This is a preliminary version of the JASIST submission.

Perspectives on Social Tagging

[1]Ying Ding (dingying@indiana.edu)

Elin K. Jacob (ejacob@indiana.edu)

School of Library and Information Science, Indiana University, 1320 E 10th, Bloomington, IN 47405 USA. Tel: +1 812 855 5388; Fax: +1 812 855 6166

Zhixiong Zhang (zhangzx@mail.las.ac.cn)

Library of Chinese Academy of Sciences, Beijing, China.

Schubert Foo (sfoo@pmail.ntu.edu.sg)

Division of Information Studies, Nanyang Technological University, Singapore.

Erjia Yan (eyan@indiana.edu)
Nicolas L. George (nlgeorge@indiana.edu)

Lijiang Guo (lijguo@indiana.edu)
School of Library and Information Science, Indiana University, Bloomington, USA

Abstract

Social tagging is one of the major phenomena transforming the World Wide Web from a static platform into an actively shared information space. This paper discusses various aspects of social tagging, including different views on the nature of social tagging, how best to make use of social tags, and how to bridge social tagging with other Web functionalities. It also discussed the facet management of tagging data to ease browsing and searching. The analogy between bibliometrics and tagometrics has been discussed and compared. Established bibliometric methodologies can be applied to analyze tagging behaviour on the Web. Based on the Upper Tag Ontology (UTO), the web crawler has been built up to crawl tag data from Delicious, Flickr and YouTube in September 2007. In total, 1.8M objects including bookmarks, photos and videos, 3.1M taggers and 12.1M tags have been collected and analysed. Some tagging patterns and variations have been identified and discussed.

Keywords: social tagging; Semantic Web; Upper Tag Ontology; Ontology alignment; Social Web

[1] Corresponding author.

1. Introduction

The World Wide Web (Web) is evolving from a static platform into an actively shared information space. Participation of the average user in the first generation Web (Web 1.0) consisted primarily of browsing online content. Although documents were connected through hyperlinks, allowing the user to move easily from one resource to another, the average Web 1.0 user was locked into a one-way process of communication, much like reading a book, and was not actively involved either in online discussions or information sharing. Because writing HTML pages was beyond the skill level of many users, it was difficult for the average user to publish data on the Web. More importantly, resources published in the Web 1.0 environment consisted of simple character strings that adhered to a prescribed syntactic format and the machine had no way to interpret the meanings of strings that did not have adequate semantics embedded in them.

The current Web environment is the second generation Web, often referred to as the Social Web or Web 2.0. The phrase "Social Web" was introduced in 1998 by Peter Hoschka to stress the social function of the Web (Hoschka, 1998). Social Web is as an open and globally distributed data sharing network that links people, organizations and concepts. The scope of the Social Web is extended here to include any Web-related technologies, phenomena and developments that enhance the social features of the Web.

Web 2.0 is a substrate of the Social Web that provides platforms and technologies (e.g., wikis, blogs, tags, RSS feeds, etc.) for online communication and collaboration. Communication on the Web has morphed from one-way communication to human-to-human communication; and the Web has become a convenient platform for users to publish and share information. The average user not only browses Web 2.0 but is also actively involved in online communication, including publishing of resources, tagging of interesting bookmarks, and sharing of images and videos. Online publishing in Web 2.0 is now so easy that anyone who can type can publish. This has spurred users of all ages, from

teenagers to seniors, to become involved in the Web communication. One of the newest ways of communicating via Web 2.0 is through the activity of tagging. Tagging is the act of adding keywords online resources. In this way, the World Wide Web is evolving from a collection of hyperlinked documents to a hyperlinked Web of Data.

What is the future of the World Wide Web? While some talk of Web 3.0, others speak enthusiastically of the Semantic Web. Regardless of what the next generation Web is to be called, it will have certain features that are foreseeable even now. For example, the average user will be able to interact with the Web just as Tim Berners-Lee, Jim Hendler and Ora Lassila described in *Scientific American* (Berners-Lee, Hendler & Lassila, 2001). The communication style of the Web will be human to machine, and it may be difficult for the user to determine if he is communicating with another human being or with a Web agent. On the current Web 2.0, human-created metadata, including tags and other social ontological data (e.g., FOAF in RDF/XML), is growing daily, and even more machine processable metadata will be available on the future Web. The trend toward introduction of additional data about Web resources will include not only the descriptive metadata of universal schemes such as the Dublin Core but also metadata generated within the Semantic Web community that is explicitly defined by an underlying ontology. Machines will also contribute data that have been generated automatically based on pre-defined ontologies. Moreover, these data will no longer be solitary annotations but will be interlinked (Miller, 2008). Based on the four principles of linking open data proposed by Tim Berners-Lee (see Link Open Data initiative available at http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData), more and more semantic data will be available to link concepts or instances using owl:sameAs or fofa:knows. These powerful semantic links will weave the current Web into a Web of semantically linked social data. Table 1 summarizes the three generations of the World Wide Web.

Table 1. Summary of three generations of the World Wide Web

	Traditional Web (Web1.0)	Social Web (Web2.0)	Semantic Web (Web3.0)
Average user	Browsing	Browsing Publishing Organizing	Browsing, Publishing, Organizing, Interacting
Communication style	One-way (e.g. reading a book)	One-way, Human-to-human (sharing)	One-way, Human-to-human, Human-to-machine (query-answering)
Data	Resources (syntactic data) - content and format mixed - documents hyperlinked	Resources, tags, metadata - content and format separated - data linked	Resources, tags, metadata - content and format separated - ontological data - data semantically linked
Data contributor	Webmaster or experienced user	Average user	Average user and web agents
Linking data	Hyperlinks	Different types of hyperlinks	Semantic links
Adding data	Composing HTML pages	Online publishing, tagging	Online publishing, tagging, machine generated data

This paper discusses various aspects of social tagging, including different views on the nature of social tagging, how best to make use of social tags, and how to bridge the phenomenon of social tagging with other Web functionalities. One main concern of this paper is how to model social tagging so as to mediate and link social data. Section 2 presents a brief review of different approaches to the study of tags and social tagging, and introduces the Upper Tag Ontology (UTO) as our approach to model social tagging data. Section 3 discusses the Universal Tag Identifier (UTI) and its important for unique identification of the tag and for being able to reference tags. Section 4 shows the potential link between social tagging and bibliometrics and provides one co-tag analysis for Delicious. Section 5 summarizes some features of three social tagging networks: Delicious, Flickr and YouTube. Section 6 concludes the paper and points out directions for future work with social tagging networks.

