GENERATION OF CLASSIFICATORY METADATA FOR WEB RESOURCES USING SOCIAL TAGS

by

Sue Yeon Syn

B.A., Ewha Womans University, Korea, 2000

M.I.S, University of Library and Information Science, Japan, 2002

Submitted to the Graduate Faculty of

School of Information Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

Sue Yeon Syn

It was defended on

December 10, 2010

and approved by

Peter Brusilovsky, PhD, Associate Professor, School of Information Sciences

Brian Butler, PhD, Associate Professor, Joseph M. Katz Graduate School of Business and College of Business Administration

Daqing He, PhD, Associate Professor, School of Information Sciences

Stephen Hirtle, PhD, Professor, School of Information Sciences

Dissertation Advisor: Michael Spring, PhD, Associate Professor, School of Information Sciences

Generation of Classificatory Metadata for Web Resources using Social Tags

Sue Yeon Syn, PhD

University of Pittsburgh, 2010

With the increasing popularity of social tagging systems, the potential for using social tags as a source of metadata is being explored. Social tagging systems can simplify the involvement of a large number of users and improve the metadata generation process, especially for semantic metadata. This research aims to find a method to categorize web resources using social tags as metadata. In this research, social tagging systems are a mechanism to allow non-professional catalogers to participate in metadata generation. Because social tags are not from a controlled vocabulary, there are issues that have to be addressed in finding quality terms to represent the content of a resource. This research examines ways to deal with those issues to obtain a set of tags representing the resource from the tags provided by users.

Two measurements that measure the importance of a tag are introduced. *Annotation Dominance (AD)* is a measurement of how much a tag term is agreed to by users. Another is *Cross Resources Annotation Discrimination (CRAD),* a measurement to discriminate tags in the collection. It is designed to remove tags that are used broadly or narrowly in the collection. Further, the study suggests a process to identify and to manage compound tags.

The research aims to select important annotations (meta-terms) and remove meaningless ones (noise) from the tag set. This study, therefore, suggests two main measurements for getting a subset of tags with *classification potential*. To evaluate the proposed approach to find classificatory metadata candidates, we rely on users' relevance judgments comparing suggested

tag terms and expert metadata terms. Human judges rate how relevant each term is on an n-point scale based on the relevance of each of the terms for the given resource.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

The World Wide Web (WWW) makes it possible for users to post resources in a distributed way and find resources by following links. Pandia Search Engine News (2007) estimates the size of the WWW to be between 15 and 30 billion pages. In a network of this size, it is difficult to locate all the resources relevant to a given topic or query by link navigation. Albert et al. (1999) demonstrated that based on the small world and power-law topology, two randomly chosen documents on the web are on average 19 clicks away from each other. As the size of the WWW grew, search engines emerged to help users search for web resources based on full-text indexing of accessible pages. Despite algorithmic improvements in ranking and clustering, full-text indexing using data such as page content, link structure, and query log data suffers from problems such as synonymy and polysemy as well as semantic connectivity. Taylor and Clemson (1996) listed the following as weaknesses of search engines:

- There are duplicate pages in the same set of retrieved hits,
- Results are unpredictable,
- Results can be quite misleading,
- Search engines do not readily disclose the contents of their databases nor do they provide a description of the criteria used to include a document in their files,
- Vocabulary is not controlled, and punctuation and capitalization rules are not standardized,

- Relationships and relevance often cannot be analyzed without actually examining each item.

Some of the weakness of search engines remain still to current search engines mainly due to the large volume of information on the Web, dynamic web pages, and spamming. Evans et al. (2005) addressed that current search engines have difficulties in indexing growing documents on the Web, and in addition, the dynamic changes of the content make the indexes stale. Moreover, increasing spamming tricks on ranking algorithms to make documents that are irrespective of user's need located high in search results list. It is also related to the weakness of the search engines in finding the content of web pages mentioned by Evans et al. (2005) since the content may change at any time and the spamming can rely on metadata information embedded in the page.

Services such as Yahoo! create directories based on content analysis done by humans. While directory services can provide more precise classification of web resources and reduce information overload, human classification is costly and does not scale well. Both full-text indexing and directory services have problems related to the churn in web pages (new pages appearing, old pages changing or being removed) and the increasing use of programmatic links (CGI programs and web services) that "hide" the content of pages. As Web 2.0 technologies such as AJAX and RSS take hold, these problems are compounded. The Semantic Web has been envisioned as a structured, machine-understandable web based upon structured resource descriptions (Berners-Lee et al., 2001). In efforts to provide a better way to find proper resources on the WWW, research has been undertaken to analyze web resource content so as to create high quality metadata automatically that is equal to or better than that generated by humans but without the cost and scalability problems.

A previous study has shown how semi-automated systems can allow novices to participate in the metadata generation process (Syn and Spring, 2008). While tools improve the quality of metadata produced by novices, in comparison with experts, novices were less stable in generating proper semantic metadata, i.e. keywords and subject classification. With the increasing popularity of social tagging systems, the potential for using social tags as a source of metadata is being explored. Social tagging systems can simplify the involvement of a large number of users and improve the metadata generation process, especially for semantic metadata. By using social tags as a type of metadata, this research aims to find a method to classify web resources. In this research, social tagging systems are considered as a source for non-professional catalogers' participation in the metadata generation process, and social tags are considered as a type of metadata for web resources. The question is whether social tags can be mined in such a way as to enable less skilled classifiers to generate classificatory metadata. Because social tags are not a controlled vocabulary, there are problems in finding high quality terms to represent the content of a resource. This research examines ways to deal with those problems to gain a better set of tags to classify and represent the resource from the user-generated tags.

1.1 Focus of the Study

Human-generated metadata developed by skilled classifiers is generally considered to be more precise than system-generated metadata. Over the last years, no system has emerged that can generate high quality metadata automatically. To reduce human effort, it seemed to be essential to sacrifice the quality of metadata.

Some studies turned their focus to the possibility of letting novice catalogers participate in metadata generation process (Syn and Spring, 2008; Trant, 2006). However, it is difficult to motivate non-professional users to create metadata. In the past few years, social tagging systems have gained popularity among web users as a method of organizing, filtering, and retrieving web resources. Social tagging systems, such as Delicious and Flickr, have successfully let users be involved in tagging by providing services to motivate and also benefit users by tagging, for example bookmarking favorite links, organizing/sharing pictures, and getting recommendations on related web pages. Quintarelli (2005) and Sen et al. (2007) indicate that social tagging systems allow ordinary users to contribute to metadata generation, out-scaling expert-maintained taxonomies. Sen et al. (2007, pp. 361) found that "in 200 years of existence the Library of Congress has applied their expert-maintained taxonomy to 20 million books. In contrast, in just four years, flicker's users have applied their adhoc tagging vocabulary to over 25 million photographs." In addition, Heymann et al. (2008) suggested that social bookmarking systems are a good source of novel and active pages in terms of information discovery. Their data comparing the Open Directory Project and Delicious showed that metadata generation by humans takes more time and therefore includes comparatively older pages. Sen et al. and Heymann et al.'s observations support the view that social tagging systems allow newer web resources to be associated with metadata information in less time.

While social tags scale well, the question arises as to whether social tagging systems can enable less skilled classifiers to generate good classificatory metadata. Much research has focused on using social tags to provide better retrieval and ranking results. When social tags are used with indexing or ranking methods, social tags are considered as a bag of terms for web documents. However, as directory services take approaches other than indexing for information retrieval, when social tags are used to classify web resources, they should be handled differently. When filtering social tag terms as classificatory metadata, it is most important to find the topics and domains of the resource content. Therefore, unlike indexing methods that consider highly specific terms appearing in the content of a document as important as frequently appearing terms, it is not always true for classificatory metadata terms to be very specific or frequently appearing to represent the contents' topics and domains. Zubiaga et al. (2009) suggested that user-generated tags and comments are actually useful (especially when used together with the content of the document) in classifying web pages than using only the content of a web document. Bischoff et al. (2008) also confirmed that tags, at least in music, are reliable and as good as expert created metadata. Syn and Spring (2009) discussed that in academic papers user-generated tags work nearly as well as author-generated keywords and suggested filtering tag noise could improve the usefulness of tags. These studies have shown that tags can be descriptive and can take part in the role of metadata for classification of web resources. In this research, social tagging systems are considered as a channel for non-professional user participation in the metadata generation process. Social tags can reduce the barrier of human metadata generation in terms of having better scalability and more contribution from users. To make the most use of the beneficial side of social tags, this study will find a way of selecting the tags for representing web resources to increase the quality from a vast amount of user-created tags by addressing the following issues and questions.

- Can the tag noise be reduced?
- Can compound tags be processed to be of use?
- Can a subset of tags be found that provide classificatory metadata?

This process will make the metadata generation process easier and faster and the quality of metadata reliable.

1.2 The Nature of Social Tagging Systems

One of the issues that has to be addressed in the design of the study is the nature of social tagging systems. This study looks to use tags as a means to generate metadata that can be used in classification. Unfortunately, this is not consistent with the design of social tagging systems. Social tagging systems are generally focused on certain types of resources and define a tag as a set of characters bounded by spaces. In addition, many systems suggest tags based on previously assigned tags. These features let users input tags in simpler and easier ways. However they may also lead users to have certain tagging behaviors. As Bischoff et al. (2008) stated, depending on the types of resources and systems, the characteristics of tags may differ. For example, there are more ‘location’ tags in Flickr (images) whereas there are more ‘type’ tags in Last.fm (music).

Furthermore, the current state of data sharing by tagging systems is such that the amount of information that can be gathered from them is limited. While it is possible with most systems to obtain user ids, tag sets, resources (URL, file, etc.), comments/notes, time of creation, etc., other important information may not be available. For example, the specific order in which tags were entered or suggested might be important, but impossible to obtain. Further, while most systems indicate when a tag set was created, they do not have information about modification actions and dates. The basic assumption on tag order is that the sooner a tag appears the more important the tag might be. Golder and Huberman (2006) argued that the position of a tag and its frequency are related – frequently used tags will appear before less frequently used tags. The first tag appearing in a tag set should be expected to be the most important tag for describing the document or at least it will be the more frequently occurring tag. The time a tag was created in a tag set may have an impact on tag frequency since users can easily accept tags that were assigned by other users already. Therefore the first user who bookmarks a document with some tags can influence other users' selection of tags. Both the order of tag input and the time of tag input may impact a decision about the importance of a tag as classificatory metadata. These studies suggest that tag data is noisy and messy and careful attention has to be paid to the process of selecting tags for any particular purpose.

1.3 Limitations and Delimitations of the Study

According to Heymann et al. (2008), the number of resources that are bookmarked is relatively small compared to the size of the web. They also observed that despite the fact that social bookmarking systems cover only a small portion of the web, it covers a high proportion of search results. Nonetheless, one limitation of this study is that it may ignore a significant portion of web resources that have not been bookmarked.

This study analyzes tags used in one bookmarking system. The results may not be generalizable to other tagging systems applied to other types of resources such as images, music, videos, etc.

The sample for this study is gathered from a single social bookmarking system, Delicious. It does not include all the bookmarks in the selected system nor does it include more than one system. While the sample data was crawled in the manner to obtain a representative sample across a number of different topic areas, the sample data of this study may not be representative of the whole population.

1.4 Definitions

1.4.1 Social Annotations

Social annotation systems provide an easy means for user involvement in describing web resources. Zubiaga et al. (2009) has defined five kinds of user-generated annotations: tags, notes, highlights, reviews, and ratings.

- **Tags** are keywords used to define and characterize a web resource. Tags are often a list of user-selected, single-word descriptions.
- **Notes** are free-text descriptions about the content of web resources. Whereas tags are one-word descriptions, notes are descriptions with multiple words or sentences. Both tags and notes are created with users' selection of words and descriptions and are commonly adopted to annotate web resources in social annotation systems.
- **Highlights** are relevant parts of a web resource. Web sites such as Diigo (http://www.diigo.com) allow users to specify the most relevant part of the web documents.
- **Reviews** are free text evaluations of the content of a web resource including both the description and opinion on the resource.

- **Ratings** are user evaluations of web resources commonly done on Likert scales. Websites such as StumbleUpon (http://www.stumbleupon.com) allow users to review and rate web pages.

In terms of how representative web resource content is, highlights and rating do not contribute much. Notes and reviews may include more personal opinions than tags. As Bischoff et al. (2008) also observed, tags for web pages cover topics of the content.

1.4.2 Social Tags

Tags may be keywords, category names, or metadata (Guy and Tonkin, 2006). Tags are collections of user-selected keywords attached to different types of web resources to describe their content. Tagging of content can allow for organization and can facilitate searching. Tags were originally designed to offer an easier method for users to manage and retrieve their own resources. More recently, tagging has allowed for the formation of social networks (John and Seligmann, 2006). Tags are useful since they can be “simple” and “easy-to-create” metadata representing the content of a resource. Social tagging systems enable users to annotate resources (e.g. web pages, images, videos, etc.) with a set of words, “tags”, which are relevant to the content of the resource according to their needs without relying on a controlled vocabulary or a previously defined structure (Specia and Motta, 2007). Social tags allow users to work together categorizing resources for future use.

1.4.3 Controlled Vocabulary

Controlled vocabulary is an established list of preferred terms from which a cataloger or indexer must select when assigning subject headings or descriptors in a bibliographic record to indicate the content of the work in a library catalog, index, or bibliographic database (Reitz, 2004). A controlled vocabulary may also be used in information organization and retrieval, especially to assist users who want material on particular subjects. A controlled vocabulary is usually carefully systematized for use in retrieval systems in the form of a thesaurus or subject heading list with synonyms (Taylor, 2004). Controlled vocabulary is used to provide a limited list of terms to describe a resource so that the problems related to synonymy and polysemy can be reduced, since only provided terms are used and the relationship among terms are defined.

The categories of controlled vocabularies include subject headings, thesauri, and ontologies (Taylor and Jourdrey, 2008). Subject headings capture the essence of topics and related concepts assigned with authoritative terms with the hierarchy, the semantic, and the syntactic relationships. Existing examples of subject headings are the *Library of Congress Subject Headings* (LCSH) and the *Medical Subject Headings* (MeSH). A thesaurus provides a list of words grouped together in a structure according to similarity of their meanings with relationships among the words defined explicitly, such as synonymy. Examples of thesauri are *Art & Architecture Thesaurus* (AAT) and *Thesaurus of ERIC Descriptors*. Ontology captures domain-specific knowledge including entities and relationships, both at a definitional level, and captures real-world facts or knowledge at an instance or assertion level (Cardoso and Sheth, 2006, pp. 13).

1.4.4 Classification

Classification is the act of organizing the universe of knowledge into some systematic order, e.g. in accord with some taxonomy (Chan, 1994). Classification makes formal, orderly access to information possible. It aims to bring related items together in a helpful sequence from the general to the specific (Taylor, 2004). Numerous classification schemes exist and usually define aspects, properties, or characteristics of specific subjects. The Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC) are the most popular classification schemes for libraries. On the web, directory services and clustering techniques are often used to provide classification of web resources. In the Semantic Web, ontologies have been proposed as a means to classify web resources conceptually.

1.4.5 Metadata

Definitions of metadata vary across research projects. Metadata is often defined as "data about data" as *meta* means "about". Burnett el al. (1999, p. 1212) defined metadata as "data that characterizes source data, describes their relationship, and supports the discovery and effective use of source data." Caplan (2003, p. 3) states "metadata is structured information about an information resource of any type or format". Other definitions emphasize the functionality of metadata. Greenberg (2003, p. 245) views metadata as "structured data about an object that supports functions associated with the designated object" with an object being "any entity, form, or mode for which contextual data can be recorded." According to the International Federation of Library Associations (IFLA), the term metadata "refers to any data used to aid the identification, description and location of networked electronic resources." Smiraglia (2005, p. 4) states "metadata are structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities." The United Kingdom Office for Library and Information Networking (UKOLN) states that "the term [metadata] is normally understood to mean structured data about digital (and non-digital) resources that can be used to help support a wide range of operations. These might include, for example, resource description and discovery, the management of information resources (including rights management) and their long-term preservation." The glossary by the Getty Research Institute (Baca, 1999, p. 37) defines metadata as "data associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation."

Concretely, metadata can be defined as a structured description of information resources. Metadata can be expressed in various formats, electronic or non-electronic, or describe certain resource types, electronic, network-accessible, or web-accessible, depending on the defined purposes of the metadata. The significant points from the various definitions of metadata include: (1) metadata is "structured" information, and (2) metadata "describes" information resources. Bibliographic metadata usually describes what, how, when, and by whom an information resource was created and collected. Metadata is described using a schema, a structured framework.

1.4.6 Classificatory Metadata

Since metadata in general can be any type of information that describes a resource, researchers have defined different types of metadata, such as descriptive, administrative, structural, syntactic, and semantic (Caplan, 2003; Cardoso and Sheth, 2006). In every case, metadata that describes the contents or context of resources is considered important for identifying the topics or domains of the resource regardless of how this type of metadata is named - whether descriptive metadata or semantic metadata. In this research, we focus on metadata that describes the context of a resource, i.e. the domain or topics of the content. The results will lead to resource classification by their topics in the collection. We name this type of metadata as *classificatory metadata*. *Classificatory metadata* refer to the types of metadata that can be used for classifying or grouping resources by topics or domains, including subject keywords, topic categories, domain names, etc.

2.0 Review of the Literature

This chapter reviews previous research on metadata generation from traditional methods and through social tagging systems. In addition, studies on using tags for retrieval and network formation are introduced along with discussion on using tags for resource classification.

2.1 Traditional Methods of Metadata Generation

Traditionally, library science has identified and located information resources by applying classification standards such as the Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC), or structured subject lists such as Library of Congress Subject Headings (LCSH) and thesauri. With the variety of resource formats and the need for interoperability in information resources demanded by the growth of WWW and digitalized information resources, a new way of identifying and locating information resources is needed. Work has been done on the use of markup languages (e.g. Machine-Readable Cataloging (MARC) and Standard Generalized Markup Language (SGML)), protocols (e.g. Z39.50), and bibliographic controls for electronic resources (e.g. International Standard Bibliographic Description (ISBD) and Anglo-American Cataloguing Rules (AACR)) to deal with the particular needs of web information resources. While many approaches for bibliographic control and cataloging on electronic resources were introduced, alternate approaches to describe Internet-based electronic resources were necessary. The concept of metadata has been suggested as a means to describe web resources.

Caplan (2003, p. 3-5) categorizes metadata as descriptive, administrative, or structural. *Descriptive metadata* facilitates discovery (how one finds a resource), identification (how a resource can be distinguished from other, similar resources), and selection (how to determine if a resource fills a particular need). It provides structured terms that enable access to resources through information retrieval systems ranging from indexes, to catalogs, to search engines (Smiraglia, 2005, p. 4). *Administrative metadata* aids in the management of resources and may include rights management metadata, preservation metadata, and technical metadata describing the physical characteristics of a resource. It can include information such as when and how an object was created, who is responsible for controlling access to or archiving the content, what control or processing activities have been performed in relation to the content and what restrictions on access or use apply. *Structural metadata* describes internal structure of complex information resources often used in machine processing, such as associating different representations of the same intellectual content. Cardoso and Sheth (2006, p. 9-12) defined types of metadata as syntactic metadata, structural metadata, and semantic metadata. *Syntactic metadata* in a simple form describes non-contextual information about content and provides very general information, such as the document's size, location, or date of creation. *Structural metadata* provides information regarding the structure of content, such as how items are put together or arranged. *Semantic metadata* describes contextually relevant or domain-specific information about content based on a domain specific metadata model or ontology, thereby capturing the meaning associated with the content. It adds relationships, rules, and constraints to syntactic and structural metadata.

Metadata is described based on a schema, a structured framework. Similar to traditional classification rules, metadata schemas are pre-established rules to organize resources in a collection, to organize entries in an index or catalog to facilitate access and retrieval, or to categorize resources into groups. Baca (1999, p. 39) defines schema as “a set of rules for encoding information that supports specific communities of users.” Formally in the library field, metadata is that information used when cataloging in accord with the AACR2/MARC standard. A wide variety of metadata schemas are being used experimentally on the WWW to describe information resources. Metadata schemas are developed based on the needs of particular fields or domains, the characteristics of resources in those fields or domains, and the types of metadata. For instance, the Government Information Locator Service (GILS) is a schema for government information; the Gateway to Educational Materials (GEM) and the Learning Object Metadata (LOM) are schemas for educational information; Categories for the Description of Works of Art (CDWA) is a schema for art information, and the Encoded Archival Description (EAD) is a schema for archival information. Many metadata schemas have been developed based on the markup languages, e.g. MARC, SGML, HTML, and XML. Other examples of metadata schemas include the Dublin Core (DC), a simple HTML-based data element set; the Encoded Archival Description (EAD), an SGML-based encoding scheme for archiving finding aids; and the Text-Encoding Initiative (TEI) Header, an SGML-based encoding scheme for complex texture structures.

2.2 Early Methods of Metadata Generation: Approaches for Automation

Along with the development of metadata schemas, research has explored how to reduce the effort to generate metadata by automating the process. There are two main approaches for automatic or semi-automatic metadata generation: extraction and harvesting (Greenberg, 2004). Extraction occurs with an algorithm automatically extracting information from the content of resources. Techniques such as information extraction (e.g. document analysis and ontology-driven extraction) and natural language processing (e.g. regular expressions, rule-based parsers, and machine learning) are often used for extraction. A number of research projects have attempted to generate metadata based on extraction. For instance, MetaExtract automatically assigns Dublin Core (DC) and Gateway to Educational Materials (GEM) metadata using natural language processing extraction techniques (Yilmazel et al., 2004). The goal of MetaExtract is to extract appropriate terms and phrases from the digital documents to populate item-level metadata. The Simple Indexing Interface is a framework for automatic metadata generation for learning objects (Cardinaels et al., 2005). Since the main resources are learning objects, the Simple Indexing Interface assigns metadata especially to the Learning Objects Metadata (LOM). GERHARD is another extraction approach that automatically classifies web resources (Möller et al., 1999). It focuses on classifying German web resources using Universal Decimal Classification (UDC). These studies demonstrate that extraction-based generation of resource descriptions can create metadata of the quality of manually generated metadata, at least when assigning the resources to specific schema such as DC, GEM, LOM and UDC. On the other hand, Han et al. (2003) extracted metadata using a machine learning method - Support Vector Machines (SVM). They classified research papers and extracted extended metadata for research papers based on the structural part of papers.

Harvesting collects metadata from existing meta-information in or associated with resources (Greenberg, 2004; Jenkins et al., 1999). For example, *meta* tags in HTML are important elements for harvesting. Many well-known HTML editor applications such as Front Page, Dreamweaver, and Microsoft Word automatically create *meta* tags with some basic bibliographic information when creating HTML files. There are tools for harvesting information for web resources such as DC-DOT. Paynter (2005) has described the factors in web resources that can be harvested in detail according to metadata element fields. For example, the potential value of the title element can be harvested from *meta* tag, *title* tag, *H1* tag, and then the sequence of words in the first 50 letters of body text. However, harvesting mainly concentrates on simple bibliographic information not considering other kinds of information, such as the semantics of the content.

Many applications combine both approaches. For example, the OCLC Scorpion project (http://www.oclc.org/research/software/scorpion/default.htm) explores the use of automatic classification with various methods for web accessible text documents (Shafer, 1997; Toth, 2002). It automatically assigns a subject using a machine-readable subject classification scheme or thesaurus by pre- and post-processing using harvesting and extraction techniques. The INFOMINE project (Paynter, 2005) is another example of using a combination of both approaches. INFOMINE exploits the fact that different metadata fields contain different types of data. It applies harvesting and extraction methods depending on the characteristics and the types of metadata fields. While these approaches are efforts to have machines understand the resource content with less human involvement, they fail to understand the semantic content since they only extract information from what is expressed in that content.

2.3 Social Tagging System Methods of Metadata Generation

The main drawback of having machines generate metadata automatically has been the quality of metadata generated. The advent of Web 2.0 technology let web users interact with systems to generate various types of information including simple metadata such as tags. Tagging systems allow users to tag or categorize different types of resources. Tags in tagging systems are generally one-word descriptions of the resource. Users benefit from tagging systems as they help users to better organize resources and find resources easily with assigned tags/keywords. Social tagging systems let users share their tags with other users. By socializing the tagging activity and tagged resources, it is not only possible to share users' resources, but also to share tags. Users can categorize or assign keywords to resources with similar content. Users can share the tags or systems can suggest tags from other users. These functions help users to use common terms within specific domains. It is also possible to form user groups with shared interests by sharing resources and collaboratively creating tags. Shared tags make it possible to use tags for better resource finding. Different types of social tagging systems are being developed, some for electronic resources (e.g. Delicious is a social bookmarking system for web pages; Flickr is a social tagging system for image sharing) and others for non-electronic resources (e.g. CiteULike is a bibliography sharing system that focuses on academic research papers; and LibraryThing is a tagging system for books and publications).

Social tagging systems first gained popularity by providing services related to digital resources, such as electronic documents, images, videos, etc. Delicious (http://delicious.com) is a social bookmarking system founded by Joshua Schachter in September 2003 and acquired by Yahoo! in 2005 ("Delicious," 2010). In September 2007, Delicious had more than 3 million registered users and 100 million unique URLs bookmarked (Arrington, 2007). It lets users store their favorite bookmarks online allowing them to be accessed from anywhere via the web. It also allows users to share their bookmarks with others and discover new web resources through the collections of others. On Delicious, users can organize bookmarks with tags. Tags in Delicious are not hierarchical in structure so that they can be more flexible and easier to manage for users. Delicious provides different methods to browse tags on a resource or a set of resources. Users can browse related tags assigned previously by other users. Delicious also provides a “tag cloud” that provides a visualization of the popular tags in the system. Flickr (http://www.flickr.com) is a photo sharing system where users can upload, organize, and share their photos with friends, family, and others. It was developed by Ludicorp and launched in February 2004 (“Flickr,” 2010). In March 2005, it was also acquired by Yahoo! and replaced the Yahoo! Photos service in 2007. Its main features include organizing images/videos with tags and building online communities based on personal or group interests. Tags in Flickr often represent the name or subject of the images including information such as location, date created, genre, name, medium, etc. Users can assign up to 75 tags to an image. Flickr has also implemented tag clouds, which provide access to images tagged with the most popular keywords. Flickr allows users to assign "sets" or groups of photos that fall under the same heading. Sets are more flexible than the traditional folder-based methods of organizing files - one photo can belong to one set, many sets, or none. The concept of “sets” is similar to categorical collection rather than hierarchical grouping.

With the popularity of social tagging systems for digital resources, some systems started to focus on non-digital resources as well. Reference information about publications can be shared online whether the resource is available digitally or not. By focusing on bibliographic information for publications (e.g. books, research papers, online publications, etc.), people can tag the targeted resource. CiteULike (http://www.citeulike.org) is one of the most popular social tagging systems for reference information for academic papers. It was developed by Richard Cameron in November 2004 in the UK[1]. According to Emamy and Cameron (2007), CiteULike is a fusion of web-based social bookmarking services (such as Delicious) and traditional bibliographic management tools (such as EndNote). It encourages researchers to “gather, collect, share” information on academic papers and “network” with others who have similar research interests. Users of CiteULike are motivated by an easier method of collecting information, especially for selected publishers[2], and better methods of discovering and sharing information. For example it lets users find related articles by author name and tags from user profiles. Since CiteULike focuses on academic areas, users can benefit by specifying semantic meaning of their tags and forming communities with other researchers with similar interests. Groups are formed by inviting friends/colleagues to join, forming research groups, and letting users create or join groups of interest. In addition, researchers can easily generate their literature library by importing or exporting a personal library in BibTex or RIS file format. Discovering new articles is also possible with CiteGeist, which lists recent popular articles posted to CiteULike. Other social tagging systems for non-digital resources include LibraryThing (http://www.librarything.com/) and Listal (http://www.listal.com/). LibraryThing was developed

[1] CiteULike FAQ, http://www.citeulike.org/faq/all.adp

[2] CiteULike supports automatic extraction of bibliographic information from major publisher sites: ACL Anthology, AIP Scitation, Amazon, American Chem. Soc. Publications, American Geophysical Union, American Meteorological Society Journals, Annual Reviews, Anthrosource, arXiv.org e-Print archive, Association for Computing Machinery (ACM) portal, BioMed Central, Blackwell Synergy, BMJ, Cambridge University Press, CiteSeer, Cryptology ePrint Archive, DBLP, EdITLib, Education Resources Information Center, HighWire, IEEE Explore, informaworld, Ingenta, IngentaConnect, IoP Electronic Journals, IUCr, IWA Publishing Online, Journal of Machine Learning Research, JSTOR, Mary Ann Liebert, MathSciNet, MetaPress, NASA Astrophysics Data System, National Bureau of Economic Research, Nature, Open Repository, Optical Society of America, Physical Review Online Archive, plos, PLoS Biology, Project MUSE, PsyCONTENT, PubMed, PubMed Central, Royal Society, Royal Society of Chemistry, Science, ScienceDirect, Scopus, Social Science Research Network, SpringerLink, Usenix, Wiley InterScience. For the publishers that CiteULike does not support, users need to input bibliographic information manually.

by Tim Spalding and went live on August 29, 2005 (“LibraryThing,” 2010). It is a service to help users easily catalog their own books with high quality information (Wenzler, 2007). Users can store and share personal library catalogs, book lists, and reviews of books, and also manage and search their library with tags. Up to 200 books per user can be entered without any fee. Book catalogs can be easily created with any qualified sources such as Amazon.com and the Library of Congress. Users can form group forums or book clubs online. Recommendations are available for related topics and genre based on tag information and/or user recommendations. Listal is similar to LibraryThing except it includes not only books but other media types such as movies, TV shows, DVDs, music, and games. Listal lets users input tags and review and share them with other users and group members. LibraryThing provides more detailed and qualified metadata and Common Knowledge provides general information on the book (e.g. name of characters, awards, etc.) in addition to user tags. On the other hand, Listal manages additional information such as people (e.g. actors, artists, authors, and directors) and platforms (e.g. Nintendo, card game, etc.).

2.3.1 Research on the Use of Social Tags

Researchers are beginning to look at ways that social tags might be used. In general, social tagging systems are based on a collection of 3-tuples consisting of users, tags, and resources (Hotho et al., 2006a, 2006b; John and Seligmann, 2006; Marlow et al., 2006; Mika, 2007; Smith, 2008, pp. 41-53; X. Wu et al., 2006) (Figure 1).

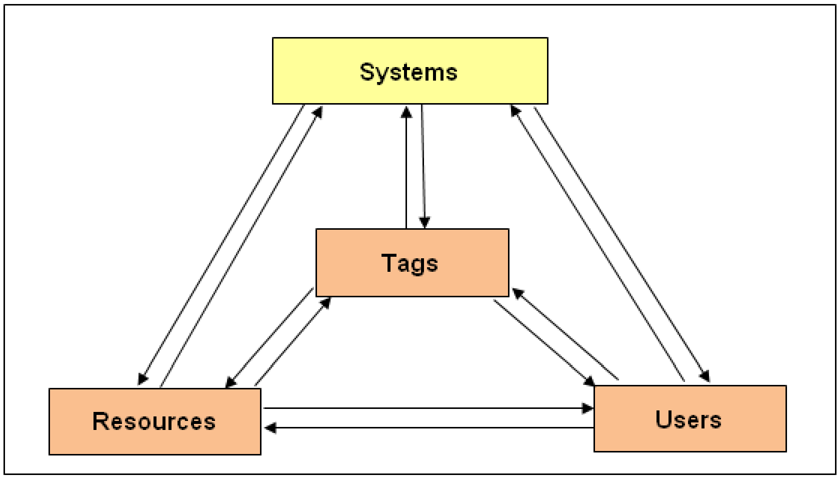


Figure 1. The Triple Model of Tags

One stream of research relates to improving information retrieval. There are many possible uses of social tags to improve search results. First of all, one may consider tags as one type of index for documents. Although using tags as an index does not fully solve linguistic problems of full text indexing, tags are expected to provide more precise semantic information with shared agreement and can be used to index or rank web resources (Choochaiwattana, 2008; Choochaiwattana and Spring, 2009; Golder and Huberman, 2006; Mika, 2007; Shirky, 2005; Trant, 2006). Second, tags can be used to build ontologies as a part of the Semantic Web (Mika; 2007, Ohmukai et al., 2005; H. Wu et al., 2006). Since tags provide semantic information about web resources, it may be possible to extend and organize tags into ontologies. In the information retrieval and Semantic Web domains, from the 3-tuples, tag-resource elements are more focused (Figure 2). Tags can also be used to form community networks. This kind of research emphasizes the social aspect of tagging systems (Marlow et al., 2006; Mika, 2007; Ohmukai et al., 2005; H. Wu et al., 2006). The ease of tag input in many social tagging systems encourages web users to participate in the tag creation process. Since tags are assigned to a resource by different users collaboratively, the triple association and networks can be formed from the linkage of elements of the triple model. While the tag-resource sets are more critical in retrieval research, the user element of the triple becomes very significant in network research (Figure 3).

