SOCIAL STRUCTURE IN TAGGING PRACTICES:
REALITY OR MYTH?

by

AMIN FANI MARVASTI

A thesis submitted to the
School of Computing
in conformity with the requirements for
the degree of Master of Science

Queen’s University
Kingston, Ontario, Canada
December 2008

Copyright © Amin Fani Marvasti, 2008
Abstract

Tagging is widely adopted in so-called “collaborative-tagging” systems which are one of the Web 2.0 applications that have achieved lots of attention lately. They provide services for users to store, manage and search web resources with the help of freely chosen keywords, called “tags”. Because of the high-volume usage of these systems and the annotations that users provide by their tags, these systems are regarded as good targets for disciplines like knowledge discovery. Roughly, two lines of research have been pursued so far on collaborative tagging: to study the structure of tags and to study their functionality in web search. In this research we investigated tagging structures in a popular collaborative-tagging system, called del.icio.us, by focusing on the relations of “tags”, “users” and “web resources”, three main components of any collaborative-tagging system. Particularly we are interested in finding whether there are social structures that could be used to increase the usability of these systems for content retrieval and navigation. Our results show that people mainly use tags for their own informational needs which are personal rather than social. Any social structure or communities around tags and users is rare and weak which suggests that collaborative tagging has not added much to personal bookmarking. However, we show some regularities in tagging behavior that could be utilized for user experience improvement.
Acknowledgments

First of all I would like to thank the Baha’i Community of Canada very much for their most great support without which I couldn’t do my Master’s. Many thanks to my family and beloved friend, Mona, for their love and encouragement. Many thanks to my supervisor Dr. David Skillicorn for his great guidance and advice. Many thanks to Queen’s for giving me a chance to pursue my graduate studies.
# Table of Contents

Abstract .................................................................................. i

Acknowledgments .................................................................. ii

Table of Contents ................................................................... iii

List of Tables ........................................................................... v

List of Figures .......................................................................... vi

Chapter 1: Introduction ............................................................. 1
  1.1 Motivation .......................................................................... 1
  1.2 State of the art ................................................................. 3
  1.3 Objective ........................................................................... 4
  1.4 Contribution ....................................................................... 5
  1.5 Organization of Thesis ..................................................... 6

Chapter 2: Background .............................................................. 7
  2.1 Collaborative-Tagging Systems ........................................ 7
  2.2 The Tagging Process ......................................................... 10
  2.3 Delicious.com ..................................................................... 19
List of Tables

3.1 Size of tag and keyword sets for dataset A ........................................ 35

4.1 Tag and content similarity results ...................................................... 44
# List of Figures

2.1 Tripartite graph structure of a folksonomy .................................. 11
2.2 The cognitive process behind tagging .......................................... 12
2.3 The cognitive process behind categorization ................................. 13
2.4 Social interaction and conceptual transmission with tags ............... 16

3.1 Four stages in calculating the similarity of tags and keywords .......... 33

4.1 Histogram of the percentage of similarity for 100 websites .............. 45
4.2 Inclusion of top tags in popular bookmarks ................................. 49
4.3 100 Top Tag co-occurrence matrix .......................................... 50
4.4 Tag frequencies ........................................................................... 51
4.5 Average and standard deviation of user similarities for the first 300 top tags .................................................................................................................................................. 53
4.6 Average and standard deviation of user similarities for the first 400 and 500 top tags .................................................................................................................................................. 54
4.7 Maximum values of user similarities for the first 500 top tags .......... 57
4.8 User similarity matrix for tag “web2.0” ........................................ 57
4.9 User similarity matrix for tag “mp3” ............................................. 58
4.10 User similarity matrix for tag “seo” ............................................. 58
4.11 User similarity matrix for tag “3d” ............................................. 58
Chapter 1

Introduction

This chapter provides the motivation for the research, describes its objective and explains the contribution of this work. It ends by outlining the organization of this document.

1.1 Motivation

After the year 2001, “Social Web” applications have achieved impressive popularity, as a general trend emerged for using the web to communicate and contribute socially to an information environment. Applications like “Wikipedia”, “Facebook” and “Youtube” are among the top 10 web destinations and are even more popular than CNN.com (according to [3]).

Websites like Delicious\(^1\) and Flickr\(^2\), two other social-web applications that have received a great deal of attention lately, have pioneered a concept that some people

\(^1\)http://delicious.com
\(^2\)http://www.flickr.com
call “folksonomy” (in contrast to taxonomy), a style of collaborative categorization of resources using freely chosen keywords, often referred to as “tags” for future navigation, filtering and search. Folksonomies, which are also known as collaborative-tagging systems or social-bookmarking sites, let users save, organize, search and manage web resources (such as web-page locations or URLs) called bookmarks that they want to remember and retrieve later with the help of tags. Users are free to choose tags that they think are useful for describing their web resources. Tags allow the web resources to be found again by browsing or searching. User bookmarks are usually public, but they can also be saved privately or shared only with specified people or groups.

Before these systems, Internet web browsers on desktop computers let users store their bookmarks for their own purposes.

The bookmarks are normally visible in a browser menu and stored on the user’s computer. Commonly a folder-based scheme, rather than a tagging approach, is used for organization. With the advent of social bookmarking, users save their bookmarks on a remote web server, accessible from anywhere. Users can benefit from tagging, instead of categorization in folders, and they have the option of sharing bookmarks with other users, which may lead to the formation of communities with similar interests.

According to recent research [18], millions of Americans are tagging every day. There are several social-bookmarking sites and they have millions of users around the world. Because of the high volume usage of these systems and the annotations that the users provide by their tags and bookmarks, these systems are good targets for disciplines like knowledge discovery and ontology learning (which aims to automatically extract relevant concepts, or ontologies, and their relations from a given corpus or other kind
of data sets). By analyzing the relation between “users”, “tags” and “bookmarks”, the three main components of folksonomies, the possible existence of interesting patterns of tagging behavior can be investigated. If there are, data mining techniques could be applied to tagging structures. Exploration of the current status of these systems contributes to a better understanding of them which is useful both for their users, and for developers, who can learn how people are utilizing these systems. As an example, “New Scientist has discovered that Pentagon’s National Security Agency, which specializes in eavesdropping and code-breaking, is funding research into the mass harvesting of the information that people post about themselves on social networks. And it could harness advances in Internet technology specifically the forthcoming “semantic web” championed by the web standards organization W3C to combine data from social networking websites with details such as banking, retail and property records, allowing the NSA to build extensive, all-embracing personal profiles of individuals” [17].

1.2 State of the art

As Zeng and Li suggest in [7], roughly two lines of research have been pursued so far on collaborative tagging:

- Studying the structure of tags. For example, Goler and Huberman [22] and Wetzker et al. [21] show some regularities in user activity, kind of tags used and bursts of popularity in bookmarking, and Halpin et al. [15] show that frequency of tags for popular sites can be described by a power-law distribution.

- Using tags to promote web search. Cattuto et al. [6] compared three measures

However, both of these lines of research remain preliminary and need to be extended.

1.3 Objective

In this research we want to analyze tagging structures in collaborative-tagging systems by focusing on the relations of three main components of these systems, that is: “tags”, “users” and “bookmarks” to reveal any underlying patterns and gain better understanding of tagging behavior. We assume that these relations matter because they signal behavior, attitude and information. Specifically we aim to answer these questions:

- What is the functionality of tags and in which ways are people using them? Are they used for content abstraction or other purposes?

- Are there any communities based on similar tags or users or bookmarks?

- What other patterns or regularities could be found in tagging practices beside those we know so far?

This research will help gain more knowledge about the value of tagging in content management and retrieval. Finding any social-tagging structure and understanding tagging behavior can help increase the usability of these systems and can be utilized for modeling user behavior and user-experience improvement.
1.4 Contribution

A number of plausible motivations and mechanisms for tagging have been advanced in the literature. We show that they are not well-supported by the way in which tagging is done.

In this research we investigated the structure of top tags and popular bookmarks in a popular collaborative-tagging system, called del.icio.us, to find the answers to questions in the previous section.

We investigated the functionality of tags and their relevance to the content they are being used for and we noticed that tags and keywords extracted from the content overlap very little. The results also show that people mainly use tags that are meaningful to themselves, which do not contribute much to finding web resources socially. However, property tags are very helpful for retrieving multimedia contents. Therefore, we found that tags provide an orthogonal search mechanism to content-based search, but one that is limited by the specificity of tags. We analyzed the dynamics of top tags as well as tag frequencies to find patterns in tagging practices. We categorized popular tags and we also realized that popular tags change very little over a time period. The findings show that popular tags are uniformly distributed among popular web documents. Top-tag co-occurrence measurement indicated that there is little social structure to top tags that might be exploited.

We also designed a model for further examining the relationship of tags, documents and users to look for communities showing similarities. We then compared the tagging behavior of active and inactive users. Finally by defining normal tagging practice we examined users’ tagging behavior.
We show that any social structure or communities around tags, users and web documents is rare and weak which confirms that collaborative tagging has not added much to personal bookmarking. So, the main contribution of what follows is that collaborative and social tagging is largely a myth at present. We found out, although collaborative-tagging systems are conjectured (by researchers and developers of these systems) to support globally navigation of organized resources, they do not actually support that in a convincing manner as there is little collaboration between users. We also find some regularities in top tags, tag frequencies and tagging behavior of active and inactive users, which can be used for user-behavior modeling and user-experience improvement.

1.5 Organization of Thesis

In Chapter 2, we explain the tagging process in detail, further introduce collaborative-tagging systems, especially del.icio.us and describe some previous work related to this research. In Chapter 3, we describe the data preparation and the experimental methods that were applied in this research. In Chapter 4, we show the results obtained from the experiments along with the discussion and in Chapter 5, we finish by providing the conclusion of this research and some of its limitations.
Chapter 2

Background

In this chapter, collaborative-tagging systems are further explained, the process of tagging is described, more details about the del.icio.us website is given and finally some related previous work in this field is reviewed.

2.1 Collaborative-Tagging Systems

The Social Web is a description of Web 2.0 technologies that are focused on social interaction and community. But what’s the buzz about Web 2.0? How has it changed the functionality of the web? It could be said that Web 1.0 was about commerce whereas Web 2.0 is about people/participation. Web 2.0 applications provide “services” rather than “products”. In Web 2.0, users are encouraged to contribute and feel that they belong to a community as well as feeling empowerment and ownership. So, according to [26]: “Web 2.0 is actually a fundamental transformation of the web into a true collaborative and social platform, which aims for facilitating information sharing, collaboration and functionality of the web”. It is used to describe a subset
of interactions that are highly social, conversational and participatory.

Recently new applications emerged under the Web 2.0 paradigm with the introduction of “participation” instead of “publishing”: currently “Wikipedia” is preferred over “Britannica Online”. Its traffic ranking, according to [3], is currently ranked as 8th while Britannica Online is not even among the top 500 mostly visited websites.