2. Approaches to Tagging

Sinha (2005) approaches tagging from the perspective of cognitive science, finding the “wisdom of crowds” in this new social phenomenon (Surowiecki, 2004). Sinha argues that taggers enjoy being

embedded in a social environment, being watched by others, and receiving feedback on their actions. Furthermore, social tagging can lead to the formation of like-minded groups, thereby enabling social discovery and the development of connections among taggers. These emergent groups can grow to reflect not only the wisdom of crowds, but also the diversity of opinion that emanates from and is supported by the independence of individual group members. Tagging helps to spread ideas, memes, trends and fashions. The act of tagging reflects an individual's conceptual associations and enables loose coordination, but it does not enforce a single interpretation of a tag or a concept. More importantly, tagging works because it strikes a balance between the individual and the social. Suchanek, Vojnovic and Gunawardena (2008) have conducted a social tagging analysis on 65,000 Delicious bookmarks and a user study over 4000 participants. They found that the more popular a tag is and the more likely it is to be meaningful. Their analysis also validates the assumption that the more users who have tagged a document, the more meaningful the popular tags assigned to that resource will be.

Hammond, Hannay, Lund and Scott (2005) have discussed user motivations for tagging, aligning them with popular tagging sites. According to Hammond et al., these motivations range from the 'selfish' to the altruistic: from tagging of one's own content for personal use (e.g., Flickr) or tagging one's content for the use of others (e.g., Technorati and HTML meta tags) to tagging another's content for one's personal use (e.g., Delicious, CiteULike, Connotea) or tagging the content of another for the use of others (e.g., Wikipedia). Lawley (2008) presents another view of tagging when she cautions that it can be easily manipulated to produce negative social consequences, such as, to make political statements, or to promote personal interests (Blood, 2005).

Tagging is one of the main ways for adding metadata to resources in Web 2.0. Mathes (2004) describes different methods used to generate metadata. In libraries, metadata has traditionally been created by trained professionals using standards such as Machine-Readable Cataloging (MARC) and controlled vocabularies such as the Library of Congress Classification (LCC), the Dewey Decimal

Classification (DDC) or the Library of Congress Subject Headings (LCSH). These professionally created metadata records form the basis for user displays in online public access catalogs (OPACs) and are generally of high quality; but they are time consuming to create and thus impractical for handling the volume of resources available on the Web. Metadata has also been generated by content creators using HTML meta tags as well as a myriad of metadata standards, including Dublin Core (DC), the Text Encoding Initiative (TEI), and the VRA Core 4.0 (VRA) developed by the Visual Resources Association, to name but a few. All too often, however, content creators using these standards have not been formally trained in metadata generation, leading to inadequate, inaccurate or "noisy" resource descriptions. More importantly, in both of these approaches, the metadata generation process is disconnected from the actual users of the resources.

A more practical approach to metadata creation is to allow the users of resources to generate metadata records online through social tagging. Social tagging reflects the vocabulary and conceptual associations of users (Pika, 2005). Unlike formal taxonomies and classification schemes, which specify explicit relationships among terms, the accumulation of tags from individual contributors is "metadata for the masses" and reveals the digital equivalent of the “desire lines” (Merholz, 2004) created in the physical landscape by purposeful foot traffic. This collective vocabulary reflects the interests of users; more importantly, this vocabulary speaks the same language as users. With their uncontrolled nature and organic growth, user-generated vocabularies have the ability to adapt quickly to changes in both the needs and the vocabulary of users. Barriers to entry and the cognitive effort required to use such vocabularies are very low. Indeed, Butterfield (2004) has claimed that the lack of hierarchy, of synonym control, and of semantic precision -- the hallmarks of controlled vocabularies -- are exactly why user-generated vocabularies work as well as they do. The freedom of natural language, with its everyday familiarity and loose associations, is more intuitive and requires less cognitive effort than making a decision about how well a pre-defined category captures the content of a resource or

represents the immediate needs of the user. In the future, it may be feasible to combine these approaches to metadata creation in order to achieve optimal representation and retrieval of resources -- it may be possible to harvest user-generated tags in order to develop controlled vocabularies that truly speak the language of specific user groups and communities.

In social tagging applications, tags are used for navigating, for browsing and for retrieving resources. These systems often provide tag recommendation mechanisms (generally based on co-occurrence) that are used both to facilitate the identification of appropriate tags for a resource and to encourage consolidation of a tagging vocabulary across users. Recent work on more specialized topics, such as structure mining of user-generated vocabularies, has attempted to visualize trends (Dubinko et al., 2006) and identify patterns (Schmitz, Hotho, Jaschke & Stumme, 2006) in tagging behaviour or to rank terms in a vocabulary (Hotho, 2006). Xu, Fu, Mao and Su (2006) introduced a collaborative tag suggestion approach that, based on the HITS algorithm (Kleinberg, 1999), computes a measure of the goodness of tags rooted in collective user authorities that are iteratively adjusted using a reward-penalty algorithm. Fountopoulos (2007) designed the RichTags system, which uses Semantic Web technologies to overcome the weaknesses of conventional social tagging systems (e.g., polysemy, synonymy, and the basic problems of spelling variations, acronyms, etc.). The RichTags system contains a pre-defined ontology of meaningful tag concepts that is collectively maintained and updated by the users of the system.

Modelling of social tagging behaviours can help to organize tagging data and to interlink it with data from other social applications. Tom Gruber (2007) has proposed the use of ontologies to model tagging data. His ontology contains tagging concepts (*object*, *tag*, *tagger*, *source*, + and -) and introduces *vote* for collaborative filtering. The SCOT (Social Semantic Cloud of Tags) ontology represents the structure and semantics of a collection of tags as well as social networks among users (Kim, Passant, Breslin, Scerri and Decker, 2008). Holygoat Tag Ontology (available at

http://www.holygoat.co.uk/projects/tags/) models the relationship between an agent, an arbitrary resource and one or more tags; and taggers are linked to FOAF and RSS using rdfs:subClassOf or rdfs:subPropertyOf, which support simple subsumption inferences. The Meaning Of A Tag (MOAT) ontology (available at http://moat-project.org/ontology) is a lightweight ontology used to represent how different meanings are related to a tag and focuses on providing unique identifiers for those tags which have an associated semantic meaning.

