Collaborative Tagging Supported Knowledge Discovery

Krešimir Zauder, Jadranka Lasić Lazić, Mihaela Banek Zorica
Faculty of humanities and social sciences, Department of information sciences
Ivana Lučića 3, 10000 Zagreb Croatia
kzauder@ffzg.hr, jlazic@ffzg.hr, mbanek@ffzg.hr

Abstract. *Web 2.0 has brought about a new sort of user centred services which rely a great deal on flexible organizational capabilities designed for user-supplied organization. Collaborative tagging is especially interesting in this context and this article explores what this kind of organization in connection with some Web 2.0 principles means for knowledge discovery in various ways. To fully explore this, the article defines collaborative tagging and gives an overview of collaborative tagging in general, of services using it and of tags themselves. It concludes with mechanisms this kind of approach to knowledge organization provides for knowledge discovery.*

Keywords. Collaborative tagging, knowledge discovery, Web 2.0, social bookmarking.

1. Introduction

A few years ago a Web user was just a user, creating and collecting content accessible to others only if he or she was willing to create, host and maintain Web pages. Today, the new user and Web service centred Web known by the much discussed and criticized buzzword Web 2.0 [11] gives power to the users in the sense that it provides services which allow users to create, collect, organize, connect and share content without much fuss and required prior knowledge. These systems in turn harness the power of these users and by pooling what the users did they are able to provide services not possible without this approach. The "user 2.0", to take the buzzword a bit further, thus can have: their bookmarks online and share them with other users, their blog, the metadata about the music they listen to, their profiles in general in order to connect with people with similar interests, the shared collection of personal photos, news items or clippings from various websites, and so on... The important fact is that any user, no matter how small his or her contributions or usage of a particular service is, provides for other users.

In order for all this to work, a new organizational paradigm had to happen. It should allow the user to personally create or gather and then organize their own collections while what they do is usable by other users. This came about in form of labelling via natural language keywords which, for Web 2.0 purposes, were named tags. Tags are created by the users that not only collect and/or provide the content but frequently also the organizational means for the content (e.g. the vocabulary used for tagging). The most interesting phenomenon that has come out of this approach is collaborative tagging which we distinguish from normal tagging in that several users may have same resource in their collections tagged differently in order to facilitate their personal organizational and other preferences. In other words: different users tag the same resources with different tags for their own purposes. This article will focus on knowledge and resource discovery techniques resulting from the use of collaborative tagging within a service which allows its users to collect and share some kind of resources.

Since the area has developed rather recently, some new terms have emerged and they yet have to be clearly defined. For example, collaborative tagging is also frequently called social tagging and distributed classification, used as a synonym for folksonomy and even confused with social bookmarking. Folksonomy, another criticized Web 2.0 term denoting a user created taxonomy, should be used for a totality of tags produced by users through a collaborative tagging process, not the process itself. Social bookmarking is a service allowing users to store and share their URL bookmarks online. While it is true that social bookmarking services are among most prominent users of collaborative tagging as organizational means, it is by no means synonymous with collaborative tagging. It is a type of service frequently employing this organizational approach. It is important to separate collaborative tagging as the knowledge organization means and process to increase its applicability to other services.

2. Collaborative tagging in general

We define collaborative tagging as the process by which users of a Web service add natural language keywords to commonly available information resources thus organizing them and creating personalised collections of these resources specific to each user. Each user's personal collection created in this manner is made available to every other user of the service. The fact that the same resource has usually been tagged by more users allows for drawing connections between various users' collections and mutually tagged resources. It thus supports knowledge discovery, tag suggestion and insight into resource popularity and interests and trends of users and communities.

