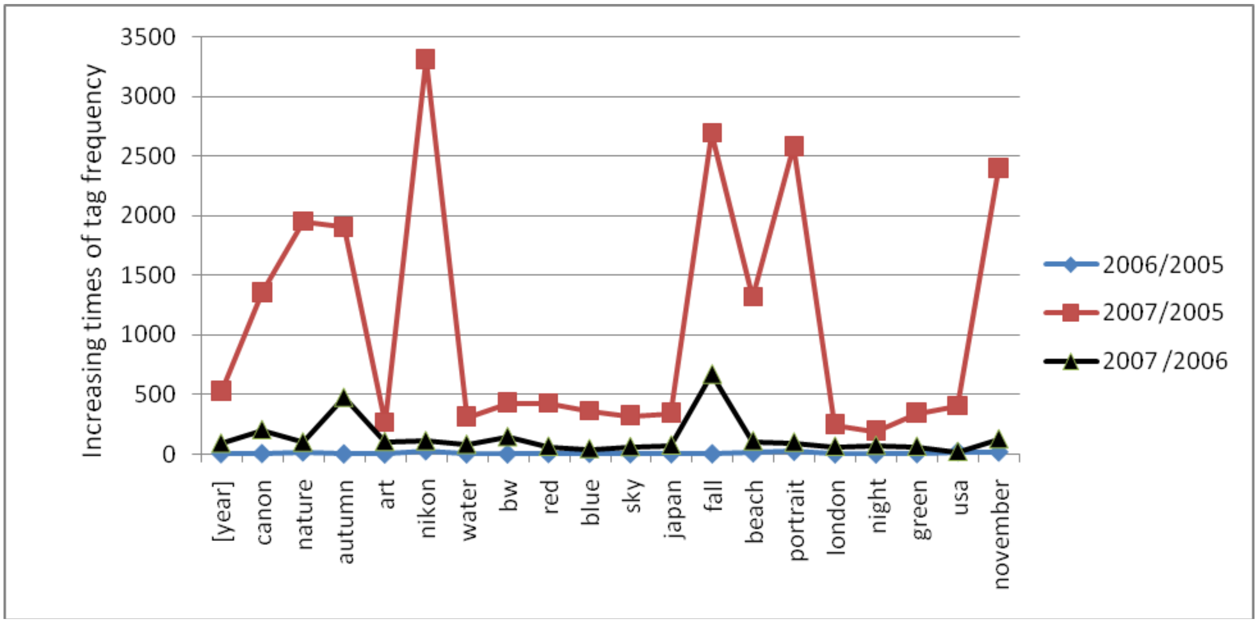


Figure 13. Evolution of the Top 20 Tags in Flickr for the Period 2005-2007.
The line with diamonds represents the increase in tag frequency from 2005 to 2006 (tag frequency for 2006/tag frequency for 2005). The line with squares represents the increase in tag frequency from 2005 to 2007 (tag

frequency for 2007/tag frequency for 2005). The line with triangles represents the increase in tag frequency from 2006 to 2007 (tag frequency for 2007/tag frequency for 2006).

Table 16. Top 20 Tags in Flickr for 2007 and Frequency of Assignment in the Years 2005, 2006 and 2007

Top 20 Tags in Flickr for 2007	2005	2006	2007	2006/2005	2007/2005	2007 /2006
[year]	21 (1)	124 (3)	11112 (1)	6	529	89.6
canon	3 (-)	20 (-)	4070 (2)	7	1357	203.5
nature	2 (-)	39 (16)	3899 (3)	20	1950	100.0
autumn	2 (-)	8 (-)	3804 (4)	4	1902	475.5
art	13 (9)	33 (-)	3416 (5)	3	263	103.5
nikon	1 (-)	30 (-)	3312 (6)	30	3312	110.4
water	10	39 (15)	3126 (7)	4	313	80.2
bw	7 (-)	21 (-)	3028 (8)	3	433	144.2
red	7 (-)	47 (12)	2988 (9)	7	427	63.6
blue	8 (-)	66 (8)	2888 (10)	8	361	43.8
sky	9 (-)	48 (10)	2878 (11)	5	320	60.0
japan	8 (-)	37 (-)	2738 (12)	5	342	74.0
fall	1 (-)	4 (-)	2690 (13)	4	2690	672.5
beach	2 (-)	24 (-)	2636 (14)	12	1318	109.8
portrait	1 (-)	26 (-)	2581 (15)	26	2581	99.3
london	10 (17)	39 (18)	2503 (16)	4	250	64.2
night	13 (11)	35 (-)	2489 (17)	3	191	71.1
green	7 (-)	38 (19)	2417 (18)	5	345	63.6
usa	6 (-)	126 (1)	2406 (19)	21	401	19.1
november	1 (-)	19 (-)	2394 (20)	19	2394	126.0

Note: Numbers in parentheses in the columns labeled *2005*, *2006* and *2007* reflect the ranking of a term for that particular year. The column labeled *2006/2005* indicates that the value in each cell is the result of dividing the value for 2006 by the value for 2005. The result indicates the increase in raw numbers of frequency of tag assignment from 2005 to 2006. This also applies to the columns labeled *2007/2005* and *2007/2006*.

Interestingly, an analysis of tagged photographs indicates that there may be two major communities of Flickr taggers. One community appears to consist of non-professional photographers who use Flickr as a platform for sharing photographs with friends and family and thus tag images so that they can be retrieved by others. The second community appears to consist of professional photographers who do not tag regularly but who frequently provide comments on photographs taken by other professionals.

YouTube

Table 17 shows the 20 most popular tags in YouTube for the years 2005, 2006 and 2007. The topics that are most frequently tagged in this social network are music, videos, humor, sex and girls, apparently reflecting the broad interests of the general Web community.

Table 17. Top 20 Tags in YouTube for the Years 2005, 2006 and 2007

Rank	2005	2006	2007
1	music	the	the
2	funny	funny	music
3	video	music	funny
4	the	video	video
5	dance	live	girl
6	crazy	of	of
7	commercial	comedy	sexy

8	dancing	dance	live
9	live	rock	dj
10	AMV	cat	2007
11	fun	Halloween	dance
12	guitar	love	hot
13	hot	girl	comedy
14	girl	movie	rock
15	japan	dj	love
16	anime	in	and
17	Halloween	sexy	sex
18	cat	and	in
19	halo	fight	new
20	of	you	cat

Tagging activity in YouTube increased dramatically between 2005 and 2007. The total number of tags assigned in YouTube increased from 4,735 in 2005, to 366,147 in 2006, to 1,073,042 in 2007. Tag use was 78.7 times greater in 2006 and 236.7 times greater in 2007 than it was in 2005. Compared with 2005, the tag [*year*] had the greatest increase in use in 2007, followed by *new* and *sex*/*sexy*, while *dance* showed the least increase between 2005 and 2007. The tag set in YouTube appears to be more stable than that of Flickr for the same time period, seemingly indicating that areas of user interest have remained fairly steady for the Social Web community as a whole (see Figure 14 and Table 18).