The Upper Tag Ontology (UTO) is an upper level ontology for social tagging that is designed to circumvent the complexity and potential redundancy inherent in user-generated tagging vocabularies. Let O be the UTO ontology, $O = (C, \Re)$ (1)

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

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

In UTO, $C = \{Tag, Tagging, Object, Tagger, Source, Date, Comment, Vote\}$,

$$\Re = \begin{Bmatrix} has\,\mathrm{Re}latedTag, hasTag, hasObject, hasSource, hasDate, hasCreator, \\ hasComment, hasVote \end{Bmatrix}$$

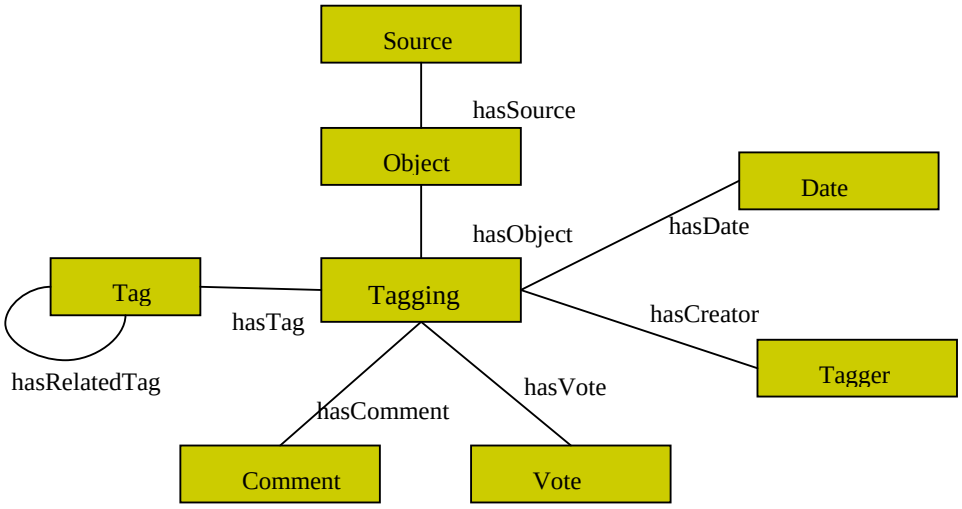

Figure 1. The Upper Tag Ontology (UTO)

UTO is based on Gruber's (2007) suggestion that an ontology can be used to model tagging data; but it extends this idea with its focus on ontology alignment and the integration of tagging data with other sources of social metadata. The emphasis in UTO is on the structure of tagging behaviours rather than the meaning of the tags themselves. By focusing on the structure of social tagging behaviours rather than tag semantics, this simple and easy-to-use ontology is able to integrate metadata from one social tagging community with metadata from other social tagging sites. More details about UTO and its role for integrating social tagging data are discussed in [Reference].

3. Universal Tag Identifier (UTI)

As social tagging becomes more popular, it is important to find efficient ways to identify, manage and reference the increasing abundance and diversity of tag data. Identification is the key enabler for data integration because identification of tags on a global level can facilitate disambiguation of tag meanings and uses. Already, efforts are under way to support the global identification of tags. The OpenID initiative (http://openid.net/) allows Web users to have one Web account for logging in to different websites: Using a Yahoo! account, for example, an individual can log in to AOL, Livejournal, DIGG, etc. The MOAT project provides a framework for taggers to produce semantically annotated content by using URLs of existing resources from DBPedia (available at http://dbpedia.org/), geonames or any other well-accepted knowledge base that offers pre-defined meanings for tags. Rather than having users type in tags, it should be possible for them to select tags from pre-defined lists of tags (descriptors) such as WordNet, LCSH, DDC, Wikipedia or any well-known social tagging network such as Delicious. These controlled tag lists would have a Universal Tag Identifier (UTI) -- a URI for each tag -- that would make reuse of or reference to tags easier.

Because pre-defined tag lists are similar to traditional controlled vocabularies (indeed, some tag lists could be provided by utilizing existing vocabularies such as LCSH or DDC), some pundits may argue that one of the primary advantages of social tagging is that tags are natural language keywords and thus uncontrolled, allowing users to select any word(s) to represent an online resource. This argument is most powerful when applied to individual taggers who are interested in representing resources for their own future use. But social networks are platforms not only for storing one's own bookmarks but for sharing those bookmarks as well. When the intent of taggers is to share resources with other users, natural language keywords can sometimes hamper retrieval.

Using an identifier for a tag or drawing tags from a controlled vocabulary facilitates easy sharing and reuse of resources as well as tags. Tagging and traditional indexing are similar in that the objective of both activities is to provide access to and support retrieval of a group of resources that share some feature. The primary difference between tagging and indexing is the source responsible for assigning a descriptor, be it a natural language tag or a term from a controlled vocabulary. Using descriptors from a controlled vocabulary ensures shared meaning and clean data, while using natural language keywords often leads to ambiguity and noisy data. Despite some well-known disadvantages of controlled vocabularies (e.g., the potential lack of currency among terms and the difficulties that can be associated with updating and maintaining a controlled vocabulary), they offer definite advantages for social tagging and could be an integral component of a combined approach: controlled vocabularies can be used to supplement natural language tagging and natural language tags supplied by users can be used to enrich and extend controlled vocabularies. For example, when a user bookmarks ski resorts with the tag *ski*, the LCSH descriptor *skiing* could be added to disambiguate the natural language tag and allow future users to evaluate the appropriateness of the resource (see Figure 2).

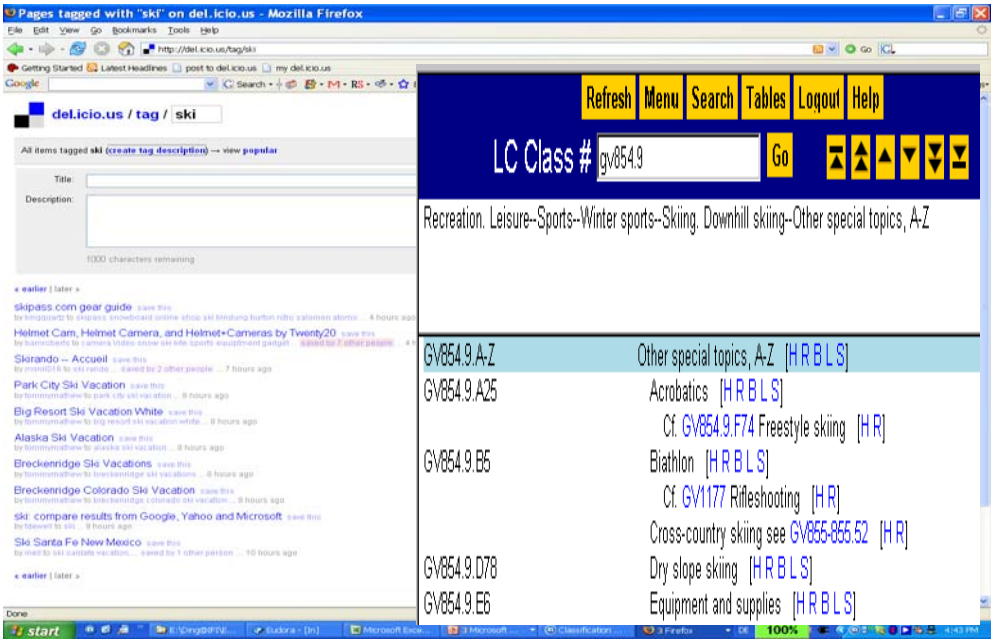

Figure 2. Subject Headings and Tags

FaceTag (available at http://www.facetag.org/) is a working prototype that demonstrates how the flat structure of user-generated tags can be combined with a faceted vocabulary to enrich an information system by building in relationships between tags (see Figure 3). If users are averse to the use of tags from controlled vocabularies because it seems to undermine individual freedom of choice, incorporating a simple but pre-defined faceted vocabulary can improve access to resources by supporting shared semantics. Through the simple addition of four basic facets (*resource types*, *themes*, *people* and *purposes*), users can supplement their own tags with descriptors chosen from the appropriate facet(s); and users searching or browsing Web resources can clearly view both the natural language tags and the faceted vocabulary that has been assigned to each resource. However, providing a unique identifier for each semantically meaningful tag would be a helpful addition to social tagging and would ultimately contribute to the development of the Semantic Web.

