Pregledni članak

Collaborative Tagging: Providing User Created Organizational Structure for Web 2.0

Mihaela Banek Zorica
Department of Information Sciences
Faculty of Humanities and Social Sciences
Ivana Lučića 3, 10000 Zagreb Croatia
mbanek@ffzg.hr

Sonja Špiranec
Department of Information Sciences
Faculty of Humanities and Social Sciences
Ivana Lučića 3, 10000 Zagreb Croatia
sspiran@ffzg.hr

Krešimir Zauder
Department of Information Sciences
Faculty of Humanities and Social Sciences
Ivana Lučića 3, 10000 Zagreb Croatia
kzauder@ffzg.hr

Summary

The Web 2.0, having the user both creating and organizing content, has changed much of how one approaches to and uses the Web. While the concept of user-submitted content is by no means new, user created organizational structure is. The article gives an overview of the organizational means and processes that enable it. To provide the general framework the article gives a short overview of Web 2.0. It then centres on the collaborative tagging process as a central organizational process and means for the Web 2.0 and provides definitions for the Web 2.0 terminology used. After describing the general process, its strengths and weaknesses and pointing out that, while useful, it cannot replace professional indexing tools and library and information science professionals the article goes on to describe collaborative tagging and its specific features in general. Some of the more common services to use collaborative tagging are then described.

Keywords: Collaborative tagging, Web 2.0, resource discovery, IR, social bookmarking

Introduction

Organization of the World Wide Web resources in the traditional sense has been a problem since its explosion. The early attempts to categorize the Web in subject directories have failed in comprehensiveness and have a significant lag in adding new resources due to the speed of Web growth and the sheer amount of resources present. Metadata, which could have enabled automatic gathering and organization, has failed for the Web as a whole due to lack of use and widely accepted standards and even due to misuse. This has left us with the search engine as the main access point to Web resource since any subject or author based access requires organizational structure that is simply not present. There was no task force large enough to provide the organization for the World Wide Web and the automatic means for providing subject based access still have to be developed. However, one of the new approaches to organization of online resources, central to the so-called Web 2.0, has managed to enlist the help of the common user in organizing vast quantities of these resources.

We can characterize the phenomenon known as Web 2.0 by its technological and design aspects but they are just a support for its conceptual nature which is what distinguishes it the most from the “old Web”. Web 2.0 is dependent on and built from user supplied content and organization in a collaborative environment. We could call this aspect of Web 2.0 “social Web” or “collaborative Web” but “Web 2.0” is far more common. It is this collaborative principle which gives it power and it has come a long way from merely linking to other pages.

Web 2.0 services constitute the backbone of collaboration. Technologically, Web 2.0 treats the Web as a platform with its own applications accessible through the browser. These applications make the services which provide the needed collaborative environment and enable one of the key Web 2.0 principles: the service automatically becomes better the more users it has [7]. In using many of these services users collaboratively create content (as is the case with Wikis) or, more frequently, create collections of either their own resources (e.g. Flickr) or publicly available ones (e.g. del.icio.us, LibraryThing).

User collections are thus one of the central concepts for Web 2.0 IR and resource discovery. One of the essential features of a collection, and the one which distinguishes it from a heap, is its organization, a topic which has plagued the web from its beginning. The collection organization, if done right, enables IR and resource discovery by providing different access points to organized resources without just resorting to full text search combined with statistical methods and ranking algorithms. While search engines are great for known item retrieval they fail somewhat on subject related searches and casual resource discovery (comparable to shelf browsing in libraries).

The central principle and means of organization, and thus the creation of access points to resources, for Web 2.0 is tagging which is most frequently implemented in its collaborative form.

Tags, tagging and folksonomies

We can define tagging as the organizational method and the very process of assigning ad hoc user-created natural language keywords to information resources thus organizing them into user specific collections. It should be noted that indexing term is here used in its broadest sense since indexing terms are traditionally used to denote the subject of the resource while the scope of tags is broader. The ad hoc created index terms used for tagging are most frequently called tags although one can also find other terms such as topics or labels. From the definition we can clearly see the three needed components in a tagging system: users, resources and tags. Indeed, we can call a basic information object of these systems a *post* and say it consists of a (user, resource, {tag}) triple [1]. In other words, to have a post we need one user, one resource and zero or more tags. Depending on a service, a resource may be an URL (or bookmark), an image, a Web clip, a bibliographic description of an academic paper, metadata about a real world object (a book or just about anything else), etc. As we can see the resource can be either just the metadata about an information object or it can include the information object a user wants to post. Tags are separated from the resource as a distinct feature supporting several functions within a service and thus a special case of metadata.

