

Document downloaded from:

<http://hdl.handle.net/10251/59784>

This paper must be cited as:

Estelles Arolas, E.; González Ladrón De Guevara, FR. (2012). Study about the different use of explicit and implicit tags in social bookmarking. *Journal of the American Society for Information Science and Technology*. 63(2):313-322. doi:10.1002/asi.21663.



The final publication is available at

<http://dx.doi.org/10.1002/asi.21663>

Copyright Association for Information Science and Technology (ASIS&T): JASIS&T

Additional Information

This is the accepted version of the following article: Arolas, E. E., & Ladrón-de-Guevar, F. G. (2012). Uses of explicit and implicit tags in social bookmarking. *Journal of the American Society for Information Science and Technology*, 63(2), 313-322. doi:10.1002/asi.21663, which has been published in final form at <http://dx.doi.org/10.1002/asi.21663>

Title: Study about the different use of explicit and implicit tags in social bookmarking.

Enrique Estellés Arolas

Universidad Politécnica de Valencia

Camino de Vera s/n – 46022 – Valencia - Spain

Phone: 689435554

kikeestelles@gmail.com

Fernando González—Ladrón—de—Guevara

Universidad Politécnica de Valencia

Camino de Vera s/n – 46022 – Valencia - Spain

Phone: 645-982-912

fgonzal@omp.upv.es

Enrique Estellés (kikeestelles@gmail.com) will receive correspondence and proofs

Abstract:

Although Web 2.0 contains many tools with different functionalities, they all share a common social nature. One tool in particular, social bookmarking systems, allows users to store and share links to different types of resources i.e. websites, videos, images, etc. In order to identify and classify these resources so that they can be retrieved and shared, fragments of text are used. These fragments of text,

Introduction

Web 2.0 has enabled the proliferation of applications such as blogs, social networks, wikis, social bookmarking systems, etc. These allow users to communicate and share resources in a collaborative way in a professional field as well as in an academic or research sphere. These web applications have 3 common features: there are user profiles, it is possible to follow other users or add them as friends or contacts, and it is possible to add comments to the generated content (Mason & Rennie, 2008).

Another feature that most of these systems share is the possibility of labeling the contents through the use of keywords called tags. The content can be a blog entry (e.g. technorati.com), a resource marked in a social bookmarking system (e.g. delicious.com), books (e.g. librarything.com), objects in a museum (e.g. www.steve.museum), user-generated videos (e.g. youtube.com), or images (e.g. flickr.com) (Bar-Ilan et al., 2010). Tags are very important in these types of systems because they make the search of these resources as well as their organizations and description easier (Oliveira et al, 2008) and they also enable the user to find similar resources (Millen et al., 2005). Social bookmarking systems are web applications that allow users to store and manage their markers or favorites not in the browser but instead in a central server, so that they can be consulted from different locations and shared by other users (Illig et al., 2009).

Regarding text resources (i.e. text found in a website or in a blog entry), two types of tags can be found: obvious, also called explicit, or non-obvious, also called implicit. Implicit tags are those that do not appear within the textual content of the resource. Explicit tags are those appearing at least once within the textual content visible for users. For example, they can appear within a web title, a paragraph, or a link of the website itself (Farooq et al., 2007, Liu et al., 2008).

Throughout this paper some questions will be answered, such as – in general, do users use the same quantity of explicit tags and implicit tags?, what about on a resource level?, furthermore, on a resource level, are explicit tags stored in a series of specific resources, or are they distributed equally among them all?, is there a difference between the lengths of the two types of tags?, and are the terms most frequently used to tag resources implicit or explicit?

This paper is divided into four sections. The first section consists of a theoretical introduction about the tags and a detailed description of some features. In the second section, the methodology that has been implemented is described and in the third section, the analysis that has been carried out. Then, the paper takes into consideration the results that have been obtained and answers the questions previously asked. Finally, conclusions are shown along with a series of suggestions about the applications and for future research.