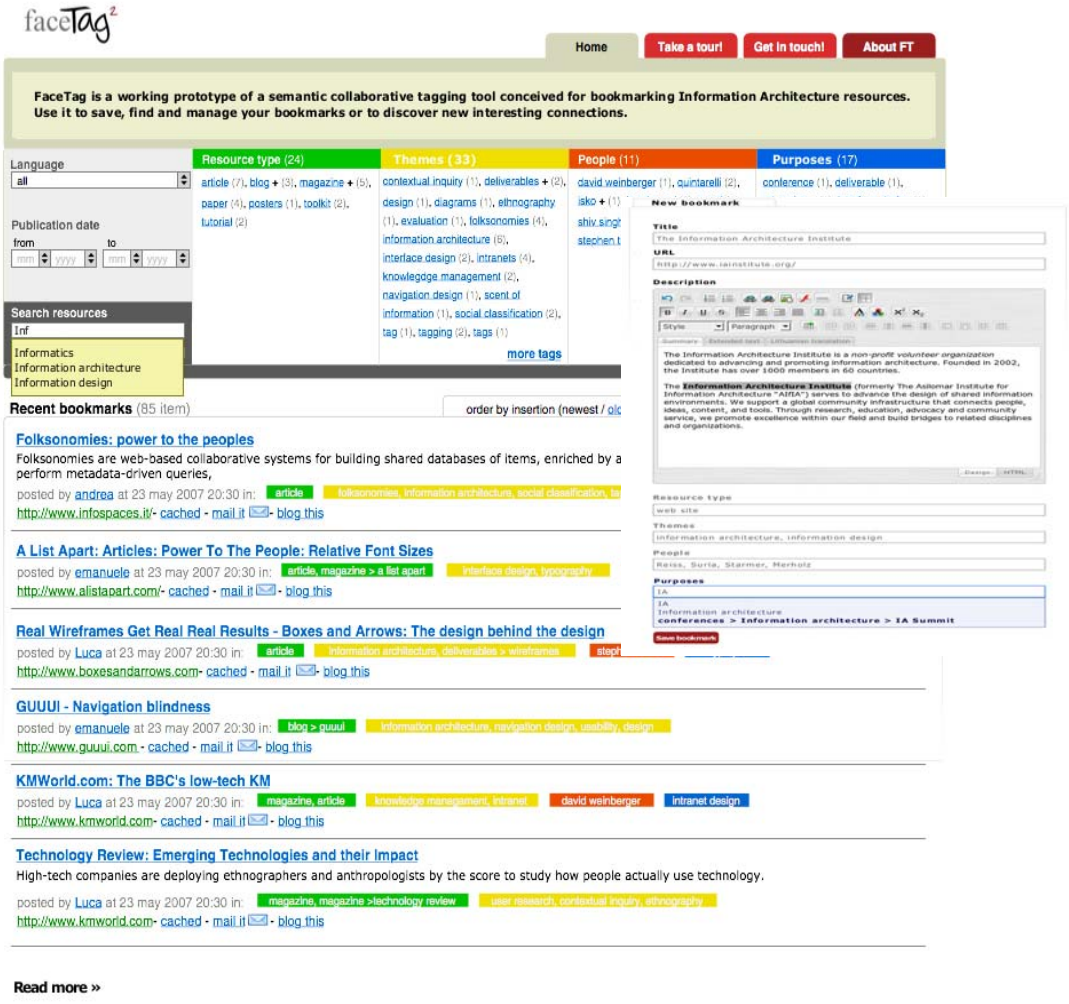

Figure 3. FaceTag Screenshot

FaceTag uses green (Resource type), yellow (Themes), red (People), and blue (Purposes) highlighting for tags assigned to resources in the retrieval set, allowing the searcher to identify tags from particular facets in a controlled vocabulary.

4. From Bibliometrics to Tagometrics

Tagging serves the function of reference and is thus one of the most important activities on the Social Web. Like the hypertext links of Web 1.0, tagging is a form of citation in Web 2.0, and all kinds of citation applications can be adapted to the tagging environment. For example, a future application

such as TagSeer could function as CiteSeer does today to rank online objects and taggers. Bibliometrics provides methodologies to map and measure scholarly communication based on citation behaviour; and similar methodologies can be applied as *tagometrics* to track and measure social communication online based on tagging behaviours. Bibliometrics focuses on the evaluation of scientific impact and influence using published documents; tagometrics would focus on the evaluation of social impact and influence using tags and votes.

Citation analysis and co-citation analysis are two bibliometric methods that can be applied to social tagging (Kipp and Campbell, 2006): Bibliometric measures used to investigate co-author relationships, co-word occurrence and co-journal citations can be extended to the analysis of co-taggers, co-tags and co-objects. Other methods commonly employed in bibliometric analysis can also be deployed to investigate tagging behaviour, including clustering, multidimensional scaling and factor analysis (see Table 3).

