

PREPRINT

Visualizing Social Indexing Semantics

Yusef Hassan-Montero¹ and Víctor Herrero-Solana²

March 2007

¹ SCImago Research Group. yusefhassan@gmail.com

² SCImago Research Group. victorhs@ugr.es

Abstract

Social tagging is a distributed indexing process where web resources are described by means of tags – freely chosen keywords or labels – by their users. The aggregated result of social tagging is usually known as Folksonomy, an index where each resource is related to different tags by different users. This paper describes an approach to visualize both the overview and detail of semantic relationships intrinsic in the folksonomy. For this purpose, folksonomy is displayed as a network: tags are represented as nodes and semantic relationships as links between nodes. The semantic relationships are measured using the Jaccard coefficient, which determines the thickness of the link between tags. The bisecting K-mean algorithm is applied to cluster the tags, and the tags in the same cluster are displayed with the same colour. In order to ensure that the overview of social tagging is comprehensible, the Pathfinder Network Scaling algorithm is used to prune the network and only show the most significant links among tags. In addition, to get a better understanding of the local semantic relationships of a tag, an interactive procedure was applied to display some important links pruned by Pathfinder when the user moves mouse over a tag. The presented approach provides a means of knowledge acquisition and understanding of the socially constructed meaning of tags, as well as a means of visual information retrieval.

Keywords

Social Indexing, Pathfinder Networks, Clustering, Visual Information Retrieval Interfaces

1. Background

With the arrival of social bookmarking tools, such as del.icio.us or connotea [1], a new approach for metadata creation has emerged. This is widely known as Tagging. Tagging-based systems enable users to add tags – freely chosen keywords – to web resources to categorize these resources in order to retrieve them later.

Tagging is not only an individual process of categorization; it is also a social indexing process and, therefore, a social process for knowledge construction. Users share their resources through the tags, generating an aggregated tag-index known as folksonomy, a term coined by Thomas Vander Wal on the AIFIA mailing list. This one-word neologism comes from the words ‘taxonomy’ and ‘folk’ [2]. There is some discussion about its accuracy, yet we prefer to use it in light of its popularity and wide acceptance.

Folksonomy enables anyone to access to any web resource that was previously tagged, on the basis of two main models for information access: Information Filtering (IF) and Information Retrieval (IR).

In IF, the user plays a passive role, expecting the system to push or send along information of interest according to some previously defined profile. Social bookmarking tools allow for a simple IF access model. The user can subscribe to a set of specific tags via RSS/Atom syndication, and thus be alerted when a new resource is indexed with any tag pertaining to this set.

Meanwhile, in IR the user seeks information actively, pulling at it, by means of querying or hypertext browsing. In tag querying, the user introduces one or more tags in the search box to obtain an ordered list of resources related to these tags. As the user scans this list, the system also provides him with a list of related tags (i.e. tags having a high degree of co-occurrence with the introduced tags), thereby allowing hypertext browsing.

Because anyone who so desires can contribute to the social indexing process, tags may suffer from ambiguity and arbitrariness [3]. Polysemy, synonymy and the ego-oriented nature of many tags (e.g. ‘todo’, ‘toread’, ‘me’), are well known problems of tagging in the context of IR. Furthermore, a large volume of tags may identify the resource’s author (e.g. ‘nielsen’), a subjective tagger’s opinion (e.g. ‘cool’), or some type of resource (e.g. ‘book’, ‘blog’) [4, 5, 6]. In other words, tags are not exclusively subject index terms.

Guy and Tonkin [7] discuss how to improve tags by means of tagging literacy. They conclude with a thought-provoking statement: “Possibly the real problem with folksonomies is not their chaotic tags but that they are trying to serve two masters at once; the personal collection, and the collective collection. Is it possible to have the best of both worlds?”

Wu, Zhang and Yu [8] demonstrate that is possible to extract tags with collective usefulness from the sum of individual and freely assigned tags, automatically solving tag ambiguity problems. From our point of view, tagging systems entail important and

exclusive strengths – described below – some of which stem precisely from the uncontrolled and free nature of tagging.

Because indexing is performed to facilitate information retrieval by users, it is necessary for index terms to reflect the requests for which a document might be relevant. In short, indexing must be user-centered [9]. Folksonomies directly reflect the vocabulary of users, conjugating users' real needs and language [2, 10]. The best way to guarantee user-centered indexing is through user-generated indexing.