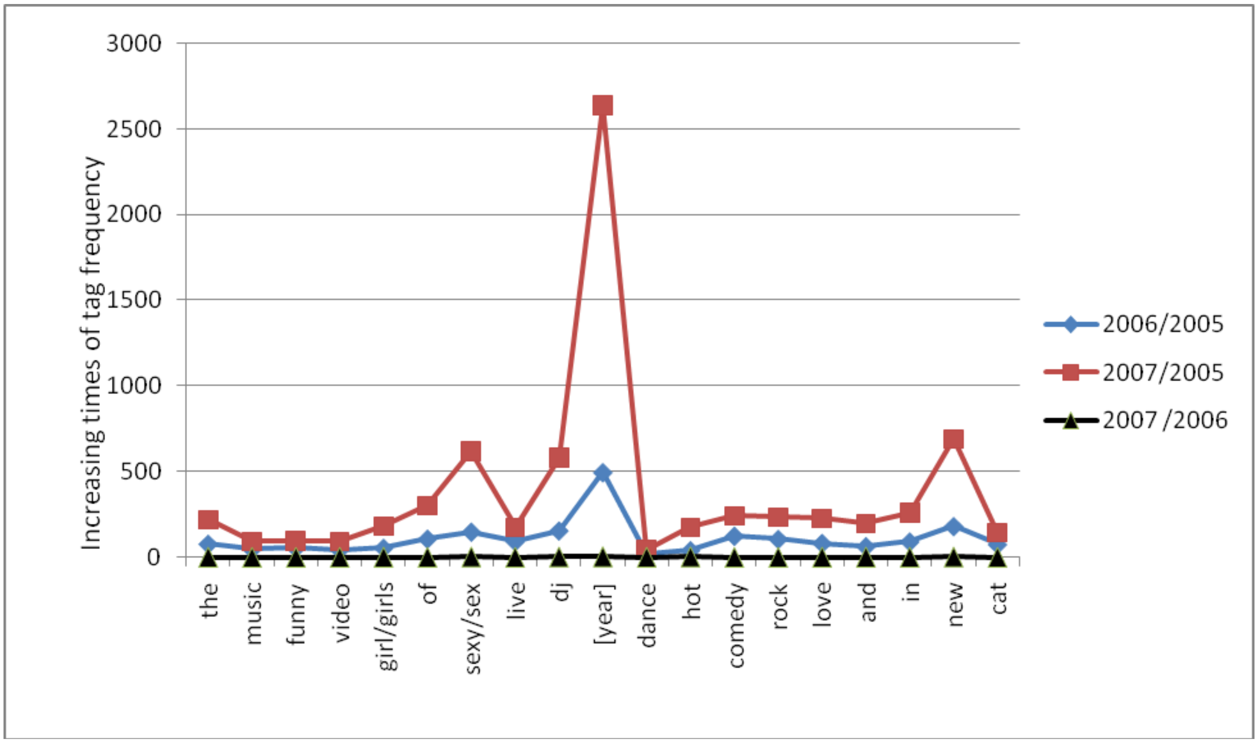


Figure 14. Evolution of the Top 20 Tags in YouTube for the Period 2005-2007.
The line with diamonds represents the increase in tag frequency from 2005 to 2006 (tag frequency for 2006/tag frequency for 2005). The line with squares represents the increase in tag frequency from 2005 to 2007 (tag frequency for 2007/tag frequency for 2005). The line with triangles represents the increase in tag frequency from 2006 to 2007 (tag frequency for 2007/tag frequency for 2006).

Table 18. Top 20 Tags in YouTube for 2007 and Frequency of Assignment in the Years 2005, 2006 and 2007

Top 20 Tags in YouTube for 2007	2005	2006	2007	2006/2005	2007/2005	2007 /2006
the	42 (4)	3240 (1)	9371 (1)	77	223	2.9
music	67 (1)	3080 (3)	6452 (2)	46	96	2.1
funny	58 (2)	3091 (2)	5784 (3)	53	100	1.9
video	53 (3)	2234 (3)	5065 (4)	42	96	2.3
girl/girls	25 (14)	1334 (13)	4647 (5)	53	186	3.5
of	13 (20)	1390 (6)	3955 (6)	107	304	2.8
sexy/sex	9 (-/-)	1338 (17/-)	5601 (7/17)	149	622	4.2
live	17 (9)	1563 (5)	3028 (8)	92	178	1.9
dj	5 (-)	777 (15)	2920 (9)	155	584	3.8
[year]	1 (-)	498 (-)	2641 (10)	498	2641	5.3
dance	56 (5)	1061 (8)	2526 (11)	19	45	2.4
hot	14 (13)	552 (-)	2467 (12)	39	176	4.5
comedy	10 (-)	1245 (7)	2461 (13)	125	246	2.0
rock	10 (-)	1059 (9)	2380 (14)	106	238	2.2
love	10 (-)	817 (12)	2294 (15)	82	229	2.8
and	11 (-)	689 (18)	2190 (16)	63	199	3.2
in	8 (-)	723 (16)	2095 (18)	90	262	2.9
new	3 (-)	544 (-)	2079 (19)	181	693	3.8
cat	13 (18)	977 (10)	1906 (20)	75	147	2.0

Note: Numbers in parentheses in the columns labeled *2005*, *2006* and *2007* reflect the ranking of a term for that particular year. The column labeled *2006/2005* indicates that the value in each cell is the result of dividing the value for 2006 by the value for 2005. The result indicates the increase in raw numbers of frequency of tag assignment from 2005 to 2006. This also applies to the columns labeled *2007/2005* and *2007/2006*.

5. UTILIZING WEB 2.0 IN WEB SERVICE RANKING

5.1 SYSTEM DESCRIPTION

In our third study of social networks and semantics, we propose a social ranking approach for Web service selection and ranking that is based on Delicious, one of the largest social networks. The system is quite simple and straightforward. Given a set of Web services, the system checks to determine if there are Web pages in the Web services domains that have been bookmarked in Delicious. When this is the case, the services are ranked based on the numbers of users in Delicious who have tagged the associated Web pages. For example, a user would like to know which of the shipping services have high social visibility. He identifies relevant keywords and runs his query against the service repository to get a set of .wsdl shipping services. For each Web service, the system then checks to determine if there are Web pages from the same domain already bookmarked in Delicious. These Web pages are retrieved from the Web and stored locally. Finally, the resulting set of Web services is ranked based on how many users have tagged the corresponding bookmarks and then displayed to the user.