In addition, according to [4] and [16], Wikipedia has received major attention as it is constantly updated. “Wikis” (like Wikipedia) were introduced to enable documents to be written collaboratively, in a simple markup language using a Web browser. “Blogging” has also become very popular. In other words, recently, Social Web applications have been widely used.

Another characteristic of Web 2.0 services, is the notion of “tagging (folksonomy)” as opposed to “directories (taxonomy)”. It led to the introduction of “Social Bookmarking”, a method for Internet users to store, organize, search, and manage bookmarks of web resources on the Internet with the help of metadata (tags). Social Web applications, named Collaborative-Tagging Systems have been developed to provide this service and are gaining popularity. In these systems, users annotate resources such as web pages, photos, videos and other contents with freely chosen keywords, called “tags”.

Tagging in these systems is defined as associating a set of words with an object (such as a web page). It allows for multiple, overlapping associations, rather than rigid categories. In an example, a Flickr photo of a puppy might be tagged both “puppy” and “cute”, allowing for retrieval based on these two tags.

Systems such as del.icio.us allow users to annotate web documents with arbitrarily chosen tags. Users can not only browse their own bookmarked web resources but also
what the other users have annotated in the system. Having more than 5 million users and 150 million bookmarked URLs according to [23], the delicious.com website (formerly known as del.icio.us) is one of the most popular social-bookmarking systems. There are many other tagging systems in existence, but here is a list of seven that are representative of the diversity:

- **delicious** (http://delicious.com): is a social bookmarking website for storing, tagging, sharing and discovering web resources. It was acquired by Yahoo! in 2005.

- **YouTube** (http://www.youtube.com): is a video-sharing website for uploading, tagging, sharing and viewing video clips.

- **flickr** (http://www.flickr.com): is a photo-sharing website for storing and tagging photos, as well as photo networking by maintaining a group of contacts and tagging other's photos.

- **Yahoo! MyWeb2.0** (http://myweb.yahoo.com): is similar to del.icio.us, but includes a social networks of contacts which allows users to privately share their bookmarks. As Yahoo! also owns del.icio.us, it recommends MyWeb to those who want a personal-bookmarking service as opposed to del.icio.us which is for those who need a social-bookmarking service.

- **citeulike** (http://www.citeulike.org): is a website for tagging citations and scientific references, such as academic papers or books.

- **last.fm** (http://www.last.fm): is an Internet radio and music community website where users can tag artists, albums and tracks.
• **Technorati** (http://www.technorati.com): is a weblog searching tool allowing blog authors to tag their posts.

General properties of social-bookmarking services (offered in collaborative-tagging systems) are as follows:

- Users can save links to web resources they want to remember as “bookmarks”.
- Users can share their bookmarks with other users in different ways: with specified people or groups, with a certain network, or any other combination of public and private domains.
- Users are able to search based on tags and then view those bookmarks which have been associated with a specific tag by any user. Based on the relationships between different tags, some clusters of tags or bookmarks could be created.
- Users can subscribe to web feeds, so they can become aware of new bookmarks which are shared and tagged by other users.
- As these services have matured and grown more popular, they have added extra features such as ratings and comments on bookmarks, the ability to import and export bookmarks from browsers, emailing of bookmarks, web annotation, and groups or other social network features.

### 2.2 The Tagging Process

Tagging is being done heavily in bookmarking systems, so it can be inferred that people enjoy doing it. But the question is: how is this done? Is it the same as categorization using directories? To answer that, a comparison must be made between
categorization in taxonomies and tagging in folksonomies:

*Taxonomy*, or the science of classification, aims to organize the objects in a *hierarchical* structure with supertype-subtype relation. For example since “car” is a subtype of “vehicle”, an object of “car” type can be placed in a “vehicle” directory.

The product of tagging is a user-oriented way of organizing information, known as a *folksonomy*.¹ Folksonomies are a kind of sense-making. In this way of organization, people label objects with freely chosen keywords, called *tags*. A formal definition of a folksonomy is as follows:

*A folksonomy $F$ is a tuple $F = (U,T,D,Y)$, where $U$ is the set of users, $T$ is the set of tags, $D$ is the set of web documents and $Y \subseteq U \times T \times D$ is the set of annotations, as shown in Figure 2.1*

![Tripartite graph structure of a folksonomy](image)

In folksonomies, users are free to choose any tags that they think are related to the object they are annotating. Tagging, in this way, helps to relax the constraint that users confront while using directories (folders) in taxonomies, especially when the object is an element of multiple categories; in which case, finding the most suitable

---

¹The word, folksonomy, is a combination of “folk” and “taxonomy”
category is a difficult task.

Rashmi [19] in his cognitive analysis of tagging explains how the lower cognitive cost of tagging versus categorizing makes it popular. He states that tagging is a 2-stage process, as shown in Figure 2.2:

The main stage is the activation of related concepts, which is based on cultural knowledge and is done by computation of similarity of the item and candidate concepts. He further continues that, in tagging, no filtering is involved in this stage, so as many of those associations as needed could be noted.

In categorizing, there is an extra stage in which a decision is made and one of the categories is selected as the right one. In the digital world this stage is more difficult, since contents are organized for future navigation, filtering or search, so not just the most likely category but the most likely place in which it might be looked for must be considered. These two questions are sometimes conflicting and lead to a problem which Rashmi calls it “post activation analysis paralysis”, a state of fear that the wrong decision will be made and the item will be lost forever - “it will land in some deep well, some hard to access branch of the tree and disappear from your view and

---

Figure 2.2: The cognitive process behind tagging, as Rashmi [19] explains.
Therefore, “search” versus “browse” is another fundamental difference between tags (in folksonomies) and directories (in taxonomies): Tags are normally used for retrieving resources, while directories are just used as a matter of organization and this degree of power makes tagging more interesting. Golder and Huberman [22] show that, in categorization, multiple checking is necessary to find an item of interest. For example, consider a hypothetical researcher who downloads an article about cat species native to Africa. If the researcher wanted to organize all her downloaded articles in a hierarchy of folders, there are several hypothetical options, of which we consider four:

1. C:\articles\cats → all articles on cats
2. C:\articles\africa → all articles on Africa
3. C:\articles\africa\cats → all articles on African cats
4. C:\articles\cats\africa → all articles on cats from Africa

In (3) and (4) the first category is the primary one and the second category is the more specific one. In this example, looking in (3) for a file in (4) will be fruitless and
checking multiple locations becomes necessary.

So, we can say that, since tagging is neither exclusive nor hierarchical, it has an advantage over hierarchical categorization. On the other hand, because of search functionality, tags must be chosen carefully in a way that helps recalling the object. To this end, some people suggest using a controlled vocabulary set (like a thesaurus) in folksonomies to avoid having a big collection of related tags describing the same concept (a simple example could be tags: “Apple”, “Mac” and “OSX”), but this is not a good idea, since it only works well with static entities. The controlled vocabulary set is too restricted to cover all the dynamic resources on the web, especially newly emerging concepts that are characteristics of the fast-growing web. In other words, using a controlled vocabulary set imposes the limitation of categorization over these systems which are fundamentally different in their basic purpose. Actually those who propose the use of controlled vocabularies underestimate the difficulty of understanding what users are thinking, and overestimate the amount to which the users will agree, either with one another or the cataloguers (creators of controlled vocabularies), about the best way to categorize. The only group that can categorize everything is everybody.

2.2.1 Why do people tag?

The reason why people use tagging can be explained from two different perspectives:

- **Personal Motivation:**
  
  - *Organizing:* an alternative way to structured filing, as already explained. Tagging is much simpler than creating folders to put items in.
– *Future retrieval*: the searching ability which was already mentioned. People create more ways to find their resources by adding more tags. Instead of remembering in what hierarchical structures they put the resource, they simply search for whatever keywords that they think might be related.

• **Social Motivation:**

– *Contribution and sharing*: while tagging is helpful in personal information management, some people think it might be also effective for social information management by letting people participate and share their interests with other people. Users of social-bookmarking systems may explore topics using the tags of other users. People might learn how to tag by watching what other people tagged for a specific resource. Golder and Huberman [22], state that “Typically such sites [collaborative tagging systems], allow users to publicly tag and share content, so that they can not only categorize information for themselves, they can browse the information categorized by others. There is therefore at once both personal and public aspects to collaborative tagging systems” which implies that they think tagging might be effective for social information management. Developers of the Delicious.com website also emphasize that the public aspect of del.icio.us “greatly improves how people discover, remember and share on the Internet.”¹ Smith [12], also believes that tagging creates a bridge between personal and community knowledge and everyone’s tags together form a kind of community consensus about the content. Chi [9], believes that

¹http://www.delicious.com/about
users often tag contents “with keywords to make them easier for themselves or others to find...The result is a collaboratively formed information structure that can be used to navigate, search and browse”. John and Seligmann [2], discussed that collaborative tagging has the potential for the formation of social networks around tags or topics. An Educause website [10] explains that how the community of users will develop a unique structure of keywords to define resources over time in social bookmarking. Thomas Wander Val, the person who coined the term “folksonomy”, in [24] discussed the social value of tagging. Rashmi in [20] discussed how tagging transforms the solitary browsing experience into a social one.

Figure 2.4: Social Interaction and Conceptual Transmission with Tags, as Rashmi [20] explains.


– *Attract Attention*: Social-bookmarking systems have the ability to show which tags are currently being used heavily, so one can find what is hot on the web by searching for the most frequent tags in these systems. Using this feature, people could attract other’s attention for a special resource on the web by posting those most frequent tags which are related to this resource.

### 2.2.2 What do people tag?

People tag web pages, photos, videos, music, news, blogs, books, scientific references and any contents that they think might worth remembering. As long as a resource can uniquely identify something, it could be tagged.

### 2.2.3 How do people tag?

For tagging, people are free to use words which they think might help them retrieve the content.

Golder and Huberman [22] believed that people consider tags at different levels of specificity to be most useful or appropriate for describing the item in question:

- *Generalized level*: a more general way of identifying the item (like “feline”)

- *Basic level*: the most common way that humans use to refer to the item (like “cat”)

- *Specific level*: a more specific way of identifying the item (like “persian cat”)

They stated that people use basic level tags more than the other levels, but they continue that the problem here is that basic levels vary in specificity for individuals.
Documents tagged by “perl” and “javascript” may be too specific for some users, while a document tagged “programming” may be too general for the others. In addition, Golder and Huberman believe that tagging process is also influenced by social factors, that is, although at the beginning people use tagging systems for their personal purposes, once they start sharing their bookmarks with the other people, they are collectively sense making, by looking at what tags other people - with the same interest - have used for their bookmarks. Golder and Huberman continue that in this way people might tune the way they make sense of items towards a common sense making. So they might be able to use tags that make their resources (web documents) more visible to other users.