In addition, given that folksonomy are rooted in collective agreement, these tags have a more accurate, truthful and democratic meaning than the ones assigned by a single person (author, professional indexer or librarian). As expressed by Wu, Zhang and Yu [8], users negotiate the meaning of tags, in an implicit asymmetric communication. If we go back to the “use” theory of meaning proposed by philosopher Ludwig Wittgenstein, the meaning of a word resides in its usage. Thus, meaning is a social event taking place among language users.

One intrinsic problem of the human indexing process is the so-called “inter-indexer inconsistency”. This happens when indexers use different index terms to describe the same document [11], a problem that is reduced when indexing is done by aggregation, as in social indexing. Golder and Huberman [4] report that when 100 or so users index the same resource, each tag's frequency becomes a nearly fixed proportion of the total frequency of all tags assigned to that resource. Given that tag proportions tend to stabilize, this means that a unique tag weight could be assigned to each resource.

Furthermore, folksonomies allow for information discovery through serendipity [2, 10] and, in most cases where tagging is used, alternative solutions such as controlled vocabularies are impossible to apply. In words of Quintarelli [2], “folksonomies are better than nothing”.

Despite the potential of tagging systems for IR, there is a lack of research focusing on the effectiveness and usefulness of folksonomies. Tagging effectiveness can be measured by means of two inter-related parameters: term/tag specificity and indexing/tagging exhaustivity. These two variables show the number of resources described by one tag, and the number of tags assigned to one resource, respectively [12]. With a broad tag, the user will retrieve many relevant items, along with a considerable number of irrelevant ones. Narrow tags retrieve few and mostly relevant resources, but may neglect some relevant resources. In traditional IR terms, broad tags entail high recall and low precision, whereas narrow tags entail low recall and high precision.

Xu et al. [5] relate broader tags to discovery tasks (i.e. find new resources), and narrow tags to recovery tasks (i.e. find resources previously discovered). Likewise, it is reasonable to connect broader tags to general-purpose searching tasks, and narrow tags to more specific and goal-oriented searching tasks – browsing and querying, respectively. The main question, then, is: do folksonomy tags tend to be broader or narrower? And thus, are folksonomies more suitable as an aid to browsing or to querying?

The relationship between the number of items and the volume of tags applied to describe them by an user are not well correlated – i.e. some users resort to large tag sets, whereas others use small ones [4]. However, Brooks and Montanez [6] reveal that taggers tend to prefer broader tags. In addition, tagging is not an exhaustive indexing process, since in 90% of cases the average user employs less than five tags per item [13]. These findings are not surprising if we bear in mind that the low cognitive effort involved in tagging is one of the main reasons for its popularity [14]. The assignment of a few broad or general tags would demand less cognitive effort than the designation of narrow and specific ones. Hence, the generic nature of tags allows them to better support browsing than querying.

2. Visual Browsing: Related work

In order to facilitate visual browsing, social bookmarking tools usually provide an interface model known as Tag-Cloud, a list of the most popular tags, usually displayed in alphabetical order, and visually weighted by font size. In a Tag-Cloud, when a user clicks on a tag, he gets an ordered list of items described by that tag, as well as a list of related tags. Whereas querying requires the previous formulation of a user's information needs, visual browsing lets the user recognize his information needs by scanning the interface directly, without previous action. Visual browsing is similar to hypertext browsing in that both allow to the user to search by browsing, yet there is one important difference: visual interfaces provide a global view of the entire tag collection.

The tag-Cloud is a simple and widely used visual interface model, but there are some restrictions regarding its utility as a visualization tool, due to:

- 1) The method used to select the core set to be displayed is based solely on term frequency. This inevitably means that the displayed tags wield a high semantic density. In terms of the discrimination value, the most frequently used terms are the worst discriminators [15]. As Begelman, Keller and Smadja [16] point out, very few different topics tend to reign over the whole cloud.

- 2) The alphabetical ordering of displayed tags neither facilitates visual scanning nor enables one to infer semantic relations among tags.

