Enhancing the Quality of Place Resources in Geo-folksonomies

Ehab ElGindy and Alia Abdelmoty

School of Computer Science and Informatics
Cardiff University, Wales, UK
{ehab.elgindy,a.i.abdelmoty}@cs.cardiff.ac.uk

Abstract. Users' interaction and collaboration on Web 2.0 via social bookmarking applications have resulted in creating a new structure of user-generated data, denoted folksonomies, where users, Web resources and tags generated by users are linked together. Some of those applications focus on geographic maps. They allow users to create and annotate geographic places and as such generate geo-folksonomies with geographically referenced resources. Geo-folksonomies suffer from redundancy problem, where users create and tag multiple place resources that reference the same geographic place on the ground. These multiple disjointed references result in fragmented tag collections and limited opportunities for effective analysis and integration of data sets. This paper, (1) defines the quality problem of resources in a geo-folksonomy (2) describes methods for identifying and merging redundant place resources and hence reducing the uncertainty in a geo-folksonomy, and (3) describes the evaluation of the methods proposed on a realistic sample data set. The evaluation results demonstrate the potential value of the approach.

Keywords: Web 2.0, Folksonomy, Geographical Similarity, Social Bookmarking, Tagging, Geo-Tagging.

1 Introduction

Web 2.0 has created a new type of Web-based interaction among Internet users by introducing social bookmarking applications, where users can publish contents to share it with others. The published contents are Web documents such as Web pages, images or PDF documents. In addition, users can provide keywords (tags) to categorize the contents/resources they publish, thus resulting in new structures of information – called folksonomies – that links users, tags and resources together.

Folksonomies directly reflects the vocabulary of users [12], enabling matching of users' real needs and language. Although folksonomies are semantically rich, they are un-controlled, unstructured, sometimes ambiguous and inconsistent. Ongoing research efforts consider the extraction of certain semantics from

L. Wang et al. (Eds.): WAIM 2011 Workshops, LNCS 7142, pp. 1–12, 2012.

folksonomies. For example, Rattenbury et al. [14] extract place and event semantics from Flickr[1] tags using the usage distribution of each tag.

Some social bookmarking applications, such as Tagzania[2], are specialized in tagging geographic places using a map-based Web interface. These applications generate a special kind of folksonomy, denoted geo-folksonomy in this work. The tagging behaviour - which generates folksonomies - allows users to choose keywords to describe/index the information in a specific web resource such as a web page. For example, a user can tag an article he read about a good cooking recipe as ("best", "recipe", "for", "making", "fajita"). However, when it comes to tag a place, users create a place resource using a map-based interface which represents a place in reality, and then tags are provided to describe the place in reality although they are attached to the place resource. For example, a user can create a place resource named "Cardiff university" and set its spatial location using a map interface such as Google maps, and then the user can attach relevant tags such as ("University", "Study", "Research").

Place resources in geo-folksonomies have some characteristics which do not exist in normal web resources:

1. Place resources are created by the social bookmarking applications to reference places in the real world, while normal web resources already exist in the web space and they are just referenced using unique URLs.
2. Although it is possible to assign a unique URI for any resource (including place resources [2]), URIs are not used to locate places as people always refer to places by spatial and thematic attributes such as location and place name respectively.
3. The values of spatial attributes - such as longitude and latitude - are acquired using a map-based applet. This method of acquiring data can be imprecise and is dependent on the user being able to identify and digitize a precise location on a map offered on the user interface of these applications. The accuracy is also related to the map scales offered to users and the difficulty in matching the precise location across map scales.
4. The values of thematic attributes - such as place names - are acquired using a free-text input. Although they add valuable semantics to the place resources, they are associated complexity, where people use non-standard, vernacular, place names [5] and abbreviations.

Most of the applications that generate geo-folksonomies aim to collect as much information as possible about places, which can be one of the reasons why such applications do not allow users to share place resources and why they require a new place resource to be created each time a user wants to tag a place. Such design can result in having multiple place resources that reference the same place in real world. We argue that, such redundancy in the geo-folksonomy structure can produce inaccurate results when using folksonomy analysis techniques such as tag-similarity methods.

[1] `http://www.flickr.com`
[2] `http://www.tagzania.com`

The combination of inaccuracies in place location and fuzziness in naming place entities complicates the task of uniquely identifying place resources and can hence lead to the presence of redundant resources in the geo-folksonomy, degrading its quality. Hence, identifying and relating those place resources can lead to more consistent and useful analysis of geo-folksonomies and support integrating place resources from different data sources.