5.1.1 Architecture

The system we propose consists of a set of loosely coupled components: an annotations search engine that retrieves relevant Delicious bookmarks given a user query; a Web service finder that crawls the domain of the bookmarks and identifies relevant .wsdl files; and, finally, a Web service ranker that performs the ranking of Web services. The system is implemented in Java and runs as a Web application hosted in the Apache Tomcat container. The rest of this section details each component of the system and describes how these components work together.

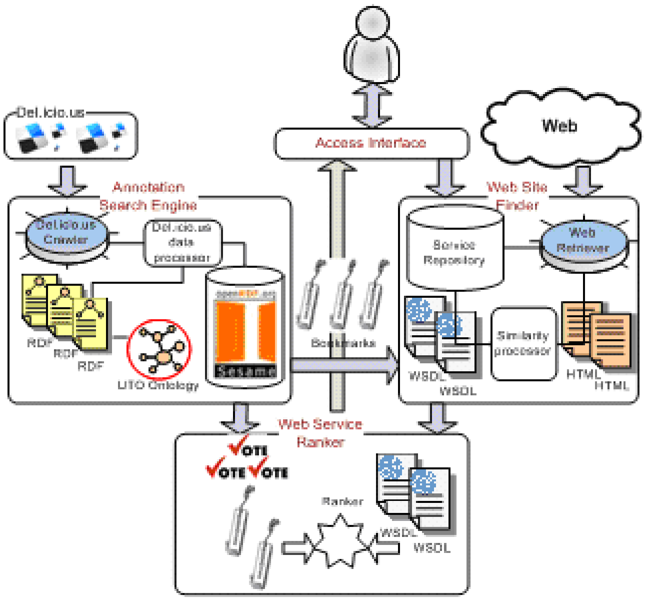


Figure 15. System Architecture

Access Interface

The access interface component is the visual gateway to the system. It allows the user to formulate a query as a set of keywords and to submit them to the system. Once a request has been processed by the system, the resulting set of ranked services is returned to the access interface, which then displays them to the user. The access interface is implemented using JSP technology.

Annotation Search Engine

The functionality of the annotation search engine can be briefly summarized as follows: Given a user keyword query, the annotation search engine returns a set of Delicious bookmarks relevant to that query. As illustrated in Figure 15, the annotation search engine includes a set of subcomponents: the Delicious Crawler, the Delicious Data Processor, and the Delicious Semantic Repository.

Delicious Crawler: This subcomponent is responsible for crawling social information from the Delicious social network. We chose an internally developed multicrawler designed for crawling major social networks, including Delicious, Flickr and YouTube (Ding, Toma et al., 2008). This crawler is based on the Smart and Simple Webcrawler. Since Flickr and YouTube are platforms for sharing photos and videos and not Web services, they were not considered in our research. Figure 15 shows the detailed class diagram of the crawler. Another option we initially considered to retrieve relevant bookmarks, given a query, was the Delicious API. This might seem a nicer approach, but it faced serious limitations. More precisely, due to authentication issues, a user can only access his own data (i.e., tags, bookmarks, updates and bundles), thus prohibiting access to all bookmarks relevant to a query.

Delicious Data Processor: This subcomponent is responsible for processing.html pages crawled from Delicious, parsing them, and generating structured, light weight semantic annotated data represented in RDF (Manola & Miller, 2004). RDF is a standard language and method for making statements about Web resources in the form of subject-predicate-object. Since RDF is schema independent, we have used the Upper Tag Ontology (Ding, Kang et al., 2008) as a model to structure RDF social data. The Upper Tag Ontology (UTO) is a light ontology that includes concepts such as *Tag*, *Tagger*, *Object*, and *Vote* and relations such as *hasTag*, *hasVote*, and *hasObject*. These ontological elements are used to build RDF representations of the data crawled from Delicious.

Delicious Semantic Repository: This subcomponent stores RDF triples generated by the data processor according to UTO. We chose Sesame (Broekstra, Kampman, & van Harmelen, 2002) as a semantic repository for storing Delicious data because it offers efficient storage for RDF (Manola & Miller, 2004) and RDF schema (Brickley & Guha, 2004), which can be deployed on top of RDBMS repositories. Although an RDBMS repository could have been used, the choice of a semantic repository was motivated by the availability of flexible data structures that can be stored without the need to extend or change the data model.

Web Site Finder

The website finder is the top-level component responsible for mining correspondences between Web services and possible Web pages in the domain of Web services that have been bookmarked on Delicious. Given a Web service description, this component checks to determine if any Web pages having the same domain have been bookmarked in Delicious. If any such pages exist, they are retrieved from the Web and their similarity with the Web service description is computed. Finally, a correspondence matrix between Web services and the most similar Web pages is built.

As illustrated in Figure 15, the website finder includes a set of subcomponents: the Web Retriever, the Service Repository, and the Similarity Processor.

Web Retriever: This subcomponent is responsible for retrieving .html files from the domain of the Web service. More precisely, the Web retriever component first makes a query to the Delicious semantic repository and receives URLs of Web pages from the same domain as the Web service. It then retrieves the content of these pages from the Web and stores them locally.

Service Repository: This subcomponent stores the .wsdl files identified by the processor subcomponent together with the set of keywords in the query. It is simply a persistent layer implemented using a classic RDBMS. The access interface component interacts with the service repository component, allowing the user to search for services based on the keywords provided as input.

Similarity Processor: This subcomponent determines the similarity between a Web service .wsdl description and each page in its domain that has been bookmarked in Delicious. To determine the similarity between a Web service and a website, we use two techniques. First, we check for co-occurrence in the set of keywords used to name Web service operations and the keywords in the Web page. Second, we check for co-occurrence in the set of keywords used to name Web service operations and the annotation keywords (or tags) that have been assigned to the Web page in Delicious.

We regard services, Web pages and the set of tags used to annotate a Web page as sets of keywords. To build the vector model for services, we consider only the names of the operations exposed by the services. For Web pages, the initial preprocessing step consists of removing all html tags. For all keyword sets representing services, Web pages, and tags, further preprocessing is performed using stopwords and stemming. We then apply the cosine similarity metric (Manning, Raghavan, & Schutze, 2008) to measure similarity between two sets of keywords representing services, Web pages, or tags.

Web Service Ranker

The Web service ranker component is the core component of the system that is responsible for the actual ranking of Web services. This component first performs a ranking of Delicious bookmarks, which translates into the ranking of associated Web services. The straightforward approach we adopt is to rank each bookmark based on how many people have bookmarked the Web page or how many votes the bookmark has. This information is already available in the Delicious semantic repository, having been crawled from the Delicious website. The ordered list of bookmarks based on the number of votes is then used to generate the ordered list of Web services that is presented to the user.