Fundamentally, Weick et al. believe that tagging is about sense making [25], a process in which information is categorized and labeled and through which, meaning emerges. On the other hand, social and psychological factors may affect users’ tag choices. By exploring the tags being used we can learn what kinds of distinctions are important to taggers. Here we list the tag functionalities as outlined by Golder and Huberman [22]:

1. **Identifying What or Who is it about**: These tags, overwhelmingly identify the topics of bookmarked items and include common nouns of many levels of specificity, as well as many proper nouns, in the case of content discussing organizations or individuals.

2. **Identifying What it is**: Tags that identify what kind of things a bookmarked item is, in addition to what it is about.

3. **Identifying Who owns it**: Some bookmarks are tagged according to who owns or created the bookmarked content.
4. **Refining Categories**: Tags that do not seem to stand alone and, rather than establish categories themselves, refine or qualify existing categories. Numbers (e.g. 25, 100), perform this function.

5. **Identifying qualities or characteristics**: Adjectives such as “scary”, “funny”, “stupid”, “inspirational” are of this type which identify the tagger’s opinion of the content.

6. **Self Reference**: Those tags beginning with “my”, like “mystuff”, “mycomments” which identify the content in terms of its relation to the tagger.

7. **Task Organizing**: When collecting information related to performing a task, that information might be tagged according to that task, in order to group that information together. Examples include “toread”, “jobsearch”.

However, Smith [12] mentions another functionality for tags:

- **To attract search engines**: This way of “machine tagging” helps the search engines to find the resource on the web more easily. Tags of this category are in this format:

  \[
  \text{namespace : key = value}
  \]

  Examples include tags like:

  “clothing:size="large””, or “blog:via="http://atomiq.org””.

### 2.3 Delicious.com

Since in this research, we applied our methods on del.icio.us website, it is going to be further introduced in this section. One of the most widely cited example of websites
using folksonomic tagging is “Delicious.com”. Users can do their bookmarking in this social-bookmarking site instead of using their browser bookmark functionality. So they can access their bookmarks from any computer on the Internet and they can benefit from other advantages of social-bookmarking services, like sharing their bookmarks with other people and seeing what other people are bookmarking.

2.3.1 Features

In order to find out what is hot currently on the web, del.icio.us has a hotlist on its home page which shows the hottest bookmarks on del.icio.us. People can also explore what is hot on any particular topic (they are most interested in) by simply checking out the most popular bookmarks for any tag.

Many features including: the website’s simple interface, human-readable URL scheme, a simple REST-like API (which allows to access information simply through URLs),
and RSS feeds for web syndication, have contributed to making del.icio.us one of the most popular social-bookmarking services.

2.3.2 Usage

Use of del.icio.us is free. Users are only required to be registered in the system to start saving their bookmarks on del.icio.us. Adding bookmarks could be done both through the del.icio.us website or by using the browser add-ons. Users can download their own data through the site’s API in an XML or JSON format.

2.3.3 Bookmarks

All bookmarks posted to del.icio.us are publicly viewable by default, although users can mark specific bookmarks as private, and their imported bookmarks are private by default. The public aspect is emphasized; the site is not focused on storing private (“not shared”) bookmark collections.

2.3.4 Networks

People can use del.icio.us to send interesting bookmarks to their friends through their inbox. To make a bookmark show up in a friend’s del.icio.us inbox, it is tagged with “for:” followed by their username in del.icio.us. To do this even more simply, a user can add people to his/her network, so the name of all their favorite users will show up and their bookmarks will also be collected in one place.
2.3.5 Tagging

When a user adds a new bookmark to del.icio.us, tags are suggested (in two forms: Recommended and Popular) based on an aggregate list of tags previously provided by other users, so the cognitive load of organizing the user’s bookmarks is reduced. “Recommended” tags are a combination of tags this user has already used and tags that other people have used for the current object. “Popular” tags are what other people have tagged this page. Based on what tags people have used for their bookmarks, this system shows the related tags and related people.

2.3.6 Subscriptions

Users can subscribe to subscriptions which allows them to watch all their favorite tags in one place. It works as a “tag aggregator”, giving the users the opportunity to watch what people are bookmarking about a favorite topic. After a user adds a tag to his/her subscriptions, del.icio.us watches for everyone’s bookmarks saved with that tag and delivers them to his/her subscriptions page.

2.3.7 Tag Bundles

There is another feature called tag bundles which allows for arranging a user’s tags into named groups. Although tag bundles, unlike tags, are not shared among the users of the system and cannot be browsed, they are a simple way for users to organize their tags. For instance, tags like: “cats”, “dogs” and “fish” could be organized into a bundle called “pets”. 
2.3.8 View Options

There are different options for a user to view his or her tags in del.icio.us:

- The “list view” which is the default, with one line for each tag, includes a count of how many times that tag has been used.
- The “cloud view” arranges all tags without new lines after each tag, and makes tags bigger or smaller depending on how many times they’ve been used.

Users can also choose to hide all of the tags that are saved with 1) only one bookmark; 2) two or fewer bookmarks; or 3) five or fewer bookmarks. They can sort out their tags either in alphabetical order (the default) or by count.

2.3.9 Search Options

Search facility in del.icio.us is comprised of three components:

- **Bookmarks**: a user can search in 1) his or her own bookmarks; 2) popular bookmarks; 3) recent bookmarks; or 4) lookup a URL.
- **Tags**: a user can search in 1) his or her own tags; 2) his or her own subscriptions; or 3) explore a specific tag. del.icio.us will also show the updated list of most popular tags.
- **People**: a user can search through 1) his or her network; or 2) go to a certain user’s bookmarks.
2.3.10 API

del.icio.us lets users have read/write access to their tags and bookmarks via its API. All del.icio.us APIs are done over https and require HTTP-Auth. It is still under development, but the current version (according to [8]) consists of methods for update, posts, tags and tag bundles.

2.3.11 Feeds

del.icio.us also provides data feeds for news readers, blogs and users’ third-party applications. These feeds come in several formats, including RSS (which is a format used by many news sites and blogs to publish content on the web) and JSON (which is a lightweight data-interchange format easily used in browser-based mashups, blog badges, and other scenarios including server-side and desktop applications).

Feeds at del.icio.us all share the following base URL prefix:

\[\text{http://feeds.delicious.com/v2/format/}\]

The format specifies the output format for the feed, which currently includes the values “rss” and “json”.

2.4 Related Work

There has been a small amount of research dealing with folksonomies:

Golder and Huberman [22] analyzed the structure of collaborative tagging to discover some regularities in user activity, kind of tag used and bursts of popularity in
bookmarking in del.icio.us. Their results showed that users exhibit a great variety in their set of tags and tags vary in frequency of use. They found out that, while many URLs reach their peak of popularity in del.icio.us very quickly, there are many other URLs which are hardly ever bookmarked for a long time until they are “rediscovered” and then experience a jump in popularity. They also empirically found that there is a stability pattern showing that the relative proportions of tags used for a given URL in del.icio.us is fixed (meaning that each tag’s frequency is a nearly fixed proportion of the total frequency of all tags used for that URL), usually after the first 100 or so bookmarks, which they attribute to imitation and shared knowledge.

Cattuto and colleagues [6] analyzed three measures of tag relatedness: tag co-occurrence, cosine similarity of co-occurrence distributions and FolkRank which is an adaptation of the PageRank algorithm to folksonomies. They mapped del.icio.us tags, when possible, to WordNet synonym sets and used well-established measures of semantic distance in WordNet to provide a semantic grounding on their finding. Their results indicated that different characteristics of these three selected measures of relatedness makes them applicable to different subtasks, i.e. cosine similarity for “synonym discovery”, both FolkRank and co-occurrence relatedness for “concept hierarchy” and FolkRank for “discovery of multi-word lexemes”.

Wetzker and colleagues [21] analyzed del.icio.us bookmarks and showed the monthly growth of del.ici.ous between 2004 and 2008 by posted bookmarks, new users, new URLs and new tags. They collected a corpus of bookmark assignments at del.icio.us and analyzed it to find out that:
Unsurprisingly, the del.icio.us community is biased toward web community and web-technology-related content by showing top 10 most frequent URL and domains at del.icio.us.

User activity follows a power-law distribution with a few users being responsible for a high number of posts. The top 1% of users proliferate 22% of all bookmarks and the top 10% contribute 62%. They assumed parts of this tendency to be triggered by spamming users.

Each bookmark is labeled with 3.16 tags on average and only 700 tags account for 50% of all assignments.

Finally, they proposed that out of the top 20 most active del.icio.us users, 19 users are of apparently non-human origin (probably software but perhaps users who are very into tagging), posting tens of thousands of URLs pointing to only few domains. So they proposed that spammers exhibit one or more of the following characteristics: high tagging rate, very low tagging rate, bulk posts, or combinations of the above.

Halpin and colleagues [15] examined the dynamics of collaborative-tagging systems. In particular, they examined and showed the distribution of usage frequency of tags for popular sites with a long history (many tags and many users) can be described by a power-law distribution. They showed empirically that this behavior depends on the number of users and to some extent on the temporal duration of the tagging process. They also empirically examined the tagging history of sites in order to determine how this distribution arises over time and to determine the patterns prior to a stable distribution.
Zeng and Li [7] analyzed the degree distribution and clustering coefficient associated with a tri-partite graph of users, tags and URLs. This graph was constructed from a tagging dataset, of four sample tags, collected from del.icio.us. They observed that user-URL and user-tag degree distributions deviate from the power-law, but not significantly, while the tag and URL distributions follow the power-law pretty closely. They showed that a URL can account for roughly 30% of a random user’s interest, by calculating the average clustering coefficient between URLs and tags that are neighbors of each top 500 users. Therefore, they drew the conclusion that tags are a great source of information representing user interests and information needs. They also investigated several web-page recommendation approaches, including: top-N most heavily tagged pages, user-based neighborhood and item-based neighborhood (the last two are the most commonly used successful collaborative-filtering algorithms). The results indicated that, under the user-based recommendation framework, tags can be fruitfully exploited as they facilitate better user similarity calculation.

Sun et al. [1] applied a statistical approach for discovering topic-specific bursts from del.icio.us. They used the $\chi^2$ model to determine if the appearance of tag $t$ in date $d$ is significant. Statistically, they adopted a threshold-based strategy: for any tag $t$ having a $\chi^2$ value greater than 7.879, they drew the conclusion that its generation process has varied and classified it as a “burst” tag of month $d$. They applied this technique to a sample of del.icio.us bookmarks containing “game” tag and detected some burst tags. In order to understand why some of these tags become popular, they proposed looking at the most frequently co-occurred tags for each burst tag. They also indicated that users and resources contributing to the bursts can be
classified into two categories: old and new, based on their past usage histories. Using this classification scheme, they found that users tend to follow the trend, rather than creating it. They learned this by observing that new users of a burst tag account for a large proportion of the burst population and they bookmark old resources (already bookmarked with this tag), more than new resources, in the burst period.

