Social Indexing: A Literature Review

Lila Sadkin

LIS 5736 Indexing and Abstracting

Professor Burnett

5 December 2008

Social indexing is exemplified on sites such as Flickr, Delicious (formerly del.ici.ous), and other photo sharing or social bookmarking sites. Social indexing is more commonly known as social tagging, or just tagging, and it is a new form of indexing that has been studied in quite some depth. The influence of the internet on indexing, just as on many other fields, is studied and written about from many different perspectives: sometimes doubt, often hope. There is always the attempt to take an emerging trend and improve upon it--this can mean making it more like the old and familiar, or it can mean taking even farther from its roots. This paper seeks to investigate some of the research and study done in the field of social indexing.

There is a general consensus about what social indexing is. Adam Mathes's 2004 paper "Folksonomies - Cooperative Classification and Communication Through Shared Metadata" doesn't actually use the term "social indexing," but it seems to be considered an important early work in the field and it provides a concise definition: "[K]eywords […] allow users to describe and organize content with any vocabulary they choose." Jakob Voss defines tagging as "end users do[ing] subject indexing instead of experts only, and the assigned tags are being shown immediately on the Web" (2007). Social indexing and social tagging are synonymous.

A folksonomy is further defined as an "organic system of organization" (Mathes, 2004) or a "type of distributed classification system" (Guy & Tonkin, 2006), or "simply the set of terms that a group of users tagged content with" (Mathes, 2004). Folksonomies have unique features as compared to other forms of index. The terms in a folksonomy have no hierarchy (Mathes, 2004) and are not a predetermined set but develop organically over time (Mathes, 2004; Rafferty & Hidderley, 2007). Voss describes the folksonomy as created by feedback. Feedback is part of any indexing system, in both creating an index and searching for information. With a controlled vocabulary, an indexer searches for the best-matching term. In a free keyword tagging system, feedback “influences tagging behaviour towards consensus” (2007).

The purpose of indexing is to create metadata that facilitates organization and access of information (Mathes, 2004). This metadata can be created by information professionals, authors, and users (Mathes, 2004). Rafferty and Hidderley expand on these categories, and distinguish between monologic and dialogic indexing practice (2007). Expert-led indexing, reliant on existing controlled vocabulary systems, and author-based indexing, where terms are chosen by the author of a text, are both monologic systems. Monologic systems face the problem of providing only one viewpoint. User-based indexing is dialogic, allowing users to add their own tags to documents, whether text, music, or images (Rafferty & Hidderley, 2007).

The question of the purpose that social indexing serves is a more complicated one. Mark Stefik discusses “sensemaking” in terms of the individual’s “information diet” as it relates to the collective consumption of information, which has a long-tail distribution and generally doesn’t match up with an individual’s personal information diet. (2008). Tagging helps people who are seeking to manage their information diet by allowing them to develop their own personal organization system (Mathes, 2004; Guy & Tonkin, 2006). This system can act as an information filtering system, sending the user information that has been filtered by a specific set of predefined tags (Hassan-Montero & Herrero-Solana, 2006).

The “social” in “social tagging” indicates that tags serve a communal purpose as well as an individual one (Hassan-Montero & Herrero-Solana, 2006). Tags are meant to be seen by others: there are various ways to view groups of tags on websites that use social tagging. One of the most common is the tag cloud, which displays tags in alphabetical order with the font size of each tag relative to the tag’s frequency of use (Hassan-Montero & Herrero-Solana, 2006). Tag listings and search functionality provide users with ways to find information that others have tagged. Heymann & Garcia-Molina name three methods by which a user can view the tags used in a system: a list of all objects tagged a certain way, a list of the most popular tags, or a list of tags that are related to the current tag (2006).

The strengths of social indexing beyond the previously mentioned use for personal organization have been described by authors in various ways. One unique strength of tagging systems is "serendipity," they way that things can be found unexpectedly while browsing the system (Mathes, 2004). A folksonomy reflects the vocabulary of its users (Mathes, 2004), while an index created by a single person reflects only the interpretation of that one individual (Rafferty & Hidderley, 2007). Merholz (2004) is cited by Mathes, as well as Rafferty and Hidderley, in the comparison of folksonomies to "desire lines" that could be used as the basis for a controlled vocabulary that reflects the language of its users.