5.1.2 ALGORITHM

The current implementation of this Web service ranking system follows a simple and straightforward approach with respect to how the actual ranking of services is determined. Basically, the number of votes received by a relevant bookmark in Delicious becomes the rank value of the service. In the rest of this section, a general algorithm to compute the rank of a Web service is proposed. Our algorithm is based on the Social Page Rank (SPR) algorithm proposed by Bao et al. (2007) and has been extended to address the Web service dimension. The SPR algorithm evaluates the popularity of a Web page based on mutual enhancement among three distinct sets of objects: (a) popular Web pages, (b) Web users, and (c) hot social annotations. We introduce Web services as a fourth dimension or set of objects, and call the new algorithm the Social Web Service Rank (SocWSRank). The algorithm is provided below.

Algorithm 1 Social Web Service Rank (SocWSRank)

```
Require: Set of association matrixes $M_{AP}, M_{PU}, M_{UA},$
$M_{PS}$
Ensure: $S^*$ the converged Social Web Service Rank
While $S_t$ not converged do
    $P_t = M_{PS} * S_t$
    $U_t = M_{PU}^T * P_t$
    $A_t = M_{UA}^T * U_t$
    $P_t' = M_{AP}^T * A_t$
    $A_t' = M_{AP} * P_t'$
    $U_t' = M_{UA} * A_t'$
    $P_{t+1} = M_{PU} * A_t'$
    $S_{t+1} = M_{PS}^T * P_{t+1}$
end while
```

If *P* is the set of Web pages having N_P elements, *U* is the set of users having N_U elements, *A* is the set of annotations having N_A elements, and *S* is the set of Web service having *NS* elements, the association matrixes can be defined between annotation and pages (MAP), between pages and users (MPU), between users and annotations (MUA), and, finally, between Web pages and Web services (MPS). The matrix elements are assigned values that capture the associations between each pair of dimensions. For example, the MAP (*pi, uj*) element is assigned the number of annotations (tags) with which a user *uj* annotates a page *pi*. Elements of the matrixes MAP and MUA are computed as described by Bao et al. (2007). Elements of the matrix MPS are assigned the value 1 if there is a correspondence between the Web page and the service or the value 0 if not.

5.2 EXPERIMENTS AND DISCUSSIONS

We performed a set of experiments using three datasets: a social dataset built using Delicious data and two Web service datasets. The social dataset was created by crawling Delicious data. Once crawled, the data was annotated using the UTO ontology and stored as RDF triples in a Sesame repository. The Delicious dataset totaled 1.64 GB of data.

The other two datasets are .wsdl datasets that were provided by seekda OG, an Austrian company. The first one is an 80 file dataset containing services providing shipping functionality that have valid .wsdl descriptions and are accessible online. The second one is a larger dataset containing 5,000 services from different domains with valid .wsdl files and available online.

For the dataset with 80 shipping services, we found that only three of the domains had pages annotated in Delicious. For each of these domains, we queried Sesame and determined the total number of Web pages annotated to be 14. The distribution of Web pages per Web service domains was between zero and seven Web pages.

For the dataset with 5000 services, 86 of them had Web pages from their domain that were annotated in Delicious. The number of Web pages per domain varied from zero to 131. In total, we found 547 Web pages annotated in Delicious that were connected to the initial dataset.

One can notice that, in both cases, the number of pages annotated in Delicious that actually have a correspondence with the Web services dataset is rather limited. However, for a small number of services we were able to find a considerable number of pages, some of them having a strong correlation/similarity with the Web services. We can conclude that, even though the link between the Web service data and Delicious data is, in general, not very strong, when such a link is visible, it can be used both to determine more information about the services and to rank them.

6. DISCUSSIONS AND CONCLUSIONS

In this chapter we have investigated the connection between Social Networks and Semantics. We have shown how synergies between these two areas can be used to solve concrete problems and we have described three approaches that show the true potential of interconnecting these technologies. The uniqueness of our three proposed approaches is given by the large-scale social data analysis, the semantic integration of data for social web and the usage of social web data for service ranking. Related studies and approaches suffer from a set of disadvantages that were mentioned in Section 2. These include: (1) a small set of the data considered in the studies, (2) lack of intelligent solutions for data integrating and cross-domain and cross-network search and (3) lack of using the social tagging features for ranking of services. In the following sub-sections we briefly summarize and discuss the contribution of each of our three approaches.

6.1 SEMANTIC PROFILING OF SOCIAL NETWORKS

We have presented the first large-scale study of MySpace in an effort to better understand this new social phenomenon. Our comparative study differs from previous work both in its scale (over 1.9 million profiles) and in its breadth. In particular, we have examined how MySpace users participate in the social network (sociability), how they describe themselves (demographics), and how they communicate their personal interests and feelings (language models). The core findings of the MySpace study are:

- Nearly half of the profiles on MySpace have been abandoned, indicating that perceptions of the overall growth and explosive rate of user interest in social networks may need to be tempered; however, we have also identified a large core of active users within MySpace who account for the vast majority of friends, comments, and group activity.
- Younger users (in their teens and 20s) are most prevalent on MySpace: Women are most prevalent in the younger age groups (14 to 20), whereas men are most prevalent for all other age groups (21 and up).
- There are clear patterns of language use based on user age, location, and gender, which is a useful observation both for text mining and for characterization of applications. We have identified class-specific distinguishing terms and language model clusters that could be used to identify deceptive users attempting to misrepresent their demographics.
- Overall, the fraction of private profiles is increasing with time, indicating that new adopters of social networks may be more attuned to the inherent privacy risks of adopting a public Web presence. We found that women favor private profiles by 2-to-1 over men, and that, perhaps counterintuitively, younger users are more likely to adopt a private profile than older users. We also found that the more connected a user was in the social network, the more likely she was to adopt a private profile.

We have identified a number of surprising and interesting features that motivate our continuing research. In particular, we are interested in augmenting and extending models of social network growth to incorporate the demographic variations we have observed. Along this line, we believe fine-grained language models that move beyond age, gender, and location to capture user interests and expectations of the social network (e.g., for business-development networking, for making friends) could be beneficial.

6.2 MEDIATING AND ANALYZING SOCIAL DATA

Another problem addressed in this chapter was mediation of and alignment between social tagging systems. We have proposed the Upper Tag Ontology (UTO) as a mediating tool and have described how UTO can be used to align metadata represented according to differing tagging ontologies.

To demonstrate the application of UTO and the Social Tagging crawler, we reported on a study of three social networks: Delicious, Flickr and YouTube. When comparing these three social networks, Delicious demonstrates the tightest connection to the use of tags as extended information about resources. In Delicious, every user can tag an object with the tag(s) of his own choice; and any object can be tagged many times and by many different users, thereby indicating that it "belongs" (or is highly relevant) to the Delicious community as a whole. Delicious exemplifies community tagging, where anyone can tag (or bookmark) any online resource (Marlow et al., 2006). Similar social bookmarking networks include CiteULike and Connotea, where tagged resources are bibliographical records, and LibraryThing, where tagged resources are books.

