EXTENDING THE BASIC TAGGING MODEL: CONTEXT-AWARE TAGGING

Christoph Trattner

Institute for Information Systems and Computer Media, Graz University of Technology
Inffeldgasse 16c, 8010 Graz, Austria

Denis Helic

Knowledge Management Institute, Graz University of Technology
Inffeldgasse 21a, 8010 Graz, Austria

ABSTRACT

Tagging systems are very popular tools for organizing and structuring information about arbitrary Web resources by assigning simple keywords (tags) to those resources in a collaborative fashion. Folksonomy is a data structure resulting from the tagging process and is characterized as a flat collection of triples of the form user-resource-tag. Research on folksonomies is only taking off and as of now a number of open research questions and problems require to be answered. One of such problems is the problem of tag ambiguity – a situation where one and the same tag might have multiple meanings depending on users, topics, or context in which the tags have been created. In this paper we present a simple solution to the ambiguity problem by extending the basic triple model of tagging systems with the fourth element: context. Thus, each tag is (automatically) put into a context. Note here that the notion of context is a general one. That is the context might be inferred from the content of a resource, external information structures existing on the top of a resource, or from properties of the user profile. Lastly, the paper presents an implementation of this principle in an online Wiki-based encyclopedia called Austria-Forum and discusses directions for the future work.

1. INTRODUCTION

Recently, with the emergence of modern Web2.0 applications such as delicious (http://delicious.com/) or Flickr (http://www.flickr.com/) social tagging systems gained tremendously in popularity (Marlow et al. 2006; Wu et al. 2006). In these systems users are allowed to add simple keywords (tags) without predefined vocabulary to different resources to describe them. A subset of tagging systems comprising the tagging bookmarking services like Delicious, Citeulike (http://www.citeulike.org/) or bibsonomy (http://www.bibsonomy.org/) have received community focus due to ease of use and information discovery mechanisms. In these systems users assign free form keywords and annotations to the addresses (URLs) of resources, e.g., web pages (Hammond et al., 2005). The weighted set of keywords (tags) assigned to a resource by all users within a system and visualized as a navigation support is called the tag-cloud.

   More generally, as Ames and Naaman (Ames and Naaman, 2007) suggest the user motivation to tag a resource might be organizational or communicational on one hand, and on the other hand the users tag resources for their personal use and/or to share it with others. For example, users that tag resources for their personal use in an organizational sense use social tagging applications to organize interesting, important, and related resources according to their current needs. The tags are applied as a support for later search and retrieval of tagged resources via search or navigating the tag cloud. Typically, the tag cloud provides an overview of defined tags showing only the tags themselves but not the actual content of the tagged resources. The resources are represented via navigable links. Another motivational aspect of using tags is to share it with other users and in such a scenario tags are typically used in a communicational sense to send signals to other users about resources that might be of interest in a more general case.

Different social tagging systems differ in the ways how users may add tags and to what resources. In general the following tagging mechanisms and strategies might be identified:

- Support of tag definition by suggestion – the system suggests to users adding a new tag to a particular resource a number of tags which were already assigned to that resource by previous users.
- Very often, tags statistics are calculated to, for instance, aggregate similar tags. These statistics might take into account type of information items that are tagged and the relations between the resources, relations between the users, and so on (Marlow et al. 2006).

Regardless of a particular tagging model all social tagging systems are characterized by a network of relations between users, information items and tags. Depending on the model this network might stabilize with time forming a so-called „folksonomy"(Marlow et al. 2006).

Currently, the research on folksonomies is only taking off and a lot of open research questions still remain to be answered. For example, the question if the convergence of a tag network to a folksonomy is a positive development that utilizes „better" organized information is still unanswered (Wu et al. 2006). Also, the ambiguity problem (when a single tag might have multiple meanings) or the synonymy problem (when different tags have the same meaning) is still not properly solved in social tagging systems (Marlow et al. 2006). Finally, folksonomies are flat data structures, i.e. the relations between tags do not exist. Some research work has been done to enrich tags with relations. For example, (Heymann and Garcia-Molina, 2006) developed an algorithm for automatic creation of hierarchical taxonomies out of a folksonomy.