Tags

Definition

Tags are generated and freely chosen by the user to form descriptive strings, which are assigned or associated with a resource (Millen et al, 2007; Koutrica et al., 2008; Farooq, 2009; Lipczak & Milios, 2010). Depending on the tag system design, these descriptive chains can be words, phrases, or a combination of symbols and alphanumeric characters (Yeung et al., 2009).

Tags can also be considered as metadata (Subramanya & Liu, 2008); i.e., data about data. The three types of metadata are administrative, structural, and descriptive (Taylor, 2003) and can be developed by

A folksonomy can be defined as a tuple $F = (U, T, R, Y)$ where U , T and R are finite sets, whose elements are called users, tags and resources respectively. Y is a ternary relation between them, i.e., $Y \subseteq U \times T \times R$. The elements $y \in Y$ are called tag assignments (TAS). A post is a triple (U, T_{UR}, r) with $u \in U$, $r \in R$ and a non-empty set $T_{UR} := \{t \in T \mid (u, t, r) \in Y\}$ (Schmitz et al., 2006).

This paper will only focus on the relationship between resources and tags used to mark them, particularly on those tags known as explicit, which will be explained later.

Functions and motivation

According to Golder & Huberman (2005), there are 7 basic non-exclusive functions that a tag can carry out: to identify what or whom the resource deals with, to identify what it is, to identify who owns it, to refine categories, to identify qualities or features, to aid in self-reference (e.g. “myStuff”), and to organize tasks (e.g. “toRead”). Körner et al. (2010) and Millen et al. (2007) are more specific and they put these functions into only two groups namely, to categorize and to describe. Users using tags in order to categorize are called categorizers and they use a more complex set of tags with the main purpose of creating taxonomies for group resources. This system enables to use multiple tags so that a given resource can belong to more than one category. On the other hand, there are users that use tags with a descriptive purpose. These are called descriptors and consider the tag as a way of accurately and precisely describing saved resources. The main goal of these users is to use the tagging for a subsequent search and retrieval. The difference between these two functions is minimal in practice and users are capable of tagging with a double purpose –categorizing and describing. Other authors, like Ding et al. (2010) argue that the principal functions of the tags are to navigate, browse, and retrieve resources. They highlight the social

- To help users build communities that share their expertise and resources
- To navigate
- To browse serendipitously
- To receive feedback on their actions

All these functions can be completed through the technique known as *pivot browsing* (Millen et al., 2007; Bateman et al., 2009). This technique enables the user to reorient the navigation view by clicking on different elements of the user interface, e.g., the name of the users or the tags. By clicking on a user's name, all the resources stored by him will be displayed. By clicking on a tag, resources marked with the same tag will be shown (Millen et al., 2005).

In regards to the motivation that compels users to mark resources through this technique, Marlow et al. (2008) highlight the following:

1. Future retrieval: users will mark resources to remember pending tasks (e.g. "toRead") or define clusters of objects that will be used later, for example, marking web resources in order to write a research paper with the tag "research_paper_1".
2. Contribution and sharing: to create clusters of resources for oneself and other users, whether they are known or not. An example of this would be marking photos of a group trip with the tag "trip_Rome_2010" so that all the members of that group can see those photos.
3. Attract attention: by using commonly used tags, as those shown by clouds of tags, the rest of users can be attracted to the resources.

Depending on their meaning, tags elaborated by users can be put into three categories determining the tag function. These categories are: content tags, which describe the content; attitude tags, which enable opinion expression; or self-reference tags, which are self-reminders (Van Setten, 2007).

Regardless of the type of tag that is being used, marking resources that are interesting for whatever reason reveals the users' interests in a specific and explicit way (Li et al., 2008). In other words, the tags posted by a user will be relevant not only to the content of the bookmark but may also be specific to that user (Zhang et al., 2009). Essentially, a single resource can be marked by different users with different tags, which will represent a varied set of topics of interest.