One use of social indexing might be bridging the "semantic gap" in image indexing (Rafferty & Hidderley, 2007). There has been a lot of research and development of content-based retrieval systems for images, which attempt to perceive and identify features such as shape and color and index images according to these features. These methods are not able to describe interpretive content. The authors refer to many researchers' works from the late 80s through the 90s in the area of concept- or text-based image retrieval. The term "semantic gap" is used to mean the gap between the information that a content-based indexing method would be retrieved and the user's query. Because of the subjectivity of image interpretation, it is difficult, if not impossible, for a single individual to choose terms by which to index an image that all users will find satisfactory (Rafferty & Hidderley, 2007). However, the authors found that Flickr's author-based indexing did not always lead to effective searches. They describe an alternative approach called "Democratic Indexing." Democratic Indexing focuses on user interpretation of images.

The collection of terms used to index images is chosen by readers, thus creating a “spectrum of connotation” (Rafferty & Hidderley, 2007).

There have been several weaknesses of social indexing noted throughout the literature. Tags are often sloppy: ambiguous or inexact. Synonyms are treated as separate tags, and homonyms are treated as the same tag. Often, users create compound tags as a way to work around the problem of not being able to use multiple-word tags. Tags can often be very broad or very specific. They are also often used as personal codes whose meanings are not clear to anyone. These qualities impede the ability to effectively search for relevant information (Guy & Tonkin, 2006; Rafferty & Hidderley, 2007).

The poor precision and recall of Internet searches is brought up repeatedly. Rafferty & Hidderley say that the author-based indexing approach used by Flickr, in which the author of a photograph is the only person who can assign tags to it, leads to a system with poor precision and recall, but the authors suggest that precision and recall my not be the best measure for such a system (2007). Heymann & Garcia-Molina claim that a tagging system is limited by the few ways of viewing the tags in the system (2006).

Bill Johncocks claims that the web will cause traditional indexing skills to become marginalized, as Internet users begin to use different techniques for information retrieval that emerge from the inefficiency of the indexing techniques that are used on the Internet. He argues that Internet users will become conditioned to poor performance from indexes and will begin to favor iterative processes such as scanning titles and tables of contents. He claims that adding keywords to documents without restraint “is faintly reminiscent of giving typewriters to monkeys” (2008).

There have been various suggestions for and implementations of improvements for tagging systems that would improve usability for people searching for information. Guy and Tonkin suggest that educating users about “best practices” for creating and using tags would help to improve the quality and usability of the systems, including plurals instead of singulars, lower case, grouping words with underscores, following established conventions, and adding synonyms (2006). These were previously written about by Ulises Mejias (2005). Mejias also recommends using tags that add generic value to an item, instead of or in addition to idiosyncratic tags that add only personal value (2005).

A different way of improving a tagging system is by improving the methods used to browse or search it. Guy and Tonkin make suggestions in this area as well, including error-checking, which could prevent misspellings, and tag suggestions when users add a resource. Systems could also make synonym suggestions for searches, providing users with a tag that is more widely used than the term they searched for. They also suggest discussion tools that allow users to share their reasons for their tag choices (2006).

Suggestions for improvements have also been made to the browsing methods of social tagging systems. Hassan-Montero and Herrero-Solana, as well as Heymann and Garcia-Molina, developed algorithms for creating a more usable way to browse through a collection of tags. Hassan-Montero and Herrero-Solana present a reworked tag-cloud. In a standard tag cloud, the tags are displayed alphabetically and the font-size of each tag increases in proportion to its frequency of use. Hassan-Montero and Herrero-Solana present some restrictions of this method of visualization: only a few topics tend to dominate the cloud with all their related tags, and the alphabetical arrangement doesn’t provide any information about the semantic relationships between tags. They claim that a similarity-based layout would be an improvement. Grouping tags by topic or semantic category allows users to quickly find a general topic that they’re looking for and then see the most frequently used tags within that topic as indicated by the font-size, as well as allow a wider range of topics to end up represented in the cloud (2006).

Heymann and Garcia-Molina developed a different method for organizing tags. They identified as a problem the difficulty of finding broader or narrower tags in a tagging system. Their algorithm creates a hierarchy from the flat folksonomy. This structure is more similar to a traditional thesaurus used for indexing, and it creates the ability to see broader or narrower categories in a way that users are often familiar with (2006).

While Hassan-Montero and Herrero-Solana, Heymann and Garcia-Molina, Guy and Tonkin, and Mejias have all suggested ways of improving existing social indexing systems, Mark Stefik's proposal is much broader in scope. In his paper "Social Indexing" (2008) he describes a system that would alter the way that people find information on the web. His proposed system makes use of both automatic and social indexing, combining the two to create overlapping and connected communities centered around topics of interest.

The first issue Stefik identifies is locating information. He shares with Johncocks the opinion that search engines are not ideal for discovering information that has a place in our personal "information diets," and identifies the need to create better approaches for tracking and discovering information, for "mining our information frontiers," and for becoming oriented in an unfamiliar subject area. He recognizes that full-text automatic indexing is not ideal for discovering information because there is no domain expertise and the important and trivial are not distinguished.