What is a relatively new concept and is interesting to note is that every user not only supplies the hand-picked resources but also the tags by which these resources are organized as well as the abstract organization of his or her collection in general. This happens simply because tags themselves as well as their use and organization (in services which allow some kind of tag organization) are conceptualised by every user for their own needs rather than prescribed by the system. The system only provides the mechanics for defining, assigning and using tags and does not provide any clearly defined mechanics, guidelines or documentation for the way tags should be used semantically or for the precise way in which organization is to be achieved through the use of tags. Advice for tag usage (i.e. what can one denote with a tag, the number of tags per object, the number of objects which should share the same tag and so on) is scarce and poorly documented due to its abstract nature and lack of widely accepted and understood terminology which serves to prove how important education for information society currently is and sets some challenges for the future.

Organizing information objects by means of assigning keywords to facilitate quality searching, browsing and filtering capabilities is by no means a new idea and predates computer use. However, this has traditionally been done for users by experts who would most probably use terms prescribed through a controlled dictionary rather than by users themselves.

This also represents one of the strengths and peculiarities of the system and the most probable reason for the relative popularity of services using this approach since users are allowed to learn the system and organize their collection as they go along rather than having to learn beforehand the system and the terms they would be using. Also, since every user has a collection more or less unique to him or her, they can conceptualise the organization which fits their needs best instead of having to fit it into already made system which can either be too complex for the task or might not fit the subject matter or the type of resources they are gathering.

Services using collaborative tagging approach provide only a flat namespace in which tags are created by the users. No starting set of tags is provided (up to now, at least) so the recommendation of which tags to use starts only after tags have been provided by a certain number of users. It is thus left to users to decide what to use their tags for (i.e. for tagging content, subject, purpose, etc.), to conceptualise, if desired, postcoordinated creation of categories from tags and so on.

3. Services using collaborative tagging

We can find the most prominent use of collaborative tagging in the social bookmarking systems such as del.icio.us but the concept of tagging also owes some popularity to Flickr, which is often regarded as an early example of a Web 2.0 application. Flickr however, does not, strictly speaking, use *collaborative* tagging since by default only content owners can tag the content they are uploading. The service does have some aspects that enable collaborative organization but most resources are tagged only by owners.

3.1. Social bookmarking

Social bookmarking services allow their users to keep their "favourites" or "bookmarks" (i.e. URL's they want to collect) online and thus use them from any computer with Web access, often to tag them via a collaborative tagging approach and discover the resources other users' have collected for themselves. It has been said that this kind of service has three major axes: users, tags and URLs [1]. These axes (as is the case with del.icio.us) are often reflected in the clean URL design used by social bookmarking sites.

There is a growing number of social bookmarking services and most of them use the collaborative tagging approach to facilitate organization, information retrieval and knowledge discovery. Some of these systems are

general in that the user can collect URL's to any type of resource and, currently, no specialised mechanisms (i.e. automatic metadata harvesting) have been provided for any type of resource. Del.icio.us is a good example of this type of service although one can find many more by looking at any list of Web 2.0 services.

Some social bookmarking services have been specialised for keeping URL's of specific resources such as scientific texts, blogs or news. This kind of services includes advanced features useful for automatic manipulation of these kinds of resources. Two best examples of this kind of service are Connotea and CiteUlike, which share a number of similar features besides using collaborative tagging and supporting some mechanisms for knowledge discovery that are derived from it. Both automatically harvest metadata from a certain number of sites meaning that when a user wants to add a URL from a supported site containing a scientific article the service will automatically gather data necessary to provide a full citation later. All that user needs to do is add tags. These services also use other identifiers besides URL that support resource access (e.g. DOI) in order to provide access to collected articles after the URL changes and also provide export to and import from popular bibliographic software such as BibTEX or EndNote. CiteUlike also goes a bit beyond being just a social bookmarking tool and allows users to add citations purely for bibliographic reasons (resources not accessible via the Internet) and to add their own personal .pdf copies of the texts. Unfortunately, user uploaded pdf-s are available only to users who uploaded them to avoid possible copyright infringement.

From the point of view of knowledge discovery social bookmarking and especially that of scientific texts is perhaps the most interesting service using collaborative tagging.