Social networks such as Delicious, CiteULike, and LibraryThing are very different from Flickr, where a resource (i.e., a photograph) is generally tagged only by the individual who has uploaded it. The major activity of other members of the Flickr community is to "comment on" or "vote for" resources by indicating that a particular photograph is a favourite image. Flickr also provides users with the ability to allow friends to tag photos they have uploaded; but this functionality limits tagging behavior -- and thus the development of a sense of community -- in that it prohibits open tagging by Flickr users at large. Because tagging a resource in Flickr is not generally open to everyone (one notable exception being the Library of Congress's photostream[12]), Flickr cannot be considered a true community-based tagging system; rather, it is better thought of as a self-tagging system for users and their close friends. YouTube operates in a manner very similar to that of Flickr, allowing an individual to tag the resources (i.e., videos) he has uploaded while limiting the participation of other Flickr users to voting for resources by assigning "stars".

These differences in tagging rights have created differences not only in the roles that tags play in each system but also in the nature of the tags that are assigned (Marlow et al., 2006). Based on an analysis of the top 20 tags in each of the three social networks, it is apparent that tags in Delicious are more content-oriented in that they are generally related to the topics of the resources bookmarked. The tags used in Flickr are more like simplified, one-word descriptions in that they are generally related to the physical features of the photographs themselves, such as colors, lighting and location. While tags in Delicious are likely to reflect the intellectual content of resources and those in Flickr generally represent the physical features of photographs, tags in YouTube tend to focus on the medium or genre of the resources (e.g., *music*, *video*, *comedy*, *movie*, *tv*) and on the affective judgments of taggers (e.g., *funny*, *sexy*, *hot*, *love*, *new*).

The role of tags in Delicious is to represent bookmarked resources not only for future retrieval by the tagger but also for sharing them with the larger community. Tags play a major role in Delicious: Without the tags assigned by users of the social network, there would be no means either to share bookmarks or to identify and retrieve resources, which are the main functions of Delicious. In contrast, tagging does not play a major role in Flickr. Because decisions as to whether or not to tag a photograph and who may tag it are left to the individual uploading a photograph, tagging in Flickr is more of a secondary activity or side effect. Furthermore, photographs on Flickr can be searched for and retrieved by their titles and are ranked by comments or votes rather than by the number of tags assigned. This is also the case with YouTube in that videos are most frequently shared based on comments and votes rather than on assigned tags. Indeed, it appears that many YouTube users may not understand the purpose of tagging: Instead of adding specific tags, users often tag their videos by enter a description in the tagging field, which accounts for the occurrence of helping words such as articles, prepositions and conjunctions (e.g., *the*, *of*, *in*, and *and*) among the more popular tags in YouTube.

[12] http://www.flickr.com/photos/library_of_congress/collections/

Social tagging behaviors are also related to the community of users that populates each social network. Delicious gathers a community interested in IT-related topics. These individuals are focused on the content of bookmarked resources and tagging provides a way for them to summarize this content. In such a situation, tagging becomes the key function of the system and plays a major role in sharing and retrieving bookmarks. Users of Flickr are more interested in commenting on and sharing their photographs with family and friends. Rather than comprising a single, cohesive community, users in Flickr appear to gravitate toward one of two primary types of community: Communities of professional photographers who upload photographs for comment and feedback from other professionals and communities of non-professional users for whom Flickr provides a place to store personal photographs and share them with family and close friends. Alternatively, the community of YouTube can be viewed as a snapshot of the larger Web community. YouTube is populated by individuals from all over the world who are of different ages and have many different interests. They come to YouTube with many different purposes and expectations, and many of them do not tag their videos because the role of tagging is overshadowed by the roles of rating and/or commenting.

After analyzing social tagging behavior in Delicious, Flickr and YouTube, it is apparent that tagging activities increased tremendously between 2005 and 2007, when evermore individuals were using online social networks to tag resources for purposes of storage, access, and retrieval, both for themselves and for the purpose of sharing those resources with others. Through tag analysis, it is possible to develop a portrait of the social culture of these networks and, in some cases, to identify trends of emerging or waning topical interests among users.

While tag sets in Delicious appeared to have become more stable across the time frame of this study, it was also apparent that collective tagging vocabularies could benefit from both syntactic and semantic normalization of tags. For example, in YouTube in 2007 there were 2,796 uses of the tag *girl* and 1,851 uses of the tag *girls*. Normalization of singular and plural forms as well as acronyms and full names would increase the effectiveness of tags for retrieval purposes, as would standardization of the syntactical formation of tags (e.g., tag phrases with or without a space between individual terms). Perhaps as important is the introduction of user education regarding the potential utility of tags in social networks and the effective choice of tags (Ackerman, James & Getz, 2007).

This study demonstrates that it is possible to profile a social network by analyzing data about the tags and tagging behaviors of that network. For example, analysis confirms the popular assumption that the Delicious community is largely comprised of individuals interested in IT-oriented topics such as design and programming. In contrast, the Flickr community appears to contain two primary groups of users: professional photographers interested in feedback and non-professional photographers interested only in sharing their photographs. In contrast to Delicious and Flickr, the YouTube community is very broad and can be best viewed as a self-selected subset of the general Social Web community. Tagging is a major activity in Delicious but not in Flickr and YouTube. Tagging in Delicious is used primarily for purposes of storing, retrieving and sharing online resources across the community; tagging in Flickr emphasizes indexing objects for retrieval by the tagger and his friends and associates; and tagging in YouTube is undertaken primarily for identifying the genre of a video and for indicating the tagger's affective reaction to it. Thus, taggers are more likely to represent the content of a resource in Delicious, but they tend to focus on the specific features of an image in Flickr and the genre of a video in YouTube.

In Delicious, changing trends in user interest can be identified and tracked by analyzing tag frequencies across time; in both Flickr and YouTube, however, such trends are not obvious, perhaps because the focus of tagging activities is not on the intellectual content of resources but on more superficial features such as color (in Flickr) or affective reactions (in YouTube). Thus, even though YouTube has been characterized as a subset of the general Web population, the results of this research indicate that Delicious is a more representative venue for analyzing social tagging vocabularies and tagging behaviors within a community of users. This conclusion is supported by the finding that the community of users in Delicious is more cohesive than those of Flickr or YouTube; by the dynamic behavior of users that supports tracking of emerging and waning interests within the Delicious community; and by the participatory focus on sharing that characterizes user tagging activity in Delicious.

However, these conclusions require better understanding of the concept of "community". How is a community defined? What do we know about how communities form (or are formed) and develop? In that respect, it would be enlightening to compare "online" or "web" communities and their respective vocabularies with communities outside the Web environment (i.e., communities of scientists) and their vocabularies. Such an analysis would be beneficial for enhancing our understanding of both.