Content

The content tag, as it has already been said, consists of a term or a set of terms freely chosen by the user.

In this regard, two types of tags can be found (Farooq et al. 2007, Liu et al., 2008):

1. Explicit or obvious tags, which can be found within the text content of the marked resource.
These types of tags, as this paper tries to show, are used very frequently by users.
2. Implicit or non-obvious tags, which cannot be found within the text content. According to Farooq et al. (2007), these types of tags have a higher intellectual value because they provide insights into the content of the paper.

Various reasons may impel users to use explicit tags. According to Lipczak & Millos (2010), users want to minimize efforts and tend to use tags that are easily available. Farooq et al. (2007) point out that the explicit tag can be just a good descriptor in spite of the fact that it does not add any extra intellectual

The results of this paper show that there are other parts in web documents that also have a great impact on the selection of explicit tags, thereby verifying the results of Eisenberg et al. (2009) and Yimming et al. (2008), which show the high percentage of explicit tags found in the title and the anchor text.

Regarding implicit tags, it is important to point out that they do not always have a higher intellectual value as Farooq et al. (2007) suggest. As has already been stated, tags can be used for different functions, including self-reference and the organizing task. In such cases, the information may be valuable for the users using them, but not necessarily for the rest of users. For example, tagging a resource referring to a book as 'owned' means that that title can be found in the user's personal library, which does not add any extra value and it is, in fact, a handicap for those users looking for books that cannot be found in their libraries (Fu et al., 2010). Other examples would be tags like 'must,' 'toRead,' or 'pendent.'

Disadvantages of tagging

As it has already been stated, one of the advantages to tagging is the possibility to create tags by combining all types of characters and signs, thereby forming a kind of open vocabulary. Other terms can also be added, which describe specific content even though it is only personally relevant for an individual user. However, that advantage involves two basic problems with regard to social tagging namely, informational redundancy (Robu et al., 2009) and the loss of general significance.

The informational redundancy problem refers to the creation of many different tags that describe the same resource, so that different users use synonyms, homonyms, and polysemes (Furnas et al., 1987).

According to Fu et al. (2010), the increasing number of vocabularies will cause the connections between tags and documents to become less direct and more confusing, making Information retrieval more

This paper is based on the data obtained from the analysis of four social bookmarking systems (SBS). In order to select them (see Table 1), some of the best-known SBSs were analyzed. Those that did not meet the following standards were dismissed: the marked resources must be a website with text and not other types of files or documents (pdf, doc, etc.); they must be marked with tags, and they must enable access to the web resource. Thus, those resources requiring a subscription or a registration were rejected, as well as those not using tags or those using fragments of texts like comments or descriptions as resource metadata. Furthermore, backFlip was also rejected because it was out of order and Gnolia, because it offered very few links due to its closure on November 30, 2010.

Insert Table 1 here

After this analysis, the four resources that better fit our needs were selected - Delicious, Diigo, Mister Wong, and Connotea (see Table 2). The four of them use tags to mark resources, they are free and enable direct access to the marked resource, and do not require registration to be able to consult available resources. The first three (Delicious, Diigo, and Mister-Wong) are general social bookmarking systems, which means that they are not specialized in specific types of content. As for Connotea, it defines itself as a 'free online reference management for all researches, clinicians and scientists,' which is why it deals with scientific content.

About the feature of suggesting tags to the users which bookmark resources, Connotea doesn't suggest any, Delicious and Mister Wong suggest tags previously used by other users to bookmark the same resource; and Diigo and again Delicious suggest the last tags used by the user who bookmarks the resource. The nature of these tags, whether they are implicit or explicit, isn't taken into account when the

Each of the stored resources in each SBS was subjected to a little analysis in order to check whether the resource was active or not, was a website, was another type of web resource (image, text document, spreadsheet, etc.), or had text content (it can be a website made with flash, in which case the language used to write the site is also relevant).