In a previous paper we addressed the issue of semantic density, proposing an alternative method to Tag-Cloud tag selection: a method based on tag weighting. We also suggested the use of clustering techniques to improve the Tag-Cloud visual layout and increase the browsing success. The results show that the improved Tag-Cloud (Figure 1) has less semantic density than traditional Tag-Clouds. Moreover, the clustered layout offers a more logical visual distribution of tags and allows one to differentiate between main topics and specific ones.



Figure 1: Improved Tag-Cloud [17]

The improved Tag-Cloud gives rise to the discovery of semantic relations from adjacent or neighbouring relationships among tags and among clusters. However, it is impossible to see the other related tags. One feasible solution would be to emphasize related tags with a mouse-over action, as shown in the ‘Revealicious Tag-Cloud’ [18]. Yet with this technique all relationships are hidden until the user interacts, and it does not indicate the degree or intensity of the relation.

The most suitable visual metaphor for representing complex associative structures is that achieved through network displays, where entities are presented as nodes and their relationships as links. Some previous studies have used network displays to represent semantic structures from folksonomies. A tag browser proposed by Shaw [19] involved the application of Semidefinite Embedding (SDE) to 200 tags used by 1,494 users; but the problem with this interface model is that it shows more links than can be comprehended. It is therefore difficult to recognize a clear linking structure.

Here we propose an alternative interface model, which affords visualization of both the overview and the details of semantic relationships inherent in a folksonomy. This interface model is represented as a network of tags connected by the most significant links.

3. Materials and Methods

In order to generate the automatic visualization of folksonomy semantics, we used a novel combination of techniques and algorithms: Pathfinder Network Scaling, along with Clustering and interactive techniques. These techniques are very familiar in the framework of information visualization studies [20, 21]. The process of generation was divided into three main steps: Data acquisition and filtering, data mining, and the visual display and interaction.

3.1. Data acquisition and filtering

Our study was performed on a large sample downloaded from ‘del.icio.us’ web site, on October 2005. This sample contains 218,063 items (URLs) tagged with 242,349 tags by 111,234 users, and was gathered by means of an ad-hoc programmed crawler. The del.icio.us’ tagging system is characterized by allowing free-for-all tagging – i.e. any user can tag any web resource – and by blind tagging – i.e. the user cannot view other

tags assigned to the same web resource while tagging [22]. For these and other reasons [23], del.icio.us is probably is the best source of data for tagging research.

Since it is not possible to put all the tags on the screen, we first selected those tags that best represented the collection as a whole. In order to determine the representation value of each tag, we applied the weighting function 1. After weighting each one of the tags using this function, the 95 most heavily weighted tags were selected to form part of the interface.

$$F(T_j) = \sum_{i=1}^{i=n} \left(\frac{\log(d_{ij})}{m_i^2} \right) \quad (1)$$

If we consider folksonomy as a vector space of items $D_i = (d_{i0}, \dots, d_{in})$, each characterized with one or more tags $T_j = (t_{j0}, \dots, t_{jm})$ weighted according to the number of tag's users, then d_{ij} represents the frequency with which tag T_j has been used to describe resource D_i . Therefore, n is the number of resources that are described by tag T_j and m_i is the number of different tags assigned to resource D_i , whereas $F(T_j)$ denotes the representation value of tag T_j .

Function 1 is a summation of a simplified version of the length-normalized tf-idf function [24], where the denominator gives better weighting to those tags that describe resources that are more poorly covered by other tags, and the square of m_i diminishes the effect of low-discriminating tags. The use of the logarithm reduces the effect of tags with high frequency per resource, but also conditions which tags have no representation value; that is, the tags have never been assigned by more than one user to the same resource (i.e. $F(T_j)=0$). It is remarkable that, in our data sample, 84.7% of tags (205,395) have no representation value, which means that only 15.3% of tags (36,954) are possible candidates to form part of the interface.

	Coverage	Overlapping average	Overlapping standard deviation
a) $F(T_j) = n$	188,761 (86.56%)	0.0503	0.0414
b) $F(T_j) = \sum_{i=1}^{i=n} (d_{ij})$	187,907 (86.17%)	0.0399	0.0425
c) $F(T_j) = \sum_{i=1}^{i=n} \left(\frac{\log(d_{ij})}{m_i} \right)$	191,567 (87.85%)	0.0329	0.0406
d) $F(T_j) = \sum_{i=1}^{i=n} \left(\frac{\log(d_{ij})}{m_i^2} \right)$	190,405 (87.32%)	0.0242	0.0372

Table 1. Selection method comparison over the 95 highest-weighted tags by each function. Coverage is the number of resources that have been described by at least one of the 95 selected tags. Overlapping refers to the relative co-occurrence between selected tags.