In this paper we will investigate in more details a special case of the ambiguity problem related to multiple meanings of a tag in different contexts. Let us illustrate this problem with a simple example. Suppose we have a digital library with articles from the High Energy Physics community such as arXiv (http://arxiv.org/). The library contains article from disparate areas of physics and mathematics such as astrophysics, experimental high energy physics, theoretical high energy physics, and so on. Suppose now that a scientist with the background in experimental physics tags an article from the field of experimental high energy physics with the tag "model". Most probably, the scientists used the tag "model" to signal that the article deals with a model or models of experiments. Suppose now that another scientist with the background in theoretical physics tags another article from the field of the theoretical high energy physics. Again, the scientist uses the same tag: "model". In this case, however, the scientist wants to point out that the article deals with a theoretical model of say sub-atom particles. Obviously, these two tags have different meanings; however the tagging system cannot make the distinction between them and shows both of the articles under the tag "model" in the tag cloud. Thus, the basic tagging model of triples with the form tag-user-resource does not offer a possibility to distinguish between one and the same tag in two different contexts.

Therefore, it is our suggestion to extend this basic tagging model to a quadruple with the form tag-context-user-resource. In the previous example the context for the first tag would be "experimental high energy physics" and in the second case the context would be "theoretical high energy physics". In this paper we discuss this idea in more details – especially we investigate a possibility for (automatic) provision of tag context on the example of a system called Austria-Forum – an online wiki-like encyclopedia with content related to Austria. The key here is to retain the simplicity of the basic tagging systems, i.e. users themselves need not to explicitly define the tag context since such an intrusion would compromise usability of the system. Rather, the system provides the tag context automatically.

The rest of the paper is organized as follows. The second section presents shortly Austria-Forum and discusses the hierarchical organization of information in the system. The third section discusses a possible solution for the automatic provision of tag context in Austria-Forum. The solution is based on a simple idea that the resource context provided by the hierarchical resource organization should be inherited by tags. The fourth section describes the first implementation of the idea and the final section discusses directions for future work.

2. AUSTRIA-FORUM

Austria-Forum (http://www.austria-forum.org or http://www.austria-lexikon.at) is a networked information system that manages a very large repository of contributions, where new contributions are easily published, edited, checked, assessed, and certified, and where the correctness and a high quality of each of these contributions is backed by a person that is accepted as an expert in a particular field. Consequently, each of

the contributions is citable as any other editorially checked content and might be used in education, scientific research, or journalism. The content of Austria-Forum is always related to Austria – as such Austria-Forum might be seen as an Austrian online encyclopedia.

In the first experimental phase of Austria-Forum the system had an editorial board of more than 20 editors and a growing community of users. The number of users who contributed with the content was more than 100. The number of unique users who have visited the site was around 4000 each month.

The current number of contributions is around 80000 (including pictures and videos as well as the content converted from the well-known Austrian cultural information system AEIOU (http://www.aeiou.at)), out of which around 6000 are user-generated contributions – approximately 8% of all contributions. Most of these user contributions are pictures and photos, with a small number of blogs, discussion forum posts, and comments.

Although these numbers are quite substantial for a site that has been online experimentally a more active community involvement is desired. Community tools and facilities are already present in the system. However, as a number of users suggested, usability and a better integration of different community tools with the main system needs to be improved.

Therefore, the original system that was technically based on an in-house developed content-management system has been replaced by an open-source wiki software called JSPWiki (http://www.jspwiki.org) in order to support the user with a popular and easy to use hypertext authoring concept (see Wikipedia for instance).

But as often criticized, wikis in general do not support by default (see original wiki invented by Ward Cunningham WikiWikiWeb (http://www.c2.com/cgi/wiki)), the possibility of structuring information by means of categories or tags.

Therefore the original JSPWiki clone was enhanced in order to support these features by means of implementing a simple categorization mechanism via structural links (Thistlewaite, 1997). Essentially, the system manages an editorially created category structure, i.e. a hierarchical taxonomy. Each contribution is a part of exactly one category from the taxonomy. Additionally, each category has exactly one parent category. The system supports user navigation by means of structural links which are automatically created from categories to the contributions that belong to these categories.

A simple hierarchical addressing schema for contribution URLs has been chosen. The following URL patterns are defined in the system:

http://www.austria-lexikon.at/<category-page>/< page>
http://www.austria-lexikon.at/<category-page>/<category-page>/< page>
http://www.austria-lexikon.at/<category-page>/<category-page>/<category-page>/< page>

Category pages (<category-pages>) normally provide an overview of the category and structural links to sub-categories and contribution belonging to that category.