In order to identify the language of the resource, NGramJ was applied (<http://ngramj.sourceforge.net/index.html>). This is a Java based library containing two types of NGram based applications, where ngrams are classical instruments in Natural Language Processing (NLP) applications. Its main function is language guessing or language recognition, providing a language identifier (es-spanish, en-english, de-denmark, etc.) starting from a piece of text.

Finally, each of the resources was checked to determine if it was marked with any tag. In this case, apart from storing tags, the text of the web resource was extracted and the quantity of explicit and implicit tags was calculated. In order to consider a tag as explicit, there must be at least one exact overlap within the text of the resource. In the case where explicit tags did appear, an accurate analysis was carried out to determine in which HTML tags the explicit tags were found and how frequently they occurred.

In order to manage web resources, Jericho HTML Parser was applied. This is a Java library, which allows analysis and manipulation of parts of an HTML document, including server-side tags, while reproducing verbatim any unrecognized or invalid HTML (<http://jericho.htmlparser.net/docs/index.html>). However, this library did not avoid those problems arising as a consequence of working with Cyrillic-like alphabets. In some of these cases, characters were written as HTML entities. For example, the character "П" is represented in the source code in its hexadecimal HTML representation: "П". On such occasions,

have appeared as well as the number of explicit tags found within the HTML tags in each corresponding resource.

Links were collected in working days, from September 1, 2010 to October 15, 2010, each crawler running individually every day. 151,699 urls were collected and analyzed through the statistics program SPSS starting with the data stored in a MySQL 5.1.37 database.

Results analysis

The results obtained are described hereafter from a double point of view. Firstly, all of the related data will be analyzed to achieve a general view of the social bookmarking systems. Secondly, the collected data will be filtered in order to analyze the features and structure of explicit tags properly.

Data about SBS

The collected data can be divided into two groups, webs or resources and tags, as seen in Table 3.

Insert Table 3 here

It is important to point out that Connotea as well as Mister Wong do not have non-marked resources. It is due to the fact that, in both cases, the user is required to introduce at least one tag in order to be able to mark a resource. Also, Connotea has fewer resources because of time delays in the process of connection to different pages of the website.

Languages of the resources

Finally in 7.74% of the resources, the language has not been identified properly due to the lack of text in the resource itself or the impossibility of entering the page because it was not possible to connect to the server or because an error 404 message bounced back showing that the requested page was not available.

Insert Table 4 here

Number of tags per resource

From among all the chosen resources, 90.79% (137,739) were marked with tags. The distribution of these tags is described in Table 5 below, where it can be observed that 94% of the urls are marked with 10 or fewer tags.

Insert Table 5 here

Generally speaking, social bookmarking system resources are marked with a mean of 4.16 tags, with a mode of 1, a median of 4, and a standard deviation of 3.34. Only the 0.73% of the resources is marked by more than 14 tags.

Depending on the different social bookmarking systems, the number of tags used per resource changes, but not significantly (see Table 6 and Graphic 1): the mode changes in Diigo and Mister Wong and the average frequency of use of each tag per resource is 4 ± 1 . In contrast, Connotea has a significantly greater number of maximum tags used compared to Diigo, Delicious, and Mister Wong with one resource marked with 157 tags. This SBS has 0.75% of its resources (32) marked with more than 39 tags. which is the highest value in Diigo.

Some specific features of the collected tags are going to be described below (how long they are, how many unique tags exist, and which are mostly used). Then those features can be compared with the same features in explicit and implicit tags, which will permit easier differentiation.

In the first place, the total number of tags (573,219) has an average length of 8.53 characters with a standard deviation of 5.73. The mode value is 4 characters, which means that most of the tags are that long. On the other hand, finding tags with many characters are not strange. This is due to the fact that users do not always introduce individual terms, but instead introduce a set of linked words or words separated by different punctuation marks like “-“, “,“, or “#”. A few examples include “registrationsingapore,” “link-building-service,” or “ufc-120-live-stream-fee-online.” In other cases, bookmarking systems allow the addition of tags consisting of various terms, such as: “bisping vs akiyama live stream” or “selling antique rings.”