Although seemingly simple, there are various ways one can think of a tag. Tag has the features of metadata (more specifically, an indexing term or a keyword but is broader in scope), a category name and a navigational tool. A tag as an ad hoc created category name is an especially important notion since modern cognitive science has clearly shown categorization is central to our thinking [5]. The ability of the user to use the categories and category types that first come to mind and are easiest to use while not having to cross-reference a controlled vocabulary of some kind contributes to the cognitive ease of use of the tagging process. Also, this kind of approach to knowledge organization does not have to be learned beforehand. The problems that arise from tagging, the skill in tagging and some advanced approaches (e.g. the faceted approach) to tagging all may be learned afterwards. This ease of use might be one of the prime reasons for the current popularity of tagging. It may also be one of the pitfalls of tagging: the perceived ease of use might backfire later when a user's collection organized in this manner grows too big to be easily navigated through and used in general. This might have been prevented by planning the tagging process beforehand but the user doing the tagging did not have the needed knowledge and experience to plan it and quite probably did not even perceive the need to do so. It remains to be seen how many users will re-organize their collection when this happens since there is quite serious amount of work involved. This might also be just a short term solution which will lead to increased awareness of the problems present and thus to ease of use of more complicated tools and processes by the average user. This in turn might help in realization of the Semantic Web understood as an extension of the current Web in which data is semantically described and automatically connected across resources thus facilitating machine to machine communication, greater reuse of data and a general shift from document to data.

We can analyse a tag in terms of its specificity, objectivity/subjectivity and in terms of the properties it implicitly presupposes. If we consider the hierarchy of categories we employ in our thinking a tag may be more or less specific (or more or less general). For example, where one user, who is especially interested in knowledge organization, might use the tag "collaborative_tagging" others might use "tagging" or "metadata". Objectivity/subjectivity refers to what a tag actually describes: the resource or the user's relation to the resource. For example a tag might be "cars" which quite obviously relates to the resource and is useful for IR and resource discovery or "to_read" which is useful primarily to the user who tagged it. Objective tags thus have a much greater value for information retrieval and discovery of other users while subjective ones primarily have value for individuals who used them and sometimes for other users of their specific collections but not for users of the system in general.

In terms of the property a tag implicitly presupposes, a tag might deal with the subject, content, author, page type, task, frequency of use etc. In the traditional approach to metadata the properties are explicitly defined and they take certain values (e.g. <meta name="DC.creator" content="John Smith" />). Tags can be likened to values of these properties but the properties themselves are left implicit.

Tagging problems

Although tagging is the most prominent knowledge organization method for Web 2.0, and, due to the user acceptance, the only one that works at this level, it is not without its problems. We can summarize them in these five crucial ones:

- lack of vocabulary control
- lack of defined relationships between tags
- lack of explicitly defined properties for tags
- lack of user education and educational material
- users tag primarily for own use

Tagging is not backed by a controlled vocabulary so many problems these tools solved for databases and traditional institutions are present. There is no synonym control or homonym discrimination which presents a barrier to IR and resource discovery. For example which jaguar does a user want? An animal or the car? If a user searches for "folksonomy" it is quite probable he also wants "tagging" and "collaborative_tagging" to be included in the search. In addition, there are no conventions for the form of word to use (e.g. singular or plural?) and for the creation of compound tags. Tags are most frequently separated by spaces so "Web 2.0" would produce two tags, "Web" and "2.0". Since there are no conventions for compounding tags, this gives rise to many possibilities of the semantically same tag: “Web_2.0”, “Web-2.0”, “Web2.0”. Some services allow usage of spaces in tags by separating the tags with another character (e.g. “;”) which alleviates the problem somewhat but does not solve it completely.

Lack of defined relationships presents a problem related to specificity of tags. Searching for a tag will not find tags its hyponyms i.e. the semantically subordinate tags. For example looking for “furniture” will not find posts tagged with ”chairs”, “tables” and similar. This is another problem that was traditionally solved with various knowledge organization tools. Perhaps the main problem is that the use of these tools is a skilled process normally undertaken by highly trained information professionals [6] so one currently cannot implement them without losing users without which tagging has no significance since its greatest quality is that “common folk” are doing it in massive numbers. This fact also clearly shows that tagging is just another organizational layer for a specific purpose which cannot replace existing knowledge organization tools and methods but can add to them. One might, for example, design a Website of a large library with traditionally organized resources but allow users to build personal collections organized by tagging. If enough users did this it would provide a new layer of access points to library collection.