Table 3. Bibliometrics vs. Tagometrics

	Bibliometrics	**Tagometrics**
Impact analysis	Scientific/scholarly impact	Social impact
Object of analysis	Citation	Tag
Ranking	Authors, journals, subjects	Taggers, objects, tags
Citing object	Scholarly papers	Online objects (websites (bookmarks))
Citing purpose	Giving credit	Organizing, sharing, retrieving
Methods	Citation and co-citation analysis (co-author, co-journal, co-word)	tag and co-tag analysis (co-tagger, co-object, co-tag)
Application	Scientific evaluation, impact analysis, retrieval	Social impact analysis, retrieval
Byproducts	Co-word → taxonomy	Co-tag → folksonomy
Hybrid	Tagging publications (Cannotea, CiteULike, BibSonomy)	

Additional approaches that appear promising for analyzing tagging behaviours include Formal Concept Analysis and Support Vector Machine. Formal Concept Analysis (FCA) is a theory of data analysis based on concept lattices that was introduced by Rudolf Wille in 1982 (Wille, 1982). FCA is used to identify conceptual structures among data sets and has been applied in many fields including, among others, medicine, psychology, musicology, linguistics, software engineering, civil engineering, and ecology. Priss (2007) provides an excellent survey of the application of FCA in information science. The primary strength of FCA is its ability to reveal hidden relationships between objects and attributes, to construct the extension and intension of formal concepts, and to generate graphical visualizations of the inherent structural relationships among data and concepts. FCA can facilitate navigation and browsing in much the same way that faceted vocabulary does and can be used to generate tag lattices that will facilitate query reformulation. Like clustering, FCA is not good at handling large data sets; however, unlike clustering, which is a purely statistical approach, FCA introduces logical subsumption and inheritance into the analysis and thus has the potential to be a powerful component of tagometrics.

Support Vector Machine (SVM) is a supervised learning technique used in machine learning for both classification and regression. SVM algorithms are based on the principle of structural risk minimization (Cortes & Vapnik, 1995). In classification, SVM maps the input space into a high dimensional feature space and constructs an optimal separating hyperplane to create a maximal margin classifier, where *margin* means the minimal distance from the separating hyperplane to its closest data points. Although co-citation is generally based on clustering methods, the result of SVM can improve both the understanding and the labelling of clusters. The basic difference between clustering and SVM classification is that the data points used in clustering are unlabelled and the resulting clusters and data in the clusters are not ordered, while the data points in SVM classification are labelled and the data in classes are ordered. However, just like clustering, SVM is not good at dealing with large data sets. For

this reason, some researchers have used hierarchical clustering algorithms to provide better training sets in order to make SVM scalable for handling large data sets (Yu, Yang & Han, 2003; Awad et al., 2004)

Traditionally, multidimensional scaling has been used in bibliometrics to visualize clustered data. Other methods such as Self Organization Maps (SOMs) and Pathfinder Networks (PFNets) can also be used to map citation data (Kohonen, 1989; Schvaneveldt, 1990; White, Buzydlowski & Lin, 2000). Combining co-citation analysis with scientific impact evaluations such as the h-index evaluation (Wan, Hua & Rousseau, 2007) might be an interesting approach to explore for application in tagometrics.

Tagometric analysis for Delicious

The co-tag analysis has been conducted using the crawled Delicious data (see Section 5). We have crawled 9.3M tags (see Table 5). Tags are normalized and cleaned using WordNet, such as combining related noun and adjective (e.g., America and American → America), taking out stop words (e.g, such as an, a, and), combining different forms of verbs (e.g., go, goes, going →go), upper case and lower case (e.g. API, api → API) and so on. After that, tags with frequency more than 90 are selected to form the co-tag matrix, which leads to around 10,000*10,000 matrix. Clustering analysis was then applied to the co-tag matrix using an X-Means algorithm to cluster these data. X-Means is an unsupervised clustering algorithm that allows the specification of the minimum and maximum number of clusters generated during training (Pelleg & Moore, 2000). Table 4 presents some interesting clusters generated in this analysis. Cluster 1 contains 11 tags referencing programming languages; and Cluster 2 has five tags representing natural language topics. Cluster 3 has 11 tags dealing primarily with entertainment and entertainment media. Cluster 4 contains seven tags relevant to currency conversion; and Cluster 5 contains 16 tags generally related to banking and economic issues. Cluster 6 includes 13 tags focused on housing-related topics, Cluster 7 gathers 19 tags referencing colours and patterns, and Cluster 8 has five tags relevant to culture. Cluster 9 has eight tags addressing topics related to geography; and Cluster

10 includes five tags relevant to formatting. Furthermore, it is possible to draw some interesting insights based on this clustering of tags:

- Taggers used adjectives (i.e., the colours and patterns in Cluster 7) to tag bookmarks.
- When tagging resources related to currency conversion (Cluster 4), banking (Cluster 5), and housing (Cluster 6), taggers tended to use similar tags.

Table 4. Tag clusters in Delicious

Cluster	Tags
1	ajax, c, code, development, html, java, library, net, python, rails, rudy
2	dictionary, English, language, literature, writing
3	comic, entertainment, film, forum, japan, Japanese, movie, radio, streaming, television, tv
4	calculator, conversion, convert, converter, currency, euro, exchange
5	account, bank, banking, bill, consumer, credit, deal, doctor, financial, healthcare, insurance, loan, medical, medicare, medicine, savings
6	air, apartment, building, cleaning, do, fire, guide, house, housing, move, rental, safety, studio
7	Black, blue, brown, fairy, flower, gratis, leather, line, neo, pink, red, skull, stripes, style, Sweden, Swedish, vintage, white, yellow
8	culture, history, philosophy, politics, religion
9	astronomy, earth, geography, german, map, nasa, space, world
10	font, illustration, inspiration, portfolio, typography

5. Tagging Features of the Popular Social Networks

In order to test the underlying concept of UTO and to figure out some patterns for tagging behaviours, a web crawler was constructed that would incorporate and apply the elements of UTO in the collection of tagging data from three major social tagging websites – Delicious, Flickr and YouTube.

Crawling and integrating tagging data

In September 2007, Delicious was crawled to retrieve data about taggers, tags and bookmarks. The crawler began with the Delicious tag cloud at http://delicious.com/tag and visited every tag contained in the tag cloud. For instance, for TagA in the tag cloud, the crawler visited http://delicious.com/tagA and parsed the HTML code to grab information about bookmarks, taggers and related tags. For each bookmark, the crawler then went to http://delicious.com/url and crawled the history of the bookmark, focusing on which taggers had tagged this bookmark on which date(s). After gathering data about all of the bookmarks on the first page of TagA, the crawler continued to visit the second and subsequent pages for TagA, performing the same tasks, until it reached the arbitrarily set threshold of 99 pages for TagA. The crawler then repeated this process with subsequent tags in the tag cloud until it had visited all of the tags in the cloud. Following the same crawling method, data was also collected from Flickr and YouTube in September 2007. In total, the data retrieved contains 21 million RDF triples for Delicious, 2.3 million RDF triples for Flickr, and 2.2 million RDF triples for YouTube; Table 5 shows the details of these different datasets.

Table 5. Tag Data for Delicious, Flickr and YouTube

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

In total, our dataset contains around 1 million bookmarks, 2.8 million taggers and 9.3 million tags from Delicious, 0.3 million photos, 0.2 million taggers and 1.4 million tags from Flickr and 0.5 million videos, 0.2 million taggers and 1.4 million tags from YouTube. The average number of tags per object ranges from 2.74 (YouTube) to 9.31 (Delicious). The average number of tags the normal tagger uses ranges from 3.33 (Delicious) to 8.79 (Flickr). The average number of objects the normal tagger tags range from 0.36 (Delicious) to 2.84 (YouTube). Since when users upload bookmarks to Delicious, tag is

not the required field to fill in (but the title of the URL is a required field). So there might be many bookmarks with titles but without tags.