6.3 UTILIZING WEB 2.0 IN WEB SERVICE RANKING

Another problem investigated in this chapter was the use of Web 2.0 social annotations in Web service ranking. More precisely, we used the annotation data from Delicious, one of the biggest social networks, to discover and rank Web services connected by Delicious bookmarks. Following this straightforward idea, we have proposed a global algorithm to compute the social rank of Web services and designed and implemented a running prototype. In future work, we plan to compare and integrate our system with other approaches for ranking services. We are mainly interested in knowing if and in what situations social-based service rankers perform better than rankers using properties/descriptions of services (e.g., non-functional properties). The focus of this proposed research will not be on the comparison itself but on defining an integrated framework for ranking Web services that is able to select and use, in relevant situations, the most appropriate results of various ranking approaches, including those based on non-functional properties as well as those using socially based ranking.

REFERENCES

Ackerman, G., James, M., & Getz, C. T. (2007). The application of social bookmarking technology to the national intelligence domain. *International Journal of Intelligence and Counterintelligence, 20*, 678-698.

Acquisti, A., & Gross, R. (2006). Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *6th Workshop on Privacy Enhancing Technologies (PET)*. Retrieved August 23, 2009 from http://privacy.cs.cmu.edu/dataprivacy/projects/facebook/facebook2.pdf

Adamic, L. A., & Adar, E. (2005). How to search a social network. *Social Networks* 27(3), 187–203.

Antoniou, G., & van Harmelen, F. (2004). *A Semantic Web primer*. Cambridge, MA: MIT Press.

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks. In *Proceedings of the 12th ACM SIGKDD International Conference on KnowledgeDiscovery and Data Mining,* August 20-23, 2006, Philadelphia, PA, USA (pp. 44-54).

Bao, S. Xue, G. Wu, X., Yu, Y., Fei, B., & Su, Z. (2007). Optimizing web search using social annotations. In *Proceedings of the 16th international conference on World Wide Web* (pp. 501-510), New York, NY: ACM Press.

Batini, C., Lenzerini, M., & Navathe, B. (1986). A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys, 18*(4), 323-364.

Barnes, S. B. (2006). A privacy paradox: Social networking in the United States. *First Monday,* 11(9). Retrieved August 23, 2009 from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1394/1312

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284*(5), 34–43.

Boyd, D. (2007). Social network sites: Public, private, or what? *The Knowledge Tree: An e-Journal of Learning Innovation*. Retrieved August 22, 2009 from http://www.danah.org/papers/KnowledgeTree.pdf

Brickley, D., & Guha, R.V. (2004). Resource Description Framework (RDF) Schema Specification 1.0. Recommendation, World Wide Web Consortium, February 2004. Retrieved August, 2009 from http://www.w3.org/TR/2004/RECrdf- schema-20040210

Broekstra, J., Kampman, A., & van Harmelen, F. (2002). Sesame: A generic architecture for storing and querying RDF and RDF schema. In *Proceedings of the 2nd International Semantic Web Conference* (pp. 54-68). Berlin, Germany: Springer.

Coleman, J. (1990). *Foundations of Social Theory*. Boston, MA: Harvard University Press.

Ding, Y., Kang, S.J., Toma, I., Fried, M., Shafiq, O., & Yan, Z. (2008). Adding semantics to social tagging: Upper Tag Ontology (UTO). In *Proceedings of the 70th Annual Meeting of the American Society for Information Science & Technology (ASIS&T)*, Oct 24-29, 2008, Columbus, Ohio, USA.

Ding, Y., Toma, I. Kang, S., Fried, M., &Yan, Z. (2008). Data mediation and interoperation in Social Web: Modeling, crawling and integrating social tagging data. *Proceedings of the Workshop on Social Web Search and Mining (SWSM2008), 17th International World Wide Web Conference, Beijing, China.* Retrieved August 23, 2009 from http://info.slis.indiana.edu/~dingying/Publication/OTM2008-UTO-CameraReady.pdf

Dwyer, C., Hiltz, S. R., & Passerini, K. (2007). Trust and privacy concern within social networking sites. In *Proceedings of the Thirteenth Americas Conference on Information Systems*. Retrieved August 23, 2009 from http://csis.pace.edu/~dwyer/research/DwyerAMCIS2007.pdf

Ellison, N., Steinfield, C., and Lampe, C. (2006). Spatially bounded online social networks and social capital. In *International Communication Association*. Retrieved August 23, 2009 from http://msu.edu/~nellison/Facebook_ICA_2006.pdf

Fried, M. (2007): Social Tagging Wrapper. Bachelor Thesis, Institute of Computer Sciences, University of Innsbruck, Austria.

Gekas, J. (2006). Web service ranking in service networks. In *3rd European Semantic Web Conference ESWC '06* (pp. 501-510). Retrieved August 23, 2009 from http://www.eswc2006.org/poster-papers/FP08-Gekas.pdf

Golder, S. A., Wilkinson, D., & Huberman, B. (2007). Rhythms of social interaction: Messaging within a massive online network. In *Third International Conference on Communities and Technologies*. Retrieved August 23, 2009 from http://www.hpl.hp.com/research/idl/papers/facebook/facebook.pdf

Gradijan, D. (2007). MySpace cracks down on 69-year-old members. *CSO*. Retrieved August 29, 2009, from http://www.csoonline.com/article/216376/MySpace_Cracks_Down_on_69_Year_Old_Members?page=1

Gruber, T. (2007). Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web & Information Systems, 3*(2). Retrieved August 23, 2009 from http://tomgruber.org/writing/ontology-of-folksonomy.htm

Hinduja, S., & Patchin, J.W. (2008). Personal information of adolescents on the Internet: A quantitative content analysis of MySpace. *Journal of Adolescence, 31*(1), 125-146.

Hoschka, P. (1998). CSCW research at GMD-FIT: From basic groupware to the Social Web. *ACM SIGGROUP Bulletin, 19*(2), 5-9.

Hwang, C.L., & Yoon, K. (1981). *Multiple attribute decision making: Methods and applications*. Berlin, Germany: Springer-Verlag.

Kumar, R., Novak, J., and Tomkins, A. (2006). Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 611 - 617). New York: ACM Press.

Lampe, C., Ellison, N., & Steinfeld, C. (2007). Profile elements as signals in an online social network. In *Conference on Human Factors in Computing Systems*. Retrieved August 23, 2009 from http://www.msu.edu/~steinfie/CHI_manuscript.pdf

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences, 102*(33),11623-11628.

Liu, Y., Ngu, A.H., & Zeng, L.Z. (2004). QoS computation and policing in dynamic web service selection. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters* (pp. 66-73). New York, NY: ACM Press.