As is shown in Table 1, in terms of overlapping between tags, our approach (d) offers better results than traditional selection methods in Tag-Clouds (a, b), and than the same function without the square of m_i (c). This means that the tags selected by function 1 (d)

are less similar among themselves than those selected by the other functions (a,b,c), and thus the tag set has less semantic density.

Differences between functions' coverage are minimal, due to the powerful coverage of the most frequently used tags (such as *web*, *design* and *programming*) which are present within all selections.

3.2. Data mining

In order to visualize the semantic structure of selected tags, it is necessary to calculate an NxN similarity matrix, where N is the number of selected tags (i.e. 95). Similarity between two tags can be calculated in different ways. The easiest method is to count the number of co-occurrences, that is, the number of times that two tags have been assigned to the same resource.

As Cattuto, Loreto and Pietronero [23] show, the non-trivial nature of co-occurrence relationships between tags might be ascribed to semantics. In this paper, tag similarity is considered a kind of semantic relationship between tags, measured by means of the relative co-occurrence between tags, also known as the Jaccard coefficient (2). If A and B are the sets of items described by two tags, relative co-occurrence is defined as:

$$RC(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Relative co-occurrence, then, is equal to the quotient between the number of items in which tags co-occur, and the number of items in which either of two tags appears.

Once we determine the semantic relationships between each pair of tags, we can reveal the main topics of the tag set by means of data clustering. Data clustering comprises a collection of unsupervised learning algorithms that aim to group a set of objects into clusters through an iterative process. The members of a cluster are similar to each other, and dissimilar to members of other clusters.

To the N-dimensional similarity matrix of tags – computed with the relative co-occurrence function (2), where N is the number of different tags (i.e. 95) – we can now apply a clustering algorithm in order to classify tags in clusters. For this purpose we chose a “bisecting” K-means algorithm, a divisive (non-agglomerative) hierarchical clustering algorithm. As Steinbach, Karypis and Kumar [25] show, this technique offers better results than the standard K-means approach, and as good or better results than the hierarchical approaches that they tested for a variety of cluster evaluation metrics. In our experiment, we used the cosine similarity function to calculate co-occurrence patterns between tags, and established the number of clusters to be obtained at 12.

3.3. Visual display and Interaction

Two types of semantic information extracted with the data mining step: the main topic clusters of tags, and the degree of semantic relation between each pair of tags. In addition, we have another useful bit of information: the number of web resources described by each tag. In order to visualize this semantic information, we represent each

tag as a node. The label and ellipse sizes reflect the number of items behind the tag; and tag relationships are shown as links between nodes, the link thickness indicating the weight of the relationship.

In a large folksonomy space like this, all the selected tags are related (to some degree) among themselves. For this reason, it is impossible to visualize all the links, because the network displayed would be incomprehensible. To solve this matter, we applied the Pathfinder algorithm [26, 25], which determines the most significant links in a network. The algorithm eliminates the links that violate “triangle inequality”, a principle that states that the direct distance between two points must be less than or equal to the distance between them when passing through an intermediate point. The Pathfinder algorithm uses two main parameters, r and q , which respectively determine how to calculate distance between nodes, and the maximum number of links in paths for which the triangle inequalities must be satisfied. In our tag network we used values $r=\infty$ and $q=N-1$. The path weight between two points would be equal to the maximum weight of the links along the path, and the entire network must satisfy triangle inequality ($N-1$ is the maximum value of q for a network of N nodes).

Then, Kamada & Kawai’s algorithm [28] was used to situate the nodes in space. This algorithm generates visual topology under aesthetic criteria such as the maximum use of available space, a minimum number of crossed links, the forced separation of nodes, and the construction of balanced maps [29].

Although the network pruned using Pathfinder is clearer and more comprehensible than the original network, the pruning removes links that may be highly informative about the semantic meaning of tags. For this reason, we added an interactive procedure that makes it possible to visualize the most significant links pruned by Pathfinder, with a mouse-over action. We considered significant pruned links to be those whose weight was greater than the 9th decile. Through this procedure we gain visualization of both the overview and the detail of semantic relationships between tags.