Thus (for example) in order to retrieve a contribution about "Konrad Lorenz" out of the category "Biographien", the following URL is constructed:

http://www.austria-lexikon.at/af/Biographien/Lorenz_Konrad

In order to retrieve all contributions from category "Biographien", the following notation is used:

http://www.austria-lexikon.at/af/Biographien

As this kind of hierarchical addressing scheme is rather popular in the Web (see dmoz.org, google directory), whenever it comes to addressing structured data items, it is assumed that users should get used to this kind of notation and information retrieval behavior within the Austria-Forum system rather quickly.

3. PUTTING TAGS IN CONTEXT

Providing the context for tags in Austria-Forum follows a simple idea that the category to which a contribution belongs already provides such a context. Thus, whenever a user tags a certain contribution the system automatically extends the basic tagging triple of tag-resource-user to a quadruple of the form tag-resource-user-category.

Thus, with this approach, tags are still applied to a single information item like a normal wiki-page or a category-page within the Austria-Forum system (see figure 1), but the way tag-clouds are computed and used within the system is differently from popular known concepts. The context information that is now available

for each tag can be used to provide users with different tag clouds and in this way with alternative navigational possibilities.
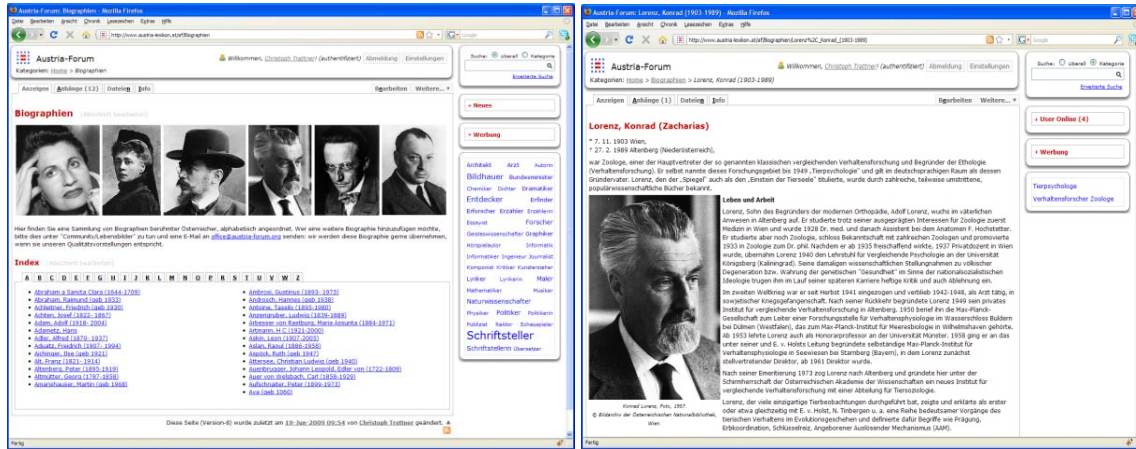

Figure 1. Left image: A typical category-page as it shows up in AF. Right image: A typical article/information-item.

Currently, the system computes tag clouds for each category page taking into account only those tags that are assigned to sub-categories and contributions belong to that particular category. In this way each tag-cloud is contextualized supporting a possible ambiguity in tags meaning.

Let us illustrate this with an example. In Austria-Forum there are several contributions about famous Austrian composer Mozart:

- Mozart biography that belongs to the Biographies category
- Several contributions on post stamps with Mozart that belong to Stamps category
- Several contributions on coins with Mozart that belong to Coins category
- Several contributions on Mozart monuments in Monuments category
- Several contributions on Mozart composing work in Compositions category and different sub-categories such as Operas, Symphonies, and so on.
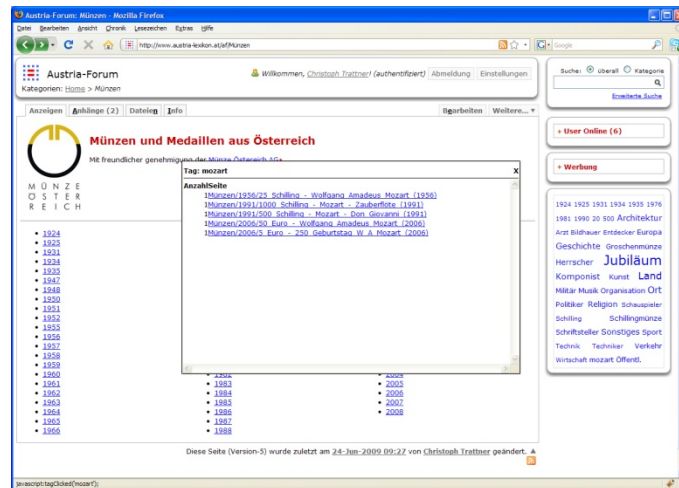

Figure 2. Simple pop-up window showing assigned URL's to a particular tag "Mozart".