3.2. Web clips

An increasing number of services, such as Clipmarks, are starting to gather not the URLs of resources but rather most interesting parts of the resources themselves in form of clips taken from Web pages. This well transcends the boundaries of traditional documents in that it creates and stores new information objects from parts of the old ones. In a way this represents a type of digest of the Web. These clips, or snippets as they are sometimes called, are constituted out of text, pictures or both taken from a Web page and are copied to the server on which the service is hosted. The URL thus only becomes a reference for where the clip came from and can be used if further reading from the same site is desired.

An important aspect of this kind of service is popularity: when viewing a clip, a user has an option to vote for it usually by clicking a conveniently placed button. When browsing other people's clips a user, except by filtering the list through tags, has an option to sort clips by popularity thus gaining a recommendation for which clips to view first.

3.3. Media sharing and discovery

A large number of services promoting media sharing such as Flickr and YouTube use tags but not *collaborative* tagging.

There is, however, a growing number of services which allow users to discover new media they might like on the basis of what they currently like (and have access to) in comparison to what other users with similar preferences like. By comparing more or less similar collections of a large number of users the service is able to recommend resources to a user based on what other users with similar interests have in their collections and he or she does not.

This is especially true of music where services such as Last.fm also connect this with Internet radio so users can get automatically generated radio station based on their profiles. Like other services specialised for a certain type of resource, this type also tries to gather as much data possible automatically so it can build a user's profile over time by being able to know and gather the titles of songs and albums as well as the names of artists a user is listening to.

Collaborative tagging is thus not a backbone of this kind of services as it is for social bookmarking but rather an "optional extra" users might use in order to support resource discovery and browsing and organization in general via tags rather than just preference comparison.

4. Knowledge discovery via collaborative tagging systems

We have mentioned before that the users supply hand-picked resources from which their collections are built. The very fact that these resources are hand-picked provides the backbone for knowledge or resource discovery via services using collaborative tagging since this fact makes sure that the service becomes a recommendation

system of sorts. This contrasts sharply with search engines which are great for *ad hoc* searches and known item retrieval but are not at all that great for new content discovery since they do very little semantic analysis and do not function as a recommendation system.

The kind of knowledge discovery supported by this kind of services is more reminiscent of Web directories since users browse the more or less specific lists of resources which are sorted and filtered in various ways and click on what catches their eye. These services, however, support a much different type of browsing, the lists of resources are often quite different in nature, origin, sorting and so on and these services are, on the whole, much more comprehensive than traditional directories. All this happens simply because they are Web 2.0: they harness the power of the users. Also, it has to be kept in mind that the tags themselves open up new possibilities.

4.1. Tags and knowledge discovery

In a larger system, a single resource is likely to and should be described by multiple tags. Generally, it is considered good practice if tags cover as many facets or aspects (i.e. subject, content, time, purpose etc.) as possible since this helps in both retrieval of the resource by the user in whose collection it is and its discovery by other users [13]. Well organized (e.g. in del.icio.us bundled according to facets) and consistently used tags in a single collection will help other users browse it and thus discover new and interesting resources. From this some facts about knowledge discovery in services using collaborative tagging may be derived. A user might discover:

1) resources he or she is interested in
2) other users/other users' collections he or she is interested in
3) tags he or she is interested in

These types of information resources a user might discover lead to one another. Let us consider a user A and a user B. User A might find a popular resource on the list of resources in the greatest number of collections (i.e. popular links), add that resource to collection and look at the list of users who also have it in their collection. Then, the user might recognise another user, user B, who has several other resources user A has also collected in his collection at which point user A might access the collection of user B whom he now identifies as sharing some similar interests. While browsing that user's collection the user might discover interesting tags and gain new ideas for organising his collection. He could then subscribe both to the tags he discovered and to the collection of user B depending on his preferences, discover other resources and so on.