The interface prototype was implemented with SVG (Scalable Vector Graphics) [30], and is currently available online at <http://www.nosolousabilidad.com/hassan/visualizious/>

4. Results

The resulting network topology (Figure 2) is a tree that shows a core of large tags around technology (*web, tools, software, programming, design, etc.*). Each branch of the tree mainly contains tags of the same cluster (displayed with the same colour). The right branch contains entertainment tags, such as *games* or *comics*; whereas the left branch contains audio-related tags, and the bottom branch contains audiovisual-related tags. There are two “computer branches”. One begins in *webdesign* and includes markup languages (*css, html, xml* and *rss*), client-side languages (*javascript* and *ajax*), and different models of interaction analysis (*usability* and *accessibility*); and the other begins in *software* and includes server-side programming languages (*java, php, perl* or *python*). The former constitutes the front-end computer branch, and the latter is the back-end computer branch. It is interesting to note that the *ruby* and *rails* tags are connected to the front-end branch, despite being back-end technologies, due to the

interface-oriented nature of the ‘ruby on rails’ framework and its native support for ajax technology.

There other branches related with web contents (i.e. *blog, blogs*; and *search, google*), but they are mixed with certain general subjects (*politics, history; religion, food, etc.*).

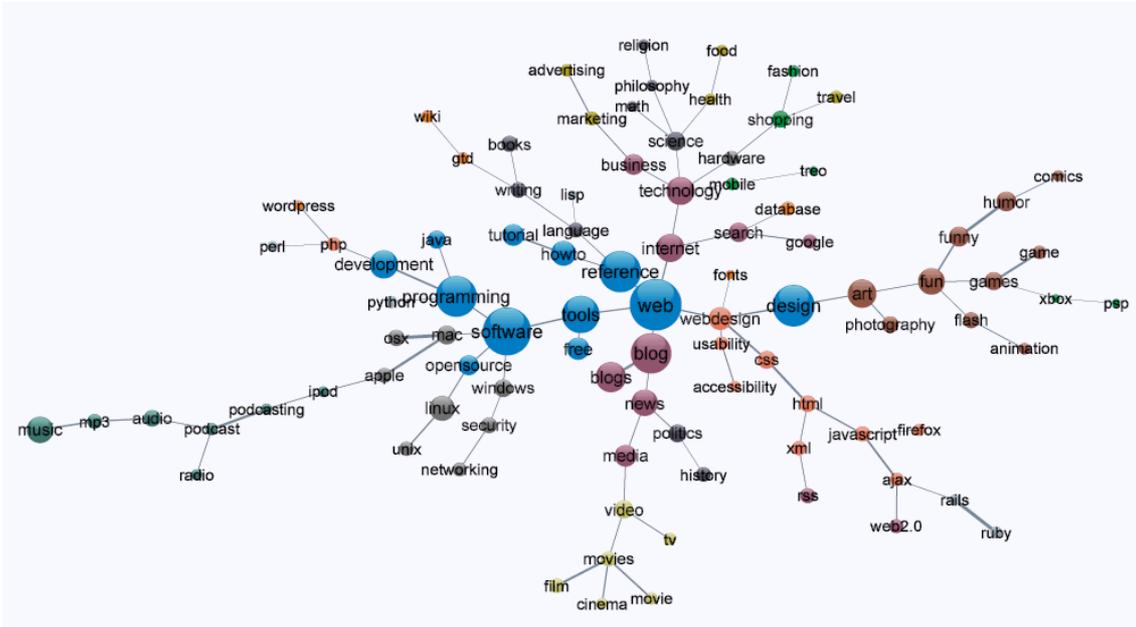


Figure 2: Interface overview

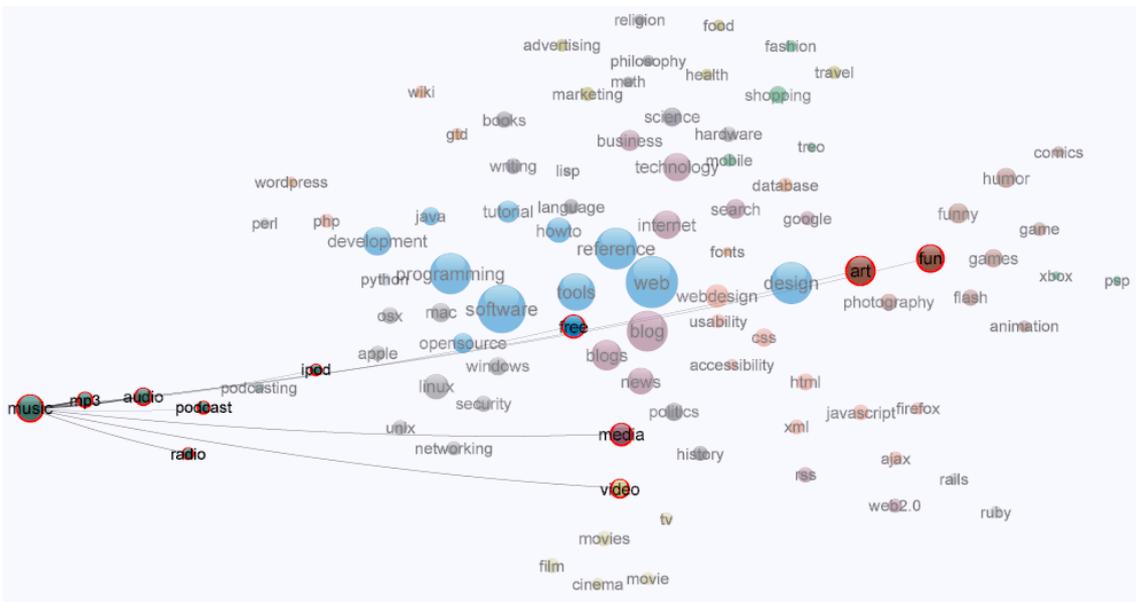


Figure 3: Interface appearance when user moves mouse over *music* tag

The mouseover feature is very interesting, because it shows us the pruned links of each tag. These hidden links sometimes connect tags having the same colour (same cluster) and located along the same branch, but they also may connect tags far away in the network. In Figure 3 we see that the links of the *music* tag are not only connected to the *ipod* neighbourhood, but also to the entertainment and media branches.