Manning, C.D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press.

Manola, F., & Miller, E. (2004). RDF primer. Recommendation, World Wide Web Consortium, February 2004. Retrieved August, 2009 from http://www.w3.org/TR/REC-rdf-syntax

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*, August 22-25, 2006, Odense, Denmark (pp. 31-40). New York: ACM.

Milgram, S. (1967). The small-world problem. *Psychology Today*, 2, 60-67.

Nussbaum, E. (2007). Kids, the Internet, and the end of privacy. *New York Magazine*. Retrieved August 23, 2009 from http://nymag.com/news/features/27341/

Page, L., & Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

Pelleg, D. & Moore, A. W. (2000). X-means: Extending K-means with efficient estimation of the number of clusters. In *Proceedings of the 17th International Conference on Machine Learning* (pp. 727-740). San Francisco, CA: Morgan Kaufmann Publishers.

Rahm, E. & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *VLDB Journal, 10*(4): 334-350.

Rosenbush, S. (2005). News Corp.'s place in MySpace. *Business Week*. Retrieved August 23, 2009 from http://www.businessweek.com/technology/content/jul2005/tc20050719_5427_tc119.htm

Spertus, E., Sahami, M., & Buyukkokten, O. (2005). Evaluating similarity measures: A large-scale study in the Orkut social network. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 678-684). New York: ACM Press.

Zeng, L.Z., Benatallah, B., Ngu, A.H., Dumas, M., Kalagnanam, J., & Chang. H. (2004). QoS: Aware middleware for web services composition. *IEEE Transaction on Software Engineering, 30*(5), 311-327.

Zhang, J., Ackerman, M., and Adamic, L. (2007). Expertise networks in online communities: Structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web* (pp. 221-230). New York: ACM Press.

Zhou, C., Chia, L.T., & Lee, B.S. (2005). Semantics in service discovery and QoS measurement. *IT Professional, 7*(2), 29-34.

APPENDIX

Core Tag Vocabulary in Social Networks: Top 1363 Tags in Delicious, Flickr and YouTube

	Core Tag Vocabulary
Numbers&others	1, 2, 3, 2005, 2006, 2007, -, .net, 3d
A	a, academia, academic, accessibility, accessories, acoustic, action, actionscript, activism, ad, admin, administration, adobe, ads, adsense, adult, advertising, advice, Africa, agency, aggregator, agile, ai, air, airline, airlines, airplane, airport, ajax, algorithm, algorithms, all, alternative, amateur, amazing, amazon, America, American, Amsterdam, analysis, analytics, and, angel, angst, animal, animals, animation, anime, anonymous, anthropology, apache, api, apple, application, applications, apps, architecture, archive, archives, argentina, art, arte, article, articles, artist, artists, arts, as3, asia, asian, asp.net, ass, asterisk, astronomy, at, atheism, atom, au, audio, audiobooks, Australia, authentication, auto, automation, autumn, awards, awesome
B	Baby, backup, bad, ball, band, bands, bandslash, bank, banking, bar, Barcelona, baseball, bass, bbc, beach, beatles, beautiful, beauty, beer, berlin, best, bible, bibliography, bicycle, big, bike, bioinformatics, biology, bird, birds, birthday, bit200f06, bit200w07, bittorrent, black, blackandwhite, blog, blogger, blogging, blogs, blood, blue, Bluetooth, boat, body, boobs, book, bookmarking, bookmarks, books, boston, boy, boys, bpm, brain, branding, brasil, brazil, bridge, Britney, Brooklyn, brown, browser, browsers, Buddhism, building, bus, bush, business, buy, bw, by
C	C, c#, c++, calculator, calendar, California, camera, cameraphone, camping, Canada, canon, car, card, cards, career, cars, cartoon, cartoons, cat, cats, cd, celebrity, cell, cellphone, celltagged, censorship, change, charity, charts, chat, cheap, cheatsheet, chemistry, Chicago, chicken, child, children, chile, china, Chinese, chocolate, chords, chris, Christian, Christianity, Christmas, church, ciencia, cine, cinema, city, class, classic, classification, climate, clip, clothes, clothing, clouds, club, cluster, clustering, cms, cocoa, code, coding, coffee, collaboration, collection, college, color, colors, colour, comedy, comic, comics, commercial, communication, community, company, comparison, competition, compiler, complexity, computer, computers, computing, concert, concurrency, conference, conferences, conspiracy, consumer, content, contest, control, conversion, convert, converter, cooking, cool, copyright, corporate, country, course, courses, cover, crack, craft, crafts, crazy, creative, creativecommons, creativity, credit, crime, crossover, cryptography, cs, css, cultura, culture, curiosidades, custom, cute, cycling
D	Daily, dance, dancing, dark, data, database, datamining, dating, david, day, dc, de, dead, deals, death, debian, del.icio.us, delicious, demo, democracy, design, designer, desktop, deutsch, Deutschland, dev, developer, development, dhtml, dictionary, diet, dig, digital, directory, diseÃ±o, Disney, distributed,