From among all these tags, 110,617 unique tags are obtained, from which 68% are used just once, 11.9% twice, and 5.3% three times. On the whole, 90% of tags are used five times or less.

On the other hand, the most commonly used tags reveal which topics are typically discussed in the SBSs and allows the analysis of terms frequently used as tags. Table 7 shows the ten explicit and implicit tags most commonly used and also that most of them deal with topics related to Internet (e.g. blog, technology, computers, online, software, etc.).

Insert Table 7

Regarding the number of resources per SBS, it depends on the response time of the different SBSs. Since crawlers ran at the same time through each SBS, if the response time proved to be short, more resources could be processed.

Altogether a total number of 524,930 tags associated to those urls were collected, from which 45.10% (236,782) are implicit tags. As it has already been stated, for a tag to be considered explicit there must be at least one overlap within the text of the resource. Through the crawlers, this condition was verified.

The selection of these tags enabled to considerate a total of 91,652 resources, which are those that were marked with at least one explicit tag. These resources are going to be used as a basis for the analysis of this type of tag.

The percentage of explicit and implicit tags that arise in the analysis of the resource is shown in Table 9. Diigo is the SBS where there are fewer explicit tags (41%) compared to Mister-Wong which has 67% of the explicit tags.

Insert Table 9 here.

Length

The average length of the tags was previously calculated. In general, there are 8.53 characters per tag.

The length obtained according to the type of tag is different from the general mean (Graphic 2). In other words, while implicit tags have a mean of 10.23 characters and a mode of 8, explicit tags have a mean of 6.84 characters and a mode of 4.

Focusing on this distribution, it can be found that 31.22% (39,618) of the resources of the SBSs are marked only with explicit tags, 27.77% (35,239) are marked only with implicit tags, and the remaining 41.10% (52,034) are marked with implicit and explicit tags. Within this 41.10% of the resources that have both types of tags, explicit tags represent 49.10% and implicit tags 50.90%.

There is a mean of 5.6 tags per resource with half of the mean being explicit tags and the other half being implicit tags.

Number of times that tags are used to mark different resources

It was proved that whether implicit or explicit, most tags are used only once.

Thus, in explicit tags (Table 10), 85% of them are used 5 times or less and the same can be said of implicit tags (Table 11).

Insert Table 10 here.

Insert Table 11 here.

Explicit and implicit tags mostly used

In Tables 12 and 13 below, it is shown which of the 110,617 unique tags available are most frequently used, making a distinction between explicit and implicit tags.

Insert Table 12 here.

Insert Table 13 here.

values tend to become equal, for example the 10th explicit tag (“web”) is used 1,120 times while the 10th implicit tag (“uploaded”) is used 861 times. This means that as far as implicit tags are concerned, there are some of them that are frequently used and others that are less frequently used, while the use of explicit tags is more consistent.

Frequency of appearance of explicit tags within the text of a marked resource

About explicit tags, it is also important to know how many times these tags appear in the resource. This data is provided in Table 14.

Insert Table 14 here.

According to this data, explicit tags normally appear only once (12.4%) or twice (11.7%) in the text. The frequency of explicit tag appearance in the text decreases gradually. It is important to note that while 24.1% represent the tags appearing once or twice, the quantity of tags appearing more than 15 times is 26.1% of the total.

Relationship between the frequency of appearance and the length of explicit tags

As Lipczak & Millos (2010) say, users want to minimize their efforts and tend to use tags more easily available. Because this assertion it could be stated that in the decision-making process the length of the potential tags and their frequency of appearance are taken into account. That is to say, a relationship between the frequency of appearance of explicit tags and their length exists, whereby the shorter the tag and the higher its frequency of appearance, the easier it will be for the user to choose it as a tag.