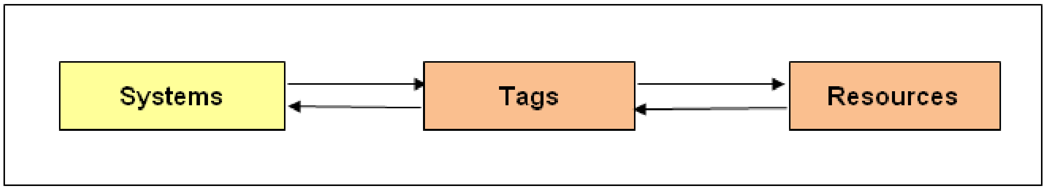


Figure 2. Information Retrieval with Tags

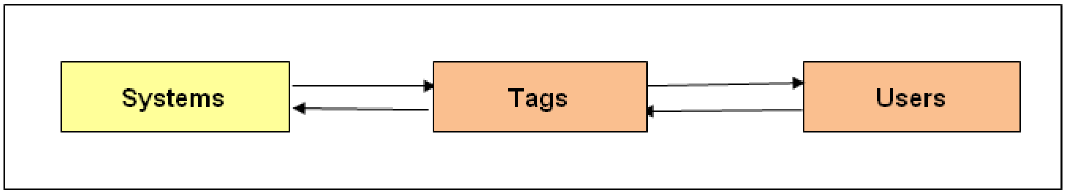


Figure 3. Community Network Formation with Tags

2.3.1.1 Tags for Indexing and Ranking

Traditionally web information retrieval focuses on building an index from the contents of web resources. This is usually done by employing full-text indexing using term-weighting. However, full-text indexing causes problems related to synonymy, polysemy and other content semantics. Although using tags as an index does not fully solve these linguistic problems, tags are expected to provide more precise semantic information with shared agreement (Golder and Huberman, 2006; Mika, 2007; Shirky, 2005; Trant, 2006). In addition to using tags for indexing, other research is using tags to improve ranking algorithms. Considering tags as an index, Yanbe et al. (2007) made use of the popularity of a web page, i.e. the total number of times a web page is tagged and measured what is called *SBRank*. They tested Delicious data to compare *PageRank* and *SBRank* and suggest that *SBRank* captures the popularity of resources among content consumers (readers) while *PageRank* is in general a result of author-to-author evaluation of Web resources. Therefore, *SBRank* is often more dynamic and quickly applied. They implemented a system that can accept different types of queries to benefit from information provided by the document index and tags. Query types include context query, metadata query, temporal query (e.g. FirstDate), sentiment query (e.g. useful), and controversial query (e.g. number of comments on pages). By providing a method to filter different types of queries using tags and enhance searches with combinations of the ranking method based on link structure analysis and social bookmarking, it was found that it is possible to provide more precise relevance estimates of documents, improve the measure of page quality, provide time-aware popularity measure, and filter pages by user impressions, sentiment characteristics, or controversy levels using user-assigned tags.

Hotho et al. (2006a, 2006b) adapted the notion of HITS for their ranking algorithm, *Adapted PageRank*. The basic idea is that the resource that is tagged with important tags by important users becomes important itself. This rule was applied to resources, tags, and users equally and used to measure similarity. Their results showed that although tags, users, and resources that are related to preference are ranked higher in the result, many of the general results still hold the top position. Therefore, in order to reasonably focus the ranking around the topics defined in the preference vector, they presented a ranking algorithm called *FolkRank*, a topic-specific ranking in a folksonomy. *FolkRank* provides ranks based on topic-specificity to user preferences. Topic can be assigned not only by assigning higher weights to specific tags, but also to specific resources and users. Therefore, *FolkRank* is a more personalized rank algorithm than *Adapted PageRank*. *FolkRank* works better for ranking within a folksonomy, since words used often globally disappear from the ranking. This study has presented the possibility of using tags for personalized ranking and recommendation.

Similar to Hotho et al. (2006a, 2006b)'s observation, X. Wu et al. (2006) explained that the semantic relatedness is embodied in different frequencies of co-occurrences among users, resources, and tags. Based on this observation, they generated a semantic index with social tags and improved search, inferring the semantic index and various retrieval models statistically. Their semantic search models include a basic search model, resource discovery model, and personalized search model. The basic search model deals with queries that are a single tag and ranks semantically related resources without considering personal user information (resource-tag). The resource discovery model can extend the basic search model by discovering semantically related resources — using tag co-occurrence to find resources tagged with related tags. The personalized search model integrates personalized information in the semantic search using users' interests represented by semantic vectors from tags. The authors considered that the significance of their search models is detecting resources that are not tagged by the query tags. In addition, they stressed that the global semantic model helps disambiguate tags and group synonymous tags together into concepts.

Bao et al. (2007) introduced two ranking methods based on the observation that social tags can benefit web search as they can better summarize web pages, and the count of tags indicates the popularity of web pages. For example, even if the page contains the tags "ubuntu" and "linux", it is not proper to calculate the similarity between the query and the document using the keyword "linux" only. They argue that an exploration of similarity between "ubuntu" and "linux" may further improve the page ranking. In fact, similar tags are usually assigned to similar web resources by users with shared interests. With this observation, they introduced *SocialSimRank (SSR)*. SSR basically uses co-occurrence of tags and semantically related resources to improve traditional full-text indexing. They also introduced *SocialPageRank (SPR)* based on the observation that popular web resources are tagged by many up-to-date users and annotated with up-to-date tags. Their preliminary experimental results show that SSR can find the semantic association between queries and tags, while SPR measures the quality of a web page from the web user's perspective.

John and Seligmann (2006) performed a similar study, proposing a ranking called *ExpertRank* that quantifies a user's expertise level in the context of a tag. The authors emphasized that by categorizing and relating content using tags, it is possible to express users' interests and thus their expertise. In *ExpertRank*, relevant factors to consider are the number of bookmarks tagged with a particular tag by a user and the age of the bookmarks. The authors calculated the rank of an expert two ways for each tag based on the number of bookmarks that the expert marked with that tag – first, assuming an unstructured tag collection (i.e., no dependencies between tags), then assuming a structured tag collection (i.e., correlations exist between tags). This study showed that by using tagging activity information, it is possible to provide better ranking of resources, especially within a community of similar interests (such as an enterprise). It is also possible to adopt the algorithm to recommendation systems such as E-Bay, as it is important to define experts in such systems.

Studies by Choochaiwattana (2008) and Choochaiwattana and Spring (2009) demonstrated two methods to integrate social tags into web search to improve users' satisfaction with search results - web resource index augmentation and search result ranking. Their study showed that the count of the number of people that used tags that matched terms in the query string normalized by the total count of all tags for a given resource ranked useful web resources higher and less useful resources lower. They also argued that social tags can provide high-level concepts about web resources and the combination of social tags and content of web resources can provide a better representation of web resources.

2.3.1.2 Tags for Folksonomy Development

Mika (2007), Ohmukai et al. (2005), and H. Wu et al. (2006) have all conducted research aimed at using tags to build ontologies or folksonomies. H. Wu et al. (2006) used tags to build a common hierarchy for a set of documents. They suggested an ontology generation algorithm using agglomerative hierarchical clustering. An ontology from tags on a large document collection allows both systematic retrieval of documents and social interactions with common reference. To mitigate the impact of polysemy, synonymy, and idiosyncratic tagging, it is necessary to have a large number of users as participants.

Unlike H. Wu et al.'s study, Mika (2007) introduced ontology generation with network analysis. Based on the triple model of a folksonomy, Mika extended the model into three ontologies based on the graph models and co-occurrence of tags. The study suggested the importance of actor (user) and concept (tag) linkage in folksonomy for ontology generation.

Ohmukai et al. (2005) proposed a social bookmarking system using several metadata and a personal network to construct a community-based ontology. Different from the work described above, Ohmukai et al. (2005) used neither clustering nor network analysis to generate an ontology. They used the community-based information with techniques such as FOAF TrackBack, matchmaker-based recommendation, and the network expansion method to build a personal ontology framework. While the results were not provided in this particular paper, the idea of making use of a folksonomy in a community-wide approach is a good example of different approaches for ontology generation.

2.3.1.3 Tags for Network Formation

The ease of tag input in many social tagging systems encourages web users to participate in the tag creation process. Since tags are assigned to a resource by different users collaboratively, the triple association, mentioned above, is defined and networks can be formed from the linkage of elements of the triple model. Research related to social tagging from different domains took users and communities into consideration and tried to benefit from this analysis. The early research in this area analyzes tags in the context of sets in communities.

Marlow et al. (2006) analyzed Flickr focusing on the impact of contact in networks. Their result showed users sharing contacts tend to have more overlap in common tags compared with overlap between random users, indicating that there is a relationship between social affiliation and tag vocabulary formation and use. Mika (2007) also evaluated the role of users in network and ontology creation. The study suggests that the actor (user) – concept (tag) association network better represented the user's or community's interests. Therefore, in ontology building, not only concepts but also users should be considered.

Some studies introduce possible implementations of communities generated based on tags. Ohmukai et al. (2005) generated a community-based ontology to solve problems with folksonomy and improve recommendations for users. They showed how an ontology can be structured using a personal network of friends and content metadata. John and Seligmann (2006) suggested using topic-based sub-communities within the social network to determine expert users and related tags for *ExpertRank*. They did not introduce the process of community network formation in detail; however, they showed how the communities can be used in ranking. H. Wu et al. (2006) used a method to identify global communities utilizing authorships and usage of tags and documents to implement their modified HITS algorithm for ontology generation. To identify communities of similar interest and information experts in a domain, linkage between tags and other knowledge sources such as contents, hyperlinks, and user behavior is considered. These studies suggest possibilities of implementing various methods for forming user or tag communities for different purposes and the need for more sophisticated structures of tag data for better tag usage.

2.3.2 Web Resource Classification Using Social Tags

Classification of web resources has evolved as one method to improve web information retrieval along with full-text indexing. Up to now, controlled vocabulary and natural language processing are the most widely used methods for web resource categorization. A controlled vocabulary can address the shortcomings of full-text indexing. However, it cannot be deployed in a scalable fashion due to a lack of qualified professionals and the sheer number of resources that need to be classified. Natural language processing, such as clustering, helps categorization done by a machine. This automates the process of controlled vocabulary generation but introduces other problems related to semantics. From the Semantic Web point-of-view, tags can play a role as a type of annotation providing semantic information about the web resources for categorization.

There is growing interest in determining if tags can be used as a type of metadata useful in web resource classification (Bischoff et al., 2008; Heymann et al., 2008; Macgregor and McCulloch, 2006; Mathes, 2004; Mika, 2007; Noll and Meinel; 2007, Quintarelli, 2005; Sen et al., 2007; Shirky, 2005; Smith, 2008, pp. 63-93; Syn and Spring, 2009; Zubiaga et al., 2009). Macgregor and McCulloch (2006) argued that social tagging systems let users participate in the organization of web resources and make it possible to lower the cost of web resource metadata creation. Sen et al. (2007) and Heymann et al. (2008) indicated social tagging systems allow users to contribute metadata for new or active pages. Heymann et al. (2008) compared the Open Directory Project and Delicious and found that metadata generation by humans takes more time and, as a result, new resources do not appear immediately; while they appear very quickly in social tagging systems. Noll and Meinel (2007) have examined tags by comparing them with web document metadata, i.e. HTML metadata tags, to define characteristics of tags in terms of metadata and web document classification. They found that tags match document content significantly better than its metadata created by the author.

Quintarelli (2005) introduced folksonomy as one type of user-generated classification that emerges through bottom-up consensus. In using tags, involvement by the public is considered important, although some trade-off between quality of metadata and metadata ecology is necessary. Since users enter tags without any restriction, terms used for tags may contain misspelled terms, compound terms, single and plural forms, personal tags, and single-use tags. Although tags may be used that have a meaning known only to their creator, there are clearly some tags that have shared social meaning (Guy and Tonkin, 2006). Shirky (2005) discusses how tags should be organized to produce meaning. Tags can be applied as raw keywords that represent the user’s resource description. Rethlefsen (2007) proposes structuring tags when representing them to users to let them benefit from it effectively. Related to concerns about tag quality when used as metadata, the results from the steve.museum study (Trant, 2006) showed that the terms provided by non-specialists for museum collections are positive. It demonstrated that using tags assigned by general users might help bridge the semantic gap between professional discourse and the popular language of the museum visitors. Zubiaga et al. (2009) suggested that user-generated annotation (tags and comments) are actually more useful in classifying web pages than using only the content of a web document. Their study showed that tags perform better for classification than content only and they perform even better when tags are used together with the content of the document. Bischoff et al. (2008) also confirmed that tags, at least in music, are reliable and as good as expert created metadata. Although tags for music resources are more structured and controlled compared to tags for other resources, Bischoff et al. still provided a possibility of using tags as metadata. Syn and Spring (2009) found that for academic papers user-generated tags worked as well as author-generated keywords and suggested filtering tag noise could improve the usefulness of tags. The results also supported using tags and folksonomies as metadata. In addition to the potential of tags as descriptive metadata, Guy and Tonkin (2006) discuss how to improve tag quality and how to educate tag creators to make use of folksonomy metadata. They suggested that providing users with helpful heuristics and introducing structure within tags might encourage users to select and create good tags.

3.0 Preliminary Studies of Social Tags as Classificatory Metadata

3.1 Introduction

This study is designed to shed light on the use of tags as classificatory metadata. To accomplish this goal, it is important that tag noise be reduced. As Guy and Tonkin (2006) found there are both noisy tags and useful tags in a tag set. Tag noise includes misspelled tags, bad combinations of words, personal tags, etc. It is expected that having tag noise filtered out will improve the quality of tags. Furthermore, ambiguous tags have to be disambiguated. There are many kinds of ambiguous tags -- personal tags, compound tags, etc. In this research, we focus on compound tags as ambiguous tags and figure out how they can be disambiguated. Once we have accomplished noise reduction and tag disambiguation, we can look for tags that will serve as classificatory metadata. This chapter begins by taking a closer look at social tagging systems. In addition, it chronicles some of the preliminary work done to set the stage for the dissertation research.

3.2 Social Tagging Systems

The goal of this study is to find good methods to generate classificatory metadata for web resources. Unlike many related studies, the goal is not to retrieve or rank resources using social tags. Instead, the goal is to select important tags (meta-terms) and remove meaningless ones (noise) from the tag set. Several preliminary observations were made to find a method to determine the better tags to use to represent a resource.

1. Social tagging systems allow users to input a term at a time. Therefore tags with multiple terms are often input as multiple single terms (e.g. “semantic” and “web”) or a compound term (e.g. “semanticweb”, “semantic-web”, “semantic_web”).
2. A user can create only one tag set for a resource.
3. A user can assign a term as a tag only once in a tag set for a resource. That is, a tag cannot be assigned multiple times by a user for a resource nor can a user explicitly weight the importance of a tag.
4. Social tags include terms that are idiosyncratic to a user. Examples include graphical tags (e.g. “*****”), personal notes (e.g. initials, “IS2000”) and compound words (e.g. “toread”). These do not provide good metadata information for the resource.

The model of social tagging systems can be described by the tuples: *users, tags, resource*s (users add tags to resources). Users are the people who create tags, add resources, and use the systems to find or organize meaningful resources. Resources are the items added into the system including documents, web pages, videos, images, etc. Resources may be associated with tags. Tags are labels users add to resources. They can be descriptions of the resources, opinions on the resources, self-referencing notes on the resources, etc. With the elements of the tuple model, we define a *tag set* as a set of tags created by a user associated to a resource. A *bookmark*, particularly in this research, represents a tag set assigned to a resource (web document) by a user.

Based on the observations made on the social tagging system and the conceptual model, we can identify characteristics of the tag data as described below.

1. The number of *users* who added a resource equals the number of *tag sets* associated with a resource or the number of *bookmarks* on a resource. It is noted that a *tag set* or *bookmark* may not contain any tags (an empty tag set).
2. The number of times a *tag* is used to describe a *resource* equals the number of *users* who used the tag term for the resource, due to the condition of social tagging systems that allows a user to add a term only once for a resource.
3. The number of times a *tag* is used in a collection is a matter of definition. It may be defined in terms of the number of times it is used by different *users* related to each resource or by the number of *resources* with which it is associated at least once.
4. The number of *resources* in the collection represents the number of unique URLs added to the system. It is noted that a web page can be represented with various forms of URLs such as http://www.pitt.edu and http://www.pitt.edu/index.html.

3.2.1 Tag Occurrence and Distribution

Social tags are a set of tags from a group of users. Social tags provide a set of positive descriptive terms identified by a group of people. Because there are no controls for inputting words as tags, there can be problems of tag quality and agreement on selection of terms as resource descriptions. Generally, researchers accept that the occurrence of any given term represents the agreement of people – the more a term occurs, the more people believe it to be a good term.

Quintarelli (2005), based on Thomas Vander Wal’s explanation, described two aspects of folksonomy – broad and narrow. A broad folksonomy, as a result of mass agreement, shows a power law curve and a long tail effect in the distribution of tags. The power law reveals that many people agree on using a few popular tags and smaller groups prefer less known terms to describe their items of interests (i.e., narrow folksonomy). A narrow folksonomy provides benefits in finding objects that are not easy using traditional tools, e.g. full-text search, as it is often described using an individuals’ own terminology.

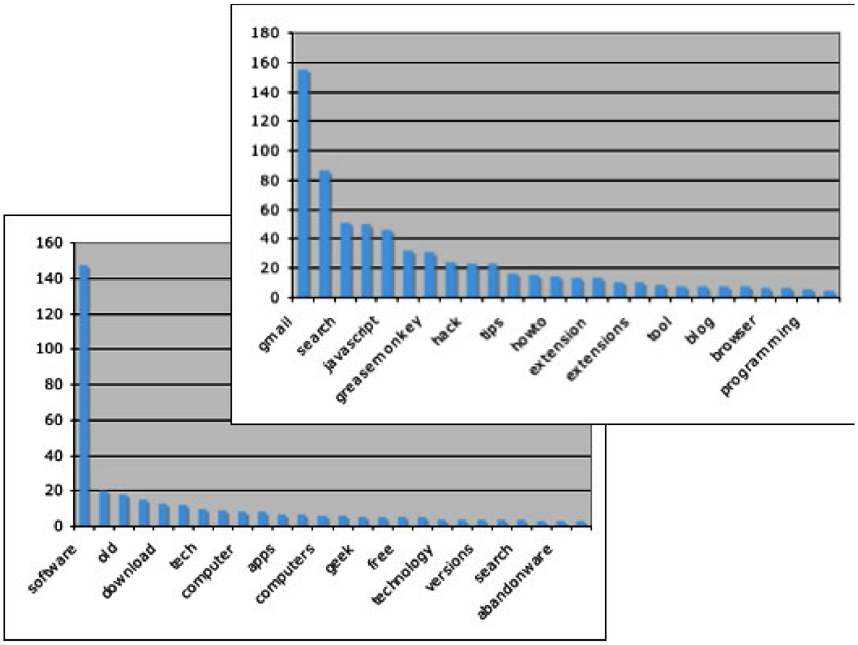


Figure 4. Different Tag Distribution for Different URLs (Shirky, 2005)

Related to Quintarelli’s observation on tag occurrence, Shirky (2005) has discussed that the frequency of tags for a URL can help determine the importance of a set of tags for each URL. The distribution and occurrence can identify the most representative set of tags for the resource. Shirky further discusses that the distribution also can cause confusion in analyzing tags of a resource. From Figure 4, the graph at the bottom left has more than 140 people tagging this URL as “software”. The next most common tag, “windows”, has only 20 occurrences. It is obvious that this resource is about software -- there is a sharp, clear fall off in tags. However, it is more difficult to determine the cutoff point for the graph at the upper right.

The observation and discussion on tag occurrence and distribution shows that finding good classificatory metadata terms from tag sets cannot be done by simply getting frequently occurring tags. In this research, not only broad folksonomy, but also tags in narrow folksonomy are considered as candidates of classificatory metadata, as broad and specific domains and topics are both important for classification. In addition, defining the cut off points from the tag distribution of a resource is an issue to consider in finding significant classificatory metadata.

3.2.2 Compound Tags

Given that most tagging systems do not allow spaces in tags, users have developed ways of specifying compound tags, combinations of words without spaces, e.g. “webdesign”, “web_design”, “WebDesign”. A number of authors have discussed compound tags as one of major characteristics of social tags (Guy and Tonkin, 2006; Lin et at., 2006; Tonkin, 2006). Guy and Tonkin (2006) indicated that the majority of tags are nouns. They observed that many tags are compound words (according to Tonkin’s sample about 16~23.5% of tags are compound words, Table 1). Tonkin (2006) analyzed tag types shown in Table 1, which indicates that compound tags are a major format for social tags. She defined types of tags as ‘words’, ‘simple compounds’, ‘known encodings’ and ‘unknown’. ‘Words’ are tags that use single terms. ‘Simple compounds’ indicate compound tags that are simple combinations of two terms or make use of strategies for indicating the word boundary such as using a separator character, e.g. hyphen, underscore, or period. ‘Known encodings’ indicate tags that imply an existing formal metadata model. ‘Unknown’ includes tags that cannot be defined in any of other forms.

Table 1. Tag Types Distribution (Tonkin, 2006)

Tag Type / %	Words	Simple Compounds	Known Encodings	Unknown
Flickr	33.8	16	9.7	40.5
Delicious	43.9	23.5	4.3	28.3

In Table 1, ‘simple compounds’ indicates compound tags with separators. In our sample data set including 1,800,651 bookmarks with 205,486 unique tags by 488,939 unique users for 7,411 resources that is much larger than Tonkin’s data set, there were 143,775 unique compound tags, that is, 69.97% of the unique tags. The average words used to form a compound tag are 2.71 words. Tonkin (2006) further analyzed the simple compounds and reported common compound separators in case of Delicious sample data: dash (39%), underscore (25%), forward slash (14%), period (14%), others (8%). Guy and Tonkin (2006) showed that there are many types of separators (Figure 5). In our sample data set, we found yet more separators. Like Tonkin’s analysis, we found that popular special characters as separators were underscore (30%), dash (19%), double quotation (15%), period (9%), comma (8%), etc (Figure 6).

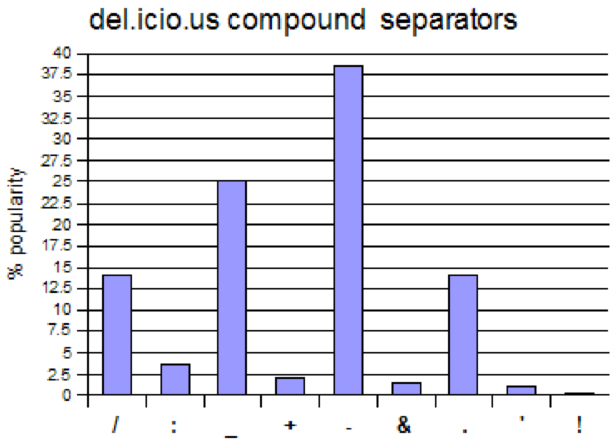


Figure 5. Delicious Compound Word Separators (Guy and Tonkin, 2006)

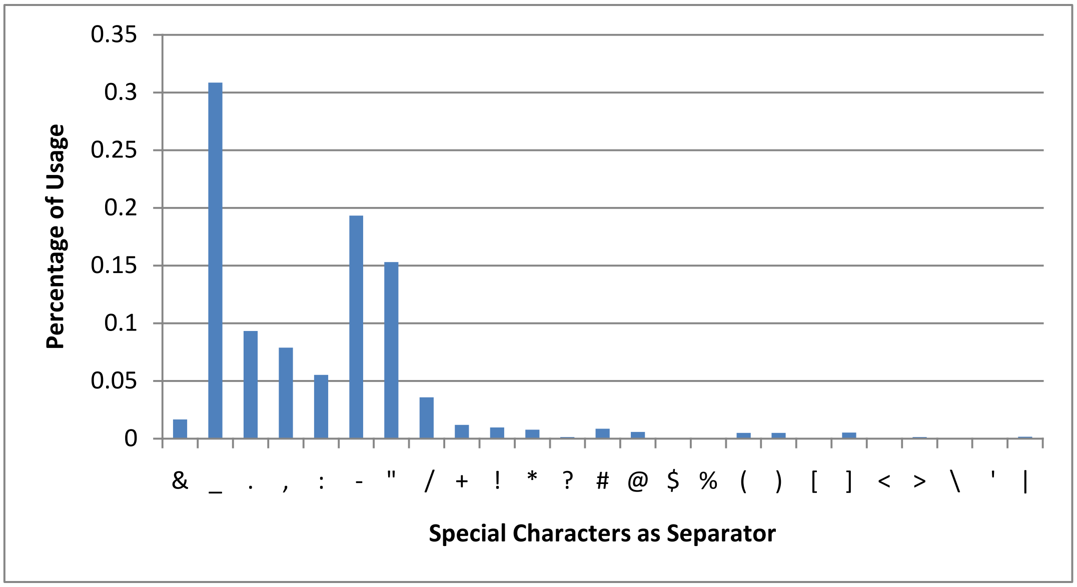


Figure 6. The Usage of Separators in Delicious Compound Tags

Observations made by Guy and Tonkin (2006), Tonkin (2006), and also by our sample data set show that although compound tags are not as structured as a formal form of metadata, it would be dangerous to make conclusions about tag information without exploiting compound tag data. Thus, in this research, we include compound tags by decomposing the words forming compound tags. Given that compound tags were formed intentionally by users, the words put together may be related in particular ways, e.g. they may be subordinate-superordinate terms or have some other relationship. As examples, “web development” specifies a kind of development and “Semantic Web” indicates a specific meaning when the words are used together that may be possible to identify the relationship among words. Compound tags could be processed algorithmically with some degree of confidence. In this research, we consider tags as a source to provide classificatory information. As one of the major formats of tags, compound tags are analyzed to be re-formed after the words in the compound tag are extracted in the way that can represent possible categories of topics.

3.3 Finding Good Terms to Use as Classifiers

3.3.1 Reflection on TF-IDF

Social tagging systems generate relationships between resources, tags and users. The 3-tuples can provide the following kinds of information. (1) The number of times a tag is associated with a resource, i.e. the frequency of a tag. The highest frequency of a tag for a resource cannot exceed the number of users who bookmarked the resource. (2) The number of resources with a tag, i.e. the portion of documents in the collection that has a tag. (3) The number of users bookmarking a resource equals the number of tag sets for a resource. (4) The number of users using a tag is the portion of users who use a term as a tag from the whole user group. These can be used in various ways to find high quality tags to use as metadata for a resource. Our goal is to find tags that will be representative of content. We will need a metric to separate good and bad tags. In beginning the exploration, we thought about term weighting in information retrieval: Could a metric similar to TF-IDF be developed to find representative terms in the tag set?

In information retrieval, a bag of weighted words from the document is often used to rank more relevant documents for a search query. TF-IDF (term frequency-inverted document frequency) weighting is a standard method to weight terms (Salton and Buckley, 1988). It provides a measure of the “importance” of a word in a document. Term frequency (TF) is a measure of the importance according to the number of times a word appears in a single document.

$$TF = \frac{Count(T_i, D_j)}{Count(T, D_j)}$$

It represents the ratio of a certain term (T_i) in a document (D_j) over the total number of terms in the document D_j (T). However, TF alone cannot ensure a word will be good for ranking, especially when high frequency terms are not concentrated in the contents of a particular topic and represent general concepts. Inverted document frequency (IDF) is used to find terms that indicate relevant resources.

$$IDF = \log(\frac{Count(D)}{Count(T_i, D)})$$

It represents the log of the count of documents containing a certain term (T_i) divided into the count of the total document set. IDF decreases in importance (weight) when the word occurs frequently in the collection and increases in weight when it appears rarely. Therefore, IDF gives more weight to the terms that are specific to a given document and gives less weight to the terms that are general. TF-IDF has been well-tested in the information retrieval domain.

3.3.2 New Measures for Classificatory Metadata

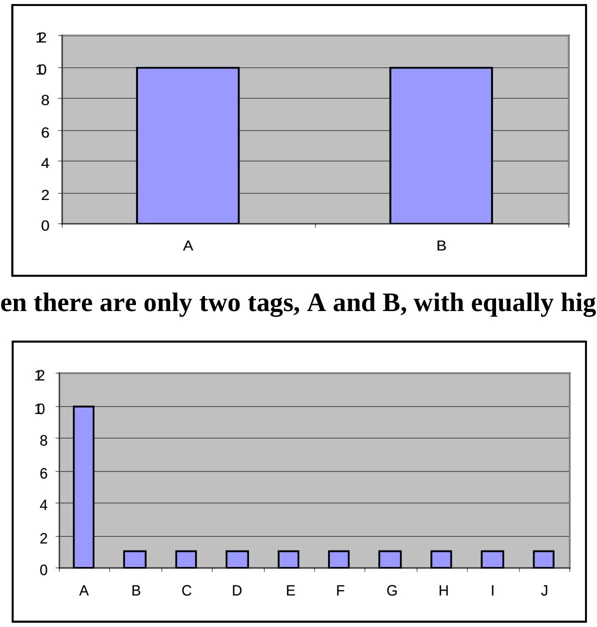
To deal with social tags that contain words that are personally created, don't have generally accepted meanings, and do not appear in the content of the resource, new measurements are needed. We introduce two measures, based loosely on TF-IDF, *Annotation Dominance (AD)* and *Cross Resources Annotation Discrimination (CRAD)*. We believe they provide measures to discriminate among tags, especially for classification purposes.

3.3.2.1 Annotation Dominance (AD)

Annotation dominance (AD) is suggested as a way to measure the importance of an annotation. Basically, *AD* is a way of measuring how often the tag is used related to a resource. Considering that a tag can be associated with a resource by a user only a single time, *AD* provides the importance of a tag in a document. AD can be formalized as Equation (1) where A_i is a certain tag and R_j is a resource.

$$\frac{Count(A_i,\ R_j)}{Count(A, R_j)} \qquad \text{--- } \textit{Equation (1)}$$

Given the observation on 3-tuples relationships, *Annotation Dominance* should reflect the difference in importance of tags when distribution of tags on a resource is different. However, Equation (1) does not reflect the difference in importance by the distribution of tags. For example, Figure 7 shows different cases of tag distribution for a document. The first case has two tags (Tag A and B) with equal frequency and the second case has ten tags with one dominant tag (Tag A) and nine other very low frequent tags (Tag B to J). In the first case, both tags are equally important, whereas, in the second case, only tag A is important. Obviously, many other situations are possible making the development of a simple yet comprehensive heuristic difficult.



Case 1. When there are only two tags, A and B, with equally high frequency

Case 2. When there are 10 tags, A to J, with dominant high frequency for A and very low for others

Figure 7. Cases of Tag Distribution

To make *Annotation Dominance* reflect this issue conceptually, we included tag set as a factor. The number of tag sets will reflect how many users have adopted a certain tag. This includes the three main factors of the tag information from the 3-tuple relationship. The *Annotation Dominance* is formulized by modifying Equation (1) as follows, where R_i represents a given resource, U is any user, and T_{Ai} is a tag set that contains tag A_i, and A_i is a given tag.

$$AD = \frac{Count(T_{A_i}, R_j)}{Count(U, R_j)} \qquad \text{--- } \textit{Equation (2)}$$

Thus, the *Annotation Dominance (AD)* is a measure of how much a tag is agreed by users to represent a given resource. In the extreme case, if every user who bookmarked the resource R_j assigned a given tag term A_i, the *AD* of A_i becomes 1. On the other hand, if nobody selected A_i as a tag for the recourse R_j, then the *AD* of A_i will be 0. Given that we are dealing with tags that appear in the tag set, the range of *AD* will be greater than 0 and less than or equal to 1.

Below we introduce examples comparing Equation (1) and *AD* (Equation (2)) for different cases of tagging patterns to show that cases of various tag distribution is considered in finding representative tags. Table 2 introduces 5 extreme cases to compare. All five cases represent 5 resources that have 1000 tag sets, i.e. 1000 users. Case 1 is when all 1000 users only include tag A. Case 2 is when tag A as a tag set with one tag dominant with 500 other tag sets, e.g. {B}, {C}, {D}, etc. Case 3 is when 4 tags (each as a tag set such as {A}, {B}, {C}, and {D}) were assigned with equal frequency of 250. Case 4 is when four tag sets contain more than 1 tag with equal frequency of 250. Note that in case 4, only tag A is included in every tag set. Case 5 is when 4 tags (A, B, C, and D) were included in 1000 tag sets.