In a traditional tagging system each of these contributions is very likely to be tagged (among other tags) with "Mozart" tag. A standard tag-cloud will show the "Mozart" tag prominently because there would be a huge number of contributions tagged with "Mozart". However, a distinction between different contexts to which particular contributions are related cannot be visualized with a standard tag-cloud. However, in Austria-Forum each category page has its own contextualized tag-cloud. Thus, a user interested in coins with

Mozart would in the first step navigate to the Coins category. At the category page a tag-cloud is shown with the tags assigned only to contributions from the Coins category. Thus, the "Mozart" tag provides only links to contributions to coins with Mozart (see figure 2).

3.1 Tag cloud calculations

In order to compute a tag cloud of a top-category holding many tag clouds of sub-categories/pages, frequencies of particular tags/sub-categories/pages are simply added to the one within in the top-category. Thus, being more precise, tag clouds within the Austria-Forum system are calculated in the following form:

$$\left\lceil \frac{f_{max} * (t_i - t_{min})}{t_{max} - t_{min}} \right\rceil = f_i$$

for $t_i > t_{min}$; otherwise $s_i = 1$

f_i	=	font size being displayed
f_{max}	=	max. font size
t_i	=	tag frequency
t_{min}	=	min. tag frequency
t_{max}	=	max. tag frequency

t_{min}, t_{max} and t_i are computed recursively in the following form,

t_{min}	=	calc_t_{min}(page)	=	calc_t_{min}(page-1)	+	t_{min}(x = page-1)
t_{max}	=	calc_t_{max}(page)	=	calc_t_{max}(page-1)	+	t_{max}(x = page-1)
t_i	=	calc_t_i(page)	=	calc_t_i(page-1)	+	t_i(x = page-1)

where the term "page-1" means one page level down, and the functions t_{min}(x), t_{max}(x) and t_i(x) represent the corresponding t_{min}, t_{max} and t_i of a given page level.

For simplicity tags are sorted by tag name at the moment. Thus a combination of tag clouds out of the sub-categories contained (for instance) in the sub-category "Community-Beiträge" (see figure 4 and 5) within the Austria-Forum system looks the following (see figure 3).
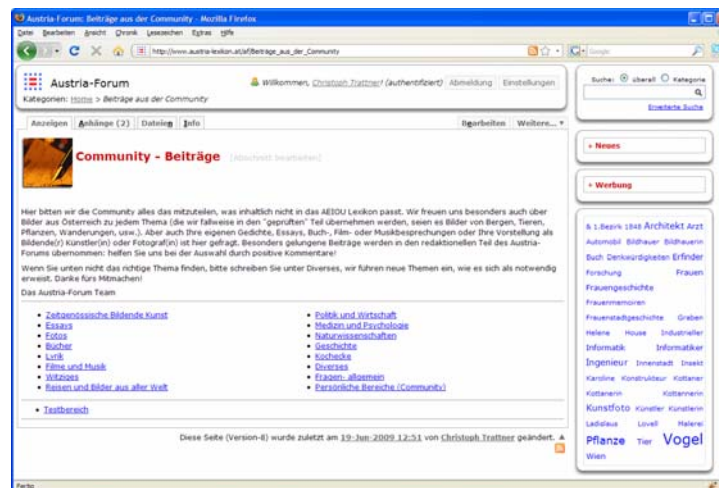