Lack of clearly defined properties of which tags are instances is a more serious problem than is immediately apparent. Traditional metadata implementations had explicitly defined properties which were then filled with instances that were quite often taken from controlled vocabularies. In tagging, however, these properties are present only implicitly. This means that while human users might sometimes recognise them (i.e. when a bookmark is tagged with “cars” one will naturally suppose it is the subject of the resource) they cannot be used automatically by the machine. Even human users will sometimes be confused: if a bookmark is tagged with “images” and “John_Smith” one cannot know if the resource contains images of John Smith, images taken by John Smith or even text about images (i.e. resolution and colours) written by John Smith. So the property of which “John_Smith” is an instance might be “author” or “subject” and “images” might be the instance of “content” or “subject”.

As in traditional environment, tagging is also a matter of knowledge and skill. The “common user” still does not possess that qualities since these subjects have yet to enter the curricula and most of the services using the tagging approach to organization still do not offer tagging tutorials, FAQs, and other documentation which would describe collection building problems and give extensive tagging advice.

It is important to keep in mind that users of services employing tagging for users’ collections organization primarily tag for themselves. This means that they will frequently use tags that make sense only to them and generally be sloppy in their tagging. They might also employ different means to overcome tagging problems which might make sense to them and even be quality solutions for their collection organization but which do not contribute to the system as a whole. A user might, for example use prefixes to define properties (such as "subject.images") or use different tags to connote different properties (such as "audio" for content and "music" for subject or vice versa).

Tagging benefits

The fact that tagging is an organizational method and process that common users are actually employing to organize massive quantities of very heterogeneous resources themselves is its most important benefit. This is the first time this has happened on such a massive scale and the first time we have the infrastructure to make it happen. Given the number and the growth of information resources and given that automatic methods still have many difficult problems to solve, this constitutes a very important aspect of Web organization. It is this fact that makes tagging worthwhile in spite of its many problems. Besides that, it is quite interesting that many things that constitute the problems of tagging are also its benefits. The lack of vocabulary control also means, as already mentioned, that users can start using it without advance preparation and that it is easy to use every time since there is no need for cross-referencing terms. Also since terms are created ad hoc there is no delay until the terms enter the vocabulary and it is quite difficult to imagine a creation of a controlled vocabulary comprehensive enough for general purpose tagging.

The fact that the users are tagging for themselves and the free nature of the tagging makes sure that the organization of every collection will make sense to its user. In other words, every user is able to use objective tags which make sense to him or her on the level of specificity which suits him or her best and subjective tags to facilitate an organizational layer tailored according to tasks, opinions, frequency of use and other personal parameters.

It should also be mentioned that this kind of approach, where users are included in the resource organization, is an important step to raising the users' organizational skills and their awareness of organizational problems. This could in turn lead to increased educational efforts and materials (both professional and available through the Web services in forms of tutorials, FAQs and other documentation) which is necessary if the users will get to the next step, i.e. the Semantic Web. It is quite probable, given the scope of the problem, that only a smaller part of the Web will ever become "semantic" if the users are not involved.

The tags taggers use might help in gathering data for controlled vocabulary creation for which the terms users use are an important input. The idea of combining the results of tagging with controlled knowledge organization systems is well perceived as a possible building block for the Semantic Web. According to different research activities [7] Web 2.0 ideas and applications can contribute to the creation of ontologies for a multitude of domains, which is essential for the development of the Semantic Web. The model of ontologies defines precise but non exhaustive semantic relationships between terms while the model of tagging associates terms into exhaustive contexts with no specific relationships.

The optimistic combination of these models would result in an exhaustive and precise model of knowledge organization.

Besides ontology creation, users might also play a significant role in actual deployment of this model. Use of ontologies and semantic description in general for Web organization has similar problems as many other approaches: the sheer amount of heterogonous data involved and the need for human intellectual input in organizing a large part of it. However, users might be brought to help overcome this problem in much the same way they are tagging, only in connection with ontology and on data, rather than document, level.