Explicit tags appear most often within the HTML tags link and title as other studies show (Eisterberg et al., 2009, Yimming et al., 2008). Analysis showed that after the HTML tags link and title; p, div, and span are the next HTML tags where explicit tags are most frequently found. P tag is used to include text in paragraphs, div tag enables the creation of layers to put inside whatever is wanted (e.g. images, text...), and span tag makes the introduction of text fragments possible.

A total number of 208 HTML tags containing explicit tags have been identified. Among them, there are obsolete tags (e.g. center, I, font) and HTML tags that do not meet the W3C standard (e.g. figcaption, title1, article_body). Table 15 shows a summary of the HTML tags containing the 90.21% of the sample.

Insert Table 15 here.

Discussion and Conclusion

From the results, it can be inferred that explicit tags (54.9%) are used just as frequently as implicit tags (45.1%). This fact suggests that the data users obtain from the resource is enough for them to mark it, describe it, or classify it, or at least as useful as the data not contained inside the resource.

Explicit tags are shorter (a mean of 7 characters) than implicit tags (a mean of 10 characters) and appear in the text 1-15 times in 74% of the cases.

According to this data, the relationship between the frequency of appearance of explicit tags and their length was studied. Due to the fact that users want to minimize their efforts and tend to use tags more

With regard to HTML tags where explicit tags appear, even though “title” and “a” labels have more explicit tags (34.8%), the most important tags are not HTML that somehow highlight the text, but instead are content-tags such as “p”, “div,” and “span,” which represent 30.15% of the remaining HTML tags. This means that when choosing explicit tags, users do not take into account the physical size of the text (such as headlines “H” or those texts highlighted as “strong”) as a reference, but rather they freely choose among the text available. These results can be very useful in tag suggestion systems based on resource content: using only the content inside the most commonly HTML tags where explicit tags appear, can suppose an improvement reducing workload and execution time because less content has to be analyzed.

As for the state of SBSs, it must be pointed out that the 9.2% are non-marked resources and 7% of the resources are offline. Due to the fact that in these systems the pivot browsing is usually performed through the tags, when these are not available in a resource, that resource will rarely be visited because of its little to no visibility. On the other hand, the percentage of offline resources shows that these types of systems need to apply mechanisms that are able to keep them updated. In this case, it is not about removing links to the resources, since they belong to the users, but instead about warning them that they own a link repository containing links to unavailable resources, which are useless.

Several SBSs have been analyzed: with regard to the percentage of use of explicit and implicit tags, Connotea and Delicious have a percentage of use close to 50%, Diigo uses implicit tags a little more (59%), and Mister Wong uses explicit tags even more frequently (67%). Generally speaking, users do not use explicit tags more frequently than implicit tags.

However, there exist several limitations due to the selection of the sample and the content analyzed.

In conclusion, although the use of explicit tags has been usually less valued than the use of implicit tags because their lack of additional intellectual power (Farooq et al., 2007), the results of this paper support the idea that explicit tags are as useful and used as implicit tags. Therefore, the use of explicit tags is a valid and an important tool for tagging web resources.

References

- Bateman, S., Muller, M. J., & Freyne, J. (2009). Personalized retrieval in social bookmarking. Proceedings of the ACM 2009 International Conference on Supporting Group Work, pp. 91-94, Sanibel Island, Florida, USA. ACM.
- Bar-Ilan, J., Zhitomirsky-Geffet, M., Miller, Y. and Shoham, S. (2010), The effects of background information and social interaction on image tagging. Journal of the American Society for Information Science and Technology, 61: 940–951.
- Delicious' Blog. *What's next for Delicious*. Available on line: <http://blog.delicious.com/blog/2010/12/whats-next-for-delicious.html>. Retrieved on 16-3-2011.
- Ding, Y., Jacob, E. K., Zhang, Z., Foo, S., Yan, E., George, N. L. and Guo, L. (2009), Perspectives on social tagging. Journal of the American Society for Information Science and Technology, 60: 2388–2401. doi: 10.1002/asi.21190
- F. Eisterlehner, A. Hotho, and R. Jäschke, editors. *ECML PKDD Discovery Challenge 2009 (DC09)*, volume 497 of *CEUR-WS.org*, Sept. 2009.
- Farooq, H., Kannampallil, T. G., Song, Y., Ganes, C. H., Carroll, J. M. and Giles, J. (2007).