Figure 3. Sub-category "Beiträge_aus_der_Community" and tag cloud of all sub-categories on the right side.
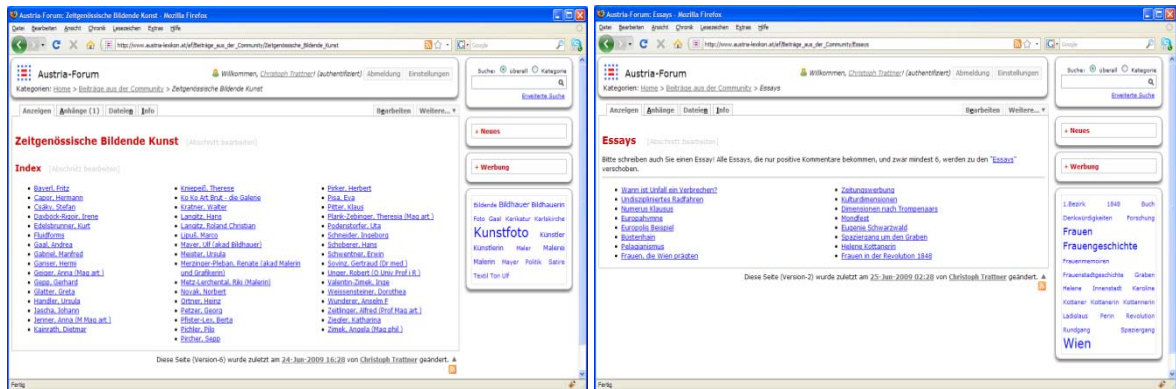
Figure 4. Left image: Sub-category „Beiträge_aus_der_Community/Zeitgenössische_Bildende_Kunst". Right image: Sub-category „Beiträge_aus_der_Community/Essays"
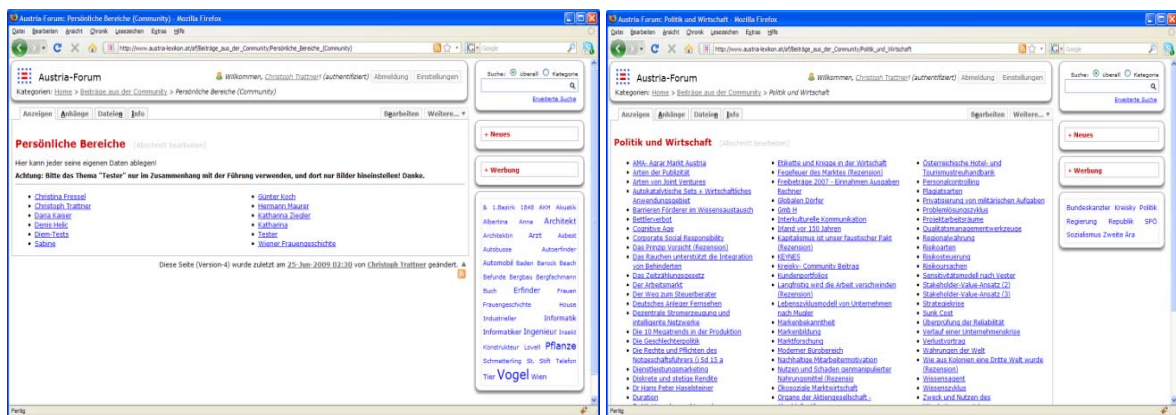

Figure 5. Left image: Sub-category „Beiträge_aus_der_Community/Persönliche_Bereiche". Right image: Sub-category „Beiträge_aus_der_Community/Politik_und_Wirtschaft"

4. IMPLEMENTATION

As shown in figure 6, the tag cloud module within Austria-Forum relies on two major parts: A XML based tag cloud service and an AJAX based tag cloud plug-In.

The tag XML based tag cloud service module can be seen as the center peace of the whole structured tagging module. It consists of three peaces which do the following:

- A mySQL database module based on the model of the well-known open source tagging software Scuttle for storing tags for particular pages within the Austria-Forum system.
- A controller unit, which deals with tag cloud calculations and authorization procedures over a authorization bridge.
- A tag cloud servlet, which handles requests of format GetTags(URL), to produce a XML based tag cloud representation within the context of URL and GetLinks(URL,Tag), to create a XML based list of all links for the tag "Tag" within the context "URL".

In order to show up a nice tag cloud within the Wiki clone of Austria-Forum a tag cloud plug-in was implemented. It self mainly acts as a simple JavaScript based rendering module, which fetches data via XMLHttpRequest and transforms the so fetched "raw" XML based tag cloud information from the tag cloud service routine, into a neat XHTML based and CSS styled tag cloud presentation (see figure 3 for instance).
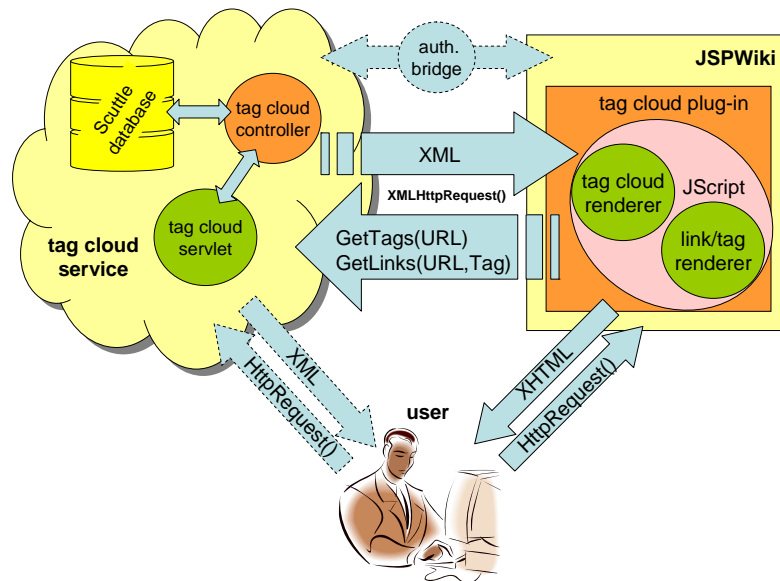
Figure 6. Architectural diagram of the tag cloud module within the Austria-Forum system.