The work presented in this paper defines and formulates a quality problem in geo-folksonomy resources. Methods are proposed for addressing this problem and for creating an enriched geo-folksonomy. The solution involves using online Web resources to first qualify place instances with identifiers that can then be used in a process of clustering and aggregation to uniquely identify related resources.

The enriched geo-folksonomy contains more certain information, as the method takes into consideration the user votes and agreements.

The rest of this paper is organized as follows. Related work is discussed in section 2. The research problem is described in section 3, followed by the proposed methods in section 4. Experimental results and evaluation are given in section 5 and the paper concludes by some discussion and outlook on future work 6.

2 Related Work

Folksonomies are user-generated data created by users' interaction and collaboration using social bookmarking applications. Typically, such applications are designed to acquire the input from users in free-text format to simplify the user interface. As a result, the generated folksonomies contain uncontrolled vocabulary of keywords (tags) with several problems such as polysemy (a word which has multiple related meanings) and synonymy (different words that have identical or very similar meanings) [6]. On the other hand, folksonomies can be considered as a rich data source that contain embedded semantics. As such, many research works targeted the problem of extracting semantics from folksonomies including the problems mentioned above [20,15,8,13,18,1,17,3]. The extracted semantics are usually represented by a simple lightweight ontology, which is a simple tree hierarchy of terms where parent terms are semantically general/broader than their children.

Folksonomies are typically modeled by a tripartite graph with hyper edges [13]. Vertices are partitioned into three disjoint sets of tags, resources and users and each edge connect three vertices (a vertex from each set). A fundamental step in extracting semantics from folksonomies is to transform the tripartite graph into a bi-graph of tags and resources to reveal their inter-relationships.

Map-based and geo-enabled collaborative applications on Web 2.0 generate geo-folksonomies - folksonomies with a geographical dimension - using geographic places as resources. Applications such as Google Maps[3], Tagzaina[4], Openstreetmap[5] and Geonames[6] allow users to create place resources and give

[3] `http://maps.google.com`
[4] `http://www.tagzania.com`
[5] `http://www.openstreetmap.org`
[6] `http://www.geonames.org`

them spatial (such as longitude and latitude) and thematic attributes (such as place name and description). These applications are becoming increasingly popular and currently store millions of references to geographical places. On the other hand, geo-enabled applications such as Flickr[7] and Wikipedia[8] allow users to create Web resources - images in Flickr and Web pages in Wikipedia - and "geo-tag" those resources by assigning them a spatial location or place reference.

Pre-processing folksonomies to enhance the results of folksonomy analysis methods [15,14,9] has been tackled in the literature on different scales. One scale was to process the tags by removing the stop words and stemming the tags such in [18]. Another scale of pre-processing was to enhance the structure of the folksonomy such in [11], which introduced four different aggregation methods to enrich the folksonomy structure, by adding weights that represent the level of users agreement on resource-tag pairs. All the above work targets the general folksonomies, which can be used in geo-folksonomies as well. However, up to our knowledge, there is no research work covers the problem of pre-processing the resources in geo-folksonomies, which is the problem covered by this research work.

3 Problem Definition

The term 'folksonomy' (from folk and taxonomy) was coined by Vander Wal in 2004 [19]. Folksonomy can be seen as a user generated index to classify and organize the Web resources. In social bookmarking applications, a folksonomy tuple, also called tag application [4], is created every time a user tags a Web resource. It can be formalized as follows:

$$F = \{S, U, R, T, \pi\} \tag{1}$$

Where S is the social bookmarking application that hosts the folksonomy tuple, U is a User, R is a Resource, T is a Tag and π is the time stamp of the creation of the tuple.

Users are usually identified by IDs. A user ID is always represented by a unique user name chosen by the user. Resources are Web documents such as Web pages, images or PDF files. Each resource can be located using a unique URI. Tags are single keywords supplied by users to describe and index the resources. The social bookmarking applications store the creation date of the folksonomy tuples which can be used later for temporal analysis. For simplicity, the folksonomy tuple can be redefined as:

$$F = \{U, R, T\} \tag{2}$$

where multiple resources and temporal analysis are not considered in this work.

Each tuple in the folksonomy represents a relation between a user, a resource and a tag. A simple query on such data can answer questions such as: what are the most used tags for annotating resources, or, who is the most active

[7] `http://www.flickr.com`
[8] `http://www.wikipedia.org`

user. These are typical data retrieval questions that can be answered by simple database queries. However, questions such as, what are the most related tags to the tag 'Cardiff', are more complicated where the answer requires co-occurrence analysis of tags to calculate tag similarity.

Web resources, e.g. documents, can be easily located and identified using URIs[9], where each document has a unique address on the World Wide Web. In social bookmarking applications, two users are considered to be tagging the same Web resource only if the resources they are tagging have the same URI.