6. Glossary (source: wikipedia.org)

AJAX (Asynchronous JavaScript and XML): A web development technique for creating interactive web applications.

Blog: A user-generated website where entries are made in journal style and displayed in a reverse chronological order.

Social Bookmarking: Web tools to store, classify, share and search links or 'bookmarks'.

Podcast: A media file that is distributed over the Internet using syndication feeds, for playback on portable media players and personal computers.

Rails: Ruby on Rails is a web application framework released in 2004 that aims to increase the speed and ease of web development. Often shortened to Rails, or RoR, it is an open source project written in the Ruby language.

RSS: A family of web feed formats used to publish frequently updated digital content, such as blogs, news feeds or podcasts.

Ruby: A reflective, dynamic, object-oriented programming language.

Web2.0: A phrase coined by O'Reilly Media in 2004, refers to a perceived or proposed second generation of Web-based services - such as social networking sites, wikis, communication tools, and folksonomies- that emphasize online collaboration and sharing among users.

Wiki: A website that allows the visitors themselves to easily add, remove, and otherwise edit and change available content, typically without the need for registration.

7. References

1. Hammond T, Hannay, T Lund B, Scott J. Social Bookmarking Tools (I): A General Review. D-Lib Magazine 2005, 11 (4).
2. Quintarelli E. Folksonomies: power to the people. ISKO Italy-UniMIB meeting (Milan, Italy, 2005), ISKO.
3. Fokker J, Pouwelse J, Buntine W. Tag-Based Navigation for Peer-To-Peer Wikipedia. WWW 2006 (Edinburgh, UK, 2006).
4. Golder S, Huberman BA. Usage Patterns of Collaborative Tagging Systems. Journal of Information Science 2006, 32(2): 198-208.
5. Xu Z, Fu Y, Mao J, Su D. Towards the Semantic Web: Collaborative Tag Suggestions. WWW 2006 (Edinburgh, UK, 2006).