Hot topics in social tagging

Table 6. Top 20 tags in Delicious, Flickr and YouTube from 2005 to 2007

	Delicious			Flickr			YouTube		
Rank	2005	2006	2007	2005	2006	2007	2005	2006	2007
1	blog	blog	blog	2005	usa	2007	music	the	the
2	programming	programming	design	d70	california	canon	funny	funny	music
3	software	software	software	tsimshatsui	2006	nature	video	music	funny
4	music	design	programming	hongkong	cameraphone	autumn	the	video	video
5	design	reference	reference	nightview	celltagged	art	dance	live	girl
6	web	music	tools	germany	zonetag	nikon	crazy	of	of
7	reference	web	Web2.0	newkie	sanfrancisco	water	commercial	comedy	sexy
8	java	tools	web	ragbrai	blue	bw	live	dance	live
9	art	art	video	art	light	red	dancing	rock	dj
10	tools	java	music	wonder	sky	blue	guitar	cat	2007
11	linux	video	art	night	urban	sky	fun	halloween	dance
12	news	Web2.0	linux	buttersweet	red	japan	AMV	love	hot
13	xml	linux	webdesign	15fav	sea	fall	girl	girl	comedy
14	science	news	howto	central	me	beach	japan	movie	rock
15	search	tutorial	free	light	water	portrait	hot	dj	love
16	games	howto	tutorial	marco	nature	london	anime	in	and
17	research	imported	news	london	marco	night	Halloween	sexy	sex
18	technology	development	development	apargioides	london	green	halo	and	in
19	security	research	opensource	orange	green	usa	of	fight	new
20	video	internet	java	ads1	music	november	cat	you	cat

Table 6 summarizes the hot topics identified in the analysis of Delicious, Flickr and YouTube. Social tagging behaviours have increased greatly between 2005 and 2007. More and more users are relying on social tagging applications to index online resources for future retrieval by themselves and by others. Analysis of tagging behaviours offers an insight into the culture of a social network and can identify emerging trends and topics of increasing interest to a community as well as those topics in which user interest is declining.

The Delicious community shows a strong orientation toward IT topics, with many users interested in the web and in programming, as suggested by Mathes (2004). In contrast, the Flickr community contains two primary groups: professional photographers and non-professional photographers. The YouTube community is very broad and can be viewed perhaps as a representative subset of the larger Social Web community; however, the limited nature of resources tagged, in terms of format, makes it a

less appropriate application for studying tagging behaviours than Delicious. Furthermore, tagging is one of the major activities in Delicious, but not in Flickr or YouTube. Tagging in Delicious is used primarily for purposes of retrieval and sharing, while tagging in Flickr is used mainly for organizing one's own photographs and in YouTube for sharing videos. Taggers tend to represent the content of a resource in Delicious and YouTube, but focus on the specific feature(s) of a photograph in Flickr. While it is possible both to profile community interests and to use tag frequency to identify emerging trends in Delicious, this does not appear to be feasible with either Flickr or YouTube.

Tagging patterns

Based on the analysis of the top 1300 popular tags selected from the integrated tag sets of Delicious, Flickr and YouTube (see Reference), we identify several explicit tagging features (see Table 7). Users are interested to tag time. They simply use the current year or month to tag their objects, especially in Flickr as the current years are always the top-ranked tags (see Table 6). Users also like to tag geological locations, such as cities, countries and continents. They tag different scientific domains, such as bioinformatics, biology, ecology and so on. Different religions also appear in these top ranked tags, which contains Buddhism, Christianity and Islam. Computer programming languages, IT topics and large IT companies are commonly represented in this tag set, which might due to the community of Delicious which is formed by mainly IT-oriented users. Opinion terms are also used as tags to tag bookmarks, photos and videos, which shows that the users have the intention or needs to express their opinions on selected objects. Colour becomes popular tags in Flickr to represent the colour of the photos and as usual black and white are the most popular colours. Interestingly, celebrities, politicians (e.g., Britney, Spears, George, Bush) and other names also appear in this tag set to convey certain meaning that these names are no long just person names, but certain social topics and social trends as well. It is also interesting to see that users either use first name or last name to tag rather than use the name as a whole. It might due to the reason that some users do not know that blank spaces are used to

separate different tags by most of the social tagging websites. Also the social networks themselves are used as tags.

Table 7. Tagging features

Tagging Features	Tag Examples
Time	2005, 2006, 2007, November,
Geo-location	Africa, America, Amsterdam, Berlin, Argentina, Japan, China, Chicago, Chile, England,
Scientific domains	bioinformatics, biology, ecology, ecommerce,
Religion	Bible, Buddhism, Christianity, islam,
Programming	ajax, algorithm, C, C++, eclipse, hp, ibm, intel, j2ee, java, javascript, lisp, perl, python, ruby,
Company	apple, amazon, ebay, google, Microsoft, nokia, Nikon, oracle, w3c, yahoo,
Opinion	bad, cool,
Animal	bird,
Colour	black, bw, dark, white,
Name or Celebrity	Britney, Spears, Bush, George, james, john, Michael, David
Social topics	facebook, flickr, secondlife, youtube, Delicious, Wikipedia, myspace,
IT	database, datamining, enterprise2.0, itunes, ontology, owl, rdf, semanticweb, semweb, wordpress, xml, xslt,

Meanwhile we also identify some variations for the usage of tags (see Table 8). There exists lots of mix of using plural or singular of terms to tag objects, such as animal or animals, girls or girl and so on. Users also tend to use conjunctives, prepositions or articles to tag objects, which are quite counter-intuitive for tagging purposes (e.g. categorizing online objects). Lots of acronyms are used for tagging, especially for common computer terms, languages, and companies. Taggers also learned to use the compound terms to tag objects otherwise these terms will be separated due to the black space in between of them. Clearly, there exist wrong spelling for tags as well and providing some spelling help for tagging can be very useful. Verbs and adjectives are used to tag objects as well and some of them act as a memory of remindings, such as todo, toread and howto. There are also different languages used for tagging, such as foto, halo, musik and so on. We identify lots of variations, such as different spelling, plurals and singulars, acronyms and full names. Tag variations reflect the way how the crowds are organizing and classifying their online resources. These tags are varied syntactically, but they do contain certain semantics. It also shows the difference between folksonomy and the controlled vocabulary and which easily leads to the long-term debate on the “free-will tagging” or “a-little-bit-controlled tagging” (see Section 3).