One way one can think of tags themselves is in terms of their objectivity or subjectivity. If a tag is objective then it describes an inherent property of a resource without relating it to the users tagging it such as subject (i.e. cats, WWW, video), content (i.e. downloads, tutorials), genre (i.e. jazz, comedy), type (i.e. homepage, blog), etc. Objective tags support knowledge and resource discovery to the greatest extent so their definition and usage should be encouraged and it is generally a good idea, from the knowledge discovery point of view, that they cover as many different facets as possible. Although there will usually be some disagreement among users relating to description of certain resources, this in a way supports the identification of users with similar interests for we are most likely to identify with those that tag in a similar manner. This is especially true of resources which posses highly subjective qualities, such as the genre of a popular music composition.

Subjective tags on the other hand describe the relationship between a user and a resource such as purpose (i.e. temp, frequent), task (i.e. to_do), user's perception of the resource (i.e. fun), etc. Subjective tags are mainly used for users' organizations of their own collections and do not support knowledge discovery so much as objective tags.

The other way of looking at tags is in terms of their specificity. This is simply described with an example: user A, who is greatly interested in cats, might tag a resource about the ocicat cats as "ocicat"; user B, who is also interested in cats but to a much lesser degree, might simply tag the same resource as "cats". When one uses the term "ocicat" from a specified taxonomy, thesaurus or a similar knowledge organization tool containing hierarchical relationships it is immediately apparent that anything labelled "ocicat" has something to do with "cats" but is more specific since the term "cats" is a parent of the term "ocicat".

However, current collaborative tagging systems produce a flat list of tags without any hierarchical relationships except which users use which tags. When one uses the term "ocicat" from such a list, one can only learn which users

are using it and not its broader or narrower categories. Some researchers have expressed their concern about the adverse effect of this on the resource discovery and retrieval in systems using collaborative tagging since "ocicat" might be too specific for some while "cats" might be too general for others [7]. While this does indeed present a problem for tag search and tag subscriptions (since a person looking for resources about "cats" will not retrieve a resource tagged "ocicat", for example) it does not really present a problem for identifying users with similar interests, which is one of the main knowledge discovery techniques in these systems, since it is obvious from the tags to which level is a user interested in a certain subject.

5. Mechanisms for knowledge discovery in collaborative tagging systems

Different services using collaborative tagging provide similar mechanisms for knowledge and resource discovery that are derived from the nature of the system itself and, sometimes, from the type of resources the service is made for.

5.1. Recommendation

Recommendation in services using collaborative tagging functions on two levels:

1) resources have been picked, described and added to their collections by human users
2) resources have their general popularity which may be reached in two ways:
    a. via the number of users who have the resource in their collection
    b. via the number of users who have "voted" for the resource

The first level can be tied to the identification of users with similar interests. Simply, if a user A has similar interests to user B and has a resource X in his collection while user B does not, then user B may treat the resource X as being recommended by user A.

The second level functions on an overall popularity of a resource: if a resource is popular then it may be treated as a recommended resource (indeed, this idea has been around since Amazon's recommendation system and, later, PageRank). In most services popularity is gained simply through inclusion of a resource in their collections by users: the more users have a resource in their collections the more popular the resource is. This is often indicated by an easy to see number attached to the title or a description of a resource.

Some resources, however, can't be pronounced popular this way simply because most of them are too unique to be in several collections at the same tame. This is characteristic of the resources that are stored on the server rather than just metadata about them. Web clips and news items are a good example. This is then circumvented by a voting system: an easy "give it my vote button" (usually called "pop it", "digg it" or something else equally catchy and specific to the service) is placed near the resource so when a user views it he or she, if she likes it, can easily give it a vote.

Whether it is the first or the second case, recent popular resources are usually displayed on the main page.