Table 2. Five Example Cases of Tag Distribution

Case 1	Case 2	Case 3	Case 4	Case 5
1000 tag sets 1000 users	1000 tag sets 1000 users	1000 tag sets 1000 users	1000 tag sets 1000 users	1000 tag sets 1000 users

1000 × {A}	500 × {A}
	1 × {B}
	1 × {C}
	: :
		250 × {A}
		250 × {B}
		250 × {C}
		250 × {D}
			250 × {A, B}
			250 × {A, C}
			250 × {A, D}
			250 × {A, E}
				1000 × {A, B, C, D}

Table 3. Comparison on *Annotation Dominance* (Equation (1) and *AD*)

Case	Equation (1)					AD (Equation (2))				
	A	B	C	D	E	A	B	C	D	E
1	1.0	-	-	-	-	1.0	-	-	-	-
2	.5	.001	.001	.001	.001	.5	.001	.001	.001	.001
3	.25	.25	.25	.25	.25	.25	.25	.25	.25	.25
4	**.5**	**.12**	**.12**	**.12**	**.12**	**1.0**	**.25**	**.25**	**.25**	**.25**
5	**.25**	**.25**	**.25**	**.25**	**.25**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**

Table 3 represents the result of Equation (1) and *AD* (Equation (2)) applied to each case. It clearly shows that *AD* applies different weights on tags depending on their distribution among users (tag sets) whereas the result of Equation (1), focusing on the frequency of tags on a resource, does not reflect the significance of tags as effectively. Cases 1 to 3 result in the same weight values since only one tag was included in every tag set, which is not a case in real tag sets. In these cases, only frequency matters to identify the importance of a tag. Cases 4 and 5 highlight the differences in the two equations. These cases reflect real tag sets better. Tag A is the important tag in case 4 whereas all four tags (A, B, C, D) should be important tags in case 5. It seems both Equation (1) and *AD* reflects the expected result; however, when case 4 and case 5 are compared, it is obvious that *AD* is a better method to measure the significance of a tag reflecting the agreement among users. Equation (1) reflected the distribution of tags within a resource; however, it failed to weight tag A in case 4 and case 5 equally. In addition, in case 5, since all 5 tags appear 1000 times, i.e. all 5 tags are included by all 1000 users and that is the total number of tag sets in case 5, *AD* seems to provide a better measurement of reflecting the importance of each tag.

3.3.2.2 Cross Resources Annotation Discrimination (CRAD)

In addition to *Annotation Dominance (AD)*, *Cross Resources Annotation Discrimination (CRAD)* is considered as a means to offset the weight of general tags since general terms are used widely as tags. Conceptually similar to IDF, it is designed to remove tags that are used broadly in the document corpus. If a tag is assigned for every document in the collection, we consider it to be a weak candidate as a tag for document classification. Related to IDF concept, Equation (3) is suggested as follows, where A_i is a tag and R is resources.

$$\log \frac{Count(R)}{Count(R, A_i)} \qquad \text{---- } \textit{Equation (3)}$$

Equation (3) gives a lower score for a general tag and gives a higher score to a specific tag, that is, it gives a high score when a tag is less used in the collection. While these high scores help in ranking, they are not exactly what we want for classification, i.e. terms that identify only one resource are not classifiers. Tags with a high value based on Equation (3) contain idiosyncratic terms or personalized terms that are not useful in representing the topic category of resource content. Thus, we modified Equation (3) to remove idiosyncratic terms and to normalize *Cross Resources Annotation Discrimination (CRAD)* as a measurement of the portion of a set of resources about a topic or domain (represented by a given tag A_i) against the resource collection. The *CRAD* is designed to discriminate tags that are used too broadly or too narrowly in the document collection. If a tag is assigned for every resource in the collection, the *CRAD* value will be 0. We consider it to be a weak candidate as a tag to identify the domain classification of the document. Similarly, if a tag is assigned for a small subset of resources, that is, a *CRAD* value close to 1, we also consider it to be a weak candidate to discriminate the subset as a category. *Cross Resources Annotation Discrimination (CRAD)* is defined as follows, where A_i is a given tag, R is resources, and U is users.

$$CRAD = \left\{ (Count(U, A_i) = 1 \rightarrow 0) \middle| (Count(U, A_i) > 1 \rightarrow \frac{\log\left(\frac{Count(R)}{Count(R,A_i)}\right)}{\log(\,Count(R)\,)}) \right\} \quad \text{--- } \textit{Equation (4)}$$

In Equation (4), *CRAD* penalizes idiosyncratic tag by giving weight 0 to the tag that appears once in only one resource in the collection by only one user. In doing so, *CRAD* removes the long tail of the tag distribution. For instance, from Figure 7, *CRAD* gets rid of tags that occur only once, i.e. in case 2, the annotations that get a 0 score by the *CRAD* measure are tags from B to J. It is divided by log(Count(R)) to normalize the numerator values. The denominator represents the maximum value of *CRAD* values. It will normalize *CRAD* by the collection size, and make the range of the *CRAD* to be greater than or equal to 0 and less than 1 regardless of the change of the collection size. However, the *CRAD* is affected by the total size of the collection. Table 4 and Figure 8 show how CRAD values change for given collection sizes and document set sizes. Table 5 and Figure 9 show how the document set coverage ratio changes for different collection sizes and the *CRAD* values.

Table 4. CRAD Values of the Difference Collection Size (rows) and Document Coverage (columns)

	0%	**10%**	**30%**	**50%**	**80%**	**100%**
1,000,000	1	0.166667	0.087146	0.050172	0.016152	0
100,000	1	0.2	0.104576	0.060206	0.019382	0
10,000	1	0.25	0.13072	0.075257	0.024228	0
1,000	1	0.333333	0.174293	0.100343	0.032303	0
100	1	0.5	0.261439	0.150515	0.048455	0

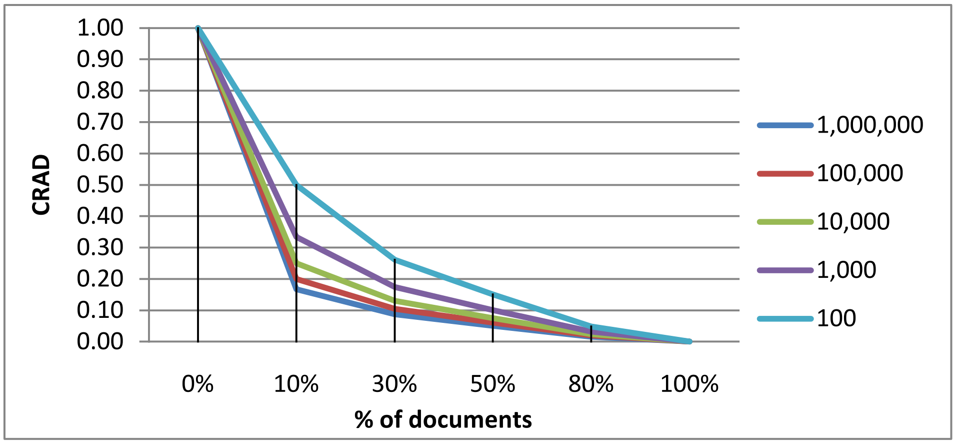


Figure 8. CRAD Values of the Difference Collection Size and Document Coverage

Table 5. Document Coverage (%) Changes for the CRAD (rows) and the Collection Size (columns)

	1,000	**10,000**	**100,000**	**1,000,000**
0	1.0000	1.0000	1.0000	1.0000
0.1	0.5006	0.3980	0.3162	0.2512
0.2	0.1436	0.0758	0.0398	0.0209
0.3	0.0315	0.0103	0.0033	0.0011
0.4	0.0071	0.0015	0.0003	0.0001
0.5	0.0018	0.0003	0.0000	0.0000
0.6	0.0005	0.0001	0.0000	0.0000
0.7	0.0001	0.0000	0.0000	0.0000

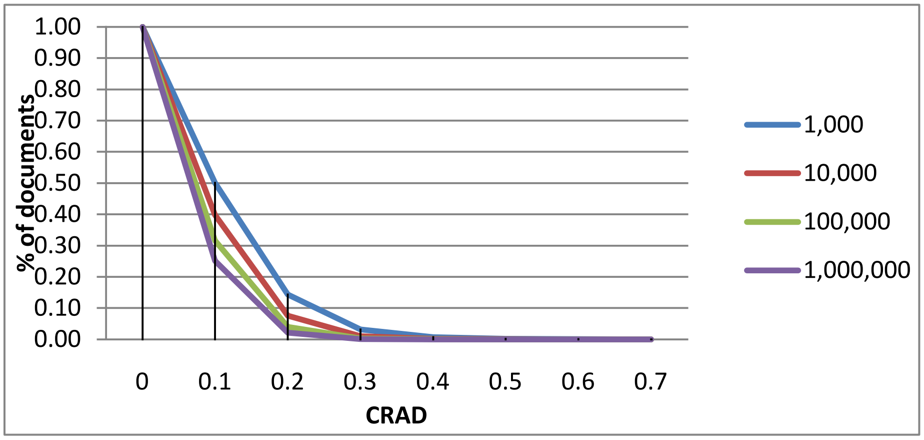


Figure 9. Document Coverage (%) Changes for the CRAD and the Collection Size

3.3.3 Exploratory Analysis on AD and CRAD

Using a sample of 1,800,651 bookmarks with 205,486 unique tags by 488,939 unique users for 7,411 resources, preliminary observations and tests were made on *AD* and *CRAD*.

3.3.3.1 Stability of Tag Pattern

OBSERVATION 1
Social tags as an uncontrolled method to develop subsets of topics will stabilize their portion in a collection after the collection reaches a sufficient size.

We expect that tags will become stable in their occurrence in the collection as the collection size grows beyond a threshold point. It has been observed that when a collection is developed without intended control, its subsets or categories manage to keep a certain portion in a collection. For example, when the library collections are developed, except when policy and controls are explicitly involved, the proportion of certain domains or topics stays the same regardless of the growth of collection size. Examples of the proportion of subjects in the collection by year for Brown University Libraries (Figure 10) and Wellesley College Library (Figure 11) shows that the proportions of subject areas stay the same in the collection unless other factors occur, e.g. in the case of a library, factors such as intentional increases/decreases in collection development due to users' needs and unexpected increase due to donations may occur.

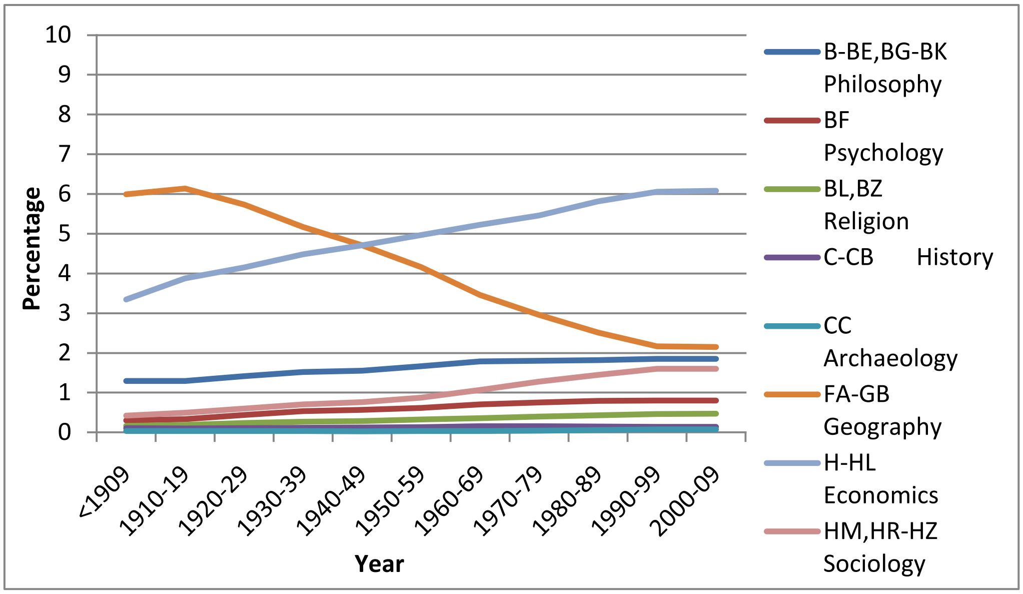


Figure 10. Proportion of Subjects by Year (Brown University Libraries)

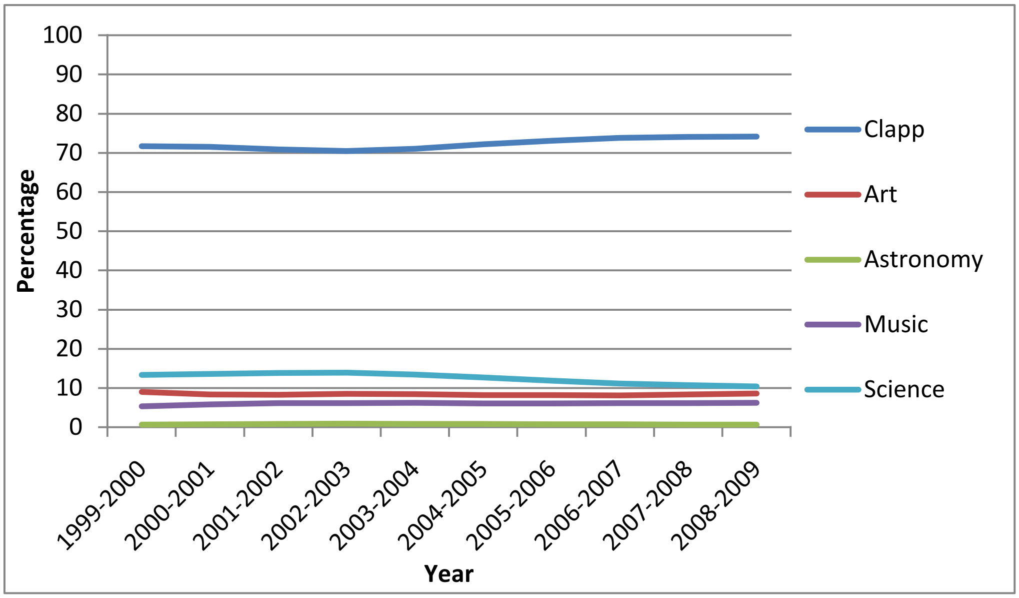


Figure 11. Proportion of Subjects by Year (Wellesley College Library)

A similar observation related to social tags was made by Golder and Huberman (2006). They explained stable patterns in tag proportions related to the dynamics of a stochastic urn model originally proposed by Eggenberger and Polya. The urn model explains the probabilistic occurrence of balls in an urn with two colors, for example red and blue. This model demonstrates that when a ball is randomly drawn from the urn and replaced, after a number of draws, a pattern emerges such that the probability of red or blue ball being drawn becomes stable over time. Based on this model, Golder and Huberman (2006) showed stability of tags emerges for a certain resource after a certain number of bookmarks are added, i.e. after fewer than 100 bookmarks (Figure 12).

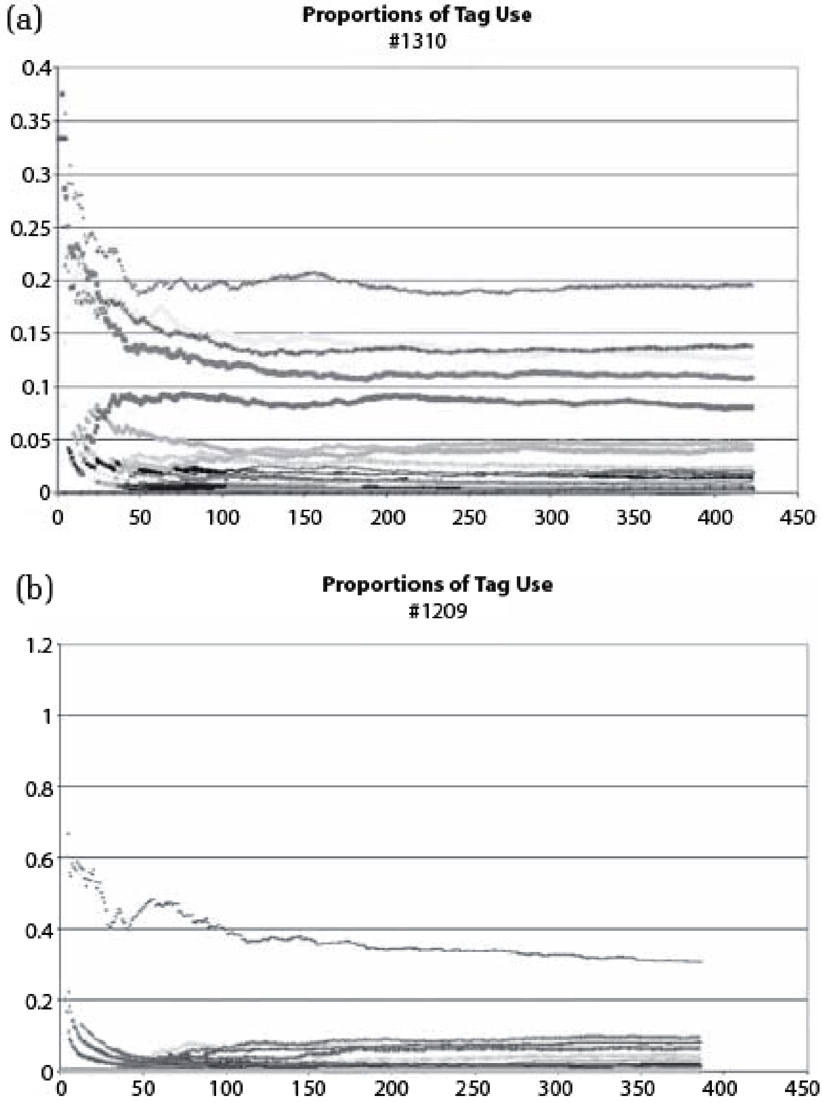


Figure 12. The stabilization of tags' relative proportions for two popular URLs (#1310 (a) and #1209 (b)). The vertical axis denotes fractions and the horizontal axis time in units of bookmarks added (Golder and Huberman, 2006).

Given the characteristics of stability of domains or topics in an uncontrolled collection, we tested the tags' stability patterns over the collection. We observed how the proportion of tags in a collection stabilize as the collection size grows from 1 to 7388 (Figure 13). Figure 13 represents that the proportion of the tags' occurrences for 30 randomly selected tags. It shows that tag occurrence stabilizes as the size of collection grows. In addition, it shows three clearly divided groups of tags – popular tags, unpopular tags (idiosyncratic tags), and often-used tags. Popular tags that occur approximately from 15% to 25% in the graph fall into the broad folksonomy; unpopular tags that occur near 0% and often-used tags that occur less than 10% in the graph can be defined as the narrow folksonomy. For classificatory metadata, we are sure that unpopular tags are not our concern. We will only filter out popular and often-used tags as candidates of classificatory metadata terms. Further analysis is needed to define the threshold *CRAD* value since it is affected by the size of collection. Finding the point at which tag proportion stabilizes will also be important. Appendix A describes further discussion on finding broad, often-used, and narrow folksonomy from the collection along with a detailed explanation on Figure 13.

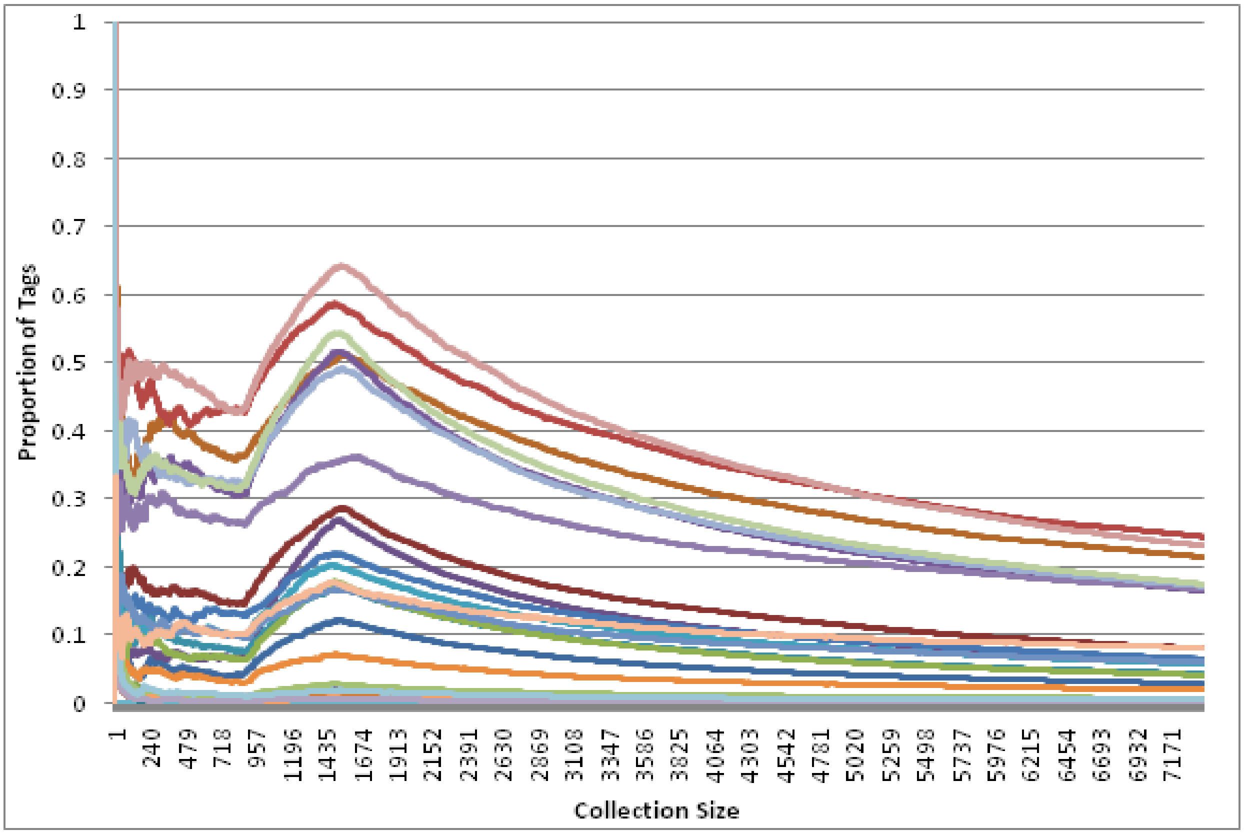


Figure 13. The stabilization of Tags' Proportion over Collection

3.3.3.2 AD and CRAD Relationship

OBSERVATION 2
There is a relationship between the Annotation Dominance (AD) and the Cross Resources Annotation Discrimination (CRAD) that represents patterns for the importance of a tag.

Since the *Annotation Dominance (AD)* is a measure of the dominance of a tag in tag sets for a resource and the *Cross Resource Annotation Discrimination (CRAD)* is a measure of the extent to which a tag defines a reasonable subset in the collection, it seemed that there should be a relationship between *AD* and *CRAD* of a tag. For example, given a tag T_i over a resource set R_j, one might suspect that AD_{TiRj} (the *AD* of term T_i over resource collection R_j) would be high for some sets and low for others. More to the point, in an ideal world T_i would be used in all or most of the bookmark sets for some set of resources R_j and in none or almost none of the annotation sets for the remaining resources. Graphically, this situation might look like Figure 14.

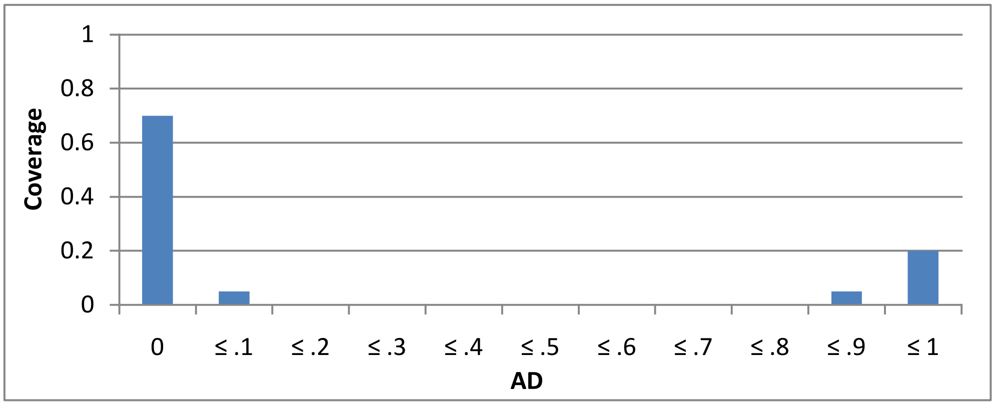


Figure 14. Expected Distribution of AD of T_i over a resource collection R_j

The test with sample data set resulted that the distribution of AD_{TiRj} (the *AD* of term T_i over resource collection R_j) did not follow the expected pattern. We expected to see changes in the pattern of tags according to their popularity, however, it turned out that the shape of the *AD* over the collection graph showed a power curve regardless of the popularity of tags (Figure 15). If the tag term is popular over the collection, the curve becomes extreme and the tail becomes short. When the tag term is very popular over the collection, the peak is low and tail is long. Figure 15 provides examples of two tags – “best” and “design”. The tag “design” is one of the most popular tags in the sample data set. It appears in 23.4% of the resources (1732 resources out of 7411 resources). The tag “best” is one of non-popular tags in the sample data set, appearing in only 5.5% of the resources (406 resources out of 7411 resources). Contrary to our expectation, there was no bump at the high *AD* when the tag was popular, meaning that the tag rarely appears in all or most of the tag sets for a resource. One explanation might be that users select different terms to represent similar concepts – not having a controlled vocabulary.

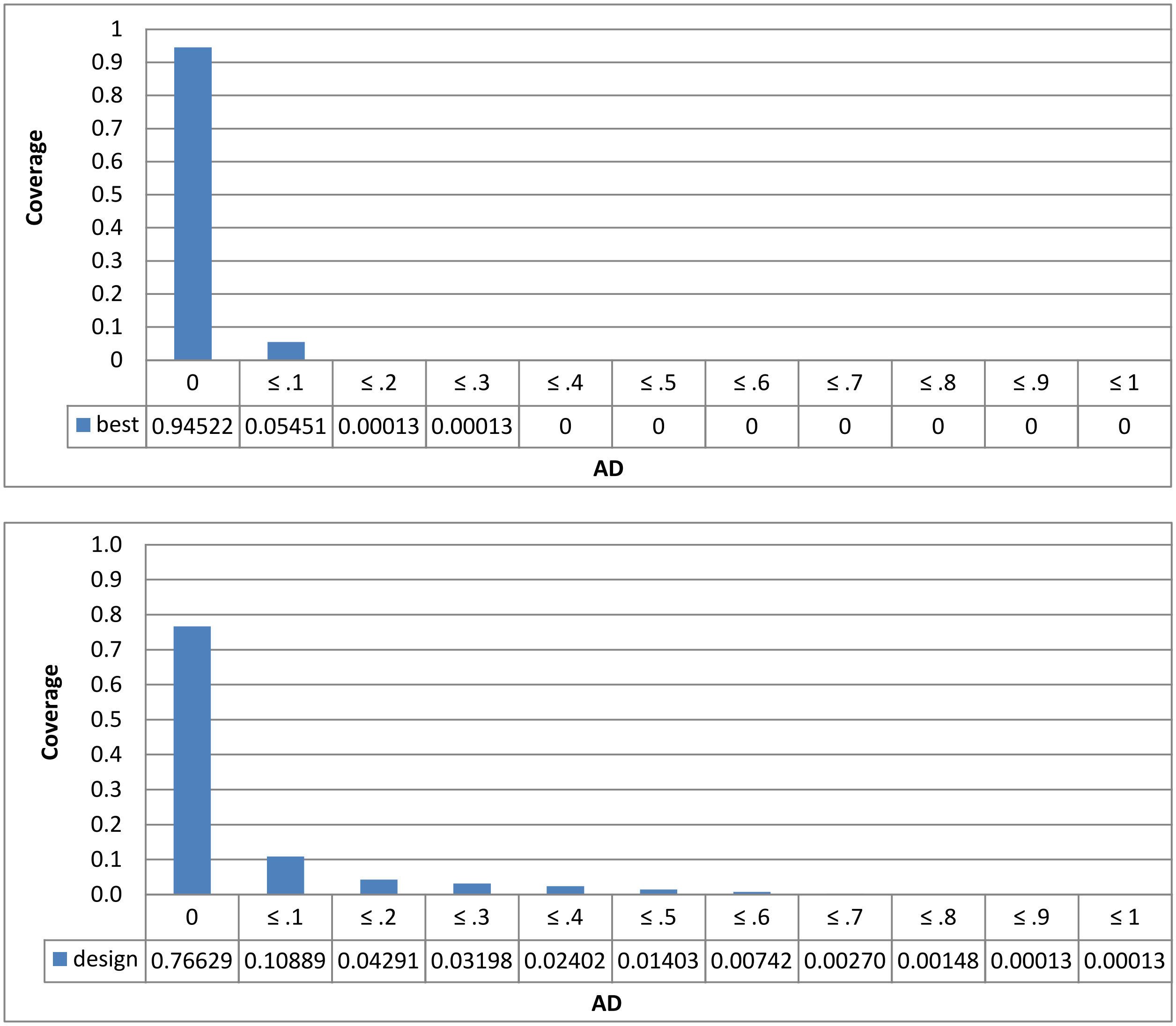


Figure 15. Distribution of AD of Tag "best" (top) and "design" (bottom) over sample collection

3.3.3.3 Optimal CRAD Values

OBSERVATION 3
Some of the Cross Resources Annotation Discrimination (CRAD) for a tag ranges from 0 – the tag is used non-discriminately – to 1 – the tag is used very infrequently. Tags with either value are less than optimal. There is some optimal range of CRAD values that highlights tags used over a subset of the collection of optimal size for classification. Tags with this value combined with AD will identify tags that will serve as classificatory metadata tags.

The *Observation 1* has shown the necessity of identifying an intermediate range of good *Cross Resources Annotation Discrimination (CRAD)* values.

Below, we show two examples of term selection from a set of tags found in a set of bookmarks for a given resource, script.aculo.us and cnn.com. In all cases, the *CRAD* values obtained are against a total collection of 7,411 resources for which 1,800,651 bookmarks exist using 205,486 distinct tags. The examples each provide three sets of tags. The set of the first column lists the 20 most dominant terms used to tag the resource. That is, the first column shows the tags that would be selected if only *Annotation Dominance* were used. The center and the right columns show the product of *AD* and *CRAD* (Equation 5) with or without the values of *CRAD* ranged. Keep in mind that the *CRAD* value will produce a value close to 0 when the tag is heavily used and value close to 1 when it is used for only one resource. The column to the right shows the top twenty terms when a weighted or ranged *CRAD* value is applied. To favor *CRAD* values that collect approximately 1.5-17% of the collection (*CRAD* values of .2 - .5), we limited the range of *CRAD* values to less than .5.

$$AD * CRAD = \left(\frac{Count(\,T_{A_i},R_j)}{Count(U,R_j)}\right) * \left(\frac{\log(\frac{Count(R)}{Count(R,A_i)})}{\log(Count(R))}\right) \qquad \text{--- } \textit{Equation (5)}$$

The yellow highlights indicate newly appearing tags and the green highlights indicate disappearing tags. In the example of script.aculo.us, the list of the 20 most dominant tags stays almost the same (Table 6). This example shows more agreement in terms of tag selection for this particular resource. When the tags are in the range of being a good candidate term for classificatory metadata, i.e. not too specific and not too general, defining the range of *CRAD* would not affect the result as much.

Table 6. Ranks of top 20 AD, AD*CRAD, and AD*ranged CRAD for script.aculo.us

AD (rank)	AD*CRAD (AD rank, rank)	AD*ranged CRAD (AD rank, rank)
javascript (1)	javascript (1, 1)	javascript (1, 1)
ajax (2)	ajax (2, 2)	ajax (2, 2)
web2.0 (3)	framework (7, 3)	web2.0 (3, 3)
programming (4)	programming (4, 4)	programming (4, 4)
webdesign (5)	web2.0 (3, 5)	framework (7, 5)
web (6)	webdesign (5, 6)	webdesign (5, 6)
framework (7)	css (8, 7)	web (6, 7)
css (8)	web (6, 8)	css (8, 8)
design (9)	scriptaculous (15, 9)	library (12, 9)
development (10)	library (12, 10)	development (10, 10)
webdev (11)	webdev (11, 11)	webdev (11, 11)
library (12)	development (10, 12)	design (9, 12)
tools (13)	scripts (14, 13)	scriptaculous (15, 13)
scripts (14)	rails (16, 14)	tools (13, 14)
scriptaculous (15)	design (9, 15)	scripts (14, 15)
rails (16)	prototype (23, 16)	rails (16, 16)
code (17)	AJAX (22, 17)	prototype (23, 17)
reference (18)	tools (13, 18)	code (17, 18)
opensource (19)	Javascript (24, 19)	AJAX (22, 19)
software (20)	code (17, 20)	opensource (19, 20)
	reference (18, 31)	reference (18, 25)
	opensource (19, 26)	software (20, 19)
	software (20, 35)	

On the other hand, in the example of cnn.com, it shows a big shift in the tag list when *CRAD* was defined with a range (Table 7). Tags with high dominance in the resource and low dominance in the collection were ranked low, since high *CRAD* tags were not being considered to be important for classificatory metadata. Therefore tags such as *cnn* (specific names), *NEWS* (un-usual form), *CurrentEvents* (not topic-specific and compounded tag) were pushed to the bottom of the rank (Note that no pre-processing on tags was done for this test). When looking closer at the top 20 ranked tags by range-defined *CRAD*, tags that represent what cnn.com is remain, such as *news* (content type of web page); *video*, *television*, *TV* (type of media); *politics*, *world*, *entertainment*, *international*, *usa* (topics in cnn.com); *english* (language provided); *daily*, *information* (characteristics of contents), etc.