	distro, diy, dj, django, dns, do, documentary, documentation, dog, dogs, dom, domain, dotnet, download, downloads, drawing, driver, drm, drugs, drunk, drupal, duesouth, dvd
E	Earth, ebay, ebook, ebooks, eclipse, ecology, ecommerce, economia, economics, economy, editing, editor, edtech, educaciÃ³n, educacion, education, effects, el, elearning, e-learning, electronic, electronics, email, embedded, employment, emulation, en, encryption, encyclopedia, energy, engine, engineering, England, English, enterprise, enterprise2.0, entertainment, entrepreneur, entrepreneurship, environment, erlang, esl, espaÃ±a, espaÃ±ol, essay, ethics, eu, europa, Europe, event, events, evolution, examples, excel, exchange, exercise, experimental, extension, extensions, eyes
F	f1, face, facebook, fall, family, fanfic, fanfiction, fantasy, faq, fashion, fat, feed, feeds, female, feminism, festival, fetish, fic, fiction, fight, file, files, filesharing, filesystem, film, films, finance, financial, fire, firefox, firefox:bookmarks, firefox:rss, firefox:toolbar, firewall, fish, fitness, flash, flex, flickr, flight, flights, florida, flower, flowers, fob, folksonomy, font, fonts, food, football, for, forms, forum, forums, foto, fotografia, fotos, framework, france, free, freedom, freelance, freeware, French, friends, from, fuck, fun, functional, funny, furniture, future
G	Gadget, gadgets, gallery, game, games, gaming, garden, gardening, gay, gear, geek, gen, gender, genealogy, generator, genetics, geo, geography, George, geotagged, german, germany, ghost, gifts, girl, girls, gis, glass, global, gmail, gnome, gnu, go, god, good, google, googlemaps, government, gps, graffiti, grammar, graph, graphic, graphicdesign, graphics, gratis, great, green, grid, gtd, gui, guide, guitar
H	Hack, hacking, hacks, hair, Halloween, halo, happiness, happy, hardware, Haskell, hci, hdr, health, healthcare, heart, Hebrew, help, het, hibernate, high, hip, hiphop, history, holiday, home, hop, horror, hosting, hot, hotel, hotels, house, housing, how, howto, hp, html, http, human, humor, humour
I	I, ia, ibm, ical, icon, icons, ict, ide, idea, ideas, identity, ie, illustration, illustrator, im, image, images, imported, in, india, indie, info, informatica, information, innovation, inspiration, install, installation, insurance, intel, intelligence, interaction, interactive, interesting, interface, interior, international, internet, interview, investing, investment, ip, iphone, ipod, iptv, iran, Iraq, irc, Ireland, is, islam, island, Israel, it, italia, Italian, Italy, itunes
J	j2ee, jabber, jack, james, japan, Japanese, java, javascript, jazz, jesus, jewelry, job, jobs, john, joomla, journal, journalism, journals, jsf, json, juegos
K	Kernel, keyboard, kid, kids, king, kiss, knitting, knowledge, korea, korean
L	La, lake, landscape, language, languages, laptop, latex, latin, law, layout, learn, learning, leaves, lectures, legal, lego, lesbian, lessons, libraries, library, library2.0, libros, life, lifehack, lifehacker, lifehacks, lifestyle, light, lighting, lights, linguistics, link, links, linux, lisp, list, lists, literacy, literature, literature, little, live, local, logic, logo, lol, London, long, los, losangeles, love, lyrics
M	Mac, macbook, macintosh, macosx, macro, Madrid, magazine, magazines, magic, mail, make, man, management, manga, manual, mÃºsica, map, mapas, mapping, maps, market, marketing, mashup, math, mathematics, maths, mckay/Sheppard, me, media, medical, medicine, memory, men, menu, messaging, metadata, metal, mexico, Michael, microformats, Microsoft, midi, military, mind, mindmap, misc, mit, mix, mobile, model, modeling, models, modern, module, money, monitor, monitoring, motion, motiongraphics, motivation, mountain, movie, movies, Mozilla, mp3, multimedia, museum, music, Musica, musik, my, myspace, mysql
N	Naked, naruto, nasa, national, nature, navigation, nc-17, Netherlands, network, networking, networks, new, newmedia, news, newspaper, newspapers, newyork, night, Nikon, Nintendo, nlp, no, nokia, nonprofit, notes, noticias, November, nptech, nude, nutrition, nyc
O	Ocean, October, of, office, oil, old, on, one, online, ontology, open, opened, openoffice, opensource, open-source, opera, opinion, optimization, oracle, orange, organic, organization, origami, os, osx, out, outdoors, outlook, owl
P	p2p, painting, palm, paper, papers, parenting, paris, park, parody, parser, parsing, party, password, pattern, patterns, paul, pc, pda, pdf, peace, people, performance, perl, personal, personality, pet, pets, philosophy, phone, photo, photographer, photography, photos, photoshop, php, physics, piano, picture,

	pictures, pink, planning, plants, play, player, plugin, plugins, pocketpc, podcast, podcasting, podcasts, poetry, poker, Poland, police, policy, polish, politics, politik, pop, porn, portable, portal, portfolio, portrait, Portugal, post, power, powerpoint, pr, presentation, presentations, print, printing, privacy, process´, processing, product, production, productivity, products, programming, project, projectmanagement, projects, property, prototype, proxy, psychology, public, publishing, punk, puppy, pussy, puzzle, python
Q	quotes
R	r, race, racing, radio, rails, random, rap, rdf, read, reading, real, realestate, recherché, recipe, recipes, recording, records, recovery, recursos, red, reference, reflection, regex, religion, remix, remote, repair, research, resource, resources, rest, restaurant, restaurants, retro, review, reviews, rights, river, road, robot, robotics, robots, rock, roma, rome, rpg, rps, rss, ruby, rubyonrails, running, Russia, russian
S	Safari, safari_export, sam/dean, san, sanfrancisco, satellite, scary, scheme, school, science, scifi, Scotland, screen, script, scripting, scripts, sculpture, sea, search, searchengine, searchengines, seattle, secondlife, security, seguridad, self, semantic, semanticweb, semweb, seo, series, server, service, services, sewing, sex, sexy, sf, sga, share, sharepoint, sharing, shell, shoes, shop, shopping, short, show, simulation, singing, site, sky, skype, slash, sleep, slideshow, smallville, sms, snow, soa, soap, soccer, social, socialmedia, socialnetworking, socialnetworks, socialsoftware, society, sociology, software, solar, song, songs, sony, sound, source, south, space, spain, spam, Spanish, spears, speech, speed, spirituality, spn, sport, sports, spring, sql, ssh, standards, star, startup, starwars, statistics, stats, stock, stocks, storage, store, stories, story, strategy, streaming, street, streetart, studio, study, stuff, stupid, style, subversion, summer, sun, sunset, super, supernatural, support, sustainability, svn, Sweden, sweet, swing, Switzerland, symbian, sync, sysadmin, system
T	Tabs, tag, tagging, tags, Taiwan, teaching, tech, techno, technology, tecnologia, telephone, television, template, templates, terrorism, test, testing, texas, text, the, theme, themes, theory, thesis, time, tips, to, todo, Tokyo, tom, tool, tools, top, toread, Toronto, torrent, torrents, tour, tourism, toy, toys, trabajo, tracking, trading, traffic, trailer, train, training, translation, transport, transportation, travel, tree, trees, trends, tricks, trip, tuning, tutorial, tutorials, tutorials, tv, twitter, type, typography
U	Ubuntu, ui, uk, uml, uni, university, unix, unread, up, upload, urban, us, usa, usability, usb, useful, usenet, utilities, utility, ux
V	Vacation, validation, Vancouver, vc, vector, vegetarian, viajes, video, videogames, videos, vim, vintage, vinyl, viral, virtual, virtualization, vista, visual, visualization, vmware, voip, vs
W	w3c, wall, wallpaper, wallpapers, war, Washington, water, weather, web, Web 2.0, webapp, webcam, webcomic, webdesign, webdev, webdevelopment, weblog, webmaster, webservice, webservices, website, websites, webstandards, webtools, wedding, weird, white, widget, widgets, wifi, wii, wiki, Wikipedia, wikis, window, windows, wine, winter, wireless, wishlist, with, woman, women, wood, word, wordpress, words, work, workflow, world, wow, writing, wysiwyg
X	X, xbox, xhtml, xml, xp, xslt, xxx
Y	Yahoo, yellow, York, you, young, your, youth, youtube
Z	Zombie, zoo