Collaborative tagging

The most frequent implementation of tagging is in its collaborative form. Collaborative tagging is tagging of resources in such a way that it is possible for different users of the same service to tag the same resource. This can manifest itself in two ways. The most frequent way (as in del.icio.us, citeUlike, etc...) is that the resources are publicly available to all users and all users may add it to their collections and tag it separately. The same resource may exist in collections of different users where it may (and frequently is) tagged with a different set of tags. The types of resources in this kind of services include metadata about: Web pages (i.e. bookmarks), academic papers, music artists, albums and tracks, books and anything else that is not unique to the individual user. They may also include the whole information object as the resource part of the post but the usual resource of these systems is constructed of just metadata. In this case just a resource is not enough to identify a post: it is identified by both resource and a user. A resource as input can provide a list of users who have that specific resource in their collections which can be useful for identification of users with similar interests and thus support casual resource discovery.

The other and much less frequent form of collaborative tagging happens when a resource is specific for a user, is located just in his or her collection and some or all of the other users are allowed to change the tags of the resource. This is the case with Flickr, for example. If other users were not allowed to add or change tags (e.g. YouTube) it wouldn't be *collaborative* tagging. For this type of collaborative tagging system it is characteristic that the user is the owner or creator of the information object and that the whole information object with the accompanying metadata constitutes a resource. The user may or may not give the permission to other users or groups of users to change the tags of his posts. Sometimes this permission is automatically given. For example, when the post is included in a meta-collection (e.g. a collection of posts based on a subject or a group) other users who have their posts in the meta-collection can change the tags of all the posts in that collection, not just their own.

Collaborative tagging is a form of tagging most prominent because by its very nature it alleviates (but doesn't solve!) some of the problems present in tagging and displays some social organizational aspects and thus supports knowledge discovery. The central phenomenon is that a same resource in most services is usually tagged by several users. Indeed, the number of users who have a same resource tagged in various ways in their collections is frequently very big and the totality of the tags used to describe the resource clearly shows which tags are most prominent among users for the description of that specific resource. This totality of tags is highly significant for information retrieval and is important for resource evaluation. It helps with the problems of objectivity/ subjectivity and specificity as well as sloppy tagging by users and some of the language related problems. If one user tagged a resource just with “to_do” others quite probably collaboratively “corrected” this by using more objective tags (e.g. subject or content related). The same thing happens with more or less specific tags (e.g “metadata” and “tagging”), synonyms (e.g. ”movie” and “film”), compound tags (e.g. “Web_2.0” and “Web2.0”), sloppy tagging (e.g. “metdata” and “metadata”) all of which increases the recall of the system and is not possible in a system without the collaborative approach. Although the approach cannot distinguish homonyms, when searching with a tag it can provide related tags (i.e. the tags users frequently used in conjunction with the tag used as query) to further specify the search.

This however provides much meta-noise since some semantically same words constitute different tags. Examples are singular and plural forms, tags compounded in different ways, shorter forms of the same word, abbreviations, etc. So a del.icio.us tag cloud for flickr will include “photo”, “photos” and “photography” all as highly popular and thus highly ranked tags for this resource. Another problem is for a service to attain the critical mass of users that tag frequently and thus support other users’ tagging and resource discovery. In the future, there will probably be few main collaborative tagging services for each type of resource and the highest challenge for new services will be attaining enough “taggers” for the system to reach its full potential. It remains to be seen just how serious are the problems derived from too many users. As things stand now a large group of diverse users seems more of a benefit than a hindrance but as new approaches are implemented and as community based Web develops both technically and socially, benefits may also be gained from tailoring a service for a highly specific community.

An interesting benefit of these services is an instant feedback. While tagging, a user can usually see the tags other users frequently used when tagging. Immediately after tagging the user can see how many other users have the resource in their collections, and look at the tag cloud of the resource and start finding other users with similar interests.

Resource popularity plays a highly significant role in this kind of services since it is used for ranking and recommendation. Also, a popular resource will be easier to find since its folksonomy will be rich and tagging a popular resource will be somewhat easier since it will be easier for the system to recommend tags. The resource popularity is not, however, gained from the collaborative tagging method and process but from the inclusion in user collections and would thus be present in the same form if some other organizational approach was used.

Folksonomies

Another term which is frequently used in conjunction with collaborative tagging is folksonomy. As is the case with other terms in this area, folksonomy is frequently ill defined. It is often used either to denote a whole variety of phenomena or defined to broadly to be of any use.