Table 7. **Ranks of top 20 AD, AD*CRAD, and AD*ranged CRAD for CNN.com**

AD (rank)	AD*CRAD (AD rank, rank)	AD*ranged CRAD (AD rank, rank)
news (1)	news (1, 1)	news (1, 1)
News (2)	cnn (3, 2)	News (2, 2)
cnn (3)	News (2, 3)	world (5, 3)
media (4)	CNN (9, 4)	media (4, 4)
world (5)	world (5, 5)	politics (6, 5)
politics (6)	media (4, 6)	daily (7, 6)
daily (7)	politics (6, 7)	tv (8, 7)
tv (8)	daily (7, 8)	BookmarksBar (11, 8)
CNN (9)	tv (8, 9)	imported (10, 9)
imported (10)	CurrentEvents (15, 10)	usa (12, 10)
BookmarksBar (11)	weather (13, 11)	television (22, 11)
usa (12)	BookmarksBar (11, 12)	video (14, 12)
weather (13)	currentevents (19, 13)	international (28, 13)
video (14)	imported (10, 14)	entertainment (18, 14)
CurrentEvents (15)	usa (12, 15)	TV (30, 15)
reference (16)	current (17, 16)	events (31, 16)
current (17)	news, (20, 17)	Media (24, 17)
entertainment (18)	NEWS (26, 18)	us (32, 18)
currentevents (19)	sports (25, 19)	information (21, 19)
news, (20)	television (22, 20)	english (29, 20)
	video (14, 21)	reference (16, 25)
	reference (16, 44)	current (17, 287)
	entertainment (18, 23)	CNN (9, 392)
		news, (20, 424)
		currentevents (19, 821)
		CurrentEvents (15, 873)
		cnn (3, 911)
		weather (13, 994)

Looking at the results, it would maximize the effectiveness of *AD*CRAD* if the analysis on compound tags and multiple form tags are combined in calculating *AD*CRAD*. For example, cnn.com examples contains multiple terms such as “news”, “News”, and “NEWS”. It also contains terms such as “news,” with a trailing comma, which might need to be included as “news”. The combination of these forms will increase the importance of “news” as well as add more terms in the top ranks, including “television”, “video”, “international”, and “entertainment” in the case of the cnn.com example.

3.3.4 Exploratory Analysis on Compound Tags

The examples of rank change comparison show that after removing idiosyncratic tags and ranging *CRAD* values, there are still some interesting tag terms, i.e. compound terms. There are efforts to relate or overlap social tags with controlled vocabularies. It was found that there is little overlap among tags, automated indexing, and controlled vocabularies (Lin et al., 2006). On the other hand, Syn and Spring (2009) have shown the relatively good potential of social tags compared with controlled vocabularies, especially when used together. Further analysis was made by Yi and Chan (2009) to link folksonomy to Library of Congress Subject Headings (LCSH). Yi and Chan (2009) suggested further analysis on compound terms would provide a better structure of social tags and provide a better link to LCSH. To find the possibility of using compound tags for further analysis such as finding relationships and relating to controlled vocabularies, we first explored the ways of decomposing compound tags using the same sample of 1,800,651 bookmarks with 205,486 unique tags by 488,939 unique users for 7,411 resources.

3.3.4.1 Decomposition of Compound Tags

Compound tags take different forms: (1) *well-delimited forms* use special characters as separators, e.g. “compound_tags”, “compound.tags”, “compound-tags”, (2) *camel case forms* use upper case for the first character of compound words, e.g. “CompoundTags”, (3) *undifferentiated* *compound tags* are simply combined multiple words, e.g. “compoundtags”. Tonkin (2006) tried to decompose English compound tags focusing on finding the longest prefix of compound word in the dictionary. We took similar steps using an English dictionary and the Wikipedia Thesaurus (http://dev.wikipedia-lab.org/WikipediaThesaurusV2/) to decompose compound tags and build a dictionary of emerging words (Figure 16). There are three major reasons for decomposing compound tags. First, when separators are used to form a compound tag, it is more likely that the words formed in between separators are single words. Therefore, if we decompose compound tags with separators, it is easier to define and disambiguate emergent words, abbreviations, and online terms that are often used as tags but not included in general English dictionaries. Second, by decomposing compound tags, the method for weighting important tags can be weighted higher since quality tags can be included in a tag set as a single word form and also a compounded word form. Third, given that compound tags are related words after decomposing them, it would be easier to define a relationship, if any, between the terms.

To include emergent words and often-accepted terms, we built a new dictionary using compound tags with separators on an assumption that when separators are used, users do not combine multiple words (undifferentiated compound tags) in between separators. The dictionary, named the *Emerging Words Dictionary*, consists of emergent words such as “blog” and “google” which were not in a regular English dictionary. We extracted all compound tags with separators from our sample dataset from Delicious with 205,486 unique tags. The separators we defined are “/”, “:”, “_”, “+”, “-“, “&”, “.”, “,”, “!”, ““”, “””, “*”, “?”, “#”, “@”, “$”, “(”, “)”, “[”, “]”, “<”, “>”, “‘” , “|” which expands what Guy and Tonkin (2006) defined as popular separators (Figure 5). After the extraction process, the Wikipedia Thesaurus was used to find likely-to-be-a-word terms and unlikely-to-be-word terms. The Wikipedia Thesaurus was selected as the thesaurus to find emerging words since Wikipedia tends to include emerging words very early on. From the words identified by using the Wikipedia Thesaurus, we made a heuristic decision to exclude the words that are not useful or considered as generally accepted words using the process shown in Figure 16. After the heuristic steps, the *Emerging Words Dictionary* was created with 2,145 words.

Criteria for words added to the *Emerging Words Dictionary*

1. Emergent words
   a. New words (e.g. semanticweb, blog, folksonomy, avatar)
   b. Names of Web sites, services, company, or applications (e.g. Flickr, Youtube, Google)
2. Commonly accepted short-hand and abbreviations
   a. Commonly accepted short-hand (e.g. ir (information retrieval, dev (development), info (information))
   b. Abbreviations (e.g. XML, IDE)
3. File extensions and file types
   a. Media types (e.g. txt, pdf, mp3)
   b. Contents (e.g. js, py)
4. Versioning (e.g. web2.0, php5)

Criteria for words not added to the *Emerging Words Dictionary*

1. Foreign words (e.g. foto, programacion)
2. Personal tags and words with no common meaning (e.g. stitch1976, ls534)
3. Misspelled terms
4. Parts of term (e.g. ish, ons (probably from “add-ons”), nt (probably from “Windows NT”)
5. Proper nouns such as name of person (e.g. Crockford, Kottke)

Figure 16. Heuristics Criteria for the *Emerging Words Dictionary*

In addition to the general English dictionary, the *Emerging Words Dictionary* is used to determine words from compound tags when decomposing them. Figure 17 provides the algorithm for decomposing compound tags.

```
SeparateTag(Tag)
    try find Tag in dictionaries
    if yes record as single word form.
    if no find special charactors.
        if yes split by special charactors
            record splitted words
        find camel cases
            if yes split by camel cases
                record splitted words
        find first possible word from Tag
            try check the last part of Tag in the dictionaries
            if found in dictionaries
                record first and last part
            if not found
                try SeparateTag(the last part of Tag)
```

Figure 17. Algorithm for Decomposing Compound Tags

3.3.4.2 Application of Decomposed Compound Tags

The tag set may include simple terms, proper terms, compound terms, and complex terms. A simple term is a single word/concept, e.g. web, java, programming. A proper term is one or more simple terms that are placed together because they refer to a named entity, e.g. “google”, “extensible markup language”, “web2.0”, “semantic web”. A compound term is two or more simple terms with no implied relationship, e.g. “airlines-fareandinfo”, “architectsandprogrammers”. A complex term is a compound term with a relationship implied between the two terms, e.g. “javaprogramming”, “webdesign”.

This implies several issues and strategies that might be used in developing classificatory metadata:

1. By breaking apart all compound and complex terms, we may change the AD and CRAD measures for simple terms.
2. By recognizing proper simple terms, we may confirm that simple terms sometimes reflect proper terms which may be appropriate for leaf node classificatory metadata.

3. Complex terms may reflect simple term order in tag sets pointing to the same resource.

One question is how modified *AD* and *CRAD* based on various algorithms for use of compound terms would impact terms that might be used for classification. For example, if a set of tag sets with four terms, *a*, *b*, *c*, and compound term *bc* where the use of compound term weighting changes the scoring significantly, we would like to determine whether the changed scoring reflects the expert user opinion of the appropriateness of the classificatory metadata. For example, if we can find a set of resources where we have two terms, *a* and *b*, such that *a* has a score that is “significantly higher” than *b*, but the application of the compound terms causes the term *b* to become “significantly higher” than *a*, we would plan to apply compound terms by asking experts which application format better describes the resource.

4.0 Research Design

The goal of this research is to find tags that have shown potential for use as classificatory metadata to group resources by topics or domains from their tag sets. It is expected that *Classification Potential* with the proposed metrics, *Annotation Dominance (AD)* and *Cross Resources Annotation Discrimination (CRAD)*, will generate a tag set optimized for classification of web resources. Based on the preliminary studies that have been done, the research will be carried forth in two phases. In phase one, we will determine (1) the appropriate range of *CRAD* for identifying classificatory metadata, and (2) the appropriate format of application of decomposed compound tags. Based on these findings the second phase will assess the quality of the generated classificatory metadata.

The major questions we want to address in this research are:

- By applying *AD*CRAD* measurements, can tag noise be reduced?
- By decomposing compound tags, can ambiguous tags be identified and disambiguated?
- By applying ranged *AD*CRAD* to tag sets, can a subset of tags be identified as classificatory metadata terms?

Using data collected from Delicious, we will first determine the range of *CRAD* and the format for applying decomposed compound tags from the subjects' judgment on the relevance of tag terms as classificatory metadata. Then for the second phase, we will use the identified range of CRAD and reformatted compound tags to assess our ability to generate metadata. We compare expert generated metadata information from Open Directory Project and INFOMINE to generated classification metadata based on tag terms from Delicious.

4.1 Delicious Data

We collected tag data from the social bookmarking system, Delicious. The current numbers of users, resources, tag, and bookmarks on Delicious is unknown. The last published figures by Arrington (2007) indicate that Delicious had "[…] 3 million registered users and 100 million unique URLs bookmarked" as of September 2007. We do know that Delicious experienced exponential growth in its user base from 2005 to 2007. In September 2006, Delicious announced on its blog that it had achieved 1 million members, about triple the number of users it had at the end of 2005 (Schacter, 2006). Hammond et al. (2005) reviewed del.icio.us, reporting that it had 50,000 users, 1 million links (resources), and 2 million tags as of April 2005. Thus, in 3 years, the number of registered users has increased by roughly 60 times, while the number of resources has increased 100-fold. The average number of bookmarks per user has also risen from twenty in April 2005 to 33.3 by September 2007.

Our data was crawled from November 2009 to February 2010. Given storage limitations we made no effort to collect a complete picture of delicious. The goals of the crawling were 1) to collect as many bookmarks as possible and 2) to build a sample Delicious dataset that was representative of Delicious as a whole. We made no attempt to filter the data by tags – all bookmarks were accepted regardless of topic, popularity, tag distribution, or language. The crawling began with a selection of several individuals at random. For each of those individuals, all of the bookmarked resources were collected. Then for each of those resources, all of the individuals who bookmarked them were collected. Given restrictions on Delicious, there is no guarantee that every user who bookmarked a web page is included. Also, given where the crawling of users and resource was terminated, we ended up with a snapshot of the users and bookmarks associated with 7,097 resources. The dataset for the experiment contains 3,077,038 bookmarks on 7,097 distinct resources by 506,341 different users using 166,379 distinct tags.

4.2 Pre-processing of the Tag Data

Although users may select the same word as a tag, since tags are created and added without any restrictions, users might enter the word in different forms, e.g. "news", "News", "NEWS", etc. Since the suggested measurements take the dominance into consideration, unifying the format of tags will affect the results of measurements. The necessary cases for pre-processing are as listed:

- **Capitalized words**: Words can be entered in lower-case, upper-case, or both. We consider all these cases to indicate the same word. For example, "news", "News", and "NEWS" are all considered and counted as "news".
- **Special Characters**: Sometimes users enter a special character mainly because they did not realize how the tagging system detects words as tags, i.e. single words separated by a space. For example, if *"Semantic Web"* is input into the system, the system will recognize this input as two tags, "*Semantic* and *Web"*, each with one double quotation mark attached. For such cases, special characters used most often are single quotations, double quotations, parentheses, and

commas. In these cases, we will ignore these special characters and consider the two words, *Semantic* and *Web*, as two unique tags.

- **Compound Tags**: Compound tags take various forms, e.g. “CompoundTags”, “Compound-Tags”, “Compound_Tags”, etc. All of the forms appearing in the collection will ultimately be converted to one standard format based on the result of the phase 1 experiment. There are two alternate forms being considered: the standardized compound form and the decomposed form. In the standardized compound form, “CompoundTags”, “Compound-Tags”, “Compound_Tags” will all be converted to “compound tags”. In the decomposed form, they would be converted to “compound” and “tags”.

It should be noted that once capitalized tags, tags with special characters (not for compounding), and compound tags are processed, the number of distinct tags is expected to decrease. As a result of pre-processing, the number of unique tags in the dataset decreased by 39.93% for the standardized compound form (from 166,379 unique tags to 99,939 unique tags) and 85.29% for the decomposed form (from 166,379 unique tags to 24,478 unique tags). After the tag data is pre-processed, the calculation of the AD and CRAD values for tags is also expected to be more accurate.

4.3 Phase 1: Finding the Range of CRAD Measurement and the Format of Compound Tags

In the first phase of the experiment, we evaluated three values for the range of *CRAD* and two formats of re-combining decomposed compound tags. The different tag sets created using various ranges of *CRAD* and formats of compound tags are selected to find the best range of *CRAD* values to apply and the best format of compound tags to identify classificatory metadata.

4.3.1 Experimental Data

From the Delicious dataset, twenty web pages are selected (Table 8). Three different *CRAD* values and two different compound tags formats are used to select tags from the selected 20 web pages. As a result, there will be six different conditions (Table 9).

The three ranges of *CRAD* values reflect the coverage of documents in the collection. Given that the main division of existing popular classification schemes ranges from 10 classes (Dewey Decimal Classification) to 20 classes (Library of Congress Classification) and that classification schemes based on web pages such Open Directory Project or Yahoo! Directory define main divisions to be around 15 classes, we decided the reasonable coverage of documents are at the threshold of 1.5–20% range. The three conditions of *CRAD* ranges include 1.5-7% coverage (*CRAD* values of 0.2999-0.4736), 7-14% coverage (*CRAD* values of 0.2217-0.2999), and 14-20% coverage (*CRAD* values of 0.1815-0.2217).

The two formats of compound tags include the decomposed form (separating compound tags as multiple single words) and the standardized compound form (re-combine compound tags in a unified format). For the standardized compound form, we will re-combine compound terms with space in between the terms since the category labels provided by human experts often include spaces to have multiple terms together. Since this form applies to category labels in Open Directory Project and subject topics in INFOMINE that we will apply in phase 2 of the experiment, we will form the application of standardized compound terms using spaces. For example, all cases of “compoundterm”, “compound_term”, and “CompoundTerm” will appear as “compound term”.

Note that although only 20 web pages are selected from the data set for this experiment, *CRAD* will be calculated on the whole data set (7,097 resources).

Table 8. List of Selected Web Pages for Phase 1

	Title	URL
1	Kayak	http://www.kayak.com/
2	Blurb	http://www.blurb.com/
3	WordReference	http://www.wordreference.com/
4	Indeed	http://www.indeed.com/
5	English-to-go	http://www.english-to-go.com/
6	10 papers you need to read	http://www.scienceforseo.com/information-retrieval/10-papers-you-need-to-read/
7	Beer Recipes and Resources for Homebrewers	http://beerrecipes.org/
8	American Hiking Society	http://americanhiking.org/
9	Prepare for Attack	http://www.thesamet.com/blog/2007/01/16/prepare-for-attack%E2%80%94making-your-web-applications-more-secure/
10	Survey System - Design Tips	http://www.surveysystem.com/sdesign.htm
11	Hulu	http://www.hulu.com/
12	50 iPhone Apps for Web Designers & Developers	http://mac.appstorm.net/roundups/iphone-roundups/50-iphone-apps-for-web-designers-developers/
13	The Cool Hunter	http://www.thecoolhunter.net/
14	WebMD	http://www.webmd.com/
15	ilovetypography	http://ilovetypography.com/
16	Layout Gala	http://blog.html.it/layoutgala/
17	70 Expert Ideas For Better CSS Coding	http://www.smashingmagazine.com/2007/05/10/70-expert-ideas-for-better-css-coding/
18	Taxonomy - Wikipedia	http://en.wikipedia.org/wiki/Taxonomy
19	MusicBrainz	http://musicbrainz.org/
20	SQUASHED PHILOSOPHERS	http://www.btinternet.com/~glynhughes/squashed/

4.3.2 Participants

Twenty participants were recruited from the University of Pittsburgh School of Information Sciences and Pittsburgh libraries[3] for phase 1. The decision on the sample size was made by power analysis (Cohen, 1988) for the Analysis of Variance (ANOVA). The power of a statistical test is used to find the minimum sample size to accept the statistical test result with certain level of confidence. The power analysis indicated that the minimum sample size to detect a large effect size (f = .5) with a significant level of p = .05 for a confidence of .95 – power of .95 indicates that there is 95% or greater chance of finding a statistical significant difference - is 16 in total, suggesting each group of between groups needs 8 participants. Therefore, based on the result of the power analysis, ten participants are recruited for each group, for a total of twenty participants.

The qualification of participants is strictly focused on their expertise in understanding the concepts of information organization and classification since the participants were expected to analyze the classificatory metadata terms as topic descriptors from the perspective of an expert cataloguer or information organization professional. Therefore, the main target groups of participants were professional librarians, Library Science degree holders (masters or doctorate), and current graduate students in the Library Science program who have taken major Information Organization courses. The listed courses for the recruitment[4] were the courses offered in the School of Information Sciences at University of Pittsburgh; however, corresponding courses

[3] The study was approved by the Institutional Review Board at University of Pittsburgh (PRO10040357).

[4] The courses appeared in the recruitment statement were Organizing and Retrieving Information (LIS2005), Introduction to Cataloging and Classification (LIS2405), Metadata (LIS2407), and Indexing and Abstracting (LIS2452), all offered from University of Pittsburgh.

from other institutions were accepted. With this condition being met, each participant was considered to be an expert and, thus, their judgment on the terms to be professional.

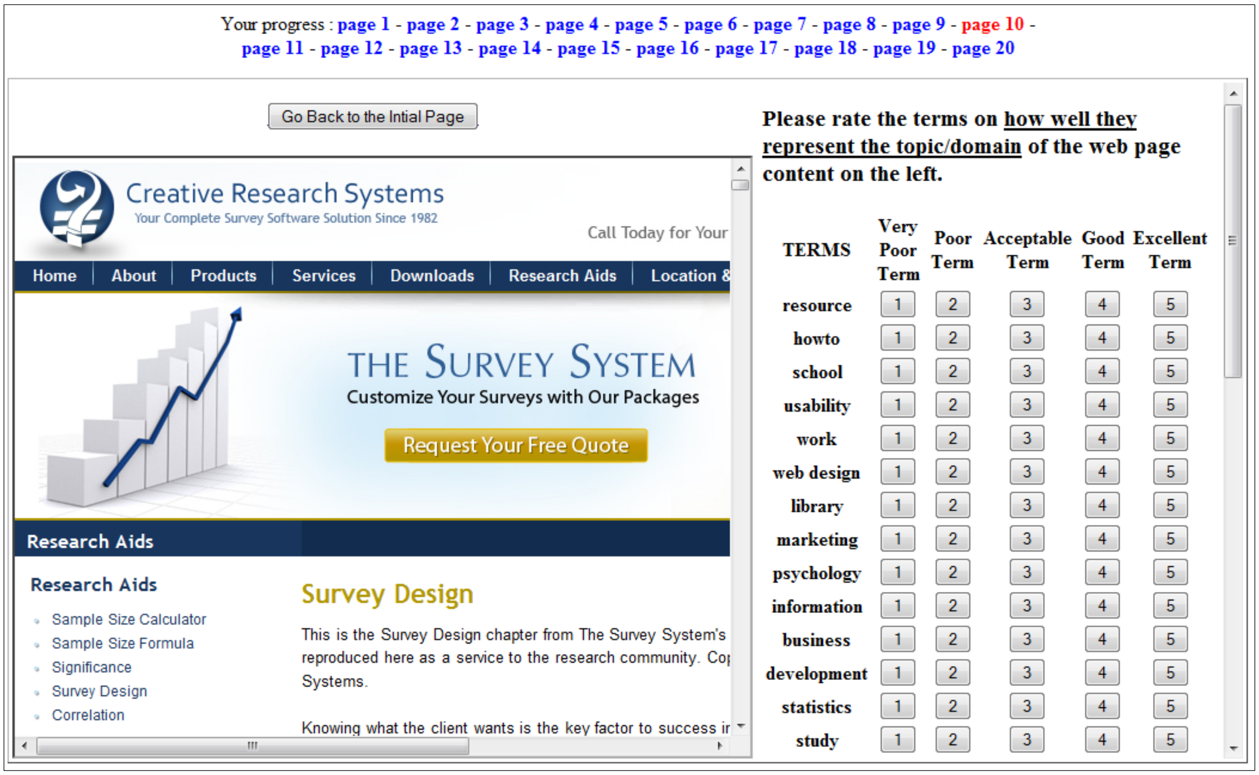
4.3.3 Experimental Design

From the dataset, 20 web pages are selected randomly for the experiment (Table 8). Twenty classification experts are recruited for the experiment. Prior to the experiment, they were given a training session and asked to take a pre-survey (Appendix B). The terms that have the *CRAD* values of the three ranges are calculated with their *AD* values as Equation 5 (*AD* CRAD*). Each subject was provided with 20 web pages and tags selected by the three ranges of *CRAD* and one format of compound tags. For example, if subject A is assigned to the decomposed form condition, subject A is assigned to all three *CRAD* conditions with the decomposed form condition. If subject B is assigned to the standardized compound form condition, subject B is assigned to all three *CRAD* conditions with the standardized compound form condition. Subjects only see one type of application format for compound tags so as not to confuse the subjects since the compound tags conditions provide different presentations of compound tags (Table 9). The conditions are as below.

Table 9. Experiment Design for conditions of CRAD and Compound Tags

Coverage	CRAD Range	Decomposed Terms	Standardized Compound Terms
1.5% - 7%	0.2999-0.4736	10 subjects	10 subjects
7% - 14%	0.2217-0.2999		
14% - 20%	0.1815-0.2217		

Tag terms from the proposed classificatory metadata candidate terms are provided for the 20 URLs in random order. Subjects were asked to rate the relevancy of terms on a five-point scale where “1” indicates a very poor term to identify the subject domain, “2” indicates a poor term to identify the subject domain, “3” is an acceptable term to identify the subject domain, “4” indicates a good term to identify the subject domain, and “5” is an excellent term to identify the subject domain. As the subjects in this experiment are experts, their relevancy ratings are considered to be perfect. The interface a subject viewed for the experiment is shown in Figure 18.



Your progress : page 1 - page 2 - page 3 - page 4 - page 5 - page 6 - page 7 - page 8 - page 9 - page 10 -
page 11 - page 12 - page 13 - page 14 - page 15 - page 16 - page 17 - page 18 - page 19 - page 20

This is the Survey Design chapter from The Survey System's reproduced here as a service to the research community. Co Systems.

Knowing what the client wants is the key factor to success ir

Please rate the terms on how well they represent the topic/domain of the web page content on the left.

TERMS	Very Poor Term	Poor Term	Acceptable Term	Good Term	Excellent Term
resource	1	2	3	4	5
howto	1	2	3	4	5
school	1	2	3	4	5
usability	1	2	3	4	5
work	1	2	3	4	5
web design	1	2	3	4	5
library	1	2	3	4	5
marketing	1	2	3	4	5
psychology	1	2	3	4	5
information	1	2	3	4	5
business	1	2	3	4	5
development	1	2	3	4	5
statistics	1	2	3	4	5
study	1	2	3	4	5

Figure 18. The Experiment Interface

The relevance ratings of terms from the proposed classificatory metadata will be compared using a two-way Mixed Analysis of Variance (ANOVA) test. The hypothesis is as follows.

H$_{1\text{-}0}$: There is no statistical difference among the means of the ratings of the proposed classificatory metadata terms for three conditions of *CRAD* (CR1, CR2, CR3). ($\mu_{CR1} = \mu_{CR2} = \mu_{CR3}$)

H$_{1\text{-}1}$: There are statistical differences among the means of the ratings of the proposed classificatory metadata terms with three conditions of *CRAD* (CR1, CR2, CR3). ($\mu_{CR1} \neq \mu_{CR2} \neq \mu_{CR3}$)

H$_{2\text{-}0}$: There is no statistical difference between the means of the ratings of the proposed classificatory metadata terms for two different application formats for compound tags (decomposed, standardized). ($\mu_{decomposed} = \mu_{standardized}$)

H$_{2\text{-}1}$: There is statistical difference between the means of the ratings of the proposed classificatory metadata terms for two different application formats for compound tags (decomposed, standardized). ($\mu_{decomposed} \neq \mu_{standardized}$)

The null hypothesis will be rejected if the results indicate a significant difference at the 0.05 level. When the null hypothesis is rejected, all pair-wise differences will be examined to find the applicable *CRAD* value range and format of compound tags.

4.4 Phase 2: Evaluation of the Generated Classificatory Metadata

There are limited methods for evaluating a controlled vocabulary. In most cases, it is done using expert analysis and user feedback. Owens (2006) stated that a thesaurus as a type of controlled vocabulary is evaluated when it is “being analyzed by an expert, criticized by users, checked against other indexing and access vocabularies, or its features compared with national or international standards.” Accordingly, Owens (2006) introduced methods of thesaurus evaluation categorized as expert evaluation, focus group, retrieval tests, observational report, and comparative methods. An *expert evaluation* is done by expert users criticizing the scope and selection of a word or category to aid the improvement of the controlled vocabulary. Evaluation on Library of Congress Subject Headings (LCSH) was often done by expert evaluation. For a *focus group method*, a focus group of potential users is asked to reveal their perspectives on the subject grouping. During Open Public Access Catalog (OPAC) studies in the 1980s, the focus group was used for several evaluations including the Library of Congress. A *retrieval test* can be done by testing a collection of documents indexed using the controlled vocabulary. Searchers phrase their queries and then experts examine every item in the collection to determine relevance. An observational report uses transaction logs or controlled tests of use. *Comparative evaluation method*, such as mapping and vocabulary switching, is to compare with existing authorized controlled vocabulary to determine the best audience for the controlled vocabulary and to generate specific suggestions for improvement.

In this study, the classificatory metadata we generate is evaluated with expert evaluation - having experts compare it with professionally generated controlled vocabularies. As is discussed below, we examined the generated metadata against two different sources.

4.4.1 Professionally Generated Data

The data generated from Delicious, if it is classificatory metadata, may provide faceted and/or hierarchical metadata. To understand and evaluate the generated classificatory metadata from Delicious, we compare it with two different sets of professionally generated classificatory metadata, one from the Open Directory Project and the other from INFOMINE. The Open Directory Project (ODP, http://www.dmoz.org/) is a web directory created by humans. The ODP's catalogue is created by humans based on their collection of web resources and ODP claims their catalogue to be a definitive catalogue of the web. The Category labels were considered as subject keywords of controlled vocabulary. The ODP data source is both a controlled vocabulary for classification and a classification scheme. A classification scheme contains particular structure, most of the time a hierarchy, to represent the broader and narrower concepts. In contrast, INFOMINE (http://infomine.ucr.edu/) provides subject keywords and Library of Congress Subject Headings on web resources, which are less of a hierarchical classification and more like a faceted classification.

The generated classificatory metadata may be more a group of descriptors, similar to subject headings or subject keywords that do not need to be pre-coordinated and are intended to describe the topics of a document with one or more authorized terms (Olson and Boll, 2001, pp. 111-152). Although we consider the terms (category labels) from classification schemes as descriptors, there still is a concern about whether participants will understand the category labels properly when the relationship is removed. Different from subject headings, category labels are meant to make sense when the path from the top category to current topic is presented together. For example, for a resource dealing with designing of the web pages, in classification, the issue becomes whether it should be in a category of “Web – Design” or “Design – Web”. Another example of confusion caused by removing the relationship can be a resource with a category label “Java” that can be clearly understood only when the top categories are presented together, e.g. “Computer – Programming – Java” versus “Food – Beverage – Coffee – Java”. At this point, it is not clear whether the generated terms will be more like ODP terms or INFOMINE terms. It is clear that the structure will not be presented explicitly as would be the case of ODP classification.

Table 10. List of Selected Web Pages for Phase 2

	Title	URL
1	Wired News	http://www.wired.com/
2	Google Maps	http://www.maps.google.com/
3	Medscape	http://www.medscape.com/
4	IMDB (The Internet Movie Database)	http://www.imdb.com/
5	W3Schools	http://www.w3schools.com
6	Encyclopedia Mythica	http://www.pantheon.org/
7	Unbound Bible	http://unbound.biola.edu/
8	NASA's Planetary Photojournal	http://photojournal.jpl.nasa.gov/
9	IMF (International Monetary Fund)	http://www.imf.org/
10	The Onion	http://www.theonion.com/
11	Magnum Photos	http://www.magnumphotos.com/c/
12	Purdue OWL	http://owl.english.purdue.edu/owl/
13	Plus Magazine	http://www.plus.maths.org
14	Section 508: The Road to Accessibility	http://www.section508.gov/
15	MIT OpenCourseWare	http://ocw.mit.edu
16	Avian Influenza, from the CDC	http://www.cdc.gov/flu/avian/index.htm
17	Internet Public Library	http://www.ipl.org/
18	Advertising Age	http://adage.com/
19	Open Directory Project: DMOZ	http://www.dmoz.org
20	SourceForge	http://www.sourceforge.net
21	Color Scheme Designer	http://colorschemedesigner.com/
22	The World Clock	http://www.timeanddate.com/worldclock/
23	The Semantic Web Roadmap	http://www.w3.org/DesignIssues/Semantic.html
24	Wikitravel	http://wikitravel.org/en/Main_Page
25	HyperStat Online	http://davidmlane.com/hyperstat/

Twenty-five resources are selected randomly from the Delicious data where it is the case that they also exist in ODP and INFOMINE (Table 10). To make the comparison with INFOMINE and ODP, the limitation was made in selecting the web pages – there are much higher level web pages (e.g. homepage of a website) than lower level web pages (e.g. a particular document or article) since many of the web pages in INFOMINE and ODP tend to be high-level web pages as a point for reference resources. Of the pages gathered from ODP, we collect web pages that are categorized in the lower level in the hierarchy to gather enough terms. For each web page, category labels for topic domain representation are collected. From INFOMINE, we gather Library of Congress Subject Headings, subject keywords, and category for each web page. From Delicious, tag information is crawled. From the collected tags, the proposed classificatory metadata (*AD-CRAD*) for each resource is selected based on the highest weight values of *AD-CRAD* and using conditions identified from phase 1. Although twenty-five web pages are selected, *CRAD* is calculated on the whole data set (7,097 resources).

4.4.2 Participants

Twenty participants were recruited from the University of Pittsburgh School of Information Sciences and Pittsburgh libraries[5] for phase 2. The decision on the sample size was made by power analysis (Cohen, 1988) for the Analysis of Variance (ANOVA). The power of a statistical test is used to find the minimum sample size to accept the statistical test result with certain level of confidence. The power analysis indicated that the minimum sample size to detect a large effect size (f = .5) with a significant level of p = .05 for a confidence of .95 – power of .95 indicates that there is 95% or greater chance of finding a statistical significant difference - is 12 participants. Therefore, based on the result of the power analysis, twenty participants are recruited for phase 2 of the experiment.

The qualification of participants is strictly focused on their expertise in understanding the concepts of information organization and classification since the participants were expected to analyze the classificatory metadata terms as topic descriptors from the perspective of an expert

[5] The study was approved by the Institutional Review Board at University of Pittsburgh (PRO10040357).

cataloguer or information organization professional. Therefore, the main target groups of the participants were professional librarians, Library Science degree holders (masters or doctorate), and current graduate students in Library Science program who have taken major Information Organization courses. The listed courses for the recruitment[6] were the courses offered in the School of Information Sciences at University of Pittsburgh; however, corresponding courses from other institutions were accepted. Upon this condition being met, each participant was considered to be an expert, and thus their judgment on the terms to be professional.