5. FUTURE WORK

After Austria-Forum system goes online in October 2009 the presented approach will be evaluated with real users. The evaluation methodology will be based on a quantitative analysis of the user navigational data. A number of tags from the complete frequency spectrum might be selected to conduct the investigation. For example, suppose we select the tag "Mozart" for such an investigation. By access log analysis and tracking of user sessions we might measure the following properties of different information retrieval possibilities:

- By searching for term "Mozart" at the starting page the user obtains a list of search results from the whole server. Thus, contributions from different categories that deal with Mozart are presented to the user. The user browses the search result and finally reads a contribution on Mozart. By tracking how much time the user spends on each of the search results we might conclude which of the contributions has been desired by the user. Suppose that this was the contribution on a coin with Mozart on it. Now we might compare the number of clicks the user needed to make to obtain the desired contribution or the time needed to find it with the same properties of using contextualized tag clouds.
- The same scenario as above might be applied to purely navigational information retrieval, i.e. to the process of browsing the categories.
- Finally, the distribution of information retrieval facilities might be calculated to see which of the possibilities is most popular with the users.

A number of research studies already investigated with possibilities to provide context to the tags. For example, (Coyle et al., 2008) discussed weighting tags according to whether a user is a leader or a follower, i.e, they discussed the possibilities of putting tags in contexts of the user profiles. We plan to investigate such possibilities in more details.

Moreover, since the quadruple model defined above does not define the kind of context that the tags might be provided with other possibilities are imaginable. For example, in the current implementation the tags are enriched with the context from a hierarchical taxonomy already presented in the system. However, the tags might be also enriched with the context from thesauri editorially or even informally created. For example, the Wikipedia categories form informal thesauri created by numerous Wikiepdia users. The thesauri from the German Wikipedia might be used to obtain a proper tag context even in an automatic way. The wikification mechanism introduced by Witten applies the Wikipedia thesauri to automatically link textual documents to Wikipedia articles. A similar approach might be applied to find a best match concept from Wikipedia thesauri for a particular tag and in this way put the tag in question in a context. If such approaches

are proved to be sufficiently efficient then the context might be also computed for the tags of any Web resource, not only contributions from Austria-Forum.

ACKNOWLEDGEMENT

REFERENCES

Ames, M. and Naaman, M. 2007. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA, April 28 - May 03, 2007). CHI '07. ACM, New York, NY, 971-980, 2007.

Coyle M., Freyne J., Brusilovsky P., Smyth B., 2008. Social Information Access for the Rest of Us: An Exploration of Social YouTube. In *Adaptive Hypermedia and Adaptive Web-Based Systems* 2008., 93-102.

Hammond T., Hannay T., Lund B., Scott J., 2005.. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine, 11(4)*, April 2005.

Heymann P., Garcia-Molina H., 2006. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical Report. Stanford.

Marlow, C., Naaman, M., Boyd, D., and Davis, M. 2006., HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia* (Odense, Denmark, August 22 - 25, 2006). HYPERTEXT '06. ACM, New York, 2006.

Thistlewaite P., 1997. Automatic construction and management of large open webs, *Information Processing & Management, Volume 33, Issue 2, Methods and Tools for the Automatic Construction of Hypertext*, March 1997, Pages 161-173.

Wu, H., Zubair, M., and Maly, K., 2006. Harvesting social knowledge from folksonomies. In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia* (Odense, Denmark, August 22 - 25, 2006). HYPERTEXT '06. ACM, New York, NY, 111-114, 2006.