We define a folksonomy as the totality of tags emerging from the process of (collaborative) tagging. According to this we can liken a folksonomy to an "open vocabulary" which currently does not include any relationships but does include other data such as the number of times a tag has been used in the whole system, for a specific resource and by a specific user. Since some services have started implementing relationships between tags (e.g. Bibsonomy), soon we might see folksonomies with some relationships included.

From this we may distinguish three types of folksonomy: a service folksonomy, a user folksonomy and a resource folksonomy. Each of these types of folksonomies may be visualised by a tag cloud. A tag cloud is a primarily navigational device that visually shows the most popular tags where the frequency of tag usage is denoted by font size: the larger a tag's font the more often it was used. It may or may not show all the tags present in a folksonomy depending on the number of tags. A tag cloud representing the folksonomy of a service shows the popularity of tags which in turn shows "hot" topics among the users of a service. A tag cloud representing all the tags of a single user shows his interests within the resources current service allows to collect, while a tag cloud representing a folksonomy of a single resource shows how the resource was tagged by all the users who have it in their collection in one service.

Another way to distinguish folksonomies is to "broad" and "narrow" types [8]. Broad folksonomies are the ones emerging from services in which all users are able to (which doesn't mean they do) tag all resources that can be included in a service. Narrow folksonomies emerge from the services in which a resource unique to one user's collection is tagged by just that user and possibly other users which were given permission to do so by the "owner" of the resource.

Conclusion and further directions

There is no doubt that the Web 2.0 is a propulsive and already widespread phenomenon. A plethora of Web 2.0 services are freely available to the common user that very frequently use collaborative tagging to facilitate organization, ranging from social bookmarking (del.icio.us), academic paper sharing (CiteULike, Connotea), media sharing (flickr, last.fm), collections of real world items (LibraryThing) etc., or a combination of those. These and other services offer new models, methods, and technologies that can be adapted to improve different domains with a strong focus on formal knowledge management and IR, like the corporate sector, information sector or education.

With the application of the above described Web 2.0 services, traditional educational institutions or information agencies like libraries could transform themselves from a place of passive information consumption to a dynamic, participative and creative knowledge production space. Besides using traditional organizational and navigational tools, users could produce and consume knowledge, create new information architectures and change the information landscape by using collaborative, social and personalized means.

An important facet of the transformational function of the Web 2.0 ideas is their potential contribution to solutions to some of the recognized existing problems within IR, although it is believed that the real strength of these services lies in their combination with more controlled knowledge organization systems.

Finally, Web 2.0 concepts, particularly collaborative tagging, could significantly enhance and facilitate the further development of the Semantic web idea, by using emerging folksonomies for the development of ontologies and thereby acknowledging the idea that even this machine oriented concept can't be realized without a strong social dimension.

References

[1] Cattuto, C., Loreto, V. and Pietronero, L. Collaborative Tagging and Semiotic Dynamics. PNAS, 104 (5), January 2007. http://arxiv.org/abs/cs/0605015v1 [15/05/2007]

[2] Golder, S. and Huberman, B.A. (2006). Usage Patterns of Collaborative Tagging Systems. Journal of Information Science, 32(2): 198-208.

[3] Guy, M. and Tonkin, E. (2006). Folksonomies: tidying up tags? D-Lib Magazine, 12 (1), January 2006. http://www.dlib.org/dlib/january06/guy/01guy/html [04/05/2007]

[4] Hammond, T. et al. (2005). Social Bookmarking Tools (I): A General Review. D-Lib Magazine, 11 (4), April 2005. http://www.dlib.org/dlib/april05/hammond/04hammond.html [03/26/2007]

[5] Lakoff, G. Women, fire and dangerous things: What categories reveal about the mind. Chicago and London: The University of Chicago Press, 1987.

[6] Macgregor G. and McCulloch, E. Collaborative tagging as a knowledge organisation and resource discovery tool. Library Review, 55 (5), 2006. http://www.emeraldinsight.com/10.1108/00242530610667558 [05/07/2007]

[7] Mesnage C. and Jazayeri, M. Towards global collaborative tagging. 2006. http://cedricmesnage.org/resources/pdf/mesnage06aSubmitted.pdf.

[8] O'Reilly, T. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. O'Reilly Media, Inc; 2005. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html [02/27/2007]

[9] Vander Wal, T. Explaning and Showing Broad and Narrow Folksonomies, http://www.personalinfocloud.com/2005/02/explaining and .html [04/05/2007]