4.4.3 Experimental Design

Twenty classification experts are recruited for the experiment. Prior to the experiment, they were given a training session and asked to do a pre-survey (Appendix B). Each subject was provided with 25 web pages. Terms from the generated classificatory metadata candidate terms and terms from INFOMINE and ODP are provided for the 25 web pages in random order. Terms from the proposed classificatory metadata are generated based on two conditions – high *AD*CRAD* and high *AD*ranged CRAD*. Terms that are from high *AD*CRAD* are calculated as shown in Equation 5 (*AD* CRAD*) and terms that are from high *AD*ranged CRAD* are calculated similarly as Equation 5 but with the terms that only fall into the determined range of *CRAD*. The terms that appear in two or more data sources appear once in the list. Subjects are asked to rate the relevancy of terms on a five-point scale where "1" indicates a very poor term to identify the subject domain, "2" indicates a poor term to identify the subject domain, "3" is an acceptable

[6] The courses appeared in the recruitment statement were Organizing and Retrieving Information (LIS2005), Introduction to Cataloging and Classification (LIS2405), Metadata (LIS2407), and Indexing and Abstracting (LIS2452), all offered from University of Pittsburgh.

term to identify the subject domain, “4” indicates a good term to identify the subject domain, and “5” is an excellent term to identify the subject domain. As the subjects in this experiment are experts, their relevancy ratings are considered to be perfect. They are also asked to identify the type of description the provided list of terms is representing. At the end of the session, the subjects are asked to answer a post-survey (Appendix C).

The relevance ratings of terms from expert generated metadata and the proposed classificatory metadata are compared using one-way within-subjects Analysis of Variance (ANOVA) test. The hypothesis is as follows:

> **H_0**: There is no statistical difference between the means of the NDCG at 10 of the proposed classificatory metadata terms (CM1 for high *AD*CRAD*, CM2 for high *AD*ranged CRAD*) and expert generated classificatory metadata terms (ODP, INFO). ($\mu_{CM1} = \mu_{CM2} = \mu_{ODP} = \mu_{INFO}$)
>
> **H_1**: There is a statistical difference between the means of the NDCG at 10 of the proposed classificatory metadata terms (CM1 for high *AD*CRAD*, CM2 for high *AD*ranged CRAD*) and expert generated classificatory metadata terms (ODP, INFO). ($\mu_{CM1} \neq \mu_{CM2} \neq \mu_{ODP} \neq \mu_{INFO}$)

The null hypothesis will be rejected if the results from the F-test indicate a significant difference at the 0.05 level. When the null hypothesis is rejected, all pair-wise differences will be examined to find the most relevant classificatory metadata.

The second phase of the experiment will determine the extent to which the generated classificatory metadata terms were deemed to be of quality by experts. The tag terms from Delicious are compared with the terms used as category labels or subject keywords in ODP and INFOMINE. This phase will allow us to understand the agreement in term selection as topic descriptor between users and experts and to find out what levels of concepts are described by tag terms, i.e. “broader term” and/or “narrower term”. It is to see the relationship among the terms from different sources – expert-generated controlled vocabulary and user-generated subject terms. This part of the experiment can also make it possible to interpret the effect of presenting the subject terms as a set with their relationships removed.

5.0 Results

This chapter presents the results of the experiments. There were two phases of the user experiments. The first phase was to determine the range of *CRAD* and the format for applying decomposed compound tags. The second phase used the range of *CRAD* and reformation of compound tags determined from the first phase and compared them with the expert generated metadata information from Open Directory Project and INFOMINE along with the high *AD***CRAD* weighted terms.

5.1 Phase 1: Finding the Range of CRAD Measurement and the Format of Compound Tags

Phase 1 of the experiment is designed to find the most applicable range of *CRAD* from the three ranges of *CRAD* and a form of compound tag from the two formats of applying decomposed compound tags. The three ranges of *CRAD* values reflect the coverage of documents in the collection. For the study, the reasonable coverage of documents is decided to be at the threshold of 1.5–20% range. The three conditions of *CRAD* ranges include 1.5-7% coverage (*CRAD* values of 0.2999-0.4736), 7-14% coverage (*CRAD* values of 0.2217-0.2999), and 14-20% coverage (*CRAD* values of 0.1815-0.2217). On the other hand, the two formats of compound tags include decomposed terms forming compound tags into multiple single words and standardized compound terms forming compound tags in a unified phrase format. For standardized compound terms, the compound terms are re-combined with a space in between terms.

For the twenty selected web pages (Table 8), each participant is assigned to one format of compound tags condition randomly. Participants are asked to rate the terms in the three ranges of *CRAD* in the assigned format of compound tags. The ratings on the terms are analyzed to find the *CRAD* range and the compound tag format to apply for the phase 2 of the experiment.

5.1.1 Participants Level of Professionalism and Reliability of Judgments

For phase 1 of the experiment, twenty participants were recruited. Among twenty participants, six participants were librarians, two participants were Library Science doctorate holders, and twelve participants were Library Science students (2 masters and 10 doctorates). The Library Science students have taken 2.57 courses in average from listed six information related courses, including Organizing and Retrieving Information, Introduction to Cataloging and Classification, Advanced Cataloging and Classification, Metadata, Indexing and Abstracting, and Thesaurus Construction. Only three of twelve have taken only one course which is the core course of Library Science, Organizing and Retrieving Information, and two of twelve indicated that they have taken all six of listed information organization related courses.

Participants were asked to self-rate on how well they perform information organization and understand classification concepts (Figure 19). The rating was on a scale of five – 1 indicating very bad, 2 indicating bad, 3 indicating fairly good, 4 indicating good, and 5 indicating excellent. In general, participants rated themselves to be good in resource classification professionally (in average 3.62). Specifically, they rated themselves at an average of 4.05 in understanding classification schemes, 3.9 in understanding a thesaurus, and 3.85 in understanding subject headings. On the other hand, they rated themselves 3.4 on average for organizing in their ordinary life, for example, organizing a personal library, personal pictures, personal computer files and folders, bookmarks, emails/mails, documents, etc.

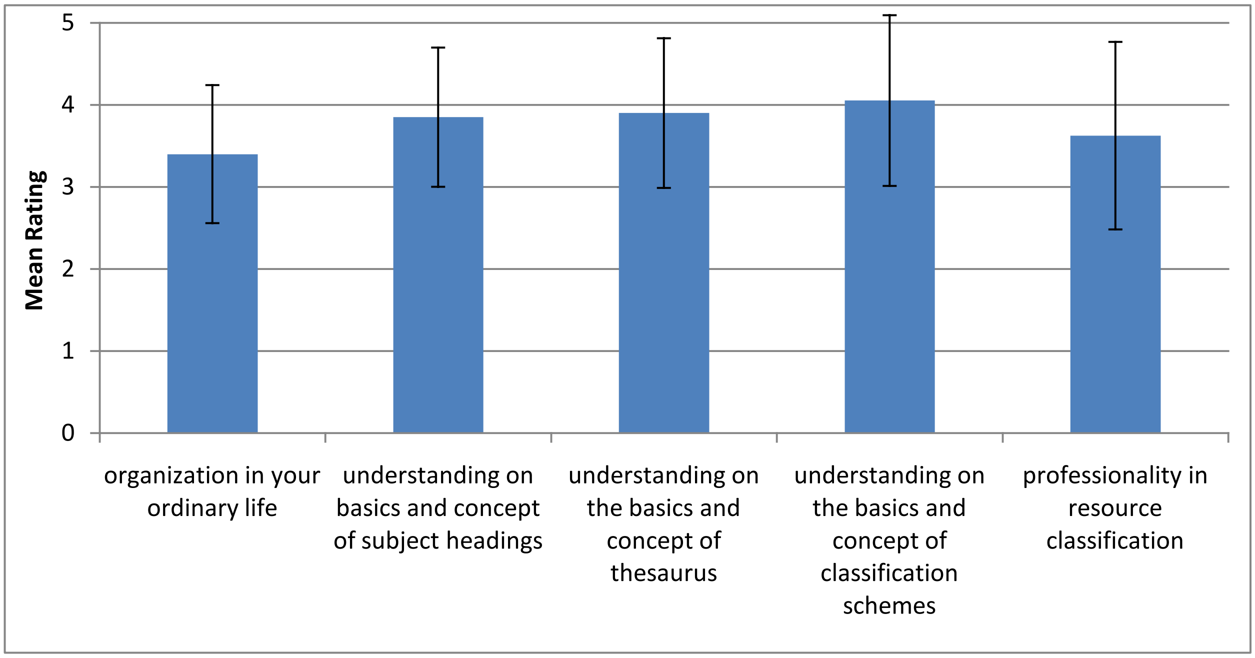


Figure 19. Self-rating on Participants Level of Understanding on Information Organization (Phase 1)

The measure the reliability of the inter-raters agreement statistically, the Fleiss' Kappa is used. Among various Kappa test methods, Fleiss' Kappa is selected since it is designed for multi-rater tests (Fleiss, 1971). The Fleiss Kappa represents the proportion of agreement among raters by chance – values between 1 and 0 indicate agreement better than chance, a value of 0 indicates a level of agreement that could have been expected by chance, values between 0 and -1 indicate levels of agreement that are worse than chance. However, Fleiss Kappa is dependent on marginal distribution that is the prevalence, which is not the case for many studies. Randolph (2005) has introduced a Free-Marginal Multirater Kappa (K_{free}) as an alternative to Fleiss' Kappa, in which raters' distributions of cases into categories are not restricted. Thus, we used K_{free} as a measurement to indicate the reliability of the agreement among participant judgments on how well the provided terms represent topics of a web page. Randolph's Free-Marginal Multirater Kappa (K_{free}) is calculated with the equation shown below where *N* is the number of cases, *n* is the number of raters, and *k* is the number of rating categories.

$$K_{free} = \frac{\left[\frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} {n_{ij}}^2 - Nn\right)\right] - [\frac{1}{k}]}{1 - [\frac{1}{k}]}$$

The K_{free} on the ratings of the provided classificatory metadata terms was 0.6068. Since the K_{free} value is a positive value, it indicates that the agreement of the participant judgments is better than what would have been expected by chance.

5.1.2 Analysis of Participants Judgments on the CRAD Ranges and Compound Tags Formats

To test the hypothesis for phase 1, a two-way mixed ANOVA was performed on the ratings of how well terms represent subject topics of a web page as a function of *CRAD* ranges (CR1, CR2, CR3) and compound tags format (decomposed, standardized). The pattern of differences on the ratings between compound tag formats among *CRAD* ranges was significantly different, *F*(2, 8914)=21.267, $p < .001$, partial $\eta^2 = .005$ (Figure 20). The standardized compound format is significantly higher in ratings than the decomposed format, *F*(1, 4457)=30.925, $p < .001$, partial $\eta^2 = .007$ (Table 11).

Table 11. Mean and Standard Deviation of the Ratings as a Function of Compound Tags Formats

	Mean	Std. Dev.
Decomposed	1.6	.017
Standardized Compound	1.75	.021

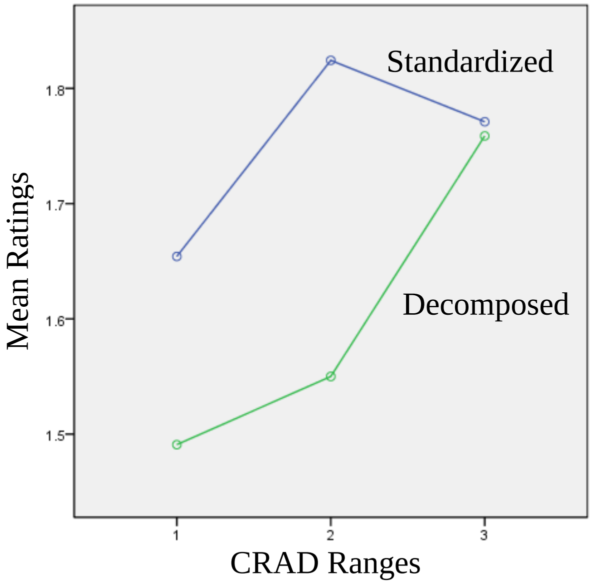


Figure 20. Estimated Marginal Means of CRAD ranges and Compound Tags Formats

Table 12. Mean and Standard Deviation of the Ratings as a Function of CRAD Ranges and Compound Tag Formats

	Standardized		Decomposed	
	Mean	Std. Dev.	Mean	Std. Dev.
CR1	1.654	0.026	1.491	0.021
CR2	1.824	0.028	1.550	0.023
CR3	1.771	0.029	1.759	0.024

In order to find the pattern of differences on the ratings among *CRAD* ranges on the standardized compound format, pair-wise differences are examined. There was a significant difference on the ratings between *CRAD* range 1 (CR1) and the average of ranges 2 and 3 (CR2 and CR3) for the standardized compound format, $F(1, 1838) = 27.619$, $p < .001$, partial $\eta^2 = .015$. Table 12 and Figure 20 represents that CR2 and CR3 have statistically higher significance in ratings for the standardized compound format.

According to the ANOVA and the pair-wise analysis on the ratings of classificatory metadata terms for the three ranges of *CRAD* values and the two formats of compound tags, it was found that the *CRAD* values in the range of 0.1815-0.2999 which covers 7-20% of the collection and the standardized compound format of decomposed compound tags, which re-combines the separated terms with a space, are the best applications of classificatory metadata terms from the tag set. As a result, the terms with *CRAD* values in the range of 0.1815-0.2999 are selected as a condition of phase 2 of the experiment. In addition, all compound tags are processed into the standardized compound format and the calculation of *AD* and *CRAD* follows accordingly.

5.2 Phase 2: Evaluation of the Generated Classificatory Metadata

The second phase of the experiment uses the tag data gathered from Delicious to propose the classificatory metadata tag terms and compares them with two different professionally generated classificatory metadata, the Open Directory Project and the INFOMINE. The Open Directory Project (ODP) is a web directory created by experts based on ODP's catalogue of the web. INFOMINE provides Library of Congress Subject Headings (LCSH) and subject keywords on web resources by experts. Twenty-five resources are selected randomly from the Delicious collection where they also exist in ODP and INFOMINE. The proposed classificatory metadata (*AD*CRAD*) for each resource is selected based on the highest *AD*CRAD* and the high *AD*ranged CRAD* with *CRAD* range of 0.1815-0.2999. For each web page, category labels and descriptions are collected from ODP and LCSH, subject keywords, and category are gathered from INFOMINE.

Twenty classification experts were assigned with 25 web pages and were asked to rate the provided terms to identify the subject domain as well as identify their familiarity of the web pages they are viewing and the type of the provided terms in representing the topics. Terms from the generated classificatory metadata tag terms and terms from INFOMINE and ODP are provided for the 25 web pages in random order. As the subjects in this experiment are experts, their relevancy ratings are considered to be perfect.

5.2.1 Participants Level of Professionalism and Consistency of Judgments

For phase 2 of the experiment, twenty participants were recruited. Among twenty participants, seven participants were librarians, one participant was Library Science doctorate holder, and twelve participants were Library Science students (4 masters and 8 doctorates). The Library Science students have taken 2.14 courses in average from listed six information related courses, including Organizing and Retrieving Information, Introduction to Cataloging and Classification, Advanced Cataloging and Classification, Metadata, Indexing and Abstracting, and Thesaurus Construction.

Participants were asked to self-rate on how well they perform information organization and understand classification concepts (Figure 21). The rating was on a scale of five – 1 indicating very bad, 2 indicating bad, 3 indicating fairly good, 4 indicating good, and 5 indicating excellent. In general, participants rated themselves to be good in resource classification professionally (in average 3.63). Specifically, they rated themselves on average 3.9 in understanding classification schemes, 3.95 in understanding thesaurus, and 3.9 in understanding subject headings. On the other hand, they rated themselves 3.55 on average for organizing in their ordinary life, for example, organizing a personal library, personal pictures, personal computer files and folders, bookmarks, emails/mails, documents, etc.

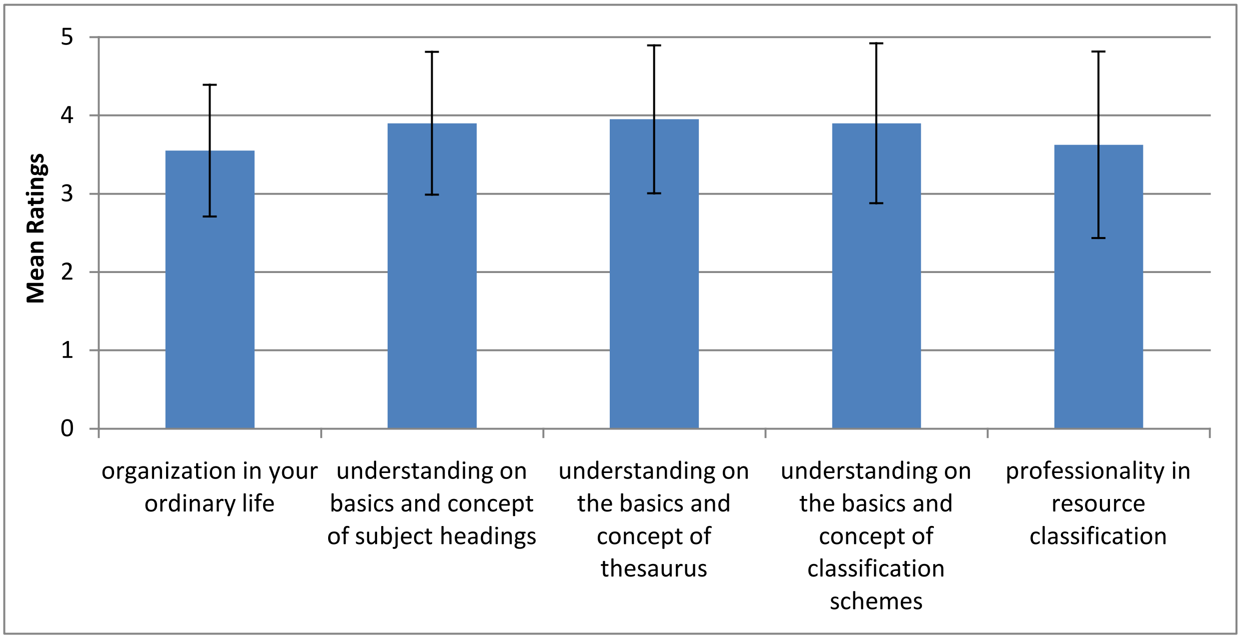


Figure 21. Self-rating on Participants Level of Understanding on Information Organization (Phase 2)

As described in 5.1.1, the Free-Marginal Multirater Kappa (K_{free}) (Randolph, 2005) is used to measure the reliability of the agreement among participant judgments on how well the provided terms represent topics of a web page. The K_{free} on the ratings of the provided classificatory metadata terms was 0.1345. Since the K_{free} value is a positive value, it indicates that the agreement of the participant judgments is better than what would have been expected by chance.

5.2.2 Relevance Measurement

For the evaluation, expert relevance judgments for each document are used. The terms from professionally created metadata and user assigned tags will be provided in a random order to the subjects. The relevance of keywords from experts and social tags will be measured using NDCG at K measurement. A group of experts as subjects of this study will rate how well each term represents the resource. The subjects' decision about relevance is considered perfect. Agichtein et al., (2006) proposed a modified Discounted Cumulative Gain (DCG) as a means to assess retrieval rating, called Normalized Discounted Cumulative Gain at K (NDCG at K). It is based on a prior work by Jarvelin and Kekalainen (2000). This metric is based on human judgments. Basically, human judges rate how relevant each retrieval result is on an *n*-point scale. For a given query *q*, the ranked results are evaluated from the top ranked down and the NDCG is computed as shown below, where *Mq* is a normalization constant calculated so that the perfect ordering would obtain NDCG of 1; each *r(j)* is an integer representing the relevancy rated by human judges (0 = "Not relevant at all" and 4="Perfect Relevant" at position *j*).

$$NDCG_q = M_q \sum_{j=1}^{K} \frac{(2^{r(j)} - 1)}{\log(1 + j)}$$

NDCG rewards relevant documents in the top ranked results more heavily than those ranked lower and punishes irrelevant documents by reducing their contributions to NDCG (Agichtein et al., 2006). We performed a similar ranking, but in this case, based on the relevance of each of the randomly proposed classificatory terms for the given resource.

5.2.3 Analysis of Participants Ratings on Classificatory Metadata Terms

The classificatory metadata terms list was created for each web resources in random order from the four conditions (high *AD*CRAD*, high *AD*ranged CRAD*, INFOMINE, and ODP). To test the hypothesis for phase 2, a one-way within-subject ANOVA was performed on the NDCG_{10} of terms to represent subject topics of a web page from Delicious tags and expert generated metadata. There was a significant difference on the NDCG_{10} depending on the proposed classificatory metadata terms (CM1 for high *AD*CRAD*, and CM2 for high *AD*ranged CRAD*) and the expert generated classificatory metadata terms (INFOMINE and ODP), *F*(3, 72) =

35.742, $p < .001$, $\eta^2 = .598$. In order to find the pattern of differences on the NDCG$_{10}$ depending on the classificatory metadata terms, post hoc pair-wise comparisons were performed. The NDCG$_{10}$ of the proposed classificatory metadata terms (including high *AD*CRAD* and high *AD*ranged CRAD*) was significantly higher than that of the expert generated classificatory metadata terms (INFOMINE and ODP), $p < .001$ (Table 14). There was no significant difference between the proposed classificatory metadata terms from high *AD-CRAD* and high *AD*ranged CRAD*. However, there was a significant difference between the expert generated classificatory metadata terms from INFOMINE and ODP, $p < .001$, INFOMINE being significantly higher (Table 13). It can be understood that since directories have defined categories of subjects, some of the pre-defined categories do not necessarily represent the topics of particular web resource.

Table 13. The Mean and Standard Deviation of the NDCG$_{10}$ for the Proposed and Expert Generated Classificatory Metadata Terms

	Mean	Std. Dev.
CM1 (high *AD*CRAD*)	.9465	.070
CM2 (high *AD*ranged CRAD*)	.8962	.131
INFOMINE	.8206	.168
ODP	.5490	.176

Since NDCG measures the effectiveness of a result list based on the position in the list, it can be interpreted from the NDCG and the ANOVA test that the *AD*CRAD* and high *AD*ranged CRAD* generates a list of the classificatory metadata based on their representativeness. However, NDCG cannot fully represent how well the proposed classificatory terms indicate the topics of web resources. For further analysis of the proposed classificatory metadata terms, one-way Analysis of Variance test was performed on the ratings of the four conditions – high *AD*CRAD*, high *AD*ranged CRAD*, INFOMINE, and ODP. There was a significant difference on the ratings of terms depending on the proposed classificatory metadata terms (CM1 for high *AD*CRAD*, and CM2 for high *AD*ranged CRAD*) and the expert generated classificatory metadata terms (INFOMINE and ODP), $F(3,\ 14937) = 779.028$, $p < .001$, $\eta^2 = .135$. It is mainly due to the difference of high *AD*ranged CRAD* since it is significantly lower that other conditions, $F(1,\ 4979) = 2291.736$, $p < .001$, $\eta^2 = .315$ (Table 14 and Figure 22).

Table 14. The Mean and Standard Deviation of the Rating for the Proposed and Expert Generated Classificatory Metadata Terms

	Mean	Std. Dev.
CM1 (high *AD*CRAD*)	3.09	1.292
CM2 (high *AD*ranged CRAD*)	2.14	1.111
INFOMINE	3.03	1.279
ODP	2.77	1.328

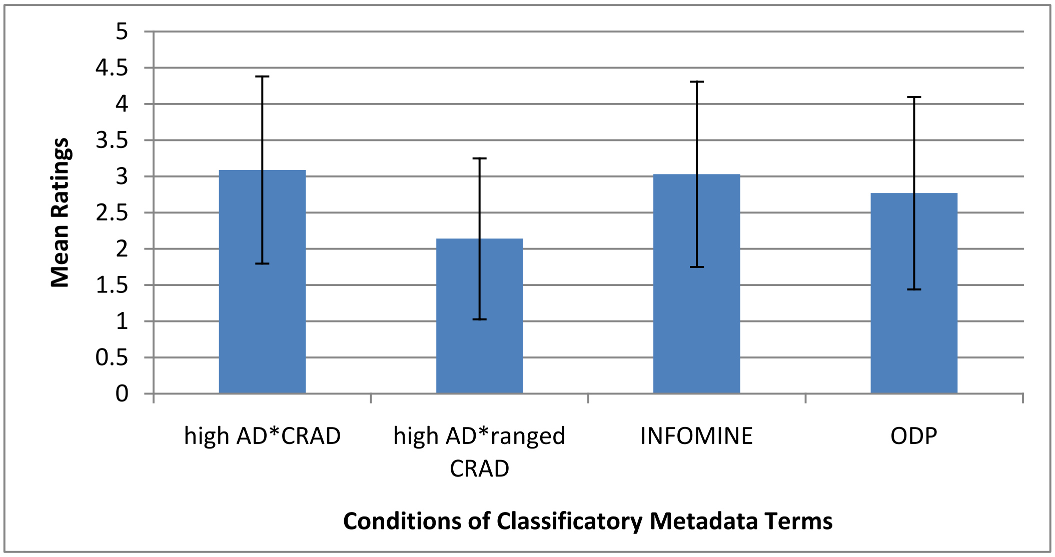


Figure 22. Mean of the Ratings on the Classificatory Metadata Terms

On the other hand, the ratings of proposed classificatory metadata terms by high *AD*CRAD* had no significant difference with the rating of the classificatory metadata terms from INFOMINE that are from Library of Congress Subject Headings and subject keywords. There still was a significant difference between the rating of the proposed classificatory metadata terms based on high *AD*CRAD* and the expert generated classificatory metadata terms from ODP that are mainly from category labels, $F(1, 4979) = 214.438$, $p < .001$, $\eta^2 = .041$. From this part of the analysis, it can be interpreted that the classificatory metadata terms proposed by high *AD*CRAD* are closer to the subject keywords and subject headings assigned to the web pages by experts.

To understand the results from Table 14 and Figure 22 further, the participants' indications on the types of information each term represents is analyzed. During the experiment, the participants were also asked to assign the types of information the terms indicate as metadata information from "Topical Subject" for subject terms, "General Category" for higher concepts, "Resource Type" for information sources and resource formats, "Others" for terms that are not topical subject, general category, or resource type, but are related to the web page, and "Not Applicable" for terms that cannot be assigned to a type and are not related to the web page at all. Figure 23 represents the percentage of each type of terms for the four conditions. It is notable that metadata terms from high *AD*CRAD*, INFOMINE, and ODP have high percentage of topical terms and general concept terms, whereas metadata terms from high *AD*ranged CRAD* have much less topical terms and relatively more general concept terms, resource type terms, and other types of terms. Since high *AD*ranged CRAD* proposes terms that covers 7-20% of the collection, Figure 23 indicates that the terms proposed by high *AD*ranged CRAD* may have potential in describing general topics and/or the resource type of web resources rather than describing the particular topics of web resource contents.

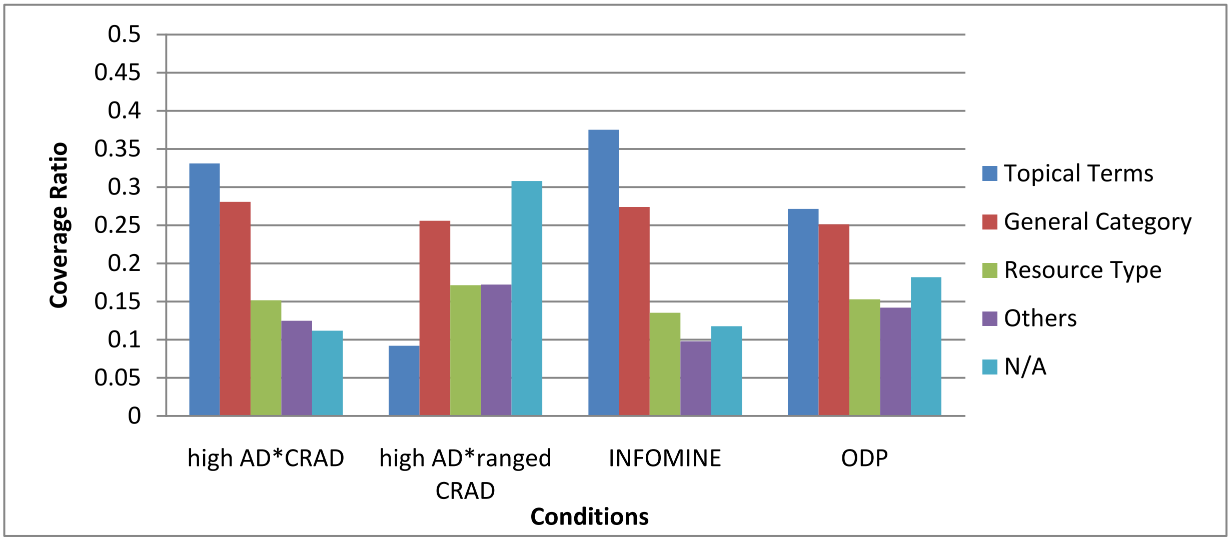


Figure 23. The Coverage of Types of Terms for the Four Conditions

The results of a two-way within-subject ANOVA test on ratings as a function of the types of the terms with Huynh-Feldt adjustment showed that the patterns of differences on the ratings on the classificatory metadata terms among the types of terms (topical terms, general category, resource type, others, and n/a) were significantly different among the four conditions (high *AD*CRAD*, high *AD*ranged CRAD*, INFOMINE, and ODP), $F(10.945, 4465.465) = 3.047$, $p < .001$, and $\eta^2 = .007$ (Table 15 and Figure 24). There was a significant difference on the ratings among the types of the classificatory metadata terms averages across the conditions, adjusted with Huynh-Feldt, $F(3.660, 1493.609) = 1328.109$, $p < .001$, and $\eta^2 = .765$. There was a significant difference on ratings among the four conditions averages across the term types, $F(3, 1224) = 68.327$, $p < .001$, and $\eta^2 = .143$. Apparently, the topical terms were significantly higher in rating measurement than other types of terms, $F(1, 408) = 2284.001$, $p < .001$, and $\eta^2 = .848$. On the other hand, the general category terms were significantly lower in rating than topical terms and resource type terms, $F(1, 2402) = 821.734$, $p < .001$, $\eta^2 = .255$ and F $F(1, 2402) = 4.984$, $p = .026$, $\eta^2 = .002$ respectively.

Table 15. The Mean and Standard Deviation of the Ratings for the Types of Terms and the Four Conditions

	High AD*CRAD		High AD*ranged CRAD		INFOMINE		ODP	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Topical Terms	3.92	1.041	3.40	1.134	3.66	1.148	3.78	1.095
General Category	3.04	1.033	2.58	0.957	2.88	1.075	2.72	1.113
Resource Type	3.14	1.166	2.62	1.000	3.09	1.093	3.05	1.152
Others	2.47	0.98	2.03	0.807	2.33	0.857	2.35	0.879
N/A	1.48	0.664	1.30	0.622	1.45	0.651	1.36	0.548

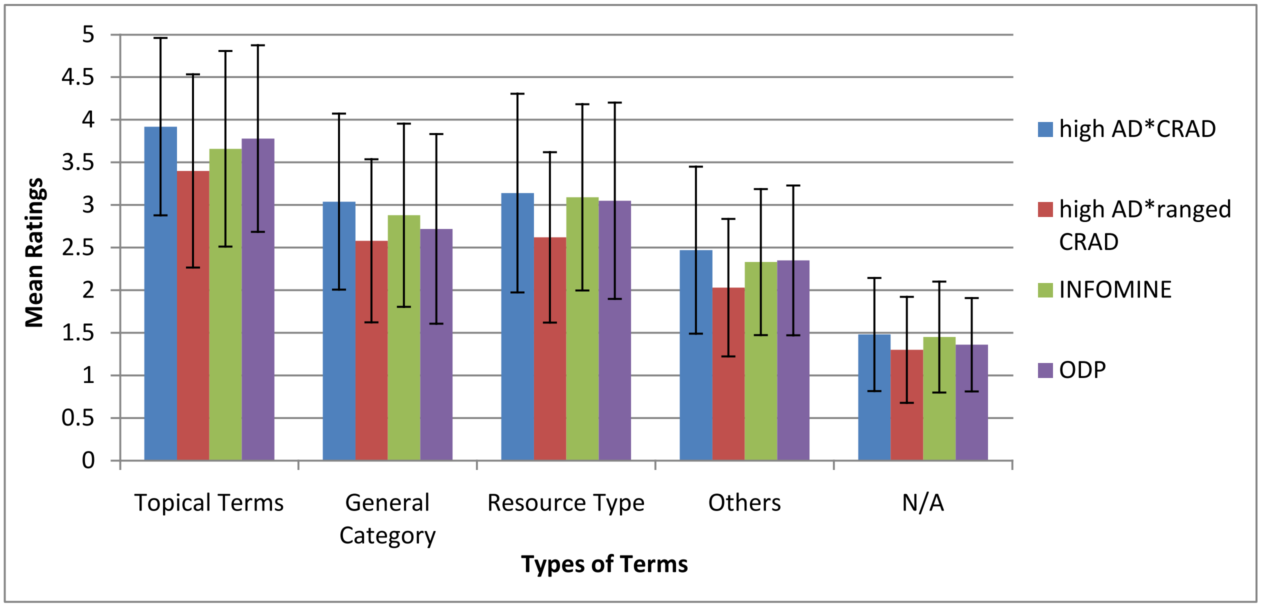


Figure 24. Estimated Marginal Means of Ratings for the Types of Terms and the Four Conditions

Figure 24 represents that participants rated topical terms highly relevant to the subject topics of web pages, but not as highly for general concept terms and resource type terms. Since the task given to the participants was to rate based on their judgment of how well the terms represent the topic of the web page, it can be understood that participants considered general concept terms and resource type terms somewhat related to web pages but did not directly represent the topics of the contents. Participants showed consistency when they answered the exit survey asking about their strategies in rating the proposed classificatory metadata terms. The most favored strategies were: title of the web page, categories of the topic concept, words used in the content and frequently appearing words, and type of the web page (all agreed to over 90% of participants). It revealed that participants concentrated more on the contents to find the topics rather than considering the classificatory structure. On the question about what to rate bad, participants answered if the term does not represent the content and/or topic and if the term describes too broad of a domain of the subject area, they rated the term to be not relevant to the topics of the web pages (all agreed to by over 90% of participants). The results from exit survey support the result from the experiment that the general concept terms did not to represent the subject topics as defined by the participants. It also explains the results from Figure 22 and 23 – as terms proposed by high *AD*ranged CRAD* did not include as many topical terms and more general concept terms and resource type terms, the ratings for the terms from high *AD*ranged CRAD* resulted to be significantly lower than other three conditions.