Gibbins and colleagues [5] demonstrated how the semantics of ambiguous tags can be discovered by analyzing the tripartite graphs of users, tags and resources. They suggested that if $Y$, the matrix of documents and users associated with the tag is created, where the entry $Y_{i,j}$ is 1 if the user $i$ has assigned that tag to document $j$, then the matrix $S = YY^T$ shows the affiliation between the users who have used the tag, weighted by the number of documents to which they have both assigned that tag and this network is likely to connect users who use the tags for the same meaning. Also matrix $C = Y^TY$ can be considered as another angle of viewing the issue of polysemous or homonymous tags. So with the edges weighted by the number of users who have assigned that tag to both documents, this network is likely to connect documents which are related to the same sense of the given tag. They believed that by studying these networks the semantics of tags could be better understood.

Begelman et al. [14] argued that clustering techniques can improve the user experience of current tagging services. They demonstrated the use of co-occurrence similarity by applying graph filtering combined with a spectral clustering algorithm to cluster tags and provided some results they obtained on the RawSugar\textsuperscript{2} database. For example, the cluster they found for tag “health” includes: “shopping”, “research”,

\textsuperscript{2}http://www.rawsugar.com

Santos-Neto et al. [11] presented the characterization of two collaborative-tagging systems, CiteULike and Bibsonomy. First, they analyzed the distribution of bookmarked items, tags and tagging actions related to each user. They found that the activity distribution is highly heterogeneous: a few active users contribute a large number of tag assignments and maintain a large number of items and tags, while the majority of users have modest tagging activity. They also found that, in these two systems, users with large number of bookmarked items tend to have a large number of tags ($R^2 = 0.98$ for CiteULike and $R^2 = 0.80$ for Bibsonomy). Additionally, they noticed that, in both communities, there is a large population of isolated users (having zero similar tag or bookmarked item with other users).

Koutrika et al. [13] proposed an ideal tagging system where malicious tags and malicious user behaviors are well defined and they described and studied a variety of query schemes and moderator strategies to counter tag spam. They introduced a coincidence factor for each user, which is high when the user agrees with other users based on the tags that he posts for his documents, and is low when the user’s postings do not agree with other people in posting tags to documents. They showed that this countermeasure can be defeated by focused spam attacks, so they proposed a focused moderator (a person that can “conceptually” identify good and bad tags for any documents) to detect the focused attacks.
Chapter 3

Experiments

This chapter presents the experiments used to gain more knowledge about tagging behavior in Delicious.com website.

3.1 Tag and content similarity

We first attempt to explore how much the tags that people choose to describe web documents represent the set of keywords extracted from the content of the documents. This will help finding out the value of folksonomy tags as a potential description of content.

To do this experiment automatically, we represented the content of each document by the most important keywords extracted from it. There are a wide variety of automated keyword extraction tools, like SEO (Search Engine Optimizer) keyword analyzer tools, or “Kea”, which is an open source tool. However, Kea requires extensive training in the specific domain of interest to be able to find reasonable results. SEO tools, on the other hand, are biased in terms of looking for popular search terms in a webpage
when extracting its keywords. Recently, Yahoo has released a new web-service feature, which returns the most significant terms or phrases in a given text. It seems that this service uses something along the lines of TFIDF, which means it is returning the most “statistically significant” terms. Actually, TFIDF (term frequency inverse document frequency) is a weight often used to evaluate how important a word is to a document in a collection or corpus. The term frequency in the given document is the number of times a given term appears in that document and is usually normalized to prevent a bias towards longer documents. The inverse document frequency indicates the general importance of the term. It is obtained by dividing the number of all documents by the number of documents containing the term, and then taking the logarithm of that quotient.

We chose to use the Yahoo! Term Extraction web service, since it needs no pre-training step and it has a simple API to work with. We also manually checked the keywords returned by this service for a sample of documents and made sure that they are the most significant words or phrases of that documents.

3.1.1 Data Preparation

We collected our dataset for this experiment, using a HTML scraping program, from the del.icio.us bookmarking service. It is called dataset A: a corpus of 100 web documents (along with their associated tags), which were bookmarked by at least 50 users. Using the notation introduced in Section 2.2:

\[
\text{Dataset } A = \{ (d, T_d) \mid d \in D \text{ and } |U_d| \geq 50 \} \text{ with } |A| = 100 \text{ where: } \\
T_d = \{ t \in T \mid \exists u \in U, (u, t, d) \in Y \} \text{ and } U_d = \{ u \in U \mid \exists t \in T, (u, t, d) \in Y \}
\]
CHAPTER 3. EXPERIMENTS

To collect this dataset, during June 2008, we randomly selected 10 users from the popular bookmarks section in del.icio.us\(^1\) and, for each user, we searched through his bookmarks to find URLs of those web documents bookmarked by more than 49 users. This task can be done straightforwardly because, for each URL, del.icio.us shows how many users have bookmarked it so far. After appending these URLs to our (initially empty) dataset, we continued the process for other users until the size of the dataset reached 100. Each URL was added just in case that it was not already included in dataset.

Then we manually reviewed the websites and deleted those that were flash-based or audio/video based. We continued the whole process to be sure that all the 100 web documents were text-based.

The reason why we added the constraint for each document to be bookmarked by at least 50 users is two fold:

- We believe that when a web document has been bookmarked by at least 50 users, there might be reasonable agreement among the tags that have been chosen by these users, whereas for a document being bookmarked by say, only 10 users, this agreement might not exist.

- On the other hand, some web documents in del.icio.us have a large number of bookmarks and going through all of them to retrieve the tags would be a time-consuming process. Since, for a given URL, del.icio.us returns 50 bookmarks in each page, by just going through that page we can access the last 50 bookmarks and their associated tags of this URL. There is no extra need to go through the other pages.

\(^1\)del.icio.us home page shows currently popular bookmarks.
The whole collection process of this dataset was automatically done by an application written in C#.NET. The statistics of dataset A are shown in the first row of Table 3.1.

### 3.1.2 Experiment

To do the experiment, we used dataset A to measure, for each document, the overlap between the set of del.icio.us tags and the set of keywords extracted using the Yahoo! Term Extraction web service. As Figure 3.1 denotes, the experiment carried out in four phases:

1. **Document Preparation**: prior to feeding the documents to Yahoo! TE web service, each document passed through a preprocessing stage in which:
   - For those web documents which used Google Adsense, their contents were manually extracted and saved as a text document, replacing the original one to be fed into Yahoo! TE service. This ensured that returned keywords were merely of the content, not the ads.
For those web documents which were weblogs, just the “tag related” part was extracted manually and saved as a text document, replacing the original document in dataset A. “Tag related” parts were discovered by manually reviewing the weblog content and its del.icio.us tags.

2. Text Conversion: using a C#.NET application, each HTML web document (all documents in dataset A, except those whose contents were extracted manually) was scraped and converted to a plain text document. This application used regular expressions to get rid of HTML tags. This step assured that all documents in dataset A are of text type.

3. Keyword Extraction: using Yahoo! TE web service API, keywords were extracted and stored for each text document. Since Yahoo! TE service tends to return significant phrases rather than words while del.icio.us tags are single words, all the returned phrases were split into single words.

4. Similarity Measurement: Tag and content similarity, was measured by calculating the overlap between del.icio.us tags and extracted keywords for each document.

For this purpose we used Jaccard Index, also known as Jaccard similarity coefficient, which measures the similarity between two sets and is defined as the size of the “intersection” divided by “union”. So the similarity was calculated in this way:

\[ \text{Sim}(T_d, K_d) = \frac{|T_d \cap K_d|}{|T_d \cup K_d|} \]

where:

\( T_d \) = the set of del.icio.us tags for document \( d \),

\( K_d \) = the set of keywords extracted from document \( d \).
Properties of these two sets are summarized in Table 3.1.

<table>
<thead>
<tr>
<th>Set</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Maximum</th>
<th>Minimum</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_A$</td>
<td>46.97</td>
<td>11.39</td>
<td>75</td>
<td>20</td>
<td>44</td>
</tr>
<tr>
<td>$K_A$</td>
<td>24.92</td>
<td>9.52</td>
<td>37</td>
<td>2</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 3.1: Size of tag and keyword sets for dataset A. $T_A$ and $K_A$ represent the set of all tags and keywords in dataset A respectively.

### 3.2 Top tags

To have a better understanding of what people use as tags we decided to look at the most popular, called top tags\(^3\). For top-tag analysis we did a couple of experiments:

- First of all, in order to find what top tags used in del.icio.us were like, we retrieved and categorized them. To do this, we gathered a collection of 140 top tags from del.icio.us, during June 2008.

- Second, to see if the top tags change in del.icio.us over a time period or remain stable we compared our dataset, described above, with a second dataset, again containing 140 top tags, but collected by the end of August 2008. We used the Jaccard index to measure the similarity between these two datasets, similar to the experiment in Section 3.1.

- Third, to find the relation between popular tags and popular web documents in del.icio.us we estimated what percentage of top tags there are in popular web

\(^3\)The terms, “top” and “popular” are used interchangeably
documents. To do this, popular tags must be compared to the tags used in popular web documents. Therefore we can check if top tags are mainly being used in association with popular web documents or if people are using them regardless of the popularity of the web page they are annotating. But while del.icio.us exposes top tags to users, unfortunately, it does not make popular web documents accessible. It just shows the most recent popular web documents on its homepage updated from time to time. However, by using del.icio.us feeds, we are able to search for a tag in popular web documents to see how many times it has been used in them. Therefore, in order to see how top tags are related to popular web documents and find tag rate of growth, we did this: for each of the 140 top tags in our dataset, already sorted based on their ranks, we found the number of popular web pages that were bookmarked using either this tag or its higher-ranked top tags and plotted the resulted diagram. This diagram shows how each popular web page contributes to adding top tags and whether or not there is a set of popular web documents which contain the majority of top tags. In other words, it shows how top tags are distributed among popular web documents.

- Finally, we did a similarity measurement on these top tags, to discover if there are any interesting clusters, indicated by blocks in their similarity matrix. The similarity measurement of these top tags was calculated based on their co-occurrence in del.icio.us. Co-occurrence of two tags counts how many times they are used together. Using co-occurrence gives us a list of statistically related tags. Using the notation defined in Section 2.2 the entries in top tag co-occurrence matrix, $W$, are defined as:
\[ W_{ij} = \text{card}\{(u,d) \in U \times D \mid i,j \in T_{ud}\} \] where:

\[ T_{ud} = \{ t \in T \mid (u,t,d) \in Y \} \]

After calculating the entries \( W_{ij} \) for this similarity matrix by counting the number of web documents which have been tagged by both tags \( i \) and \( j \), we applied a method to visualize the result, so we can find if any groups of mostly co-occurred tags are present.