Table 8. Tag variations

Tag variation	Tag Examples
Plural vs. Singular	animal(s), application(s), band(s), boy(s), car(s), cartoon(s), cat(s)
Conjunctives	and, at, all, for, from, of, on, one, out, the, to,
Acronym	api, au, ads, apps, bbc, bw, cms, cs, css, de, dc, dhtml, dj, diy, dns, dom, drm, dvd, el, en, fag, fic, gps, gui, hci, hdr, hp, nyc, rss, rpg,
Compound words	audiobooks, blackandwhite, cameraphone, celltagged, creativecommons, datamining, dotnet, enterprise2.0, filesharing, filesystem, firefox:bookmarks, firefox:rss, firefox:toolbar, googlemaps, graphicdesign, howto, losangeles, newmedia, newyork, projectmanagement, rubyonrails,
Wrong spelling	cultura, economica, educacion, fanfic,
Verb	computing, cooking, do, shopping, singing, todo, toread,
Adjective	cool, creative
variations	blog-blogger-blogging-blogs, bookmarking-bookarmks, barsil-brazil, color-colors-colour, conversion-convert-converter, el-elearning-e-learning, hack-hacking-hacks, lifehack-lifehacker-lifehacks, humor-humour, newyork-nyc, opensource-open-source, podcast-podcasting-podcasts, process'-processing, product-production-productivity-products, san-sanfrancisco-sf, tag-tagging-tags,
Other languages	foto, fotographia, fotos, halo, het, musik, musica,

Tagging features in different social networks

When comparing these three social networks, Delicious demonstrates the tightest connection to the use of tags as extended information about resources. In Delicious, every user can tag an object with the tag(s) of his own choice; and an object can be tagged many times and by multiple users, thereby indicating that it "belongs" (or is more relevant) to the community as a whole. So Delicious is a kind of community-tagging where anyone can tag any available online resources (here in Delicious are bookmarks) (Marlow, et.al., 2006). Similar social networks also include CiteULike and Connotea (online resources are bibliographical records), LibraryThing (online resources are books) and so on.

This is very different from Flickr, where content is mainly tagged by the user who uploads the photograph; the major community activities of other users can just "comment" or "vote" for resources by indicating that a particular photograph is a favourite image. Flickr also provides functions to allow users to tag photos uploaded by their friends. But this limits the community-tagging only to the users and their closed friends. So Flickr is not a fully community-based tagging system rather a kind of self-tagging system for the users and their closed friends. YouTube has a system similar to that of Flickr. A user can tag the content (videos) he has uploaded and the public is able to vote for them by assigning "stars".

So the different tagging rights have created the difference among the nature and types of resultant tags and the role of tags in the systems (Marlow, et.al., 2006). Based on the analysis of the top 20 tags in each social network, we found out that the tags in Delicious are more content-oriented which are related either to topics of the bookmarks. While the tags in Flickr are more annotation-oriented which are related to the features of the photos, such as colour, year and location. While the tags in YouTube are content and feature oriented which are somewhere related to the content and the feature of the videos. The role of tags in Delicious is to organize the bookmarks and help to retrieve and share the bookmarks. Tags play the major role in Delicious as Delicious counts on them for the users to share and find bookmarks which are the major functions of Delicious. While the tags in Flickr are a kind of side-effect which means that it is not necessary to tag your photo and it is up to the choice of the users. Photos can be searched via title of the photos and ranked by comments and votes. Tagging does not play a major role in Flickr. The same for YouTube, many YouTube users do not actually tag their videos. Videos are shared through comments and votes.

Social tagging behaviours are also related to the community of the social networks. Delicious gathers a community interested in IT-related topics. These people are interested in the content of the bookmarks and tagging provides a good way for them to summarize the content of the bookmarks. Naturally, tagging becomes the key function of the system and plays a major role for sharing and retrieving. While in Flickr, part of the community are professional photographers who would like to share their pieces of arts for comments and feedbacks and other users who just use the Flickr as a space to manage their own personal photos and share with their closed friends. Searching photos in Flickr is based on the title and tag of the photos. The community of Flickr is interested in commenting and sharing. While in YouTube, its community can be viewed as a snapshot of the whole community of the Web. They are people from all of the world with all kinds of different interests and with different age ranges. Many of them do not tag their videos. They come to YouTube with different purposes and

expectations. The role of the tagging is shadowed by the rating and commenting. Searching videos in YouTube is mainly based on the titles of the videos.

6. Conclusion and Future Works

Tags are an emerging form of inductive (bottom-up) social metadata generated by the synergy of collectives of users. Social tagging is a new way of storing and retrieving online content and of sharing that content with others. Tagging not only creates new data about online resources, but it also generates new fields for exploration. The link between social tagging and bibliometrics offers potential for the application of well-tested bibliometric methodologies to extend and enhance current research efforts in the area of tagometrics.

The Upper Tag Ontology can be used to model unstructured tag data in the social web; and alignment of UTO with other systems of social semantics can help to ease the way for data integration, data management and query formulation. Using integrated data about tags and tagging behaviour that has been modelled with UTO, new applications can be developed to enhance social communication, to extend the capabilities of social computing, and to facilitate online resource selection and ranking. Furthermore, by applying citation and co-citation analysis for analysing tagging behaviour, it will be possible to build more effective recommender systems.

Social tags are user generated metadata. In popular social networks with millions of users -- networks such as Delicious, Flickr and YouTube -- tagging can lead to the emergence of a social vocabulary that reflects the communication features of the social network. Traditional, professionally created metadata (e.g., the taxonomies used in libraries and other organizations) is not scalable and thus impractical for large collections of resources such as those found on the Web; and author generated metadata that uses schemes such as the Dublin Core Metadata Element Set often leads to inadequate or

inaccurate descriptions. In comparison, user generated metadata in the form of social tags enables individual users not only to organize resources for their own use but also to share and communicate across a community of users (Mathes, 2004).

Although social tags do not constitute a controlled vocabulary, the feedback in social tagging (e.g., not using the vocabulary of a social network will make it difficult for other community members to find you or the resources you have tagged) creates a communicative loop between users and metadata, suggesting that users may actually be negotiating the evolving meaning of a term through their individual choices of tags assigned to online objects. Social tagging in large social networks may also lead to the creation of a local social culture. This is aptly demonstrated by the history of the tag *flicktion*, which was created by Andrew Lowosky in 2004. Although *flicktion* was originally used by Lowosky to tag his personal photographs, the referent of the term evolved as other members of the Flickr community members began to use *flicktion* to tag images that had a short fiction (a “fiction in Flickr” or "flickr fiction") attached to it.