Unlike Web resources, place resources in geo-social bookmarking applications can't be easily identified and located on the World Wide Web, as such resources are not represented as Web documents and consequently don't have URIs. Typically, place resources are associated with spatial attributes for representing the place location and thematic attributes, e.g. a place name and a place type, encoded as free text. Hence, two users can be considered to be tagging the same place resource only if the resources they are tagging are 'spatially close' and have similar names.

The spatial location of place resources is acquired via a map-based user interface. Users click on the location of the place they want to tag and the mouse location on the applet is translated to the corresponding longitude and latitude. While tagging a new place, the map interface does not reveal any places created by other users in the same area and thus a place resource can be created and tagged a multiple of times by different users. The same place may be given different names. For example, both "Cardiff University" and "Cardiff uni." is used to refer to the same place by different users. Also, both instances may not be digitized at the exact same spatial location.

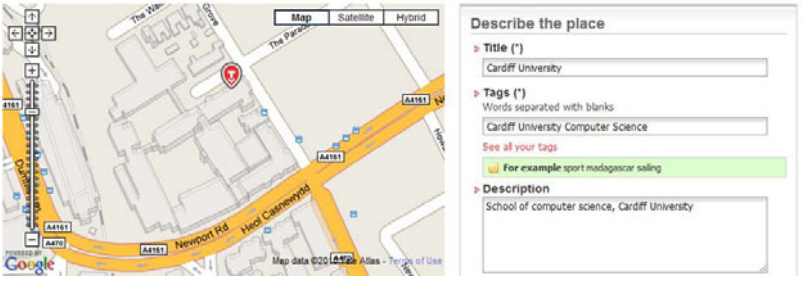

Fig. 1. User interface for creating a new place resource in Tagzania

Figure 1 shows the map-based user interface of Tagzania.com used for tagging new place resources. The map-based interface allows the current user to click on the map to locate the place and add required attributes such as the place name, tags and description in free-text from.

As discussed above, a real-world place entity can be referred to using more than one place resource/instance in the geo-folksonomy. These redundant place

[9] Unique Resource Identifier.

resources are not linked and can thus lead to an increased uncertainty in the information content of the folksonomy and will adversely affect the result of any co-occurrence analysis applied on it.

4 Identifying Redundant Place Resources

Generally, two place instances r_1 and r_2 refer to the same real world place entity if (1) they have the same spatial location and (2) they have the same place name. In this work, exact matching methods are not appropriate and fuzzy similarity matching is used. To identify the redundant place resources in a folksonomy, two stages of analysis should be used:

- Spatially cluster places that are in close proximity to each other.
- In each cluster, identify resources that have similar place names.

4.1 Spatial Clustering

The main objective of using a spatial similarity measure is to find place instances that are in close proximity to each other. This can be achieved by using cluster analysis algorithm or by consulting external reverse geo-coders to assign a unique area code for each place resource, and then area codes can be used as clusters identifiers.

Cluster analysis methods are unsupervised learning methods which aim to group a set of observations into subsets if they are similar in some sense. The feasibility of using cluster analysis is tested in this work by testing Quality Threshold (QT) Clustering [7] on a subset of the folksonomy data. QT is seen as the best candidate algorithm for this work as it does not require the number of clusters to be priori defined.

The Yahoo Where on Earth ID (WOEID) and postcode reverse geo-coders are the external data sources considered here to cluster the place resources. The WEOID web service provides a unique identifier, by reverse geo-coding APIs, for every location on earth. It represents the closest street to any given spatial coordinate. Hence, place instances with the same WOEID are spatially close as they are close to the same street.

Table 1 shows the details of a subset of place resources that represent the place "Big Ben" in London. Each resource is shown with its WOEID, postcode and the calculated QT cluster ID. As shown in the table, all the "Big Ben" instances are grouped into one WOEID while the postcode divides the resources into two groups. Postcode failed as each postcode value represents a very tight area of buildings while the resources in the dataset are not that close. The table also shows the place resources are grouped into one group by using the district level of the postcodes. Also, it shows that QT clustering algorithm could successfully cluster the place resources in this dataset.