However, unlike Johncocks, who seems to dismiss full-text indexing altogether, saying "When mass storage costs meant we only had 'enriched' titles, full-text natural language retrieval looked like the Holy Grail. Now we have it, its shortcomings are all too apparent" (2008), Stefik presents a reworked type of automated indexing that may lessen its shortcomings significantly. This system is called “index extrapolation,” which uses an existing index as training material for an automated system to use to create a new index that is updated with new information. The program would learn from the entries in the existing index and index new material accordingly.

The second issue that Stefik identifies is determining degree of interest for each item in the index. He examines Digg, a popular social media website where people vote for articles and the most popular articles get featured more prominently on the site. He identifies problems with Digg’s system: easy manipulation by cliques and and a relatively narrow topical focus.

He presents an approach that creates separate communities for people with similar interests, providing an individual index for each community, and users in that community could rank items as they see fit. The same topics may appear in different communities, although perhaps with different classifications, depending on the semantic relationships with the other topics in the community. Expanding outward, Stefik goes on to describe how these separate communities could be linked by degrees of similarity, providing community members with access to their “information frontiers.” This emphasis on semantic relationships reflects the same kind of focus shown by both Hassan-Montero and Herrero-Solana (2006) and Heymann & Garcia-Molina (2006).

Stefik explains how user-based indexing and machine-based indexing would work together: experts would create the original index and maintain the community index; users would identify new information and vote on existing information; computers would match patterns from the index against new pages, keeping the index up-to-date, combine votes to update the training sets for pattern matching, and generate search spaces to faithfully model the subtopics in the community.

This combined approach is reminiscent of Rafferty and Hidderley's approach for bridging the "semantic gap" in image indexing: combining automated techniques with democratic input from users (2007).

The literature on social indexing is from a short period of time, seeming to only really come into existence after the year 2000. Nevertheless, a lot of research has been done in the field, and it encompasses a broad range of topics. Much attention is paid to ways that the information retrieval functions social tagging systems can be improved (Guy & Tonkin, 2006; Hassan-Montero & Herrero-Solana, 2006; Heymann & Garcia-Molina, 2006). There is some criticism about the current state of Internet indexing in general and social tagging in particular, mostly related to the inefficiency of searching for information (Johncocks, 2008). There seems to be promise in social indexing for helping to overcome subjectivity problems by creating a scale of consensus (Rafferty & Hidderley, 2007).

It seems that research in social indexing may be heading in the direction of investigating the integration of traditional monologic and new dialogic indexing methods. The amount of information is increasing drastically. Computers are good at creating concordances, and can deal with large amounts of information quickly, but a human is able to create "socially valuable information" by making a careful association between a resource and a tag (Mejias, 2005). With the quantity of information available, distributing the "sensemaking" (Stefik, 2008) process to many individuals and "tying it to the individual interest of the user" (Mejias, 2005). Traditional indexing methods have changed with the growth of the Internet, and social indexing is a new form of indexing that may be uniquely suited for dealing with the vast amount of information available.

References

Guy, M., & Tonkin, E. (2006). Folksonomies: Tidying up Tags?. *D-Lib Magazine*. Retrieved from http://www.dlib.org/dlib/january06/guy/01guy.html

Hassan-Montero, Y., & Herrero-Solana, V. (2006). Improving Tag-Clouds as Visual Information Retrieval Interfaces. Retrieved from http://www.nosolousabilidad.com/hassan/improving_tagclouds.pdf

Heymann, P., & Garcia-Molina, H. (2006). Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Retrieved from http://labs.rightnow.com/colloquium/papers/tag_hier_mining.pdf

Johncocks, B. (2008). *Web 2.0 and users' expectations of indexes. The Indexer [0019-4131], 26*(1), 18-24.

Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Retrieved from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Mejias, S. (2005). Tag Literacy. Retrieved from http://blog.ulisesmejias.com/2005/04/26/tag-literacy/

Merholz, P. (2004) Metadata for the Masses. Retrieved from http://www.adaptivepath.com/ideas/essays/archives/000361.php

Rafferty, P., & Hidderley, R. (2007). Flickr and Democratic Indexing: dialogic approaches to indexing. *Aslib Proceedings, 59*(4/5), 397-410. doi:10.1108/00012530710817591

Stefik, M. J. (2008). Social indexing. Sensemaking Workshop @ CHI 2008; 2008 April 6; Florence, Italy. Retrieved from http://www.parc.com/research/publications/details.php?id=6398

Voss, J. (2007). Tagging, Folksonomy & Co–Renaissance of Manual Indexing?. Retrieved from http://arxiv.org/abs/cs.IR/0701072