Social tagging has the potential to improve traditional solutions for organizing and browsing information as well as monitoring trends; and user incentives can play an important role in the design of tagging systems. The different social tagging features of Delicious, Flickr and YouTube point to corresponding differences in the dynamics of interaction and participation. Delicious supports community tagging where anyone can tag any resource on the web, while Flickr supports personal tagging where only the owner of the image or his close friends can assign a tag. This leads to a very different design model for each of these social networks. It also leads to different social features in the tagging vocabularies that predominates in each of these social networks. System designers should be aware of these differences and take these factors into consideration when planning the architecture of a tagging system.

Social tagging generates massive collections of data that reflect the wisdom of crowds and, creatively managed, can lead to the development of a variety of interesting applications. The potential exists for social tagging to build bridges between disciplines and enhance social communication as well as social computing. Future research should focus on the investigation and implementation of these new applications.

7. Acknowledgements

The authors would like to thank the University of Innsbruck, where data collection and analysis were conducted, and Ioan Toma and Michael Fried, of the University of Innsbruck, who provided invaluable technical support. The authors are also very grateful to Blaise Cronin for his insightful comments on an early draft of the paper. The authors also give their special thanks to Mike Thelwall on his valuable comments and support.

8. References

Awad, M., Khan, L., Bastani, F. & Yen, I. L. (2004). An effective support vector machines (SVMs) performance using hierarchical clustering. In: Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence.

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

Blood, R. (2005). Rebecca's Pocket. Available at: http://www.rebeccablood.net/archive/2005/01.html#11technorati (accessed 22 Sep 2008).

Butterfield, S. (2004). Sylloge. Available at: http://www.sylloge.com/personal/2004/08/folksonomy-social-classification-great.html (accessed 29 Sep 2008)

Cortes C. & Vapnik, V. (1995). Support-Vector Networks, Machine Learning, 20(3), 273-297.

Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P. & Tomkins, A. (2006). Visualizing tags over time. In: Proc. of the 15th International WWW Conference (Edinburgh, Scotland).

Fountopoulos, G. I. (2007). RichTags: A Social Semantic Tagging System (Master Thesis, School of Electronics and Computer Science, University of Southampton).

Gruber, T. (2007). Ontology of Folksonomy: A Mash-up of Apples and Oranges. International Journal on Semantic Web & Information Systems, 3(2).

Hammond, T., Hannay, T., Lund, B. & Scott, J. (2005). Social Bookmarking Tools (I): A General Review. D-Lib Magazine, 11(4). Available at: http://www.dlib.org/dlib/april05/hammond/04hammond.html (accessed 29 Sep 2008).

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

Hotho, A., Jäschke, R., Schmitz, C. & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In: Y. Sure and J. Domingue (eds), The Semantic Web: Research and Applications (Springer: Heidelberg).

Kim, H.; Passant, A.; Breslin, J.; Scerri, S and Decker, S. (2008), Review and Alignment of Tag Ontologies for Semantically-Linked Data in Collaborative Tagging Spaces. In; Proceedings of the 2nd International Conference on Semantic Computing, San Francisco, USA.

Kipp, M. E. & Campbell, D. G. (2006). Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices. In: Proceedings Annual General Meeting of the American Society for Information Science and Technology (Austin, Texas, USA).

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment, Journal of the ACM, 46(5), 604–632.

Kohonen, T. (1989). Self-organization and associative memory. New York: Springer-Verlag.

Lawley, L. (2005). Social consequences of social tagging. Available at: http://many.corante.com/archives/2005/01/20/social_consequences_of_social_tagging.php (accessed 29 Sep 2008).

Losowsky, A. (2004). The doorbells of Florence project. The Prandial Post, September 21, 2004.

Mathes, A. (2004). Folksonomies – Cooperative Classification and Communication Through Shared Metadata. Computer Mediated Communication, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign. Available at: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html (accessed 29 Sep 2008)

Merholz, P. (2004). Metadata for the Masses. Available at:

http://www.adaptivepath.com/publications/essays/archives/000361.php (accessed 29 Sep 2008)

Mika, P. (2005). Ontologies Are Us: A Unified Model of Social Networks and Semantics. In: Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen (eds), Proceedings of International Semantic Web Conference 2005 (Galway, Ireland).

Miller, P. (2008). Sir Tim Berners-Lee talks with Talis about the Semantic Web. Avaiable at: http://talis-podcasts.s3.amazonaws.com/twt20080207_TimBL.html (accessed 31 Oct, 2008)

Pelleg, D. & Moore, A. W. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In: Seventeenth International Conference on Machine Learning.

Priss, U. (2007). Formal Concept Analysis in Information Science. In: B. Cronin (ed.), Annual Review of Information Science and Technology, 40.

Rainie, L. (2007). 28% of Online Americans Have Used the Internet to tag content. Pew Internet. Available at: http://www.pewinternet.org/pdfs/PIP_Tagging.pdf (accessed 31 Oct, 2008)

Schmitz, C., Hotho, A., Jaschke, R. & Stumme, G. (2006). Mining association rules in folksonomies. In: V. Batagelj, H. H. Bock, A. Ferligoj and A. Ziberna (eds), Data Science and Classification: Proc. of the 10th IFCS Conf., Studies in Classification, Data Analysis, and Knowledge Organization.

Schvaneveldt, R. W. (1990). Pathfinder Associative Networks: Studies in Knowledge Organization. Norwood: Ablex.

Sinha, R. (2005). A social analysis of tagging. Available at: http://blog.jackvinson.com/archives/2005/10/01/a_cognitive_analysis_of_tagging.html (accessed 29 Sep 2008).

Suchanek, F. M.; Vojnovic, M. and Gunawardena, D. (2008). Social tags: Meaning and suggestions. In Proceedings of the ACM 17th Conference on Information and Knowledge Management, Oct 26-30, 2008, California, USA.

Surowiecki, J. (2004). The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Large Print: Random House.

Wan, J., Hua, P. & Rousseau, R. (2007). The pure h-index: Calculating an author's h-index by taking co-authors into account. COLLNET Journal of Scientometrics and Information Management.

White, H. D., Buzydlowski, J. & Lin, X. (2000). Co-Cited Author Maps as Interfaces to Digital Libraries: Designing Pathfinder Networks in the Humanities. In: Proceedings of the IEEE International Conference on Information Visualisation (London, UK)

Wille, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In: I. Rival (Ed.), Ordered Sets (Reidel, Dordrecht-Boston).

Xu, Z., Fu, Y., Mao, J. & Su, D. (2006). Towards the semantic Web: Collaborative tag suggestions. In: Proceedings of the Collaborative Web Tagging Workshop at the WWW 2006 (Edinburgh, Scotland).

Yu, H., Yang, J. & Han, J. (2003). Classifying Large Data Sets Using SVM with Hierarchical Clusters. In: Proc. 2003 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'03) (Washington, USA).