5.3 Summary of the Results

The first phase of the experiment explored issues related to the preliminary studies on *CRAD* values and compound tags. From the preliminary study on compound tags, it was found that a large portion of the tag set collection included various forms of compound tags. Thus, it was expected that when the compound tags were standardized, the importance of the phrase as a tag would be increased and used as a significant description of the targeted web resources. At the same time, the preliminary observations on *CRAD* values showed a possibility of finding better classificatory metadata since *CRAD* values represent the coverage of a tag on the collection.

Since the topical domains have to be covered by some portion of the collection to represent the topics, it was one of the main objectives of phase 1 to find the most applicable range of the *CRAD* to find terms for the classificatory metadata.

The results of phase 1, as expected, showed that the standardized format of compound tags were considered to represent the topics significantly better than the decomposed terms represented in the single term format. When the compound tags were standardized, the analysis in phase 1 suggested that for the size of test collection (7,097 resources), the *CRAD* values that cover 7-20% of the collection represent the topics of web pages significantly better than *CRAD* values that cover 1.5-7% of the collection. Based on the result of the phase 1 analysis, the format of compound tags were standardized and the terms that were in the range of *CRAD* values of 0.1815-0.2999 covering 7-20% of the collection were included as a condition for the second phase of the experiment.

The second phase was designed to examine how well the proposed *AD* and *CRAD* measurements produce good topic descriptors from the tag set. We proposed four conditions to compare – high *AD**CRAD* weighted terms, high *AD*ranged CRAD* weighted terms, expert generated subject terms from INFOMINE and expert generated subject terms Open Directory Project (ODP). The hypothesis was made to find whether the high *AD*ranged CRAD* would work to find the classificatory metadata, and either *AD*CRAD* or high *AD*ranged CRAD* would work better or as well as the expert generated classificatory metadata. The simple comparison between the proposed classificatory metadata terms and the expert generated classificatory metadata terms showed that there is some overlap in the term selection between experts and non-experts in describing the web resources as previous studies have shown (Lin et al., 2006; Syn and Spirng, 2009; Yi and Chan, 2009).

The NDCG$_{10}$ was measured to evaluate the relevance to the topics. The analysis represented that both high *AD*CRAD* and high *AD*ranged CRAD* performed well in presenting the relevance as evaluated by the expert participants. In addition, the terms from high *AD*CRAD* were evaluated to represent the topics as well as expert generated subject descriptions (INFOMINE). However, even though the high *AD*ranged CRAD* values represented the relevance well among the terms selected, as a selected set the participants' ratings in judging their representativeness of the topics were significantly lower compared to other conditions. Since the high *AD*ranged CRAD* was expected to represent topic domain categories in the collection, we further analyzed how participants identified the type of terms proposed by high *AD*ranged CRAD*. The categorization of terms by participants indicated that terms from high *AD*CRAD*, INFOMINE, and OPD were topical terms and general concept terms. On the other hand, as expected, the high *AD*ranged CRAD* included fewer topical terms and more of other types of terms – general concept terms, resource type terms, and others. The analysis on rating by the types of terms showed that the ratings by participants for general concept terms are significantly lower than that of topical terms and resource type terms. The exit survey also revealed that when expert participants make judgments on a term about its relevance of the topic of the resource, they rely on the relationship of the term with the content mostly and consider the terms that represent broader concepts to be bad terms to represent the topic. The results from analysis of ratings by the types of terms and the feedback from the exit survey would seem to explain the devaluation of the terms from high *AD*ranged CRAD*. They also explained how well high *AD*CRAD* performed in emphasizing the terms that participants considered to be a good description of the topics of a resource.

Different from the high *AD*ranged CRAD* terms, the high *AD*CRAD* terms were evaluated to represent the topics better than the expert generated classificatory metadata terms and the *AD*CRAD* values are evaluated to represent the relevance well. Similar to what was observed for the terms proposed from high *AD*ranged CRAD* from the relationship with the portion of types of the terms, it can be explained that one of the reasons for high *AD*CRAD* performing well is because it consists of what expert participants considered to be topical terms. In fact, the ANOVA test results showed that the participants rated the high *AD*CRAD* terms higher than terms from expert generated classificatory metadata.

6.0 Discussion

6.1 Contributions and Implications

This dissertation analyzed social tags to determine the potential of using them in metadata generation based on tags provided by non-professional users. Given the creation process, user-generated tags for web resources tend to include a lot of noise (Guy and Tonkin, 2006). One goal of this study was to find a way of selecting the tags that can represent the subject topics of the web resource, i.e., the classificatory metadata. The major issues were:

- Can the tag noise be reduced?
- Can compound tags be processed to be of use?
- Can a subset of tags be found that provide classificatory metadata?

As a way to address the issues, two metrics, *Annotation Dominance (AD)* and *Cross Resources Annotation Discrimination (CRAD)*, were proposed. *AD* and *CRAD* measures might be used to filter tag noise out and generate a tag set optimized for classificatory metadata. In addition, efforts were made to process compound tags by creating an emerging term dictionary and decomposing compound tags based on observations made on the test data set and as suggested by other researchers (Guy and Tonkin, 2006; Tonkin, 2006). The emerging term dictionary helps in identifying emerging terms frequently used as tags. It also helped decompose compound tags. The process of decomposition for compound tags was necessary since there was a large number of compound tags in the tag set that were clearly composed of good terms. From preliminary studies, it was observed that the *CRAD* values represent the coverage of tag terms in the collection. Our assumption was the *CRAD* values would help find the better classificatory metadata since classificatory metadata includes domain categories.

Based on the preliminary studies, we evaluated the standardized format of decomposed compound tags and found the range of the *CRAD* values that would help find the better classificatory metadata terms. Although several studies have suggested disambiguating compound tags to meaningful terms better (Guy and Tonkin, 2006; Lin et al., 2006; Tonkin, 2006; Yi and Chan, 2009), the format of decomposed compound tags was not defined in the previous research. The result of the first phase showed that the adoption of the standardized format for decomposed compound tags represents the topics of web resources better than representing them in a single word format. In addition, it was suggested that terms that covers 7-20% of the collection best represented topics for the web resources.

A controlled experiment on *AD* and *CRAD* measurements compared with the expert generated classificatory metadata was performed. The high *AD**CRAD* terms performed well both in representing the subject topics and indicating the relevancy of topics. The high *AD**ranged CRAD* terms represent general concepts, resource types, and other types of information, and thus were evaluated to be less applicable for describing the subject topics. However, there is still a suspicion that high *AD**ranged CRAD* terms may help describe other types of information for a resource and may be useful as classificatory metadata.

Although social tagging systems opened a method to involve users in metadata generation (Heymann et al., 2008; Macgregor and McCulloch, 2006; Quintarelli, 2005; Sen et al., 2007; Trant, 2006), due to the large amount of the tag noise it was often asked how social tags can be used as metadata. This dissertation presents a method for finding the classificatory metadata from social tags of web resources. From the evaluation made for high *AD*CRAD*, the quality of the proposed classificatory metadata as the subject descriptor could be considered to fit to the expectation of the expert cataloguers.

6.2 Future Work

This research confirms the potential of using social tags as classificatory metadata by proposing metrics to filter tag noise. However, there are more research questions that need to be explored related to using social tags in finding metadata information.

First, since the high *AD*ranged CRAD* appears to represent other types of terms rather than topical terms, the quality of the high *AD*ranged CRAD* as a representation of different types of terms needs to be conducted. It is worth investigating whether classificatory metadata can include other types of information such as general concept, resource types, etc (Caplan, 2003; Cardoso and Sheth, 2006; Smiraglia, 2005). Once the quality of the high *AD*ranged CRAD* terms is studied, the high *AD*CRAD* terms and the high *AD*ranged CRAD* terms may be able to generate general and specific concepts of a web resource.

Second, to increase the quality of the proposed classificatory metadata, supplementation or adjustment with existing subject headings and the thesaurus can be studied. Studies have indicated the potential in using existing controlled vocabularies to find useful tags (Lin et al., 2006; Syn and Spring, 2009; Yi and Chan, 2009). From the classificatory metadata from the second phase, it was observed that, overall, 9.89% of the terms from Delicious (high *AD*CRAD* and high *AD*range CRAD*) overlapped with the expert generated terms from INFOMINE and

ODP. Table 16 represents that, although small in portion, the existence of overlapping classificatory metadata terms between the tag exported metadata and the expert generated metadata opens possibilities for expanding the vocabulary and relating general-specific concepts to the current proposed classificatory metadata. Related to the first future work suggestion, adding information from existing controlled vocabulary may help improve the classificatory metadata proposed by *AD* and *CRAD*.

Table 16. Overlap Ratio between the Classificatory Metadata Terms from Delicious and the Experts

	Overlapping with	Overlap Ratio
High *AD*CRAD* Terms	INFOMINE	0.0441
	ODP	0.0703
High *AD*ranged CRAD* Terms	INFOMINE	0.0072
	ODP	0.0203

Third, in improving the two measurements, the third element of the tuple (users) can be included as a factor into the measurement. The current measurements include users as a factor; however, the effect is minor. Since the user is one of the tuple and plays an important role in social tagging systems (Hotho et al, 2006a, 2006b; John and Seligmann, 2006; Mika, 2007; Ohmukai et al., 2005), it can be considered as a significant factor to improve the effect of the two measurements. For example, by identifying affinity networks of users, it might be possible to identify more consistent sets of terms.

Fourth, the measurements can be applied and tested to other types of resources with tags, such as images, video, blogs, etc. As Bischoff et al. (2008) indicated, the types of information provided by tags depends on the type of resource, e.g. tags for music include terms to indicate genre, tags for picture include terms for location, etc. In this study, we observed and evaluated *AD* and *CRAD* for web documents (mainly text possibly with images and multimedia). However, whether *AD* and *CRAD* are general measures that can be applied to other types of resources will require additional study.

Appendix A. Tag Proportion Stability

We tested tags stability patterns over the resource collection. We observed how the proportion of tags in a collection stabilize as the collection size grows from 1 to 7388. Figure 25 represents the proportion of tags' occurrences over the resource collection for 30 randomly selected tags. The selected tags are: ajax; app; ayudas; biblioteca; bookmark; Bookmarks; desarrollo_web; design; design,; Design.Style; free; GraphicResources; Great; images; javascript; javascripts; links; music; nonflash; Program; programming; Programming.js; programming.languages.javascript; snippet; socialmedia-tools; software; tagging; tools; web2.0; webdesign. They include both popular and non-popular tag terms. The figures show that tag occurrences stabilize as the size of collection grows.

Before they stabilize, the appearance of tags varies depends on the resources added to the collection (see the red boxes in the Figure 25). Since the collection is incremented with randomly selected resources, the proportion of tags changes at different iterations depending on the order of the resources added. However after the collection reaches to certain size, the proportion of tags stabilizes and represents a similar pattern.

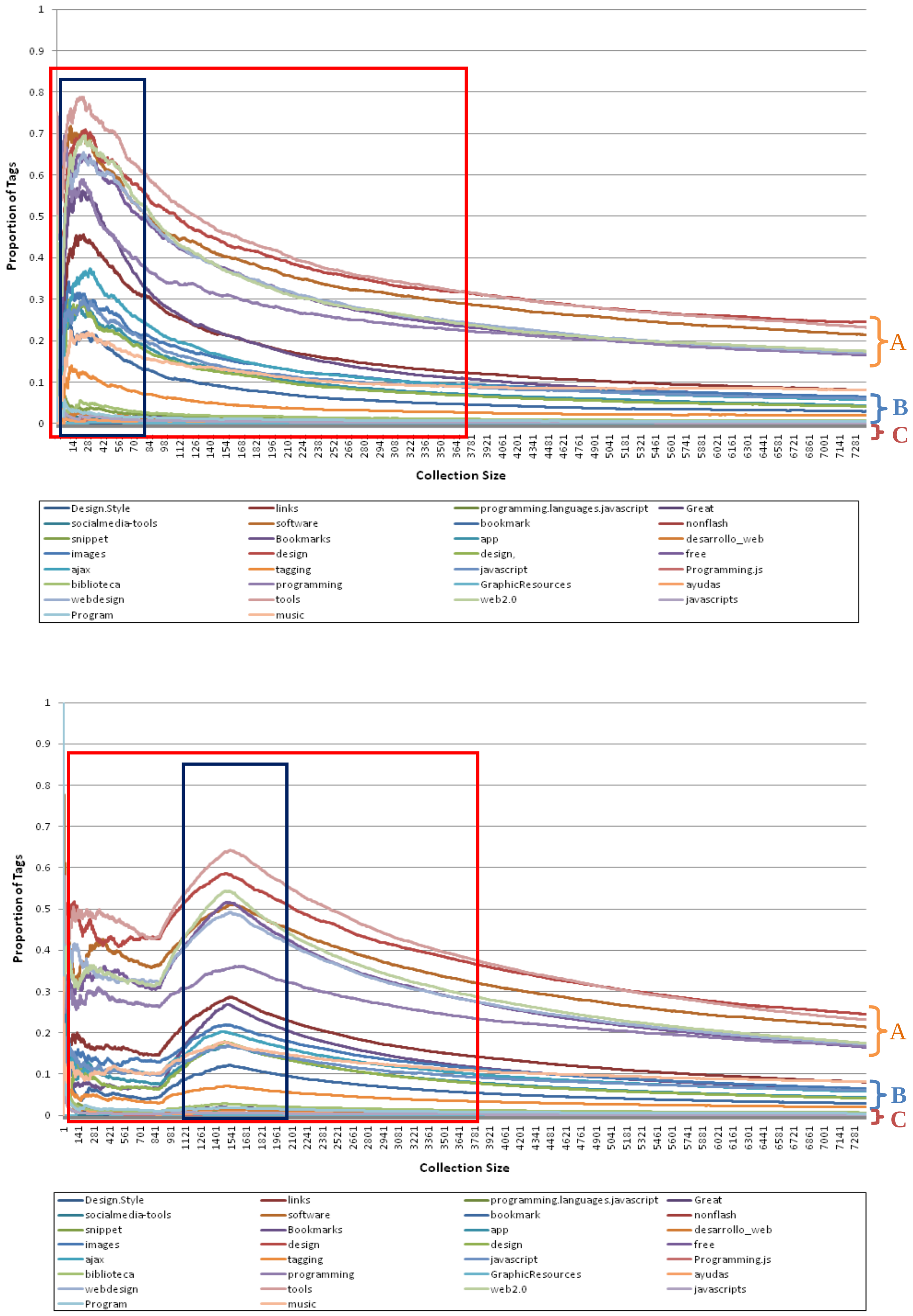
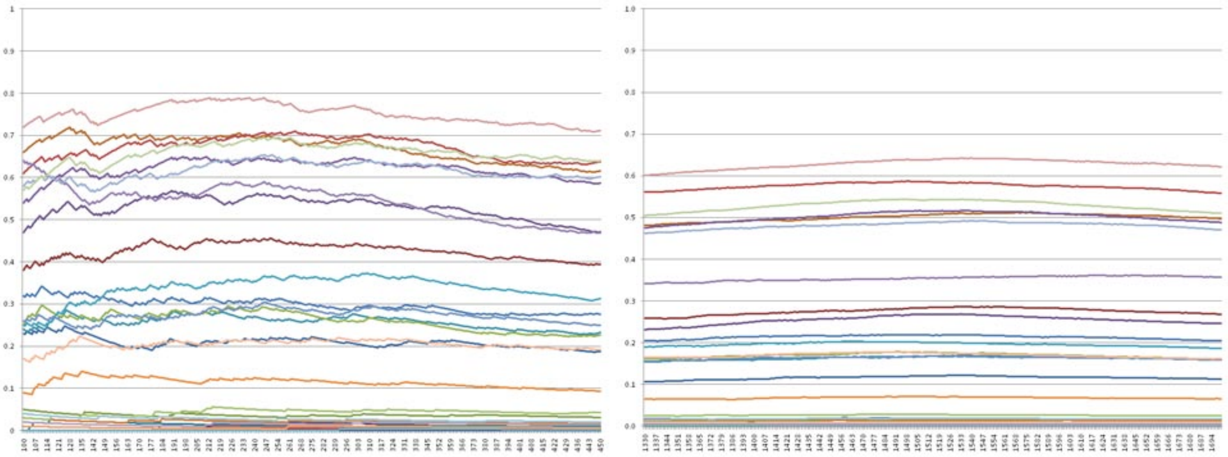


Figure 25. Proportion of Tags for Sample Collection in Different Iteration

After the proportion of tags starts to show stability, it shows three clearly divided groups of tags – popular tags, unpopular tags (idiosyncratic tags), and often-used tags. Popular tags that occur approximately from 15% to 25% in the graph fall into the broad folksonomy (‘A’ in Figure 25), unpopular tags that occur near 0% (‘C’ in Figure 25), and often-used tags that occur less than 10% in the graph (‘B’ in Figure 25) can be defined as the narrow folksonomy. Regardless of the order that the document is added to the collection, the groups were formed identically after the stabilization occurred. It is important that the observation of the three groups were clearly detected in this analysis. Our concern in identifying classificatory metadata for certain resource is how to discern popular and often-used tags as the candidates of classificatory metadata terms and how to exclude un-popular tags from the candidates of classificatory metadata terms.



a. Top: Ranging 100-450 Resources **b. Bottom: Ranging 1330-1700 Resources**

Figure 26. Proportion of Tags in Peak Area for Sample Collection in Different Iteration

Additional observation is made on the identical pattern of a peak on both graphs (shown in the blue boxes in Figure 25). Figure 26 shows a closer look at the peak area of the graph. The graphs in Figure 26 represents the increase in proportion of tags made in the growth of resources in about 351-370 documents is not extremely large as it appears in Figure 25. In addition, the increase in the proportion of tags is not made on particular resources. Nonetheless, the interesting phenomenon is that all of 30 selected tags tend to become high at the peak area (blue boxed area) regardless of the iteration. The particular documents added into the collection for the two iterations in the peak area are compared to provide a clear reason. There were 93 web pages overlapping in both peak areas (about 26%). Table 17 shows the list of 628 URLs in both peak areas that includes a relatively large amount of technical related documents. Considering the selected tags for this analysis include many technical terms such as ajax; design; javascript; programming; snippet; software; tools; web2.0; webdesign, it somewhat explains why the peak appears in both graphs. Therefore, the possible interpretation of this pattern is that, although the document is added to the collection in a random manner and since there are so many technical-related documents, and thus more tags, at some point, those resources were added closer together and formed the peak in the graph.

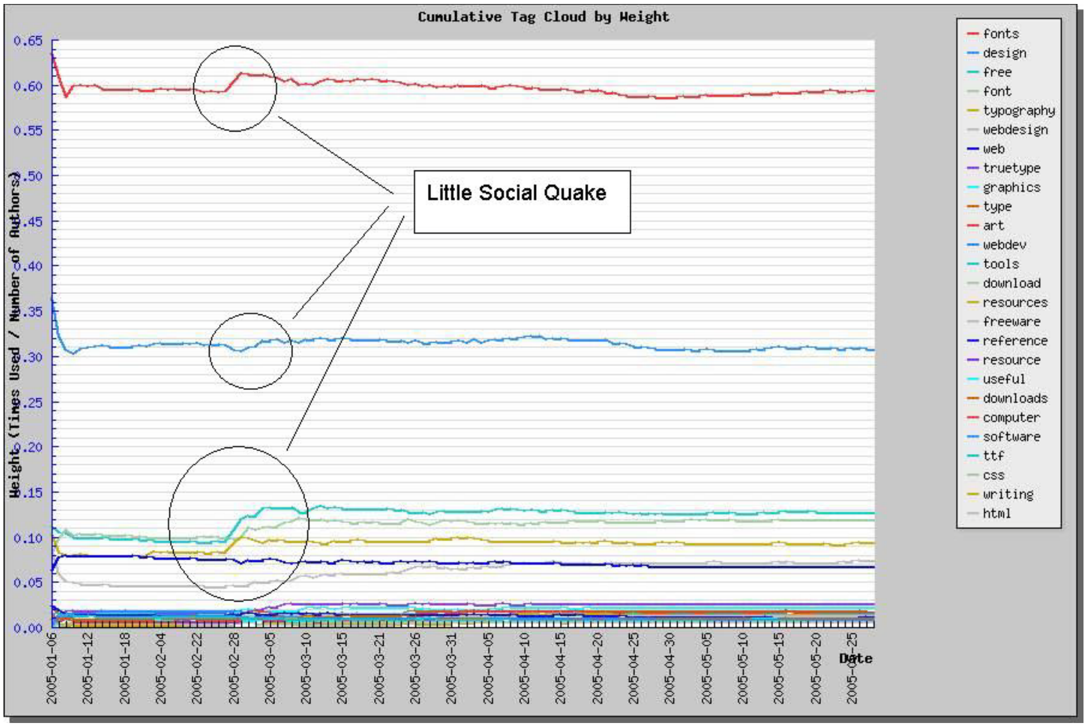


Figure 27. Cumulative Tag Cloud Over Time Showing a Social Quake for Webpage "Essential Fonts for Designers" (http://www.goodfonts.org/) (Di Fenizio, 2005)

Di Fenizio (2005) describes this type of pattern related to "cultural changes." He observed the agreement on the tags by users over time (Figure 27). With the observation we made above, we can expect that the proportion of agreement stabilizes too. Similar to what we have seen in our observation, there was a rise in the pattern at certain point. His two possible explanations are: 1) the bookmark became popular (it was already public before, but not well known), and people started to use more tags, 2) the link was handed to a subculture which tended to use on average more tags for each post. Since this observation was made on a particular web page over time, Di Fenizio's explanation cannot be directly applied to our case. However, we could consider the "cultural changes" as another possible cause assuming this pattern would also appear in a collection growing in real settings.

Although different observations were made based on the analysis on the proportion of tags over the collection size, our focus here is to understand that there were three groups of tags – popular tags, often-used tags, and unpopular tags (idiosyncratic tags). For classificatory metadata, we are sure that unpopular tags are not our concern. We will only filter out popular and often-used tags as the candidates of classificatory metadata terms.

Table 17. List of Resources in Peak Area

URL	Title
http://ya.ru/	Яндекс
http://www.topcoder.com/tc	TopCoder
http://www.google.ru/	Google
http://python.net/~goodger/projects/pycon/2007/idiomatic/handout.html	Code Like a Pythonista: Idiomatic Python
http://python.net/%7Egoodger/projects/pycon/2007/idiomatic/handout.html	Code Like a Pythonista: Idiomatic Python
http://nant.sourceforge.net/	NAnt - A .NET Build Tool
http://www.mozilla.com/products/firefox/central.html	Firefox Central
http://en-us.start.mozilla.com/firefox	Firefox Start Page
http://www.spoj.pl/	Sphere Online Judge (SPOJ)
https://msdn.microsoft.com/en-us/subscriptions/securedownloads/default.aspx	Download - Home page
http://www.facebook.com/inbox/	Facebook | Inbox
http://developer.mozilla.org/en/docs/XUL_Reference	XUL Reference - MDC

http://developer.mozilla.org/en/docs/Main_Page	Main Page - MDC
http://developer.mozilla.org/en/docs/Gecko_DOM_Reference	Gecko DOM Reference - MDC
http://developer.mozilla.org/en/docs/Building_an_Extension	Building an Extension - MDC
http://drupal.org/node/193318	Zen
http://drupal.org/handbook/customization/tutorials/beginners-cookbook	The Drupal Cookbook
http://www.randsinrepose.com/	Rands In Repose
http://www.w3.org/2001/03/webdata/xsv	XSD Validator
http://icpcres.ecs.baylor.edu/onlinejudge/	UVa Online Judge - Home
http://www2.toki.or.id/book/AlgDesignManual/BOOK/BOOK/BOOK.HTM	The Algorithm Design Manual
http://acm.timus.ru/	Timus Online Judge
http://www.jair.org/	JAIR
http://www.microsoft.com/isapi/redir.dll?prd=ie&amp;pver=6&amp;ar=CLinks	Customize Links
http://www.microsoft.com/isapi/redir.dll?prd=ie&amp;ar=windowsmedia	Windows Media
http://www.microsoft.com/isapi/redir.dll?prd=ie&amp;ar=hotmail	Free Hotmail
http://www.microsoft.com/isapi/redir.dll?prd=ie&amp;ar=windows	Windows
http://go.microsoft.com/fwlink/?LinkId=30857&amp;clcid=0x409	Windows Marketplace
http://clien.career.co.kr/	□□□□ □□□□ □□□□□!!!
http://www.voidtools.com/	Everything Search Engine
http://www.faceyourmanga.com/faceyourmanga.php?lang=eng	FaceYourManga.com | Shake Yourself!
http://www.bugzilla.org/	Home :: Bugzilla :: bugzilla.org
http://www.worldwidefred.com/home.htm	Fred&#039;s Home
http://etl.stanford.edu/	MS&amp;E 472 - Entrepreneurial Thought Leaders Seminar Series
http://www.egofoto.net/site.html	egofoto / Şenol Zorlu
http://www.ruby-toolbox.com/	The Ruby Toolbox: Know your options!
http://www.exampledepot.com/egs/index.html	Examples from The Java Developers Almanac 1.4
http://www.microsoft.com/DOWNLOADS/details.aspx?familyid=22E69AE4-7E40-4807-8A86-B3D36FAB68D3&amp;displaylang=en	Download details: Consolas Font Pack
http://hivelogic.com/articles/view/ruby-rails-leopard	Hivelogic - Installing Ruby, Rubygems, Rails, and Mongrel on Mac OS X 10.5 (Leopard)
http://drnicwilliams.com/2008/01/31/get-ready-for-the-textmate-trundle-to-rails-20-bundle/	Dr Nic ’s Get ready for the TextMate “Trundle to Rails 2.0 Bundle”
http://rubyosx.rubyforge.org/	rubyosx - Ruby One-Click Installer for OSX
http://mac.appstorm.net/roundups/iphone-roundups/30-iphone-apps-with-sexy-interfaces/	30 iPhone Apps with Sexy Interfaces « AppStorm
http://java.sun.com/blueprints/corej2eepatterns/Patterns/ServiceLocator.html	Core J2EE Patterns - Service Locator
http://www.cyberciti.biz/faq/mysql-change-root-password/	MySQL Change root Password
http://www.iphoneos.co.kr/	KIDG :: iPhone □□□ □□□□
http://www.smashingmagazine.com/2008/09/03/40-creative-design-layouts-getting-out-of-the-box/	40 Creative Design Layouts: Getting Out Of The Box | Design Showcase | Smashing Magazine
http://allseeing-i.com/ASIHTTPRequest/	ASIHTTPRequest Documentation - All-Seeing Interactive
http://lifeonrails.org/2007/8/30/netbeans-the-best-ruby-on-rails-ide	Netbeans THE best ruby on rails IDE
http://thinkvitamin.com/features/20-steps-to-better-wireframing/	20 Steps to Better Wireframing | Think Vitamin
http://www.markforster.net/autofocus-system/	Autofocus System - Get Everything Done
http://www.sony.jp/products/Consumer/handycam/camwithme/main.html	Cam with me□□□ □□□ □ー）｜デジタルビデオカメラ Handycam “□□□□□□” | □□ー
http://icpcres.ecs.baylor.edu/onlinejudge/index.php	UVa Online Judge - Home

http://www.livemocha.com/	Language Learning with Livemocha | Learn a Language Online - Free!
http://mind42.com/	Mind42.com - Collaborative mind mapping in your browser
http://www.photoshoplady.com/	Photoshop Lady : Best Photoshop Tutorials Around the World
http://labs.ideeinc.com/multicolr/	Multicolr Search Lab - Idée Inc.
http://www.gliffy.com/	Gliffy.com - Create and share diagrams online.
http://www.findsounds.com/types.html	FindSounds - Sound Types
http://tides.ws/2007/10/15/most-powerful-and-unforgettable-images-from-around-the-world/	Most Powerful and Unforgettable Images from around the World
http://tutorialblog.org/free-vector-downloads/	» Free Vector Downloads
http://zenhabits.net/	Zen Habits | Simple Productivity
http://www.pdf-mags.com/	pdf-mags.com - Your PDF mag’s magazine
http://posterous.com/	Posterous - The place to post everything. Just email us. Dead simpl...
http://ilovetypography.com/	Typography. I Love Typography, devoted to fonts, typefaces and all ...
http://www.alvit.de/handbook/	Web Developer&#039;s Handbook | CSS, Web Development, Color Tools, SEO, ...
http://www.alextrochut.com/	Alex Trochut - Creativity, Type &amp; Illustration.
http://search.twitter.com/	Twitter Search
http://www.behance.net/	Behance Network :: Gallery
http://tweetdeck.com/beta/	TweetDeck
http://www.jamendo.com/en/	Jamendo : Open your ears
http://www.bittbox.com/	BittBox
http://wordle.net/	Wordle - Beautiful Word Clouds
http://www.brusheezy.com/	Free Photoshop Brushes at Brusheezy!
http://www.fullyillustrated.com/	Fully Illustrated - The Portfolio of Michael Heald
http://www.apple.com/quicktime/tutorials/texttracks.html	Apple - QuickTime - Tutorials - Text tracks
http://torrentz.com/	Torrents Search Engine
http://twitter.com/	Twitter: What are you doing?
http://mozy.com/	Mozy Online Backup: Free. Automatic. Secure.
http://www.ohloh.net/	Ohloh, the open source network
http://www.ipl.org/	Internet Public Library:
http://www.zimbra.com/	Zimbra offers Open Source email server software and shared calendar...
http://www.pocketmod.com/	PocketMod: The Free Disposable Personal Organizer
http://javimoya.com/blog/youtube_en.php	Download videos from Youtube, Google, iFilm, Metacafe, DailyMotion,...
http://keepvid.com/	KeepVid: Download videos from Google, Youtube, iFilm, Putfile, Meta...
http://10minutemail.com/10MinuteMail/index.html	10 Minute Mail
http://www.techmeme.com/	Techmeme
http://www.43things.com/	43 Things
http://musicovery.com/index.php?ct=us	Musicovery : interactive webRadio
http://www.speedtest.net/	Speedtest.net - The Global Broadband Speed Test
http://www.livejournal.com/	LiveJournal.com
http://www.cafepress.com/	CafePress.com : Create, Buy and Sell Unique Gifts, Custom T-Shirts ...
http://www.michaelbach.de/ot/	Optical Illusions and Visual Phenomena