To group tags based on their similarities, we applied a method which is outlined below and tries to push the larger values to the top right corner of the matrix:

Starting with \( N \times N \) similarity matrix:

1 - Sort the rows of the matrix in descending order, based on the maximum values in the rows, so that first row will contain the largest value in the matrix, the second row will contain the second largest value and so on.

2 - Sort the columns in ascending order, based on the maximum values in each columns, so that very first column on the left will have the lowest maximum among other columns and the rightmost column will contain the biggest maximum.

3 - Ignoring the first (topmost) row and last (rightmost) column, repeat the first and second steps for the remaining \( N - 1 \times N - 1 \) matrix, until \( N = 1 \)

Using MATLAB, the above method was applied on the top tag co-occurrence matrix created in this experiment and the image of resulted matrix was displayed using the \texttt{imagesc()} function.
3.3 Tag frequencies per user

Exploring the tag frequency distributions for users is a good way to find their tagging behavior. We can examine whether people follow a long tail (power-law) distribution in their tag frequencies or if there are cases where their tag-frequency distribution shows a noticeable decrease in the frequencies of a small proportion of their tags.

For this purpose we picked 100 random users having at least 100 tags on del.icio.us. We used the del.icio.us home page and, for each popular bookmark listed there, we searched to find its users. Those users who have at least 100 bookmarks were added to the dataset. This data collection process finished when the size of dataset reached 100.

At second step, the list of tags for each user were retrieved and added to the dataset. The complete dataset, which we call dataset B, is defined as:

\[ \text{Dataset B} = \{ (u, T_u) \mid u \in U \text{ and } |D_u| \geq 100 \} \text{ with } |B| = 100 \]

where:

\[ T_u = \{ t \in T \mid \exists d \in D, (u, t, d) \in Y \} \]
\[ D_u = \{ d \in D \mid \exists t \in T, (u, t, d) \in Y \} \]

Using dataset B, we plotted the tags and their frequencies for each user. For each user, the frequency of each of his/her tags shows how many times that tag occurred in this user’s tag list. The tags were sorted in descending order of their frequencies, for each user.

In the resulting diagrams, we considered the frequency of the most-frequent 100 tags for each user, since all of the diagrams had a similar tail after the first 100 top tags. This shrunk the diagram and let us focus on the top tags of users to discover their differences.
3.4 User similarities per tag

The next similarity measurement experiment explores the similarity between people who all use the same tag. This helps us discover if the users in del.icio.us have common interests when they happen to use a specific tag. In this way we can know more about the formation of communities around tags.

In order to do this experiment we collected 500 top tags from del.icio.us along with 50 users associated with each tag. Using the notation defined in Section 2.2, this dataset, which is called dataset C, is defined as:

\[ \text{Dataset } C = \{ (t, U_t) \mid t \in T \text{ and } |U_t| = 50 \} \text{ with } |C| = 500 \text{ where:} \]

\[ U_t = \{ u \in U \mid \exists d \in D, (u, t, d) \in Y \} \]

The set of all users returned for a certain tag in del.icio.us is already sorted according to the date they have bookmarked a URL using this tag. We picked the first 50 users returned while searching for each 500 tags.

To measure the similarities between the users of each tag, there are different candidate features of the users that could be chosen, including: (1) the URLs of the web documents bookmarked by users, (2) the content of the documents bookmarked by users, or (3) the set of tags used by each user.

The first option is not a good choice, since some people use tiny URLs in their bookmarks while others use the original URL. This makes these similar URLs look different while they are actually the same. Using the second feature is expensive because it takes a long time to download each web document and then extract the keywords to measure their similarity. As a result we selected the third option as a feature for
measuring the similarity between the users of each tag. The set of all tags were downloaded for each user and then the Jaccard index was used as a similarity measure. Doing so, a $50 \times 50$ similarity matrix $S_t$ was generated for each tag $t$, in which the rows and columns are users of tag $t$ and $S_{tij}$, the entry of row $i$ and column $j$ in matrix $S_t$ is defined as:

$$S_{tij} = Sim(T_i, T_j) = \frac{|T_i \cap T_j|}{|T_i \cup T_j|} \text{ where:}$$

$i$ and $j$ denote the $i$th and $j$th users of tag $t$ respectively and

$T_i = \text{The set of del.icio.us tags for } i\text{th user}$,

$T_j = \text{The set of del.icio.us tags for } j\text{th user}$

To see how the user similarities change, we did this experiment in 5 steps: starting with the first top 100 tags and adding the next 100 top tags at each step. We measured the average, standard deviation and maximum of user similarity values for each tag (i.e. its similarity matrix) at each step, but since the similarity matrix is symmetric, we just considered the upper triangular part of these matrices to measure the statistical values.

### 3.5 Web document similarities per user

So far we have explored the “web document-tag” and “tag-user” relations. Now we consider the “user-web document” relation.

The structure of this section follows exactly the structure of the last section.

The dataset used in this experiment is: a group of 500 users, using different numbers of tags, collected from del.icio.us, along with 50 documents, for each user, bookmarked
in this system. Using the mathematical notation, this dataset, named $E$ (not $D$, since $D$ represents the set of documents), is defined as:

$$
\text{Dataset } E = \{ (u, D_u) \mid u \in U \text{ and } |D_u| = 50 \} \text{ with } |E| = 500 \text{ where:}
$$

$$
D_u = \{ d \in D \mid \exists t \in T, (u, t, d) \in Y \}
$$

del.icio.us returns the URLs bookmarked by a user, sorted based on the date they bookmarked them and we selected the first 50 URLs returned for each of the users. Like the previous experiment, we measured the similarities, but this time the similarities of web documents bookmarked by each user.

To measure the similarities among the documents of each user, the only option to use as a feature is the content of each web document. Therefore we decided to extract keywords for each of the 50 web documents (per user) and enrich them with all the tags used for this web document in del.icio.us. The ratio of the average number of extracted keywords to the average number of tags returned for web documents was $1/2$. It means that on average, for each user, the number of extracted keywords for his/her web documents is half of the number of tags returned for his/her bookmarked web documents.

By calculating the similarity of web documents for each user, we created the similarity matrix $S_u$, for each user $u$, as a $50 \times 50$ matrix in which rows and columns are the web documents bookmarked by user $u$. The entry $S_{uij}$, is defined as:

$$
S_{uij} = \text{Sim}(KT_i, KT_j) = \frac{|KT_i \cap KT_j|}{|KT_i \cup KT_j|} \text{ where:}
$$

$KT_i = \text{the union of extracted keywords and del.icio.us tags for } i\text{th web document}$, $KT_j = \text{the union of extracted keywords and del.icio.us tags for } j\text{th web document}$
3.6 Tag similarity on consecutive days

In this experiment, we investigated another factor which is also important while analyzing social systems: time. In order to see the change in tagging behavior of users, we decided to compare the tagging practice of two sets of users: those who have used del.icio.us heavily, by posting a large number of bookmarks and those who have a small number of bookmarks.

To do this experiment, we randomly collected a group of 20 users, 10 who have fewer than 100 bookmarks and 10 who have more than 1500 bookmarks. The data-gathering process was done by considering the users who appeared in the del.icio.us popular bookmark list and then retrieving their bookmark count, as this is the only way to access the users’ data in del.icio.us when user IDs are not previously known.

Afterwards, for each user, we retrieved his or her 100 most-recent consecutive bookmarks and saved all the tags for each day. Then, we calculated the Jaccard Index for the tags of each day and the previous bookmarking day, which shows the similarity of the tags a user uses in each consecutive bookmarking day. In this way we can understand if there is any agreement between the tags that people use in their consecutive bookmarking days and find the difference between active and inactive users in this behavior.

3.7 User bookmarks vs. Tags

As the last experiment, we examined whether any relation could be found between the number of tags and the number of bookmarks a user annotates within del.icio.us. To this end, we used dataset E (used in Section 3.5). In this dataset, the number of
bookmarks for all users is in range $[1, 53116]$ with the average of 4592.92 bookmark per user, while the number of tags is in range $[6, 23887]$ with the average of 2009.26 tags per user. The R-squared value for the number of bookmarks and tags is 0.05 showing that knowing one does not help knowing the other. We plotted each user in a two dimensional diagram, where the X-axis represents the number of bookmarks and the Y-axis identifies the number of tags.

Generally, we believe that since tagging is a mapping function between a resource (web document) and the set of tags, i.e. $f : resource \rightarrow \{tag\}$, a rational and applicable way of tagging will be one in which the size of both the $\{resource\}$ and $\{tag\}$ sets are small, since mathematically in that case there are a small number of choices for the reverse function, which is performed while trying to retrieve the resource, so that people could remember them easily. In this model, we expect to see almost all of the users at the left bottom corner of the plotted diagram where the number of both bookmarks and tags are small. Those users that are either at the right hand side of the diagram (having large number of bookmarks) or top left side (having large number of tags) do not fit this model of rational tag use well.
Chapter 4

Results and Discussions

4.1 Tag and content similarity

In this experiment we utilized dataset A (a corpus of 100 web documents along with their associated tags) to measure, for each document, the similarity of tags and keywords extracted using the Yahoo! Term Extraction web service. The results of this experiment can be seen in Table 4.1. Although at first it might be hypothesized that tags are mainly keywords of the content used for abstraction, the outcome contradicts this. It can be seen that there is neither complete nor zero similarity between the tag and keyword sets. From this low similarity, we infer that people are not using the keywords to describe page contents, in other words, tagging is not a way of “content abstraction”. This indicates that tags provide new and different information about the objects.

Table 4.1: Tag and content similarity results

<table>
<thead>
<tr>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Maximum</th>
<th>Minimum</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2%</td>
<td>3.7%</td>
<td>16.2%</td>
<td>1.4%</td>
<td>3.8%</td>
</tr>
</tbody>
</table>
Figure 4.1 shows a histogram of the frequency of results which displays the distribution of the percentage of similarities using intervals 1.5 percentage points wide. The shape of this histogram suggests that a greater part of the tag and content similarity results are located at the left-hand side of the diagram between 4% and 6% similarity, while the average is greater than 7%.

![Histogram of the percentage of similarity for 100 websites](image)

Figure 4.1: Histogram of the percentage of similarity for 100 websites

After finding that folksonomy tags do not match with significant keywords we continued the experiment to understand why this happens:

We removed all the tags that could be found in keyword sets, to find those tags that were not among the extracted keywords. By manual analysis, we categorized them as listed below:

1. **Purpose/Category words**: This type of tags are 50% of the non-keywords tags. Users tend to use heavily words like: “reference”, “howto”, “tips”, “tools”, “technology”, “tutorial”, “freeware”, etc to describe higher levels of abstraction of the content. Top tags in del.icio.us contain many of these kind of words which confirms their popularity.
2. Single/Multiple strings: This category contains 19% of the non-keywords tags. Using singular/plural form of a keyword as a tag is an instance for this type of tags. On the other hand, since del.icio.us only allows for one-word tags, some people prefer to join words together to make them look like one word. Tags like “wordpressplugins”, “wordpress-plugin” are two examples.