- Golder, S. A. and Huberman, B. A. (2005) The Structure of Collaborative Tagging Systems. HP Labs technical report, 2005. Available from <http://www.hpl.hp.com/research/idl/papers/tags>
- Illig, J., Hotho, A., Jäschke, R., and Stumme, G. (2009). A Comparison of content-based Tag Recommendations in Folksonomy Systems. In *Postproceedings of the International Conference on Knowledge Processing in Practice (KPP 2007)*.
- Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G. (2007). Tag recommendations in folksonomies. In *Knowledge Discovery in Databases: PKDD 2007*, pages 506-514. Springer-Verlag, Berlin, Heidelberg.
- Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G. (2008). Tag recommendations in social bookmarking systems. *AI Communications*, 21(4):231-247.
- Körner, C., Benz, D., Hotho, A., Strohmaier, M., and Stumme, G. (2010). Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 521-530, New York, NY, USA. ACM.
- Koutrika, G., Effendi, F.A., Gyöngyi, Z., Heymann, P. & Molina H.G. (2008) Combating spam in tagging systems: An evaluation. *ACM Trans. Web*, Vol. 2, No. 4., pp. 1-34.
- Lipczak, M. and Milios, E. (2010). The impact of resource title on tags in collaborative tagging systems. In *HT '10: Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 179-188, New York, NY, USA. ACM.
- Liu, Y., Kumar, R., & Lim, K. (2008). Taggers versus Linkers: Comparing Tags and Anchor Text of Web Pages. UC Berkeley: School of Information. Report 2008-020. Retrieved from: <http://escholarship.org/uc/item/8b40q59k>

- Mathes, A. (2004) Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Retrieved from: <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>
- Melenhorst, M., Van Setten, M. (2007). Usefulness of tags in providing access to large information systems. In: Proceedings of the IEEE Professional Communication Conference.
- Millen, D., Feinberg, J. and Kerr, B. (2005) Social Bookmarking in the enterprise. *Queue*, 3(9):28-35
- Millen, D., Yang, M., Whittaker, S., and Feinberg, J. (2007). Social bookmarking and exploratory search. *ECSCW 2007* (2007), pp. 21-40.
- Oliveira, B., Calado, P., Pinto, H.S. (2008) Automatic Tag Suggestion Based on Resource Contents. *EKAW '08: Proceedings of the 16th international conference on Knowledge Engineering*, 255-264. Springer-Verlag, Berlin / Heidelberg
- Robu, V., Halpin, H., and Shepherd, H. (2009). Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Trans. Web*, Vol. 3, No. 4., pp. 1-34.
- Schmitz, C., Hotho, A., Jäschke, R., Stumme, G. (2006) Mining association rules in folksonomies. *Data Science and Classification: proceedings of the 10th IFCS Conference, Studies in classification, Data Analysis and Knowledge Organization*
- Smith, G. (2004) Atomiq: Folksonomy: social classification. Aug 3, 2004. Retrieved from: http://atomiq.org/archives/2004/08/folksonomy_social_classification.html
- Subramanya, S. B. and Liu, H. (2008). Socialtagger - collaborative tagging for blogs in the long tail. In *Proceeding of the 2008 ACM Workshop on Search in Social Media* (Napa Valley, California, USA, October 30 - 30, 2008). *SSM '08*. ACM, New York, NY, 19-26.