http://torrent-finder.com/	Torrent Search :: Torrent Finder :: Torrent Search Engine
http://www.geocities.jp/iwamitsujp/	RYU&#039;S FORM SITE
http://www.nitroplus.co.jp/pc/	Nitroplus Net
http://www.youtorrent.com/	YouTorrent.com (BETA) - Your Torrents. Real Time.
http://www.ted.com/	TED: Ideas worth spreading
http://wordpress.com/	WordPress.com » Get a Free Blog Here
http://www.twenty120.com/	20/120 FILM COLLECTION
http://www.huddletogether.com/projects/lightbox/	Lightbox JS
http://feels.ru/pixel/pixel.html	Ïèêñåëüíûé ãîðîä / Pixel City
http://www.wired.com/images/article/magazine/test2007/st_infoporn_f.jpg	Consumer prices tech: cost of current technology
http://www.howtoforge.com/amfphp_adobe_flex2_sdk_p4	Using Amfphp 1.9 with the Adobe Flex 2 SDK - Page 4 | HowtoForge - Linux Howtos and Tutorials
http://www.debreuil.com/FrameworkDocs/UnitTestingOverview.htm	ASUnit : Unit Testing in Actionscript - DDW Framework Library
http://www.smashingmagazine.com/2007/01/19/53-css-techniques-you-couldnt-live-without/	53 CSS-Techniques You Couldn’t Live Without | Smashing Magazine
http://www.nytimes.com/2008/09/27/nyregion/27wars.html?_r=1&amp;pagewanted=all&amp;oref=slogin	The Shadowy, Wet World of StreetWars’ Squirt-Gun Assassins - NYTime...
http://labb.dev.mammon.se/swfupload/	SWFUpload
http://juixe.com/techknow/index.php/2006/08/12/top-13-ruby-on-rails-presentations/	TechKnow Zenze » Top 13 Ruby on Rails Presentations
http://www.coolrunning.com/engine/2/2_3/181.shtml	Cool Running :: The Couch-to-5K Running Plan
http://haveamint.com/	Mint: A Fresh Look at Your Site
http://www.alistapart.com/articles/slidingdoors	A List Apart: Articles: Sliding Doors of CSS
http://www.glish.com/css/	glish.com : CSS layout techniques
http://www.webstandards.org/	The Web Standards Project
http://www.positioniseverything.net/	/* Position Is Everything */ — Modern browser bugs explained in det...
http://37signals.com/papers/introtopatterns//index	37signals: An Introduction to Using Patterns in Web Design
http://www.boxesandarrows.com/	Boxes and Arrows: The design behind the design
http://gizmodo.com/	Gizmodo, the Gadget Guide
http://www.engadget.com/	Engadget
http://www.timeanddate.com/worldclock/	The World Clock - Time Zones
http://getvanilla.com/	Get Vanilla!
http://www.lifehack.org/	lifehack.org : Productivity, Getting Things Done and Lifehacks Blog
http://gnome-look.org/	GNOME-Look.org
http://cleancss.com/	Clean CSS - A Resource for Web Designers - Optmize and Format your CSS
http://www.ubuntu.com/	Ubuntu Home Page | Ubuntu
http://www.metacafe.com/	Metacafe – Best Videos &amp; Funny Movies
http://www.businessweek.com/	BusinessWeek: Daily &amp; Breaking News, Top Stories from BusinessWeek ...
http://ajaxian.com/	Ajaxian
http://www.getdeb.net/	GetDeb - Software for Ubuntu Linux

http://www.howtoforge.com/	HowtoForge - Linux Howtos and Tutorials | The Open Source Howto Dev...
http://www.dmoz.org/	ODP - Open Directory Project
http://thinkfree.com/common/main.tfo	ThinkFree Online beta
http://www.olacinc.org/	olacinc.org
http://www.spiegel.de/	SPIEGEL ONLINE - Nachrichten
http://dict.leo.org/	LEO Deutsch-Englisches Wörterbuch
http://www.blurb.com/	Make your own book with Blurb
http://htmldog.com/	HTML and CSS Tutorials, References, and Articles | HTML Dog
http://www.josbuivenga.demon.nl/index.html	exljbris :: Free Quality Font Foundry
http://www.wikihow.com/Main-Page	wikiHow - The How-To Manual That Anyone Can Write or Edit
http://www.maxpower.ca/free-icons/2006/03/05/	Free! Icons for your website or application at MaxPower
http://ajaxwrite.com/	www.ajaxwrite.com
http://www.pixel-peeper.com/	Pixel-Peeper -- More than 100,000 full-size sample photos from lenses, SLR cameras and digicams.
http://www.gutenberg.org/wiki/Main_Page	Main Page - Gutenberg
http://dictionary.cambridge.org/	Cambridge Dictionaries Online - Cambridge University Press
http://www.docjar.com/	DocJar: Search Open Source Java API
http://sourceforge.net/	SourceForge.net: Welcome to SourceForge.net
http://cssmania.com/	CSS Mania
http://www.magentocommerce.com/	Magento - Home - Open Source eCommerce Evolved
http://btjunkie.org/	btjunkie - the largest bittorrent search engine
http://interfacelift.com/wallpaper_beta/downloads/date/any/	InterfaceLIFT: Wallpaper sorted by Date
http://www.picnik.com/	Picnik - edit photos the easy way, online in your browser
http://www.degraeve.com/color-palette/	Color Palette Generator
http://www.widgetbox.com/	Widgetbox › World&#039;s largest widget directory and gallery - web widg...
http://bgpatterns.com/	Tiled backgrounds designer
http://hundredpushups.com/	one hundred push ups
http://960.gs/	960 Grid System
http://www.cadastre.gouv.fr/scpc/accueil.do	cadastre.gouv.fr
http://www.webconfs.com/search-engine-spider-simulator.php	Search Engine Spider Simulator
http://www.poignantguide.net/ruby/	Why’s (Poignant) Guide to Ruby
http://copypastecharacter.com/	Copy Paste Character
http://www.wpthemerkit.com/	WP Themer Kit - WordPress
http://www.emanuelblagonic.com/2007/07/19/how-to-use-photoshop-to-create-product-box/	EmanuelBlagonic.com - Something about web design » Blog Archive » H...
http://osliving.com/index.php	Open Source Living
http://www.webupon.com/Security/10-Extremely-Useful-Web-Sites-to-Stop-Big-Brother-From-Snooping-on-You.62616	10 Extremely Useful Websites to Stop Big Brother From Snooping on You
http://ninjahideout.com/blog/2007/05/16/ruby-on-rails-megapost-awesome-resources/	NinjaHideout » Blog Archives » Ruby on Rails megapost - Awesome Resources
http://www.jasonbartholme.com/2007/04/02/101-css-resources-to-add-to-your-toolbelt-of-awesomeness/	101 CSS Resources to Add to Your Toolbelt of Awesomeness » Jason Bartholme’s SEO Blog

http://forums.programming-designs.com/viewtopic.php?pid=3338	Programming Designs Forums / Five Great Programmers Fonts
http://gnome-look.org/	GNOME-Look.org
http://mvm.therealadam.com/articles/2006/03/24/down-the-rails-rabbit-hole	Down the Rails Rabbit Hole
http://www.tonyyoo.com/protolize/	Protolize | Essential web tools in one place
http://channel9.msdn.com/wiki/default.aspx/Channel9.DesktopSearchIFilters	Channel9 Wiki: DesktopSearchIFilters
http://www.colorschemer.com/online.html	Color Schemer - Online Color Scheme Generator
http://www.philb.com/iwantto.htm	I want to - a page of utilities that help you do stuff you want to
http://www.presentationzen.com/	Presentation Zen
http://userscripts.org/	Userscripts.org - Universal Repository
http://www.barelyfitz.com/screencast/html-training/css/positioning/	Learn CSS Positioning in Ten Steps: position static relative absolu...
http://37signals.com/svn	A design and usability blog: Signal vs. Noise (by 37signals)
http://www.makemylogobiggercream.com/	Make My Logo Bigger Cream
http://www.extjs.com/	Ext JS - JavaScript Library
http://www.videolan.org/	VideoLAN - Free Software and Open Source video streaming solution f...
http://dev.mysql.com/doc/refman/5.0/en/resetting-permissions.html	MySQL :: MySQL 5.0 Reference Manual :: B.1.4.1 How to Reset the Roo...
http://www.fwbuilder.org/	Firewall Builder
http://www.youtube.com/watch?v=6gmP4nk0EOE	YouTube - Web 2.0 ... The Machine is Us/ing Us
http://www.codeplex.com/AppArch	patterns &amp; practices: App Arch Guide 2.0 Knowledge Base - Home
http://www.dnsqueries.com/en/	The complete toolset for every network admin - DnsQueries
http://flowplayer.org/index.html	Flowplayer - Flash Video Player for the Web
http://www.zazzle.com/	Zazzle | Custom T-Shirts, Posters, Art and more...
http://www.devlisting.com/	The Web Developer&#039;s List of Resources
http://www.tripadvisor.com/	Reviews of vacations, hotels, resorts, vacation and travel packages...
http://www.site.uottawa.ca:4321/oose/index.html	Object Oriented Software Engineering Knowledge Base
http://www.joelonsoftware.com/articles/Unicode.html	The Absolute Minimum Every Software Developer Absolutely, Positivel...
http://liveplasma.com/	liveplasma music, movies, search engine and discovery engine
http://adaptivepath.com/publications/essays/archives/000385.php	adaptive path » ajax: a new approach to web applications
http://blogpulse.com/index.html	Nielsen BuzzMetrics&#039; BlogPulse
http://css.maxdesign.com.au/floatutorial/	Floatutorial: Step by step CSS float tutorial
http://www.washingtonpost.com/	washingtonpost.com - nation, world, technology and Washington area ...
http://www.msnbc.msn.com/	MSNBC
http://www.cnn.com/	CNN.com - Breaking News, U.S., World, Weather, Entertainment &amp; Vide...
http://www.colr.org/	colr.org
http://www.torrentreactor.net/	TorrentReactor.Net - The most active torrents on the web
http://searchenginewatch.com/	Search Engine Watch: Tips About Internet Search Engines &amp; Search En...
http://www.hvf.jp/	□ □ □ □ □ □ □ □ □ □ ー□
http://dlanham.com/	David Lanham

http://metaatem.net/words/	Spell with flickr
http://readymech.com/	Fwis • Readymech Series 002
http://www.poignantguide.net/ruby/	Why’s (Poignant) Guide to Ruby
http://www.alistapart.com/articles/slidingdoors/	A List Apart: Articles: Sliding Doors of CSS
http://www.dezwozhere.com/links.html	CSS, Accessibility and Standards Links
http://www.cssplay.co.uk/	Stu Nicholls | CSSplay | Experiments with cascading style sheets | ...
http://css.maxdesign.com.au/index.htm	css.maxdesign.com.au - CSS resources and tutorials for web designer...
http://www.instructables.com/id/Build_a_furniture_quality_Laptop_Stand_and_TV_Tray/	Build a furniture quality Laptop Stand and TV Tray
http://howto.wired.com/wiki/Get_Better_Genius_Recommendations_in_iTunes	Get Better Genius Recommendations in iTunes - Wired How-To Wiki
http://www.hulu.com/	Hulu - Watch your favorites. Anytime. For free.
http://www.drudgereport.com/	DRUDGE REPORT 2006®
http://www.slide.com/	Slide - slideshows, slide shows, photo sharing, image hosting, widg...
http://www.oneandother.co.uk/	One &amp; Other
http://www.cubeecraft.com/	cubeecraft.com
http://handbrake.fr/	HandBrake
http://thedailywtf.com/	The Daily WTF
http://edge.org/	Edge
http://arcade.itch.com/games/boomstick/	BoomsticK - the game @ itch.com
http://www.southparkstudios.com/	South Park Studios
http://drupal.org/	drupal.org | Community plumbing
http://mochikit.com/	MochiKit - A lightweight Javascript library
http://www.google.com/webmasters/	Google Webmaster Central
http://lifehacker.com/	Lifehacker, the Productivity and Software Guide
http://allnew6.com/	□ □ □ □ □
http://www.clapclap.se/	Clapclap Design
http://paperforest.blogspot.com/	Paper Forest
http://cakephp.seesaa.net/	CakePHP □ □ □ □ □ □ □ □
http://captchas.net/sample/php/	Sample PHP Implementation
http://www.txtnation.com/	txtNation :: Creating Mobile Interaction between Businesses and Consumers | Mobile Solutions
http://isohunt.com/	isoHunt - World&#039;s largest BitTorrent and P2P search engine
http://www.mint.com/	Free Personal Finance Software, Online Money Management, Budget Pla...
http://architects.dzone.com/news/common-rest-design-pattern	Common REST Design Pattern | Architects Zone
http://www.dvd43.com/	DVD43 v3.9.0 - Download Sites
http://www.openstudio.fr/jQuery-Multimedia-Portfolio.html	jQuery Multimedia Portfolio - OpenStudio Communication sur Internet
http://freesound.iua.upf.edu/	freesound :: home page
http://www.google.com/reader/view/	Google Reader
http://mactechnotes.blogspot.com/2005/10/controlling-webkit-and-safari-through.html	MacTechNotes: Controlling WebKit and Safari through Preferences
http://www.walkscore.com/	Walk Score - Helping homebuyers, renters, and real estate agents fi...

http://www.seriouseats.com/	Serious Eats: A Food Blog and Community
http://www.pendrivelinux.com/	Boot and run Linux from a USB flash memory stick | USB Pen Drive Linux
http://www.getpaint.net/	Paint.NET - Free Software for Digital Photo Editing
http://www.pentoo.ch/-PENTOO-.html	NETwork Security Consortium
http://usernamecheck.com/	Where is Your Username registered
http://www.linux.com/feature/126186	Linux.com :: Five fun ways to use a Linux webcam
http://www.btinternet.com/~glynhughes/squashed/	Squashed Philosophers- Condensed Plato Aristotle Augustine Descarte...
http://www.webmd.com/	WebMD - Better information. Better health.
http://www.lonelyplanet.com/	Lonely Planet: the world&#039;s best guidebooks, travel advice and infor...
http://www.dailylit.com/	DailyLit: Read books by email and RSS.
http://www.biblegateway.com/	BibleGateway.com: A searchable online Bible in over 50 versions and...
http://strangerthings.tv/	Stranger Things - iPod (640×480)
http://scrapetorrent.com/	Torrent Search - ScrapeTorrent.com
http://en.wikibooks.org/wiki/Main_Page	Main Page - Wikibooks, collection of open-content textbooks
http://www.opendesigns.org/	Open Design Community - Download Free Web Design Templates - OpenDe...
http://www.merbivore.com/	Merb | Looking for a better framework?
http://www.splashup.com/	Splashup
http://www.ajaxrain.com/index	Ajax Rain
http://www.gotapi.com/	gotAPI.com - quick developer reference for CSS, HTML, JavaScript, P...
http://www.ntwind.com/software/utilities/visual-subst.html	Visual Subst
http://www.writely.com/	Writely - The Web Word Processor
http://wufoo.com/	Wufoo - HTML Form Builder - Free Contact Forms &amp; Online Surveys
http://moofx.mad4milk.net/	moo.fx - the next small thing
http://www.worldmapper.org/	Worldmapper: The world as you&#039;ve never seen it before
http://www.strobist.blogspot.com/	Strobist
http://pageflipgallery.com/	FlippingBook Wordpress Gallery
http://flowplayer.org/index.html	Flowplayer - Flash Video Player for the Web
http://www.schillmania.com/projects/soundmanager2/demo/360-player/	360° MP3 player UI demo (SoundManager 2)
http://cow.neondragon.net/stuff/reflection/	Reflection.js
http://philrenaud.com/156	The Next 35 Sexiest Designed Websites You&#039;ve Forgotten - PhilRenaud.com
http://www.smashingmagazine.com/2009/04/27/the-mystery-of-css-sprites-techniques-tools-and-tutorials/	The Mystery Of CSS Sprites: Techniques, Tools And Tutorials | CSS |...
http://thedesignsuperhero.com/2009/01/80-free-retrovintage-style-wallpapers-the-ultimate-list/	The Design Superhero » 80+ Retro/Vintage Style Wallpapers: The Ultimate List!
http://www.noupe.com/icons/50-most-beautiful-icon-sets-created-in-2008.html	50 Most Beautiful Icon Sets Created in 2008 | Noupe
http://www.javascriptkit.com/script/script2/tengcalendar.shtml	Cut &amp; Paste Date Time Picker
http://marqueetool.net/examples/changing-of-shroud-color-and-opacity/	Rectangular Marquee Tool. Changing of shroud Color and Opacity
http://www.webmonkey.com/blog/Fring_Turns_Your_iPhone_into_a_Free_Skype_Phone	Fring Turns Your iPhone into a Free Skype Phone - Webmonkey

http://www.mukurtuarchive.org/index.html	Mukurtu Wumpurrarni-kari Archive :: An Indigenous Archive Tool
http://upcoming.yahoo.com/	Home - Upcoming
http://www.surveymonkey.com/	SurveyMonkey.com - Powerful tool for creating web surveys. Online s...
http://news.bbc.co.uk/1/hi/sci/tech/6616651.stm	BBC NEWS | Science/Nature | Power station harnesses Sun&#039;s rays
http://www.susanmeiselas.com/	Susan Meiselas
http://blogs.techrepublic.com.com/10things/?p=919	10 low-cost, high-value Web 2.0 strategies
http://www.theonion.com/	The Onion
http://www.pdf-search-engine.com/	Ebook Search - Pdf Search Engine
http://trac.manent-backup.com/	Manent – Trac
http://www.netvibes.com/#General	Netvibes
http://supercook.com/	Supercook: recipe search by ingredients you have at home
http://mashable.com/2009/09/10/openbox-mobile/	Box.net Brings Cloud Storage to iPhone Apps
http://www.scottrobertsweb.com/scoville-scale.php	Official Scott Roberts Web Site - Scoville Scale for Hot Sauces and Hot Peppers
http://www.smashingmagazine.com/2009/05/26/20-time-saving-tips-to-improve-designers-workflow-part-1/	20 Time-Saving Tips to Improve Designer&#039;s Workflow | How-To | Smashing Magazine
http://www.searchfreefonts.com/	Search Free Fonts - over 13,000 free fonts available for download
http://www.dreamcss.com/2009/05/jquery-and-ajax-based-tag-cloud.html	8 jQuery and Ajax based tag clouds for web developer
http://video.google.com/	Google Video
http://www.evernote.com/	Remember everything. | Evernote Corporation
http://www.bustedtees.com/	BustedTees - Funny T-Shirts - New T-Shirt designs every week - Craz...
http://www.gmail.com/	Gmail
http://mail.yahoo.com/	Yahoo! Mail
http://jp.reuters.com/	□□□ー.co.jp | □□□□ース, □□□□, 経済　金融ニュース, &amp; More
http://www.chromasynthetic.com/scripts/jibberbook/	JibberBook 2 - Free AJAX Guestbook
http://yotophoto.com/	Yotophoto | Find free photos... fast!
http://fonts500.com/	Fonts 500
http://codylindley.com/Webdev/335/im-not-an-interaction-designer-i-did-however-stay-at-a-holiday-inn-last-night	Webdev Entry - Cody Lindley: I&#039;m not an Interaction Designer, I did...
http://lifehacker.com/software/calendar/download-of-the-day-magical-mac-247838.php	Download of the Day: MagiCal (Mac) - Lifehacker
http://www-csli.stanford.edu/~schuetze/information-retrieval-book.html	Introduction to Information Retrieval
http://www.usshortcodes.com/	CSCA
http://www.thewednesdaychef.com/the_wednesday_chef/2008/11/chez-panisses-w.html	The Wednesday Chef: Chez Panisse&#039;s Winter Squash, Onion and Red Wine Panade
http://www.retailmenot.com/	Coupon codes for thousands of online stores - RetailMeNot.com
http://www.someecards.com/	someecards.com | ecards for when you care enough to hit send | home
http://lii.org/	Librarians&#039; Internet Index
http://allrecipes.com/	All recipes – complete resource for recipes and cooking tips
http://maps.google.com/	Google Maps

http://www.nba.com/wizards/index_main.html	WashingtonWizards.com - The official website of the Washington Wizards
http://www.processlibrary.com/	ProcessLibrary.com - The online resource for process information!
http://pipl.com/	Pipl - People Search
http://blogsearch.google.com/	Google Blog Search
http://www.velocityaircraft.com/	Velocity Aircraft
http://www.workingforchange.com/activism/index.cfm	ActForChange
http://gethuman.com/	gethuman 500 database
http://geektechnique.org/projectlab/797/openbsd-encrypted-nas-howto	OpenBSD encrypted NAS HOWTO :: projects :: geek technique
http://www.dailykos.com/	Daily Kos: State of the Nation
http://www.trulia.com/	Trulia - Real Estate, Homes For Sale, Sold Properties, Real Estate ...
http://www.wireshark.org/	Wireshark: Go deep.
http://www.wordreference.com/	English to French, Italian &amp; Spanish Dictionary - WordReference.com
http://www.zabasearch.com/	Free People Search by ZabaSearch!
http://springwise.com/	Springwise: new business ideas for entrepreneurial minds.
http://www.winpwn.com/index.php/Main_Page	Main Page - WinPwn
http://www.tomshardware.com/	Tom&#039;s Hardware
http://www.flattvpeople.com/tutorials/lcd-vs-plasma.asp	Flat TV People : LCD TVs versus Plasma Televisions
http://www.viruspool.net/	viruspool.net is THE database to index virus descriptions
http://www.archive.org/details/bbs_documentary	Internet Archive: The BBS Documentary Video Collection
http://code.google.com/support/bin/answer.py?answer=81101&amp;topic=11982	Google Code FAQ - GearsMonkey: Google Gears + Greasemonkey to take ...
http://www.kaply.com/weblog/	Mike’s Musings
http://www.onjava.com/pub/a/onjava/2006/05/17/standardizing-with-ejb3-java-persistence-api.html?page=1	Standardizing Java Persistence with the EJB3 Java Persistence API
http://dist.leetsoft.com/api/paypal/	Paypal library for rails
http://helptutorservices.com/blog/the-32-most-commonly-misused-words-and-phrases/	The 32 Most Commonly Misused Words and Phrases
http://www.uesp.net/wiki/Oblivion:Items	Oblivion:Items - UESPWiki
http://jontangerine.com/silo/html/placeholder/	Placeholder HTML Markup with Lorem Ipsum — Jon Tan □
http://www.dreamstime.com/	High Resolution Stock Photography: Download Free Stock Photos and R...
http://www.navicat.com/	Navicat - the World&#039;s Best MySQL GUI for Windows, Linux &amp; Mac OS X
http://www.mozillaonline.com/	谋智网络，火狐浏览器中国唯一官方网站 | Mozilla, Firefox, and China
http://www.microsoft.com/downloads/details.aspx?familyid=2fcde6ce-b5fb-4488-8c50-fe22559d164e&amp;displaylang=en	Download details: Windows XP Service Pack 3 - ISO-9660 CD Image File
http://www.vimeo.com/	Vimeo, Video Sharing For You
http://www.1pixelout.net/code/audio-player-wordpress-plugin/#podcasting	Audio Player Wordpress plugin
http://30boxes.com/	30 Boxes | it&#039;s your life
http://www.youtube.com/watch?v=Wnexu_eGyYs	YouTube - Henry Rollins &quot;America is under attack.&quot;
http://pixelgirlpresents.com/	Pixelgirl Presents Free Icons, Desktops and Gallery Shop!
http://www.gen-x-design.com/index.php	Gen-X-Design | Ian Selby

http://www.rubycentral.com/book/	Programming Ruby: The Pragmatic Programmer&#039;s Guide
http://www.blinkx.com/	Video Search Engine - Blinkx
http://blogcritics.org/	Home @ Blogcritics.org
http://www.mysqlperformanceblog.com/2007/02/25/pitfalls-of-converting-to-innodb/	MySQL Performance Blog » Pitfalls of converting to InnoDB
http://www.npr.org/blogs/money/	NPR: Planet Money
http://jan.kneschke.de/2007/8/1/mysql-proxy-learns-r-w-splitting	~jk MySQL Proxy learns R/W Splitting
http://bridge.kshep.net/	http://bridge.kshep.net/
http://www.phoronix.com/scan.php?page=article&amp;item=983&amp;num=1	[Phoronix] Virtualization Made Easy In Ubuntu 8.04
http://css.maxdesign.com.au/floatutorial/	Floatutorial: Step by step CSS float tutorial
http://www.cssplay.co.uk/	Stu Nicholls | CSSplay | Experiments with cascading style sheets | ...
http://www.tastespotting.com/	TasteSpotting
http://www.tineye.com/	TinEye Reverse Image Search
http://www.ovguide.com/index.html	OVGuide Online Video Guide: Watch Free Movies, Streaming Videos, Wa...
http://www.tipmonkies.com/2005/10/04/disposable-e-mail-address-services	TipMonkies » Blog Archive » Disposable e-mail address services
http://www.onelook.com/	OneLook Dictionary Search
http://omgili.com/	Omgili - Find out what people are saying
http://blog.guykawasaki.com/	How to Change the World
http://www.hostgator.com/	HOSTGATOR WEB HOSTING - cPanel, Reseller, and Dedicated Website Hosting
http://www.paulgraham.com/submarine.html	The Submarine
https://github.com/	Your Dashboard - GitHub
http://code.google.com/p/xinc/	xinc - Google Code
http://www.youtube.com/watch?v=9hIQjrMHTv4	YouTube - History of the Internet
http://radiofly.to/nishi/cvs/	□ ージョン管理システム CVS □ □ □
http://www.designboom.com/eng/index.xtml	industrial design courses ? designboom
http://www.methods.co.nz/popup/popup.html	DOM Popup Kit
http://www.befunky.com/	BeFunky.com - Photo effects with one click, Turn your photos into a...
http://www.adherents.com/Religions_By_Adherents.html	Major Religions Ranked by Size
http://www.blog.spoongraphics.co.uk/tutorials/edit-an-image-in-photoshop-to-add-some-pazazz	Edit an Image in Photoshop to Add Some Pizazz! | Blog.SpoonGraphics
http://www.bluevertigo.com.ar/bluevertigo.htm?bvresources.htm~content	BLUE VERTIGO | Web Design Resources Links | Last update JAN.28.2008
http://anond.hatelabo.jp/20071106010842	□ □ □ □ □ □ □ □ □ □ □ □ □ □ 実
https://www.google.com/analytics/home/	Google Analytics
http://www.yahoo.com/	Yahoo!
http://www.jungledisk.com/	JungleDisk - Reliable online storage powered by Amazon S3 ™ - Jungl...
http://get-shorty.com/	Shorty
http://pixelgirlpresents.com/	Pixelgirl Presents Free Icons, Desktops and Gallery Shop!
http://www.jambor-ee.com/welovefood/day-1	Day 1: What are we doing over the next 24 days? | Jambor-ee
http://www.webcreme.com/	Web Creme | Web design inspiration
http://babynamewizard.com/namevoyager/lnv0105.html	The Baby Name Wizard: NameVoyager

http://www.technologyreview.com/Infotech/18650/	Technology Review: Help Me Redesign the Web
http://office.microsoft.com/en-us/help/default.aspx	Help and How-to Home Page - Microsoft Office Online
http://www.princeton.edu/~rvdb/JAVA/election2004/	Election 2004 Results (gradient by county)
http://www.pocketcalculatorshow.com/	Vintage Electronics Have Soul - The Pocket Calculator Show Website
http://forum.libspark.org/	□ □ ーラム - Spark project -
http://feb19.jp/blog/archives/000123.php	feb19.jp blog - AS3□ 読み込んだ外部画像にスムージングを適用す□
http://www.1-click.jp/	1-click Award by □ □ 会社リクルートメディアコミュニケーション□
http://www.feedss.com/	□ □ RSS□ □ □ □ -□ □ □ 闻,□ □ ,blog,网志,论坛搜索服务
http://www.quickonlinetips.com/	Quick Online Tips - Technology news, blogging tips, best computer software and web services
http://developer.yahoo.com/ypatterns/	Yahoo! Design Pattern Library
http://lcmm.qc.ca/	LCMM - Bienvenue au Club Macintosh de Montréal
http://northtemple.com/1608	NorthTemple.com : The Accessibility Checklist I V...
http://secondlife.com/	Second Life: Your World. Your Imagination.
http://everystockphoto.com/	everystockphoto.com - your source for free photos
http://www.rossoneri.jp/2009/01/18_23215.php	□ □ □ □ □ □ □ □ □ □ □ □ □ □ | □ □ 黒
http://jmdoudoux.developpez.com/java/eclipse/	Développons en Java avec Eclipse
http://www.aepap.org/	Asociación Española de Pediatría de Atención Primaria
http://cooltext.com/	Cool Text: Logo and Graphics Generator
http://www.printrates.com/	Digital photo printing prices and reviews
http://www.letterform.net/	Letterform | Chicago
http://www.lovelycharts.com/	Lovely Charts | Free online diagramming application
http://www.fmylife.com/	F*** My Life - FML : Your everyday life stories.
http://www.mashedjobs.com/	All Design &amp; Development Jobs from MashedJobs.com
http://www.phatfusion.net/sortabletable/	phatfusion : sortableTable
http://pixlr.com/	Online image / photo editor pixlr free
http://www.uncrate.com/	Uncrate | The Buyer&#039;s Guide For Men
http://www.floorplanner.com/	Create and Share Floorplans Online with Floorplanner.com
http://www.campaignmonitor.com/	Email Newsletter Software for Web Designers - Campaign Monitor
http://typefacts.com/	Typefacts | Typografie verstehen
http://paulgraham.com/highres.html	The High-Res Society
http://veerle.duoh.com/	Veerle&#039;s blog 2.0 - Webdesign - XHTML CSS | Graphic Design
http://www.fileqube.com/	Free Online Storage - File Qube
http://www.chinaelections.org/	□ 国选举与治理网
http://www.google.com/analytics/	Google Analytics
http://code.google.com/p/django-rosetta/	django-rosetta - Project Hosting on Google Code
http://dojotoolkit.org/offline	The Dojo Offline Toolkit | The Dojo Toolkit
http://sethgodin.typepad.com/	Seth&#039;s Blog
http://www.solitude.dk/archives/embedquicktime/	Embed QuickTime | jQuery Plugin
http://www.mindomo.com/	Mindomo - Web-based mind mapping software
http://www.psdtuts.com/	Photoshop Tutorials - PSDTUTS
http://gethuman.com/	gethuman 500 database

http://www.journler.com/	Journler - Wherever Life Takes You
http://www.warninglabelgenerator.com/	Warning Label Generator
http://www.treehugger.com/index.php	TreeHugger
http://www.zoho.com/	Online Office, Word Processor, Spreadsheet, Presentation, CRM and more
http://www.conversion-rate-experts.com/articles/understanding-your-visitors/	14 free tools why people abandon your website
http://dougscripts.com/itunes/itinfo/ituneslibrarymanager.php	Doug&#039;s AppleScripts for iTunes ♫ iTunes Library Manager v5.2.1
http://rules.gonna.jp/webapp/home/	AJAX□ □ □ ! ajax□ □ □ □ web2.0□ □ □ □ □ □ □ □ □ □ □ □ □ □ !
http://www.imdb.com/	The Internet Movie Database (IMDb)
http://php-java-bridge.sourceforge.net/	Integrate PHP &amp; Java - PHP / Java Bridge
http://gawker.com/	Gawker, Manhattan Media News and Gossip
http://www.bbc.co.uk/worldservice/learningenglish	BBC Learning English | Home page
http://extjs.eu/	Saki&#039;s Extensions, Plugins and Know-How
http://www.meevee.com/	MeeVee - TV Guide, TV listings, TV Full Episodes, News &amp; Gossip, Online Videos, Message Boards, TV Blog
http://jendryschik.de/wsdev/einfuehrung/	Einführung in XHTML, CSS und Webdesign
http://code.google.com/intl/de-DE/speed/page-speed/	Page Speed - Web Page Performance Tests
http://www.roytanck.com/	Roy Tanck&#039;s weblog
http://www.edutopia.org/	Edutopia: What Works in Public Education
http://www.bamagazine.com/	Before &amp; After magazine
http://briancray.com/2009/04/16/target-ie6-and-ie7-with-only-1-extra-character-in-your-css/	Target IE6 and IE7 with only 1 extra character in your CSS / Brian ...
http://www.megalab.it/	Megalab.it - Aperiodico gratuito di informatica e tecnologia
http://mashable.com/2009/09/07/facebook-smarter-twitter-dumber/	Psychologist: Facebook Makes You Smarter, Twitter Makes You Dumber
http://www.cspinet.org/nah/10foods_bad.html	Ten Worst and Best Foods
http://blogs.howtogeek.com/tuxgeek/2008/09/14/10-things-you-wanted-to-do-with-ubuntu-but-didnt-know-how/	Ubuntu 10 tips
http://blog.guykawasaki.com/2007/08/on-the-other-ha.html	How to Change the World: On the Other Hand: The Flip Side of Entrepreneurship by Glenn Kelman
http://zenhabits.net/2008/09/21-easy-hacks-to-simplify-your-life/	21 Easy Hacks to Simplify Your Life | Zen Habits
http://gigazine.net/index.php?/news/comments/20070616_company_font/	□ - GIGAZINE
http://www.wotanserver.com/en/	Online de-branding and software upgrade service SonyEricsson (SE) m...
http://www.makezine.com/blog/archive/2005/06/make_ebooks_for_1.html	MAKE: Blog: MAKE ebooks for your iPod guide!
http://www.borders.com/	Borders.com
http://carlogiovani.com/	carlogiovani.com
http://www.meiosepublicidade.pt/	Meios &amp; Publicidade
http://en.wikipedia.org/wiki/Learned_helplessness	Learned helplessness - Wikipedia, the free encyclopedia
http://my.yahoo.com/	My Yahoo!
http://www.proprofs.com/forums/index.php?showtopic=8478	Free Linux+ Study Guide : CompTIA : IT Certification :
http://linuxlock.blogspot.com/2008/12/linux-stop-holding-our-kids-back.html	Blog of helios: Linux - Stop holding our kids back
http://en.beijing2008.cn/	The Official Website of the Beijing 2008 Olympic Games
http://www.theworldismycanvas.com/	The world is my canvas
http://www.foodnetwork.com/	Food Network : Cooking, Recipe Collections, Party Ideas, Quick &amp; Ea...
http://www.hometrainingtools.com/	Home Science Tools
http://www.cooper.com/content/insights/newsletters/2004_issue04/Ten_ways_to_kill_design.asp	Ten Ways to Kill Design
http://www.guut.de/guut/shop/	guut.de - Jeden Tag ein Produkt in begrenzter Stückzahl