3. Acronyms/Non-English words: 11% of non-keywords tags were of this type. Example are: “ux” (stands for user experience), “tech”, “aussie” or “juego” (meaning game in spanish) tags. Some other tags did not exist neither in English nor in a widely known language (like French, Spanish, German, etc) dictionary and some tags were unfamiliar abbreviations or some weird numbers.

4. URLs/Format words: including 7% of non-keywords tags. Examples are using domain name of the website as a tag or file extensions (formats) like “pdf” or “mp3” referring to a reference website for some articles or music.

5. Time/Date: 4% of the set of non-keywords tags represent the date or time mainly referring to an event. Using the year as an indication to show when a product was released is another example.

6. Process words: 3% of the set were tags like “todo”, “toread”, etc. which are chosen by people for a temporary task.

7. Adjectives: 3% of the tags which were not in the keywords were adjectives like “interesting” or “cool” showing the attitude of the people toward the content.
to the problem that some people don’t know that each separated word is considered a tag in del.icio.us) or characters like “**”, “@”, “!!”, “,”, “&” “...”, etc.

9. Misspellings: 1% of the whole set were misspelled words being used as tags.

From this experiment we learn that people mainly use tags that are meaningful to themselves and used for their own personal needs. Tags like: “reference”, “howto”, “todo” or “***” do not contribute much to finding a web document socially and are only useful for their own taggers. Low similarity between tags and keywords indicates that there are lots of web documents bookmarked with tags that are not content-extracted keywords, so retrieving these documents based on their keywords is difficult. However, property tags that are used for a multimedia content are very helpful for retrieving it, because machine-extracted keywords often do not work with these kind of contents. Therefore, tags provide an orthogonal search mechanism to content-based search, but one that is limited by the specificity of tags.

4.2 Top tags

In this experiment we analyzed top (popular) tags to find out what they are about, whether they are stable over time or not, how they are distributed among popular web documents and whether there are any groups of mostly co-occurring top tags or not:

By manually categorizing top tags we found out that they fall into these categories:

1. Computer-related tags: 45% of the top tags were tags like: “blog”, “programming”, “ajax”, “.net”, “jquery”, “api”, etc. that are computer-related.
2. **General-topic tags**: 45% of the top tags were tags about general topics such as: “business”, “food”, “health”, “magazine”, “movies”, “news”, etc.

3. **Graphic-related tags**: 7% of the tags were those like: “graphics”, “images”, “photos” that people use both for annotating a computer-related web document (where the content is about, for example an image-editing software, or a web page for downloading icons) or a web document that is generally about photography.

4. **Adjective tags**: 2% of the tags were like: “funny”, “cool”, “interesting” which are adjectives.

5. **Process tags**: 1% of the tags were user-related process tags like: “todo”, “toread”.

This finding tells us that using general tags that represent the higher level topic of the content is popular and it confirms what we noticed in our previous experiment. On the other hand, a big group of del.icio.us users are computer domain users. We also calculated the percentages of these kind of tags in dataset A, used in the previous experiment, and we found that 35.5% are computer-related tags, 52% are general topic tags, 11% are graphics-related, 1% are adjectives and 0.5% are process tags.

The outcome of second experiment (top tag similarity over a time period) showed that the similarity is 94.42%. This indicates that, over two and a half months, top tags did not change a lot. Observation showed that the change all happened in general topic-related tags. That is probably because general topic tags represent broader topics as opposed to computer related tags and there are more events, news or breakthroughs for them compared to computer-related happenings.

Figure 4.2 shows the result for the third experiment which was finding how top tags
are distributed among popular web documents:

This figure shows that the growth rate of popular bookmarks containing top tags is linear, so each popular web document contributes the same in adding new top tags. In other words, top tags are distributed uniformly among popular web documents.

In the next experiment we calculated the co-occurrence matrix of top tags. Figure 4.3 shows the result of top-tag co-occurrence measurement after applying the technique described in Section 3.2. This figure shows that there are no groups of highly co-occurring top tags. In other words we cannot find any community of top tags that occur together frequently. This result indicates that there is little social structure to top tags that might be exploited.

4.3 Tag frequencies per user

The results of plotting tag frequencies for each user showed that the user tags fall into two or three different categories:

The first category, which is a small part of the whole tag list, includes those tags that
CHAPTER 4. RESULTS AND DISCUSSIONS

Figure 4.3: 100 Top Tag co-occurrence matrix: Brighter colors indicate higher values.

have fairly large frequency while the second category contains a bigger set of tags with considerably lower frequency. Figure 4.4 shows samples of tag frequencies for four different users. (We tried many more and all are the same.) Those points, where a noticeable change in tag frequencies (actually a sharp change in the derivatives) happens, have been shown with circles.

The empirical result of this experiment tells us that for every user we can find 2 or 3 sets of tags with higher frequencies compared to his or her remaining tags that probably shows this user’s topics of interest. It actually suggests a set of very small tags that users remember and reuse properly and then a set that they can’t remember. This result could be exploited for page recommendation to users and personalization which del.icio.us currently lacks. Recommendation of popular web documents per user can be done by recommending those popular web documents that have been bookmarked with any one of his/her most frequent tags.
Figure 4.4: Tag frequencies. X-axis and Y axis represent the tags and their frequencies respectively. Tags are sorted in descending order of their frequencies.
4.4 User similarities per tag

In this experiment we utilized dataset C, a corpus of 500 top tags along with 50 users associated with each tag, to explore the similarity of the users who all use the same tag. We generated a user-user similarity matrix for each tag. Step by step average and standard deviation of the user similarities per user can be seen in Figures 4.5 and 4.6. In all of these figures, the tags are sorted in ascending order based on the average similarity values between their users.

As can be seen from the diagrams, there is neither complete nor zero similarity between all the users of each tag. Since all the users have at least one tag in common (i.e. the one for which the similarity matrix is created), zero similarity will not occur. Complete similarity between all the users of a tag is very unlikely.

For the first 100 top tags (Figure 4.5a), the average similarity between users is fairly low, ranging from 4.6% to 13%. This shows that people who use the same tag do not have the same tagging behavior and have a wide range of different topics of interest. It also indicates that people are under no constraint for choosing other tags even if they have already used a specific tag.

The standard deviation of the first 100 top tags (shown in Figure 4.5b) sorted in accordance with the average similarities, indicates that as the average similarity increases the standard deviation is increasing overall (according to the trend line shown for the standard deviation diagram). This characteristic remains the same for the next 400 top tags. (Figures 4.5d to 4.6d)

Figure 4.5c suggests that the tags fall into two categories:

1. a large group (96% of the users) having user similarities less than 12%
CHAPTER 4. RESULTS AND DISCUSSIONS

(a) The average of user similarities for the first 100 top tags

(b) The standard deviation of user similarities for the first 100 top tags

(c) The average of user similarities for the first 200 top tags

(d) The standard deviation of user similarities for the first 200 top tags

(e) The average of user similarities for the first 300 top tags

(f) The standard deviation of user similarities for the first 300 top tags

Figure 4.5: Average and standard deviation of user similarities for the first 300 top tags. The tags are sorted in ascending order based on the average values. The trendline is shown for standard deviation diagrams.
CHAPTER 4. RESULTS AND DISCUSSIONS

(a) The average of user similarities for the first 400 top tags

(b) The standard deviation of user similarities for the first 400 top tags

(c) The average of user similarities for the first 500 top tags

(d) The standard deviation of user similarities for the first 500 top tags

Figure 4.6: Average and standard deviation of user similarities for the first 400 and 500 top tags. The tags are sorted in ascending order based on the average. The trendline is shown for standard deviation diagrams.
2. a small group (4%) having user similarities greater than 12%.

The average and standard deviation patterns for top 300, 400 and 500 tags look the same, except that the maximum average similarity for users goes up from 35% for the first 300 top tags to 42% in the next 400 and 500 top tags. The previously mentioned two categories for the tags still remain, except that group sizes are 90% and 10%.

From the fact that 90% of the top tags in the study have less than 12% user similarity, it could be inferred that one tag by itself does not indicate strong communities of people having much the same topics of interest.

The maximum of average similarity values, as indicated in Figure 4.6c is 42% which is for the tag “weblog” indicating that people tagging “weblog” have 42% of their tags (topics of interest) the same as each other, while the minimum of average similarity between users is for the tag “utilidades”, the Spanish word for “utilities”, indicating those people that use “utilidades” have very few tags in common or low agreement on their topics of interest.

Figure 4.7 shows the maximum of user similarities for the first 500 top tags. Since the diagrams were not different, just the ones for the first and last steps are shown here.

The diagrams in this figure indicate that for each tag there are at least two users with completely the same set of tags (except for the tag “books”, ranked as 76th top tag, where the maximum similarity between its users is 27%). Although this might be not expected at first glance, considering that there are many users with small set of tags, the likelihood of two users having same set of tags is high.

We also decided to visualize the top tag co-occurrence matrices, but this time on the similarity matrices for the top 500 tags to show if any communities are present (although we already know that, if any community exists, the similarity between its
people is very low).

To group users based on their similarities, we applied that technique on the similarity matrices created during the experiment. Figures 4.8 to 4.19 show the similarity matrix for the chosen tags: “web2.0”, “mp3”, “seo”, “3d”, “idea”, “literature”, “cool”, “howto”, “2.0”, “gtd”, “folksonomy”, “diy”, “rubyonrails”, “voip”, “book”. Some of these tags are general-topic tags, while some of them are computer related. The rows and columns represent users of the tag.

There are rarely blocks of similar users except for some tags like “book” or “voip” (although all the clusters of similar users are weak) and people using a common tag do not have similar topics of interest. It seems that, since users are under no restriction to use any particular tags and observing that they do not care what other tags they have used so far, normally, blocks of similar users could not be seen in bookmarking systems.
(a) The maximum of user similarities for the first 100 top tags

(b) The maximum of user similarities for the first 500 top tags

Figure 4.7: Maximum values of user similarities for the first 500 top tags. The tags are sorted in ascending order based on the maximum similarity values.

Figure 4.8: User similarity matrix for tag “web2.0”. The rows and columns are users of the tag and entries show the pairwise user similarities. Three different structures can be seen: red (a group of two users having relatively higher similarities), blue (a block of users having lower similarities), green (a small block of users with relatively lowest similarities)
CHAPTER 4. RESULTS AND DISCUSSIONS

Figure 4.9: User similarity matrix for tag “mp3”. The rows and columns are users of the tag and entries show the pairwise user similarities. 2 different structures can be seen: red (a small group of users with higher similarities), blue (a big group of users with low similarities).