6. Brooks CH, Montanez N. Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering. WWW 2006 (Edinburgh, UK, 2006).
7. Guy M., Tonkin E. Folksonomies: Tidying up Tags?. D-Lib Magazine, January 2006, 12 (1).
8. Wu X, Zhang L, Yu Y. Exploring Social Annotations for the Semantic Web. WWW 2006 (Edinburgh, UK, 2006).
9. Fidel R. User-Centered Indexing. Journal of the American Society for Information Science 1994, 45 (8): 572-576.
10. Mathes A. (2004). Folksonomies - Cooperative Classification and Communication through Shared Metadata. [WWW document] <http://www.adammathes.com/academic/computer-mediatedcommunication/folksonomies.html> (accessed 12 January 2007).
11. Olson HA, Wolfram D. Indexing Consistency and its Implications for Information Architecture: A Pilot Study. ASSIS&T Information Architecture Summit 2006, (Vancouver, Canada, 2006).
12. Spärck-Jones K. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 1972, 28 (1): 11-21.
13. Voss J. Collaborative thesaurus tagging the Wikipedia way. [WWW document] <http://arxiv.org/abs/cs/0604036> (accessed 12 January 2007).
14. Sinha R. A cognitive analysis of tagging. [WWW document] http://www.rashmishinha.com/archives/05_09/tagging-cognitive.html (accessed 9 January 2007).
15. Salton G, Wong A, Yang CS. A Vector Space Model for Automatic Indexing. Communications of the ACM 1975, 18 (11): 613-620.
16. Begelman G, Keller P, Smadja F. Automated Tag Clustering: Improving search and exploration in the tag space. WWW 2006 (Edinburgh, UK, 2006).
17. Hassan-Montero Y, Herrero-Solana V. Improving Tag-Clouds as Visual Information Retrieval Interfaces. In: Guerrero-Bote, VP (Ed.). Current Research in Information Sciences and Technologies. Open Institute of Knowledge: Badajoz (Spain); 2006. Vol II, 422-427.
18. Ivy. Revealicious TagsCloud [WWW document] <http://www.ivy.fr/revealicious/demo/tagcloud.html> (accessed 14 January 2007).
19. Shaw B. Utilizing Folksonomy: Similarity Metadata from the Del.icio.us System. Project Proposal. [WWW document] <http://www.metablake.com/webfolk/web-project.pdf> (accessed 10 December 2006)

20. Börner K, Chen C, Boyak KW. Visualizing Knowledge Domains. *Annual Review of Information Science and Technology* 2004, 37: 179-255.
21. Herrero-Solana V, Hassan-Montero Y. Metodologías para el desarrollo de interfaces visuales de recuperación de información: análisis y comparación. *Information Research* 2006, 11(3).
22. Marlow C, Naaman M, Boyd D, Davis M. Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. WWW 2006 (Edinburgh, UK, 2006).
23. Cattuto C, Loreto V, Pietronero L. Collaborative Tagging and Semiotic Dynamics. [WWW document] http://arxiv.org/PS_cache/cs/pdf/0605/0605015.pdf (accessed 15 January 2007).
24. Salton G, Allan J., Buckley C. (1994). Automatic Structuring and Retrieval of Large Text Files. *Communications of the ACM* 1994, 37 (2): 97-108.
25. Steinbach M, Karypis G, Kumar V. (2000). A Comparison of Document Clustering Techniques. Technical Report 00-034. [WWW document] <http://glaros.dtc.umn.edu/gkhome/node/157> (accessed 15 January 2007).
26. Schvaneveldt R. (Ed.) *Pathfinder associative networks: studies in knowledge organization*. Norwood, NJ: Ablex; 1990.
27. Guerrero-Bote VP et al. Binary Pathfinder: An improvement to the Pathfinder algorithm. *Information Processing & Management* 2006, 42: 1484-1490.
28. Kamada T, Kawai S. An algorithm for drawing general undirected graphs. *Information Processing Letters* 1989, 31(1), 7-15.
29. Moya-Anegón F et al. A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics* 2004, 61 (1):129-145.
30. W3C. Scalable Vector Graphics (SVG) [WWW document] <http://www.w3.org/Graphics/SVG/> (accessed 14 January 2007).
31. Granovetter MS. The Strength of Weak Ties. *American Journal of Sociology* 1973, 78 (6): 1360-80.
32. Granovetter MS. The Strength of the Weak Tie: Revisited. *Sociological Theory* 1983, 1: 201-33.