5.2. Identification of users with similar interests

Much of the knowledge discovery in services using collaborative tagging depends on users' ability to identify and connect with other users who share the same interests. Because of this, when a user adds a resource to his or her collection he or she can see which other users have also added it to their collections. It helps if one notices a user with whom one shares more than one resource. By browsing these users' collections one can pretty soon find users with similar interests by browsing through:

1) the tags a user is using
2) the resources a user has collected

Browsing through the tags is more prominent simply because it requires much less time and effort. There will usually be much less tags than resources although the number of tags compared to the number of resources is by no means uniform among the users [7]. Also, tags are usually more clearly defined than the titles.

5.3. Users' collections subscriptions and user networks

After a user has found other users with similar interests she can either subscribe to their collections, meaning he gets all their links and tags in one place without having to navigate much, or add them to his or her network, depending on what the service supports. Network usually allows some advanced features such as tagging a resource for a user in one's network.

This way one user bookmarks a URL not for himself but for another user. The other user has a special list of bookmarks others have tagged for him from which he can either delete them or add them to his collection.

5.4. Tag search and subscription

One can use tags or simply words he assumes others used as tags to search all users collections. If one is repeating the same search frequently one can subscribe to a tag, meaning that he gets an easily accessible and updated list of every resource that has been tagged by any user with the selected tag. This can be useful if one is trying to follow what is happening in a specific area (i.e. one could subscribe to the "Web2.0" tag). Tag discovery plays an especially important role for this aspect of collaborative tagging systems.

5.5. Tag clouds

Tag clouds are interesting navigational devices made possible by user tagging, not necessarily collaborative. Tag cloud visually shows the most popular tags where the frequency of tag usage is denoted by font size: the larger a tag the more often it was used. From this it is more or less obvious that we can have two kinds of tag clouds. In the first kind the size shows the number of times a single tag has been used for any resource in the whole system. This shows the general popularity of subjects and what is currently "hot" in the community using a particular service. The other kind is tied to a single resource: it shows the totality of tags that have been applied to the same resource by all the users of the service who tagged it. This helps in recommending tags. It is also useful for resources which defy objective tagging like genre of popular music. It might, for example, show what the greatest number of users thinks about the genre of a popular artist which could help reach a consensus.

6. Conclusion

As we have seen, collaborative tagging is an important organizational method for Web 2.0 which opens up possibilities for knowledge discovery previously impossible. However, this method is not without its flaws: the problems of objectivity/subjectivity and specificity as well as complete lack of vocabulary control will hinder organization and discovery and lead to increased levels of meta-noise. Although the very nature of the system might circumvent some problems (i.e. the very fact that the same resource is tagged by more users might overcome the specificity problems) the exact nature of the means by which these problems might be overcome should be investigated. Also, it is clear that education for information society should include abstract organizational principles to support these kinds of efforts since chances are these are some of the core competencies for the future. This being said, the services themselves should provide clear guidelines to help users understand the problems and possibilities of the system and to support this kind of education.

7. References

[1] Biddulph, M. (2004). Introducing Del.icio.us. XML.com; 2004. http://www.xml.com/pub/a/2004/11/10/delicious.html [02/27/2007]

[2] CiteULike. http://www.citeulike.org/

[3] Clipmarks. http://clipmarks.com/

[4] Connotea. http://www.connotea.org/

[5] Delicious. http://del.icio.us/

[6] Flickr. http://www.flickr.com/

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

[8] 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[02/26/2007]

[9] Last.fm. http://www.last.fm

[10] Lund, B. et al. (2005). Social Bookmarking Tools (II): A Case Study - Connotea. D-Lib Magazine, 11 (4), April 2005. http://www.dlib.org/dlib/april05/lund/04lund.html [02/23/2007]

[11] 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]

[12] YouTube http://www.youtube.com

[13] Xu, Z. et al. (2006). Towards the Semantic Web: Collaborative Tag Suggestions. WWW2006, May 22–26, 2006, Edinburgh, UK. http://www.rawsugar.com/www2006/13.pdf [02/25/2007]