Figure 4.10: User similarity matrix for tag “seo”. The rows and columns are users of the tag and entries show the pairwise user similarities. 3 different structures can be seen.

Figure 4.11: User similarity matrix for tag “3d”. The rows and columns are users of the tag and entries show the pairwise user similarities. 3 structures are shown.
CHAPTER 4. RESULTS AND DISCUSSIONS

Figure 4.12: User similarity matrix for tag “idea”. The rows and columns are users of the tag and entries show the pairwise user similarities. 4 different structures are shown.

Figure 4.13: User similarity matrix for tag “literature”. The rows and columns are users of the tag and entries show the pairwise user similarities. 2 structures can be seen.

Figure 4.14: User similarity matrix for tag “cool”. The rows and columns are users of the tag and entries show the pairwise user similarities. 2 structures are shown.
Figure 4.15: User similarity matrix for tag “howto”. The rows and columns are users of the tag and entries show the pairwise user similarities. 4 structures are shown.

Figure 4.16: User similarity matrix for tag “2.0”. The rows and columns are users of the tag and entries show the pairwise user similarities. 2 structures can be seen.

Figure 4.17: User similarity matrix for tag “gtd”. The rows and columns are users of the tag and entries show the pairwise user similarities. 2 structures are shown.
Figure 4.18: User similarity matrix for tag “folksonomy”. The rows and columns are users of the tag and entries show the pairwise user similarities. 2 structures can be seen.

Figure 4.19: User similarity matrix for tag “voip”. The rows and columns are users of the tag and entries show the pairwise user similarities. 4 structures are shown.

Figure 4.20: User similarity matrix for tag “book”. The rows and columns are users of the tag and entries show the pairwise user similarities. 2 structures can be seen.
4.5 Web document similarities per user

In this experiment we utilized dataset E, a corpus of 500 users, using different number of tags, along with 50 web document for each of them to measure the similarities of web documents bookmarked by each user. Since the users were randomly chosen, the overall statistics for the whole dataset are shown here as opposed to the step-by-step method we conducted in the last section.

The average and standard deviation of the results can be seen in Figures 4.21a and 4.21b. In both of these figures, the users are sorted in ascending order based on the average similarity of their documents.

As can be seen from the average diagram (Figure 4.21a), the similarity between

![Figure 4.21: Average and standard deviation of document similarities for 500 users. The users are sorted in ascending order based on the average values. The trendline is shown for standard deviation.](image)

the documents of users are rather low, having a minimum and maximum average similarity of 0.3% and 12% respectively.

This indicates that users in del.icio.us have a diverse range of topics of interest and we would not expect to find users focusing on closely related topics to bookmark.
The average diagram in Figure 4.21a suggests that users fall into two categories: (1) those for whom the average similarity of their web documents is less than 4%, which contain 90% of the whole dataset, and (2) those for which the average similarity of their web documents is greater than 4%, which contain 10% of the dataset.

Figure 4.21b shows a small increase in standard deviation as the average increases (according to the trend line).

Figure 4.22 shows the maximum of web document similarities for 500 users. It indicates that there is a group of users having at least two completely similar documents. Actually, three groups of users can be seen in this diagram: (1) a block of 40 users with maximum document similarity of 33%, (2) a block of 22 user with maximum document similarity of 66% and (3) a group of 50 users with maximum document similarity of 100%. As it was mentioned in Section 3.5, we calculated the similarity of web documents based on the aggregation of tags and keywords extracted for each document where, on average, the tags were two-thirds of the whole set of tags and keywords. This suggests that probably, for the first group, complete similarity happened for the extracted keywords of at least two of their documents.
but zero similarity in the tags of these two documents, meaning that there are documents in del.icio.us having the same set of keywords but totally different set of tags, while for the second group the story is the opposite, i.e. probably complete similarity for the set of tags for two documents but no similarities between their extracted keywords. It shows that users who have bookmarked these two documents consider them similar by using the same tags for both, although their extracted keywords are totally different. In the third block the keywords and tags of at least two documents were the same, showing that there are web documents having the same significant keywords and complete agreement upon their tags.

After observing that 90% of the users in the study have less than 4% of agreement between their documents, we decided to visualize blocks of similar documents (if any) for the users. Using MATLAB, the previously used technique for visualization was applied on the document similarity matrices created in this experiment and the image of resulted matrices was displayed.

Figures 4.23 to 4.27 show the result matrix for 5 chosen users. The rows and columns represent the web documents of each user.

From this experiment, we learn that users bookmark a variety of different topics and the similarity between their documents in terms of folksonomy tags and keywords is very low. We showed that there aren’t any blocks of similar documents for user and this indicates that clustering of web documents based on their folksonomy tags and keywords is not useful. The result of this experiment shows that tags are not robust ways to retrieve related documents.
CHAPTER 4. RESULTS AND DISCUSSIONS

65

Figure 4.23: Document similarity matrix for 54th user. The rows and columns are documents and entries show the pairwise document similarities. A wide spectrum of similarities can be seen between all the documents.

Figure 4.24: Document similarity matrix for 128th user. The rows and columns are documents and entries show the pairwise document similarities.

Figure 4.25: Document similarity matrix for 209th user. The rows and columns are documents and entries show the pairwise document similarities. 2 structures are shown.
Figure 4.26: Document similarity matrix for 376th user. The rows and columns are documents and entries show the pairwise document similarities. 1 block is shown.

Figure 4.27: Document similarity matrix for 495th user. The rows and columns are documents and entries show the pairwise document similarities. 2 blocks can be seen.
4.6 Tag similarity on consecutive days

In this experiment we compared the similarity of tags that different users use on consecutive bookmarking days. We noticed that active and inactive users show different tagging behavior. In Figure 4.28 we have displayed a typical active user (Figure 4.28b) and a typical inactive user (Figure 4.28a). The figures suggest that, for an inactive user, we could see consecutive days that the same tags were used while bookmarking, something which cannot be seen in active-user behavior. Furthermore, there are lots of consecutive bookmarking days, in which an inactive user uses completely different set of tags comparing to the previous bookmarking day, while for an active user there are hardly consecutive days with zero similarity of tags.

The reason why we could not see consecutive days with 100% similarity of tags for an active user was found by observation: there are lots of redundancies in the tags of active users on consecutive days. Tags like: “addon”, “addons”, “add-on” and “addons” or tags like: “answer”, “answers”, “answers,” and “answers) all show that they
do tagging rather on the fly and they do not care or are not able to use a minimized set of tags, while for an inactive user there is less redundancy in his/her tags. Inactive users have a small set of tags which they may remember better and use repeatedly. On the other hand we speculate since the inactive user is finding new interests, there are days that he/she posts totally different bookmarks, while an active user’s topics of interests are more stabilized and he/she hardly ever bookmarks with totally new tags.

What we learn from this experiment, is that there is different tagging behavior between active and inactive users. Active users typically have a larger set of tags with more redundancy, but they are more stable in their tagging practices in terms of not totally different or totally the same tags on consecutive bookmarking days, while inactive users typically have smaller set of tags but they are not stable in their tagging practices on consecutive days by posting totally different or completely similar tags. This result agrees with a model of a small set of memorable tags and a big set of “random” tags which we already observed in Section 3.3. This finding, while showing the tagging behavior of different users, also indicates that typically people do not manage their set of tags when they use these systems frequently. This makes difficult both knowledge discovery based on tags and also retrieving content.
4.7 User bookmarks vs. Tags

In this experiment we utilized dataset E to find whether any relation could be found between the number of tags and the number of web documents a user bookmarks. We plotted each user in a two dimensional diagram, where the x-axis and y-axis represent the number of bookmarks and tags respectively. Figure 4.29 shows the resulting diagram. There are outliers having either a very large number of bookmarks or a very large number of tags.

![Diagram showing number of bookmarks vs. number of tags for 500 users.](image)

Figure 4.29: Number of bookmarks vs. number of tags for 500 users.

By removing these outliers, we get Figures 4.30 and 4.31. 78% of the remaining users have fewer than 7000 bookmarks and 4000 tags, while 50% of them (39% of the total), have fewer than 3000 bookmarks and 2000 tags. The diagram suggests that, for most of the users, the number of their bookmarked web documents is greater than the number of tags they use. Observations also indicates a great deal of redundancy in tags for users with large number of bookmarks.

In order to find out whether the users with very large number of bookmarks are
CHAPTER 4. RESULTS AND DISCUSSIONS

Figure 4.30: Number of bookmarks vs. number of tags for 500 users (zoomed).

Figure 4.31: Number of bookmarks vs. number of tags for 500 users (further zoomed).
actually spamming users or not, we defined the following characteristics for spamming users:

1. There is at least one popular tag in some of their bookmarks which has nothing to do with the content.

2. These suspicious bookmarks are for certain limited domains.

3. Their number of suspicious bookmarks is high.

Then, we picked those users with more than 10,000 bookmarks for this purpose. Our observation suggests that these users are not actually spam users, because of the following reasons:

• In each of their bookmarks, the tags are related to the content of the web document.

• They are not bookmarking certain limited domains.

There were also 15 users with a large number of tags and a small number of bookmarks. Manually observing their bookmarks suggested that:

• their personal tags, acronyms and numbers play big role in their tag list.

• they post large number of tags to their bookmarks including the domain or the URL of the web document

• most of their tags look like redundant and not necessary for content retrieval such as “image”, “photos”, “pictures” for one bookmark or “weblog”, “blogs”, “blog”, “blogging”, “blogger”, “weblogs” all for one bookmark.
• their tag lists include personal tags like “—”, “@queue” or “toread” as well as redundancies (such as “animation”, “animations”, “animatios” or ”desing-patterns”, “designpattern”, “designpatters”, “design” and “patterns”, etc.) or typos such as “bloggig”, “bloggin”, “blogginc” or “bloggiong” along with lots of acronyms and numbers which make their tag lists big.

• almost all of them use non-English words as well as English words in their bookmarks.

![Figure 4.32: A sample of tag redundancies for a user with many tags](image-url)
CHAPTER 4. RESULTS AND DISCUSSIONS

Figure 4.33: A sample of bookmarks for a user with many tags
Figure 4.34: A sample of bookmarks for another user with many tags
CHAPTER 4. RESULTS AND DISCUSSIONS

The above observation confirms that these people do not follow the widely accepted model of tagging by not managing their tag list and having higher rate of personal tags.

In this experiment we learned that although there is not a tight relationship between the number of bookmarks and tags, it seems that people have fairly large number of tags and bookmarked web documents. Unusual tagging behavior could also be observed. From this we learn that people do not follow the presumed model of tagging in which the number of tags and bookmarks is kept small to make the retrieving easier. This suggests that people mainly use tags for arranging their web documents but they do not pay much attention to using tags in a way that helps finding the document easier, since there are lots of redundancies in their set of tags. It seems that people have not yet learned the functionality of tags for searching.