http://www.pagat.com/	Pagat
http://www.nitropdf.com/pdfdownload/welcome.asp	Thank you for installing PDF Download
http://www.adelaider.com/google-cheat-sheet/?cheatsheet&amp;page=2	Google Cheat Sheet - Page 1 - Page 2
http://www.frogdesign.com/	Frog Design
http://www.nokia.com/betalabs/locationtagger	Nokia - Location Tagger
http://drnicwilliams.com/2008/01/04/autotesting-javascript-in-rails/	Dr Nic » Autotesting Javascript in Rails
http://www.synchroedit.com/	SynchroEdit (Alpha version) - online shared wordprocessor
http://dzineblog.com/2009/01/packaging-design-inspiration-part-3.html	Packaging design inspiration - 45 Really Nice Packaging Designs | Dzine Blog
http://www.southparkzone.com/	Watch and Download every single South Park episode
http://www.windowclippings.com/	Window Clippings - High quality screen capture for Windows
http://haha.nu/creative/how-to-make-shadows-on-the-wall/	Shadows
http://thinkingforaliving.org/	Thinking for a Living™
http://meyerweb.com/eric/tools/s5/	S5: A Simple Standards-Based Slide Show System
http://labs.systemone.at/retrievr/	retrievr - search by sketch / search by image
http://www.humanmetrics.com/cgi-win/JTypes3.asp	personality test
http://www.google.com/webhp?complete=1	Google Suggest
http://sanfrancisco.menupages.com/	San Francisco Restaurants, San Francisco Menus, Ratings, Reviews, SF Restaurants Guide
http://www.designobserver.com/archives/025896.html	Everything I Know About Design I Learned from The Sopranos
http://mydreamapp.com/	My Dream App
http://seotools.jp/	SEO TOOLS β(SEO□ ールズ) - 気になるあのサイトのアクセス・SEO 対策を無料でチェック□
http://www.peters1.dk/webtools/conversion.php?sprog=en	Online converter
http://www.alvit.de/blog/article/20-best-license-free-official-fonts	Vitaly Friedman&#039;s Notebook: 25 Best Free Quality Fonts
http://tech.cybernetnews.com/2006/03/26/this-may-help-your-firefox-memory-leak/	This May Help Your Firefox Memory Leak
http://www.gonomad.com/	GoNOMAD.com--Alternative Travel, Ecotourism, Destination Guides, Travel Blogs, Volunteering Vacations
http://www.seatguru.com/	Welcome to SeatGuru! Your Guide to Airplane Seats and In-flight Ame...
http://www.diigo.com/	Diigo - Web Highlighter and Sticky Notes, Social Bookmarking and An...
http://www.textually.org/ringtonia/	ringtonia.com
http://www.instructables.com/id/When_a_Phillips_is_not_a_Phillips/	When a Phillips is not a Phillips!
http://danga.com/memcached/	memcached: a distributed memory object caching system
http://slashdot.org/	Slashdot: News for nerds, stuff that matters
http://www.marumushi.com/apps/newsmap/newsmap.cfm	newsmap
http://www.nliteos.com/index.html	nLite - Deployment Tool for the bootable Unattended Windows install...
http://www.jpb.com/index.php	Wonderful World of jpb.com
http://zenhabits.net/	Zen Habits | Simple Productivity
http://www.pbs.org/wnet/savageearth/	SAVAGE EARTH Online
http://www.microsoft.com/windowsxp/Downloads/powertoys/Xppowertoys.mspx	Microsoft PowerToys for Windows XP
http://www.motiont.com/craigslistreader.aspx	CraigsList Reader - Free tool to search craigslist
http://www.youtorrent.com/	YouTorrent.com (BETA) - Your Torrents. Real Time.
http://simile.mit.edu/httptracer/index.html	SIMILE | HTTPTracer
http://www.indeed.com/	Job Search | one search. all jobs. Indeed
http://www.chapter3.net/	CHAPTER THREE digital creations

http://www.creamundo.com/index.php?lang=en	9800 Free Fonts, fonts for free, font finder, download free fonts, ...
http://www.flickr.com/photos/pantufla/sets/72157594489508934/	19thc Shipping Posters - a photoset on Flickr
http://www.cycas.de/	CYCAS CAD 2D + 3D + ARCHITECTURE
http://www.cubeecraft.com/	cubeecraft.com
http://www.realsolve.co.uk/site/tech/easymock.php	Realsolve - Mock Object Testing With EasyMock 2
http://www.astro.umd.edu/~marshall/smileys.html	Canonical Smiley List
http://tutorialblog.org/free-vector-downloads/	» Free Vector Downloads
http://www.searchme.com/	Searchme Visual Search - Beta - rev. 2.0.2
http://cinemassacre.com/Movies/Nes_Nerd.html	Angry Video Game Nerd
http://www.pendrivelinux.com/	Boot and run Linux from a USB flash memory stick | USB Pen Drive Linux
http://sourceforge.net/	SourceForge.net: Welcome to SourceForge.net
http://www.webmd.com/	WebMD - Better information. Better health.
http://www.webmonkey.com/	Webmonkey: the Web Developers Resource
http://ocw.mit.edu/	MIT OpenCourseWare | OCW Home
http://wiki.developer.mindtouch.com/Wik.is/EC2_Infrastructure	EC2 Infrastructure - MindTouch Developer Center
http://www.jaiku.com/	Jaiku | Your Conversation
http://www.lifehacker.com/software/isight/take-isight-pics-of-invalid-logins-221262.php	Take iSight pics of invalid logins - Lifehacker
http://www.swedesignz.com/	SweDesignz - Photoshop, PHP, HTML/CSS Tutorials
http://dlatwork.com/	Download at Work
http://tv-links.co.uk/	TV Links
http://www.kayak.com/	Cheap Flights, Airline Tickets, Cheap Airfare &amp; Discount Travel Dea...
http://www.threadless.com/	Threadless T-Shirts - Designer Clothing Submissions - Tees, Tshirts...
http://www.techbargains.com/	Techbargains - discount computer sale buy cheap digital camera revi...
http://www.linuxcommand.org/learning_the_shell.php	LinuxCommand.org: Learning the shell.
http://www.fatwallet.com/	Online Coupons | Cash Back
https://addons.mozilla.org/firefox/2324/	Session Manager | Firefox Add-ons | Mozilla Corporation
http://www.acmqueue.com/modules.php?name=Content&amp;pa=showpage&amp;pid=98	Silicon Superstition
http://www.bittbox.com/	BittBox
http://www.behidden.com/	BeHidden anonymous surfing
http://www.torrentreactor.net/	TorrentReactor.Net - The most active torrents on the web
http://blog.dopplr.com/	Dopplr Blog
http://pitaschio.ara3.net/index.htm	Pitaschio
http://www.macosxhints.com/article.php?story=20060622090404212	macosxhints.com - Change Parallels Desktop 1.0&#039;s caching strategy
http://www.maxmind.com/app/city	MaxMind - GeoIP City Geolocation IP Address to City
http://bakery.cakephp.org/articles/view/simple-form-authentication-in-1-2-x-x	Simple Form Authentication in 1.2.x.x (Articles) | The Bakery, Everything CakePHP
http://www.getafreelancer.com/	Custom Web Design and Programming. Freelance Programmers. Outsource...
http://www.sturgesreps.com/	FrankSturgesReps
http://www.sampaist.com/	Sampaist
http://www.123di.com/	123di: The Most Complete, Comprehensive, Authoritative Digital Phot...
http://www.torrentreactor.net/	TorrentReactor.Net - The most active torrents on the web
http://boxesandarrows.com/	Boxes and Arrows: The design behind the design
http://tech.nitoyon.com/hatebu_nenkan/	□ □ □ □ □

http://www.gutenberg.org/	Free eBooks - Project Gutenberg
http://www.refdesk.com/	Refdesk.com ... Reference, Facts, News ... Free and Family-friendly...
http://www.econsultant.com/i-want-freeware-utilities/index.html	I want a Freeware Utility to ... 450+ common problems solved : eCon...
http://www.djangoproject.com/documentation/newforms/	Django | The newforms library | Django Documentation
http://www.ariadne.ac.uk/issue54/tonkin-et-al/	Main Articles: &#039;Collaborative and Social Tagging Networks&#039;, Ariadne...
http://www.w3schools.com/default.asp	W3Schools Online Web Tutorials
http://www.freshbooks.com/	FreshBooks - Online Invoicing, Time Tracking and Expense Service
http://www.dapper.net/	Dapper: The Data Mapper
http://www.netbeans.org/kb/articles/mysql-client.html	A simple MySQL client in NB
http://mayang.com/textures/	Mayang&#039;s Free Texture Library
http://www97.intel.com/education/	Intel® Innovation in Education
http://www.gotoandlearn.com/index	gotoandlearn.com - Free video tutorials by Lee Brimelow on the Flas...
http://www.sideshowtoy.com/cgi-bin/category.cgi?category=0	Movie, Television and Proprietary Collectible Figures - Sideshow Co...
http://projecteuler.net/	Project Euler
http://www.webdesignerdepot.com/2008/12/designing-outside-your-comfort-zone/	Designing Outside Your Comfort Zone | Webdesigner Depot
http://cssmania.com/	CSS Mania
http://www.thecoolhunter.net/	thecoolhunter.net
http://nvu.com/	Nvu - The Complete Web Authoring System for Linux, Macintosh and Wi...
http://www.codeplex.com/sushi	SharePoint SUSHI - Home
http://davidwalsh.name/php-google-analytics	Retrieve Google Analytics Visits and PageViews with PHP
http://www.niksoftware.com/index/en/entry.php	Nik Software, Inc. | Welcome
http://www.w3schools.com/css/default.asp	CSS Tutorial
http://www.shutterstock.com/	Stock Photos | Shutterstock: Royalty-Free Subscription Stock Photog...
http://freelanceswitch.com/general/101-essential-freelancing-resources/	» 101 Essential Freelancing Resources
http://www.google.com/webmasters/	Google Webmaster Central
http://msdn.microsoft.com/en-us/vstudio/default.aspx	Microsoft Visual Studio on MSDN
http://www.ncrel.org/sdrs/areas/issues/students/atrisk/at400.htm	Using Technology to Enhance Engaged Learning for At-Risk Students
http://www.freesound.org/	freesound :: home page
http://www.tour-eiffel.fr/index.html	Le site officiel de la Tour Eiffel
http://www.typorganism.com/asciiomatic/	t.y.p.o.r.g.a.n.i.s.m : ASCII-O-Matic
http://oeffentlicher-dienst.info/	Öffentlicher-Dienst.Info
http://www.readingterminalmarket.org/	Reading Terminal Market › Home
http://www.voki.com/	Voki Home
http://www.sitelutions.com/	Domain Names, Web Hosting, Free DNS, Free Dynamic DNS, Free Redirec...
http://www.smashingmagazine.com/	Smashing Magazine
http://www.musicovery.com/	Musicovery : interactive webRadio
http://metafilter.com/	Metafilter | Community Weblog
http://www.huddletogether.com/projects/lightbox2/	Lightbox JS v2.0
http://projecteuler.net/	Project Euler
http://www.angryalien.com/0604/titanicbunnies.html	Titanic in 30 seconds with bunnies.

http://ebin.wordpress.com/2007/03/21/how-to-turn-your-photo-into-movie-like-effect-using-photoshop/	How to turn your photo into movie-like effect using Photoshop? « ebin
http://www.smashingmagazine.com/2008/01/10/adobe-photoshop-tutorials-best-of/	Adobe Photoshop Tutorials - Best Of | Tutorials | Smashing Magazine
http://www.trulia.com/	Trulia - Real Estate, Homes For Sale, Sold Properties, Real Estate ...
http://homokaasu.org/rasterbator/	The Sect of Homokaasu - The Rasterbator
http://www.gigamonkeys.com/book/	Practical Common Lisp
http://sims.ambertation.de/	SimPE - The Sims 2 Package Editor
http://video.google.com/	Google Video
http://www.gutenberg.org/wiki/Main_Page	Main Page - Gutenberg
http://blog.vodkaster.com/2009/06/25/the-top-250-best-movies-of-all-time-map/	The top 250 best movies of all time Map | Vodkaster - Le Blog de la...
http://www.xtranormal.com/	Xtranormal | Text-to-Movie
http://www.blogger.com/start	Blogger: Create your Blog Now -- FREE
http://video.stumbleupon.com/	StumbleVideo
http://failblog.org/	FAIL Blog: Pictures and Videos of Owned, Pwnd and Fail Moments
http://www.xtube.com/warning.php	X Tube - What Channel Are You On?!
http://www.afr.com/	Australian Financial Review
http://wikitravel.org/en/Main_Page	Free Worldwide Travel Guides - Wikitravel
http://www.pocketmod.com/	PocketMod: The Free Disposable Personal Organizer
http://video.stumbleupon.com/	StumbleVideo
http://belleandburger.blogspot.com/2009/06/panty-tutorial-how-to-make-your-own.html	belle and burger: Panty Tutorial: How to make your own drawers
http://www.pricegrabber.com/	PriceGrabber.com - Comparison Shopping Beyond Compare
http://video.stumbleupon.com/	StumbleVideo
http://video.stumbleupon.com/	StumbleVideo
http://readymech.com/	Fwis • Readymech Series 002
http://www.ruby-lang.org/en/	Ruby Programming Language
http://factcheck.org/	FactCheck.org
http://www.starfall.com/	Learn to Read at Starfall - teaching comprehension and phonics
http://satucket.com/lectionary/	The Lectionary
http://www.bubbl.us/	bubbl.us - free web application for brainstorming online
http://video.stumbleupon.com/	StumbleVideo
http://www.joelonsoftware.com/articles/Unicode.html	The Absolute Minimum Every Software Developer Absolutely, Positivel...
http://www.winemag.com/homepage/index.asp	Wine Enthusiast Magazine
http://www.angryalien.com/	Angry Alien Productions: 30-Second Bunnies Theatre and other cartoons.
http://www.kartoo.com/	KartOO visual meta search engine
http://www.filtermusic.net/#Lounge	□ FilterMusic □ Internet radio stations, electronic &amp; house music, ...
http://www.cineol.net/	.:: CINeol ::.
http://www.nextbrick.net/	Nextbrick
http://javadude.com/articles/passbyvalue.htm?repost	Java is Pass-by-Value, Dammit! - Scott Stanchfield
http://www.bigideagroup.net/	Big Idea Group: Home Page

Appendix B. Entry Survey

Introduction of the Study

The purpose of this research study is to find methods to identify topics or domains of research content. For this purpose, we will be asking participants to make judgment on how relevantly terms represent the topics of web resources. Participants who have specialty in classification and cataloguing will be recruited from Pittsburgh area libraries and the graduate school of Library and Information Sciences. Participants will be asked to completed approximately two hours long session which will include having a training on the experimental system, answering pre-questionnaire (for the first session only), and performing experiment.

Prior to the research experiment, please provide answers to following questions.

1. I am a
   - (1) librarian at ________________________
   - (2) MLIS degree holder
   - (3) MLIS student
   - (4) PhD Student in LIS

2. If you are a librarian, what is your <u>specialty (major tasks)</u> in your library?

   ____

3. If you are a graduate student, what is your <u>specialty (track or research interest)</u>?

   ____

4. If you are a graduate student in LIS, please <u>check all</u> of the course(s) you have taken.
   - ___ Organizing & Retrieving Information (LIS2005)
   - ___ Introduction to Cataloging and Classification (LIS2405)
   - ___ Advanced Cataloging and Classification (LIS2406)
   - ___ Metadata (LIS2407)
   - ___ Indexing and Abstracting (LIS2452)
   - ___ Thesaurus Construction (LIS2453)

5. How would you rate yourself as a professional in resource classification?

Very Bad	Bad	Fairly Good	Good	Excellent
1	2	3	4	5

6. How well do you understand the basics and concept of **classification schemes**?

Very Poor	Poor	Fairly Good	Good	Excellent
1	2	3	4	5

7. How well do you understand the basics and concept of **thesaurus**?

Very Poor	Poor	Fairly Good	Good	Excellent
1	2	3	4	5

8. How well do you understand the basics and concept of **subject headings**?

Very Poor	Poor	Fairly Good	Good	Excellent
1	2	3	4	5

9. How would you rate yourself in your ordinary life in organization?

Very Bad	Bad	Fairly Good	Good	Excellent
1	2	3	4	5

10. What do you organize for yourself in ordinary life? (Check all applies)

___ Personal Library (Books)
___ Personal Pictures (Albums)
___ Personal Computer Folders and Files
___ Web Pages (e.g. Favorites, Bookmarks)
___ Emails/Mails (e.g. Folders)
___ Important Documents (e.g. Contracts, Receipts, etc.)
___ Others: ____

Appendix C. Exit Survey

1. Do you think the terms listed represent the topics of the web pages enough?

Very Poor	Poor	Fairly Good	Good	Excellent
1	2	3	4	5

2. Do you think the terms you rated as 3-5 represent the topics of the web pages enough, 3 as an acceptable term to represent the topic, 4 as a good term to represent to topic, and 5 as an excellent term to represent to topic?

Very Poor	Poor	Fairly Good	Good	Excellent
1	2	3	4	5

3. What were your strategies in rating the topic terms of the web pages? Please rank them by the importance.

___ Title of the web page
___ Type of the web page (e.g. newspaper, magazine, etc.)
___ Format of the web page (e.g. text, image, video, etc.)
___ Publisher of the web page
___ Categories of the topic concept of the web page
___ Words used in the content of the web page
___ Words appear frequently in the web page
___ Words represent the subjects of the web page
___ Words that may appear in any subject headings or thesaurus
___ Words that may appearing in any classification schemes
___ Others : explain ____

4. For the terms that you thought to be bad ones to represent the topic of the web pages, what were the main reasons? Please check all that apply.

___ The term does not represent to content of the web pages
___ The term does not represent the topic of the web pages
___ The term is not the term used in the web pages
___ The term describes too broad domain to represent to subject area of the web page content

___ The term describes too specific domain to represent to subject area of the web page content

___ The term is not a word.

___ The term is not understandable.

___ The term is misspelled/misused.

___ The term is not a noun/gerund.

___ Others : explain ____

5. Based on the terms you rated for the study, what would suggest further in finding topics of a web resource? (e.g. Possible types of terms, possible metadata elements, etc.)

Bibliography

Agichtein, E., Brill, E., and Dumais, S. (2006). *Improving Web Search Ranking by Incorporating User Behavior Information*. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. August 06-11, 2006, Seattle, WA, USA.

Albert, R., Jeong, H., and Barabási, A. (1999). Diameter of the World-Wide Web. *Nature,* 401, pp. 130-131.

Arrington, M. (2007). Exclusive: Screen Shots and Feature Overview of Delicious 2.0 Preview [Web Posting]. Retrieved February 23, 2010, from http://techcrunch.com/2007/09/06/exclusive-screen-shots-and-feature-overview-of-delicious-20-preview/

Baca, M. (Ed.). (1999). *Introduction to Metadata: Pathways to Digital Information*. Los Angeles, CA, USA: Getty Information Institute.

Bao, S., Wu, X., Fei, B., Xue, G., Su, Z., and Yu, Y. (2007). *Optimizing Web Search Using Social Annotations*. In Proceedings of the 16th International World Wide Web Conference (WWW2007). May 8-12, 2007, Banff, Alberta, Canada, pp. 501-510.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001, May). The Semantic Web. *Scientific Ametican, 284*(5): 35.

Bischoff, K., Firan, C.S., Nejdl, W., and Faiu, R. (2008). *Can All Tags be Used for Search?* In Proceedings of Conference on Information and Knowledge Management (CIKM '08). Nap Valley, California, USA, October 26-30, 2008.

Burnett, K., Ng, K. B., and Park, S. (1999). A Comparison of the Two Traditional Metadata Development. *Journal of The American Society For Information Science*, *50*(13), pp. 1209-1217.

Caplan, P. (2003). *Metadata Fundamentals for All Librarians*. Chicago, IL, USA: American Library Association.

Cardinaels, K., Meire, M., and Duval, E. (2005). *Automating Metadata Generation: the Simple Indexing Interface*. In Proceedings of 14th International World Wide Web Conference, May 10-14, 2005, Chiba, Japan, pp. 548-556.

Cardoso, J. and Sheth, P. (Eds.). (2006). *Semantic Web Services, Processes and Applications*. New York, NY, USA: Springer.

Chan, L. M. (1994). *Cataloging and Classification: An Introduction*. (2nd ed.). USA: McGraw-Hill.

Choochaiwattana, W. (2008). Using Social Annotations to Improve Web Search. Doctoral Thesis. University of Pittsburgh.

Choochaiwattana, W. and Spring, M. B. (2009). *Applying Social Annotations to Retrieve and Re-rank Web Resources*. In Proceedings of International Conference on Information Management and Engineering (ICIME 2009). April 03 - 05, 2009, Kuala Lumpur, Malaysia.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. (2nd Ed.). Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates.

Delicious. (2010). In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/wiki/Delicious_(website)

Di Fenizio, P. S. (2005). *Tagclouds and cultural changes* [Blog Posting]. Retrieved May 26, 2010, from http://blog.pietrosperoni.it/2005/05/28/tagclouds-and-cultural-changes/

Emamy, K. and Cameron, R. (2007). Citeulike: A Researcher's Social Bookmarking Service. *Ariadne, Issue 51*. Retrieved October 29, 2009, from http://www.ariadne.ac.uk/issue51/emamy-cameron/

Evans, M., Newman, R., Rutnam, T., and Griffiths, D. (2005). Search Adaptations and the Challenges of the Web. *IEEE Internet Computing, 9(3)*, pp. 19-25.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, *76(5)*, pp. 378-382.

Flickr. (2010). In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/wiki/Flickr

Golder, S. A. and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, *32*(2), 198-208.

Greenberg, J. (2003). Metadata and the World Wide Web. In Kent, A. (Ed.), *Encyclopedia of Library and Information Science* (Vol. 72, pp. 244-261). New York: Marcell Dekker, Inc.

Greenberg, J. (2004). Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Applications. *Journal of Internet Cataloging*, *6*(4), pp. 59-82.

Guy, M. and Tonkin, E. (2006). Folksonomies – Tidying up Tags? *D-Lib Magazine, 12*(1), Retrieved October 29, 2009, from http://www.dlib.org/dlib/january06/guy/01guy.html

Hammond, T., Hannay, T., Lund, B., and Scott, J. (2005). Social Bookmarking Tools (I). *D-Lib Magazine, 11(4),* Retrieved February 23, 2010, from http://www.dlib.org/dlib/april05/hammond/04hammond.html

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z. and Fox, E. A. (2003). *Automatic Document Metadata Extraction using Support Vector Machines*. In Proceedings of the Joint Conference on Digital Libraries (JCDL 2003), May 27-31, 2003, Houston, TX, USA, pp. 37-48.

Heymann, P., Koutrika, G., and Garcia-Molina, H. (2008). *Can Social Bookmarking Improve Web Search?* In Proceedings of Web Search and Web Data Mining (WSDM '08). Palo Alto, California, USA, February 11-12, 2008.

Hotho, A., Jaschke, R., Schmitz, C., and Stumme, G. (2006a). *Information Retrieval in Folksonomies: Search and Ranking*. In Proceedings of the 3rd European Semantic Web Conference, June 11-14, 2006, Budva, Montenegro, pp. 411-426.

Hotho, A., Jaschke, R., Schmitz, C., and Stumme, G. (2006b). *Trend Detection in Folksonomies*. In Proceedings of First International Conference on Semantics and Digital Media Technology (SAMT 2006), December 6-8, 2006, Athens, Greece, pp. 56-70.

International Federation of Library Associations (IFLA) Metadata (2005), Retrieved from http://www.ifla.org/II/metadata.htm

Jarvelin, K. and Kekalainen, J. (2000). *IR Evaluation Methods for Retrieving Highly Relevant Documents*. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development on Information Retrieval. July 24-28, 2000, Athens, Greece.

Jenkins, C., Jackson, M., Burden, P., and Wallis, J. (1999). *Automatic RDF Metadata Generation for Resource Discovery*. In Proceedings of the 8th International World Wide Web Conference (WWW8), May 11-14, 1999, Toronto, Canada.

John, A. and Seligmann, D. (2006). *Collaborative Tagging and Expertise in the Enterprise*. Collaborative Web Tagging Workshop in the 15th International World Wide Web Conference (WWW2006). May 23-26, 2006, Edinburgh, Scotland.

LibraryThing. (2010). In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/wiki/LibraryThing

Lin, X., Beaudoin, J. C., and Desai, K. (2006). *Exploring Chracteristics of Social Classification*. In Proceedings of the 17th ASIS&T SIG/CR Classification Research Workshop. November 4, 2006, Austin, TX.

Macgregor, G. and McCulloch, E. (2006). Collaborative tagging as a knowledge organization and resource discovery tool. *Library Review, 55*(5), pp. 291-300.

Marlow, C., Naaman, M., Boyd, D., and Davis, M. (2006). *HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead*. In Proceedings of the17th Conference on Hypertext and Hypermedia 2006. August 22-25, 2006, Odense, Denmark, pp. 31-40.

Mathes, A. (2004). *Folksonomies – Cooperative Classification and Communication Through Shared Metadata* [Online Report]. Retrieved October 29, 2009, from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.pdf

Mika, P. (2007). Ontologies Are Us: A Unified Model of Social Networks and Semantics. *Journal of Web Semantics*, *5*(1), pp. 5-15.

Möller, G., Carstensen, K., Diekmann, B., and Watjen, H. (1999). *Automatic Classification of the World-Wide Web using the Universal Decimal Classification*. In Proceedings of the Gesellschaft für Klassifikation, Bielefeld, Heidelberg: Springer.

Noll, M. G. and Meinel, C. (2007). *Author vs. Readers – A Comparative Study of Document Metadata and Content in WWW*. In Proceedings of the 2007 ACM symposium on Document Engineering (DocEng 2007). August 28-31, 2007, Winnipeg, Manitoba, Canada, pp. 177-186.

Noll, M. G. and Meinel, C. (2008). *Exploring Social Annotations for Web Classification*. In Proceedings of the 2008 ACM symposium on Applied Computing (SAC 2008). March 16-20, 2008, Fortaleza, Ceara, Brazil, pp. 2315-2320.

Ohmukai, I., Hamasaki, M., and Takeda, H. (2005). *A Proposal of Community-based Folksonomy with RDF Metadata*. In Proceedings of the ISWC 2005 Workshop on End User Semantic Web Interaction. Galway, Ireland, November 7, 2005.

Olson, H. A. and Boll, J. J. (2001). *Subject Analysis in Online Catalogs* (2nd ed.). Englewood, CO, USA: Libraries Unlimited.

Owens, L. A. (2006). Thesaurus Evaluation. *Cataloging & Classification Quarterly*, 37(3), pp. 87-102.

Pandia Search Engine News. (2007). The size of the World Wide Web. Retrieved from http://www.pandia.com/sew/383-web-size.html

Paynter, G. W. (2005). *Developing Practical Automatic Metadata Assignment and Evaluation Tools for Internet Resources*. In Proceedings of Joint Conference for Digital Libraries (JCDL05), June 7-11, 2005, Denver, Colorado, USA, pp. 291-300.

Quintarelli, E. (2005). *Folksonokies: power to the people* [Online Report]. ISKO Italy-UniMIB meetings: Milan. June 24, 2005. Retrieved October 29, 2009, from http://www.iskoi.org/doc/folksonomies.htm

Randolph, J. J. (2005). *Free-Marginal Multirater Kappa (multirater K_{free}): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa*. The Joensuu Learning and Instruction Symposium 2005, October 14-15, 2005, Joensuu, Finland.

Reitz, J. M. (2004). *Dictionary for Library and Information Science* [Electronic version]. West Port, CT, USA: Libraries Unlimited.

Rethlefsen, M. L. (2007). Tags Help Make Libraies Del.icio.us: Social bookmarking and tagging boost participation. *Library Journal*. Retrieved October 29, 2009, from http://www.libraryjournal.com/article/CA6476403.html

Salton, G. and Buckley, C. (1988). Term-Weighting Approaches In Automatic Text Retrieval. *Information Processing & Management, 24(5),* pp. 513-523.

Schacter, J. (2006). now serving: 1,000,000 [Web Posting]. September 25, 2006. Retrieved February 23, 2010, from http://blog.delicious.com/blog/2006/09/million.html

Sen, S., Harper, F.M., LaPitz, A., and Riedl, J. (2007). *The Quest for Quality Tags*. In Proceedings of Conference on Supporting Group Work (GROUP '07). Sanibel Island, Florida, USA, November 4-7, 2007, pp. 361-370.

Shafer, K. (1997). Scorpion Helps Catalog the Web. *Bulletin of the American Society for Information Science*, *24*(1). Retrieved October 29, 2009, from http://www.asis.org/Bulletin/Oct-97/shafer.htm

Shirky, C. (2005). *Ontology is Overrated: Categories, Links, and Tags* [Blog Posting]. Retrieved October 29, 2009, from http://www.shirky.com/writings/ontology_overrated.html

Smiraglia, R. P. (Ed.). (2005). *Metadata: A Cataloger's Primer*. Binghamton, NY, USA: Haworth Information Press.

Smith, G. (2008). *Tagging: People-Powered Metadata for the Social Web*. Berkeley, CA, USA: New Riders.

Specia, L. and Motta, E. (2007). Integrating Folksonomies with the Semantic Web. *The Semantic Web: Research and Applications* (Proceedings of the 4th European Semantic Web Conference, June 3-7, 2007, Innsbruck, Austria), Berlin Heigelberg, Germany: Springer.

Syn, S.Y. and Spring, M.B. (2008). Can a system make novice users experts?: Analysis of metadata created by novices and experts with varying levels of assistance. *Int. J. Metadata, Semantics and Ontologies, 3(2).* pp.122–131.

Syn, S.Y. and Spring, M.B. (2009). *Tags as Keywords – Comparison of the Relative Quality of Tags and Keywords*. In Proceedings of ASIS&T 2009 Annual Meeting. Vancouver, BC, Canada, November 6-11, 2009.

Taylor, A. G. (2004). *Wynar's Introduction to Cataloging and Classification*. (9th ed.). West Port, CT, USA: Libraries Unlimited.

Taylor, A. G. and Clemson, P. (1996). *Access to Networked Documents: Catalogs? Search Engines? Both?* OCLC Internet Cataloging Project Colloquium Position Paper, American Library Association Midwinter Conference, January, 1996, San Antonio, TX, USA. Retrieved October 29, 2009 from http://www.worldcat.org/arcviewer/1/OCC/2003/07/21/0000003889/viewer/file9.html

Taylor, A. G. and Jourdrey, D. N. (2008). *The Organization of Information*. (3rd ed.). Westport, CT, USA: Libraries Unlimited.

The United Kingdom Office for Library and Information Networking (UKOLN) Metadata (2008), Retrieved from http://www.ukoln.ac.uk/metadata/

Tonkin, E. (2006). *Searching the long tail: Hidden structure in social tagging*. In Proceedings of the 17th ASIS&T SIG/CR Classification Research Workshop. November 4, 2006, Austin, TX.

Toth, E. (2002). Innovative Solutions in Automatic Classification: A Brief Summary. *Libri, 52*(1), pp. 48-53.

Trant, J. (2006). Exploring the potential for social tagging and folksonomy in art museums: proof of concept. *New Review of Hypermedia and Multimedia, 12*(1), pp. 83-105.

Wenzler, J. (2007). *LibraryThing and the Library Catalog: Adding Collective Intelligent to the OPAC*. A Workshop on Next Generation Libraries CARL North IT Interest Group, September 7, 2007, San Francisco, CA, USA.

Wu, H., Zubair, M., and Maly, K. (2006). *Harvesting Social Knowledge from Folksonomies*. In Proceedings of the 17th Conference on Hypertext and Hypermedia 2006. August 22-25, 2006, Odense, Denmark, pp. 111-114.

Wu, X., Zhang, L. and Yu, Y. (2006). *Exploring Social Annotations for the Semantic Web*. In Proceedings of the 15th International World Wide Web Conference (WWW2006), May 23-26, 2006, Edinburgh, Scotland, pp. 417-425.

Yanbe, Y., Jatowt, A., Nakamura, S., and Tanaka, K. (2007). *Can Social Bookmarking Enhance Search in the Web?*. In Proceedings of the 7th Joint Conference on Digital Libraries (JCDL07). June 18-23, 2007, Vancouver, BC, Canada, pp. 107-116.

Yi, K. and Chan, L.M. (2009). Linking folksonomy to Library of Congress Subject Headings: An Exploratory Study. *Journal of Documentation, 65(5).* pp. 872-900.

Yilmazel, O., Finneram, C. M., and Liddy, E. D. (2004). *MetaExtract: An NLP System to Automatically Assign Metadata*. In Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL04), Tuscon, AZ, USA, pp. 241-242.

Zubiaga, A., Martinez, R., and Fresno, V. (2009). *Getting the Most Out of Social Annotations for Web Page Classification*. In Proceedings of Document Engineering (DocEng '09). Munich, Germany, September 16-18, 2009.