4.8 Discussions

Low similarity between tags and content-extracted keywords indicates that:

- Tags are a potential source of extra information about content. Currently machine-extracted keywords used in search engines represent the content more specifically than tags. However, those folksonomy tags that represent the properties of the content (such as “video”, “music”, “funny”, “political”, etc.) represent some features of the content that machine-extracted keywords cannot.

- Although clustering of similar web documents based on their associated tags may give different results compared to clustering based on extracted keywords, they might be interesting since they show how users see them in similar ways.
• Although del.icio.us is a social bookmarking system, people are mostly using it for their own personal purposes. Categorization of tags (that are not found in keyword sets) show that people use tags in a way that is meaningful for themselves without considering whether other people might understand anything from their tags. This does not help with finding web documents since they have been bookmarked by either general tags like: “howto”, “tips”, “reference” or personal tags like: “todo” or “funny”.

• People may not even use their own tags much. Actually they have a small set of tags which they remember and use frequently as well as a big set of “random” tags that they do not properly remember and use very rarely.

Examination of top tags in del.icio.us revealed some other characteristics of tagging structure. We noticed that computer-related and general-topic tags are two dominant popular tags in del.icio.us. This shows that the society on which we are doing experiments and our results apply to, includes a group of people with different topics of interest.

We also found out that the set of popular tags are mainly stable over time. We show that the growth rate of popular bookmarks containing top tags is linear, meaning that there is not a set of popular bookmarks that are pointed to by most of the top tags. In other words the top tags are distributed uniformly among the popular web documents. By considering the top-tag co-occurrence matrix we showed that there is no large or popular community of co-occurring or statistically related tags. This result indicates that clustering of top tags based on their co-occurrence is not useful for tag recommendation.

By analyzing the tag frequencies of users we noticed that all users have one or two
sets of tags that they use with high frequency, which definitely represent their most important topics of interest. This finding could be used for web-page recommendation and personalizing the system for each user by showing popular web documents related to his/her topics of interest.

Analyzing user similarities per tag, based on what other tags they have used, showed that there is low similarity between users of the same tag. However, there are at least two users with the same set of tags for almost every tag. This indicates that there are not significant communities of similar users around tags. There are partial clusters but they are weak.

The investigation of web document similarities per user, showed that there are very small similarities between the web documents each user bookmarks. For some users there is complete similarity between the extracted keywords of two documents but no similarity between their tags, while for other users it is the opposite, and for some users we see complete similarity of both tags and extracted keywords for at least two of their documents. The reason why two documents with different keywords have been bookmarked with similar tags (or the opposite case) is probably because people mainly use general tags (such as “reference”, “howto”, “tips”) as well as personal tags (such as “toread” or “funny”) which make either two different web documents look similar (based on their tags) or two similar documents look different (again based on their tags). From the similarity matrices, we found out that there are hardly ever blocks of closely related documents for each user, meaning that users are annotating a large number of different topics. This characteristics shows that clustering of similar web documents for users cannot be exploited as a model for predicting user behavior. Although both the similarities between the user tags and documents of a user are low,
we noticed that users tend to have more commonalities between their tags rather than the documents they bookmark. This tells us that similarity of tags between users is a better heuristic for web-page recommendation as opposed to using similarity of web documents. This agrees with Zeng and Li [7] result which showed that tags, under user-based framework, can be exploited for web-page recommendation.

We show the different tagging behavior of active and inactive users by comparing the similarity of tags they use on a sequence of bookmarking days. We noticed that inactive users either use totally different or very similar set of tags on consecutive bookmarking days while active users hardly ever use either the same or completely different tags on consecutive days. This finding showed another regularity in tagging behavior. The observation also showed that there are lots of redundancies in the set of tags of active users which indicates that these do not manage their tag sets and do the tagging in a rather ad hoc way.

Finally the tagging behavior model usually assumes that the size of tags and bookmarks per user will be reasonably small so that users can remember what tags they have used for a web document. We examined the number of tags and bookmarks per user. The result showed that there is a big group of users having fewer than 7000 bookmarks and 4000 tags while half of them have at least 3000 bookmarks and 2000 tags which are really big numbers. We also noticed that some people who have large number of bookmarks also have redundancies in their tags which make the size of their tag set large. These are large numbers indicating that people are not tagging in a practical way. This suggests that they do the tagging without considering or remembering the tags that they have already used. By using redundant tags they do not pay attention to making the retrieval easier. If they could manage their number
of tags and bookmarks by keeping them small and tried to be consistent in using tags, it would help them find their bookmarks easier. So it’s questionable whether tags ever get used for retrieval as the model assumes - either socially and collaboratively or individually - or tags are “write only”. However, people do tagging because they can add how many tags they like, according to any rules they want. They have this choice and they use it mainly in a self-communicative manner. They do this for adding personal notes, helping themselves recall the content or adding non-obvious details about the content.
Chapter 5

Conclusions and Limitations

In this chapter we draw some conclusions and mention some of the limitations of this research. We end this chapter by proposing future work.

5.1 Conclusions

Collaborative-tagging systems, also known as social bookmarking systems or folksonomies, are one of the applications that emerged under the Web 2.0 paradigm. In these systems, users annotate resources (such as web documents) with freely chosen keywords, called tags.

In this research we explored the value of folksonomy tags by investigating the bookmarking and tagging behavior within del.icio.us to discover how tagging works and if there are any regularities in tagging practices. Specifically we investigated the relation between tags, users and web documents to see whether there are any social structures or communities in this system than can be exploited or not.

First of all, we analyzed folksonomy tag functionalities by measuring their relevance
to machine-extracted keywords. The low similarity between tags and extracted keywords showed that people are not using tags for content abstraction, i.e. the tags are actually split across a number of different groups that are not content-related; and those tags that are content-related are mainly descriptive of the category into which the web document belongs, rather than describing the content itself. This is an interesting aspect of folksonomies since it represents another contextual dimension of the content which machine-extracted keywords cannot add. A folksonomy tag, like “political speech” reveals an aspect of the web document that is hard for machines to discover. We also found out a group of folksonomy tags which identify the properties of the web document have another value which machine-extracted keywords could not achieve at all: tags like “funny”, “video” or “music”.

We also noticed that people mainly use tags according to their own personal needs rather than in a social manner which is useful for other people too, i.e. folksonomy tags are rather personal and do not have a global semantics. One recommendation to make these systems more useful for information retrieval would be to indicate how relevant the tag is to the content, as well as how recent this tag for the document is, which is currently displayed. This could be done in the same manner we measured the similarity of tags to the content. This reduces the users’ need for manually evaluating the relevance of the tag to the content, to be confident of the reliability of tags. In this case, the user can still decide about the property tags.

We also show some regularities about tagging structures:
- That top tags tend to be stable over the time
- Top tags are uniformly distributed among popular web documents, that is, there is no group of popular bookmarks that contain a majority of top tags, which could be
exploited (because if these kinds of web documents existed they were successful web documents that have been bookmarked by many users with popular tags). We relate this finding to the equality in size of general-topic and computer-related tags, the two dominant categories existing in del.icio.us top tags.

- Tag frequency analysis, showed that users typically have two or three categories of tags with higher frequency of use related to their topics of interest which could be exploited for web-page recommendation and personalization.

- Considering time, we found that active users (having large sets of bookmarks) show more balanced behavior, using more stable tags in consecutive bookmarking sessions, while inactive user tagging shows oscillation between zero and complete similarity of tags in consecutive days. We related this to the fact that their topics of interests are still being formed in del.icio.us (causing zero similarity) and that they have smaller set of tags which helps better remembering and reusability (causing complete similarity). In contrast, active users have lots of redundancies in their tags.

To examine the formation of communities, we investigated the relationships between tags, users and web documents. We noticed that there is not a significant community of popular tags, based on their co-occurrence, so we conclude that there are hardly ever statistically related tags and little associative relationships among top tags. Similarity of users per tag is fairly low, showing that there are hardly ever communities of similar users around tags and in all cases the clusters are weak. In the same sense, we examined the relationships between users and their bookmarked web documents to find that the similarity of web documents bookmarked by a particular user are even lower than the similarities we observed in a previous experiment. There are hardly ever recognizable groups of similar web documents for a user. We relate both of these
results (for tag-users and user-documents relations) to the users’ freedom of choice in using tags and bookmarking variety of topics. As there are not communities of similar tags, any clustering of these three components yields little useful information. These results indicate that people do not manage their bookmarks, which makes data mining techniques or even logical inferences of little use since they support very few social formations.

Finally, we compared user tagging behavior with the widely-held bookmarking and tagging model and show that users do not follow this model at all closely as they have a large set of tags and bookmarks. This suggests that people mainly use folksonomies for storing their bookmarks rather than searching and retrieving them later. Most of the users show similar behavior but there are unusual users having either a large set of bookmarks or large number of tags. However, observing their bookmarks show that they have used related tags to their bookmarked documents while there are redundancies in their tags which altogether indicate they are not spammers.

All in all, the results of this research indicate that collaborative and social tagging is largely a myth at present since tagging is being done in a way that does not support the idea of social bookmarking. It suggests that, although folksonomies are conjectured to support global navigation of organized resources, they do not actually support that in a convincing manner. That’s because people do not actually share their tagging patterns and do not see what other people have tagged. This indicates that either people need more time to work with these systems to have their tags converged or these systems should better support them, for example by showing to what extent a user tags similar to other people. So far, we can say that social bookmarking has not yet supported emergent classification systems that could be utilized for
CHAPTER 5. CONCLUSIONS AND LIMITATIONS

information navigation. Calling these systems as “collaborative-tagging systems” is not correct, when there is little collaboration between users in tagging practices. The users should know that the organization of information in these systems is currently personal rather than social.

However, the findings of this research revealed a few regularities in tagging which can be exploited for user-experience improvement.

5.2 Limitations and Future work

The main challenge we encountered in this research was in collecting the datasets which is related to the del.icio.us limitations. Top tags displayed in this system are small in size, popular bookmarks are completely temporal which just indicates that they’ve been popular recently and this list is continuously updated. There is also no way to find top users. The feeds do not cover all the information we would have found useful or they provided it in a limited way. Also, an unexpected delay during the experiment happened because of del.icio.us being upgraded to a new server with new user interface and newly changed feeds without prior notice, which required us to restart the data collection. Also, del.icio.us uses IP throttling to block an IP address when it sends high rates of requests for data. This was an obstacle which required us to use a proxy for sending our requests, but using the proxy slowed down the process. Therefore we could not work with larger datasets, which are likely to be more descriptive.

Gathering large sets of data in a timely manner and then comparing the results of the experiment with the results of re-conducted experiments after a time period and also considering other social bookmarking sites than del.icio.us will reveal dynamics
of collaborative tagging systems better, both of which are recommended for future work.
Bibliography