Although using district level of postcodes and WOEIDs can produce the same results, the usage of postcodes is only limited to UK. In addition, although the

Table 1. Postcodes and WOEIDs of Big Ben place resources

ID	WOEID	Postcode	District Level PC	QT cluster ID
31758	44417	SW1A 0AA	SW1A	ID0
31759	44417	SW1A 0AA	SW1A	ID0
31760	44417	SW1A 2JR	SW1A	ID0
31761	44417	SW1A 2JR	SW1A	ID0
31762	44417	SW1A 0AA	SW1A	ID0
49775	44417	SW1A 2JR	SW1A	ID0
49776	44417	SW1A 0AA	SW1A	ID0
49777	44417	SW1A 0AA	SW1A	ID0

QT clustering algorithm also can produce the same results of WOEID, the time complexity of running this algorithm limits using it on large datasets. Thus, WOEIDs were found to be more suitable in the scope of this work as the geo-folksonomy dataset used for the experiments is not limited to UK.

4.2 Textual Clustering

After grouping place instances that are spatially similar, a further similarity check can be applied to find place instances with similar names within that group. A simple text similarity method based on "Levenshtein Distance" [10] is used here to find similar place names. The Levenshtein Distance between two strings is the minimum number of edits (insertion, deletion, or substitution) needed to transform the first string to the second string. The text similarity method can be defined by the following equation:

$$\sigma_t(n(r_1), n(r_2)) = 1 - \frac{LD(n(r_1), n(r_2))}{Max((n(r_1), n(r_2)))} \tag{3}$$

where LD is the Levenshtein Distance function and Max is the maximum length of the names of the two place instances.

4.3 Clustering Place Resources

Figures 2 and 3 show two views of an area around "Big Ben" in London. Figure 2 shows the place resources, grouped in colour-coded clusters, after applying the spatial clustering method. Figure 3 shows the same place resources, in different clusters, after identifying similar resources using both the spatial and textual clustering methods. The box in Figure 2 bounds the place resources with a unique WOEID including the place Big Ben in the first view. In Figure 3 the smaller box identifies the place resources which all refer to the place Big Ben. The first box spans an area of 750 m. across its diagonal, where as in second box the area shrinks to around a 1/3 of this size. This demonstrates the quality and accuracy of the location of these place resources.

By identifying redundant place resources, resources that references the same place in the real world are grouped into place clusters and the enriched

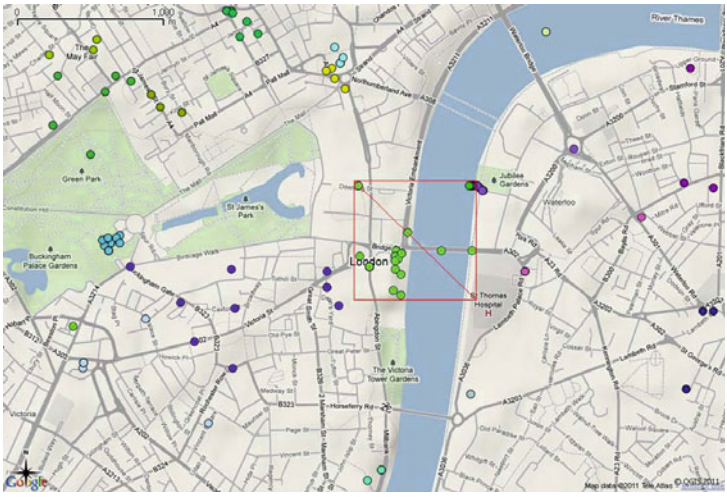

Fig. 2. place resources spatially clustered using WOEID

geo-folksonomy tuples can be defined as $Geo - F = \{U, R, PC, T\}$ where PC is a cluster of similar place resources, of which R is one.

The spatial and thematic attributes for the place clusters can be defined using the instances in those clusters. Different methods can be applied. For example, the spatial location of a place cluster can be computed as either the location of the most central place instance in the cluster, or the centroid of the polygon enclosing the set of place instances in the cluster. Similarly, the place name associated with the cluster can be chosen as the most commonly used name in the cluster, etc.

5 Experiment and Evaluation

5.1 Experiment

The dataset used for evaluation is a geo-folksonomy collected using a crawler software - developed for this work - designed to scan pages on Web 2.0 mapping sites and to index the geo-folksonomies stored on those pages. In this experiment, the crawler was set to process the site: www.tagzania.com. The collected geo-folksonomy dataset includes 22,126 place instances in the UK and USA, 2,930 users and 11,696 distinct tags. The number of geo-folksonomy tuples collected is 65,893. In addition, 10,119 unique WOEID values - cover the entire place instances in the dataset - were obtained by calling Yahoo's reverse geocoding APIs which are exposed via Flickr's Web service.

The method proposed in section 4.3 was used to enrich the collected geo-folksonomy. The text similarity threshold β was set to 0.8 (this was found to be sufficient for this experiment). After applying the method, the number of clusters (unique places) decreased to 19,614. Hence, the method resulted in merging 2,512 place instances (around 11% of the total number of place resources).

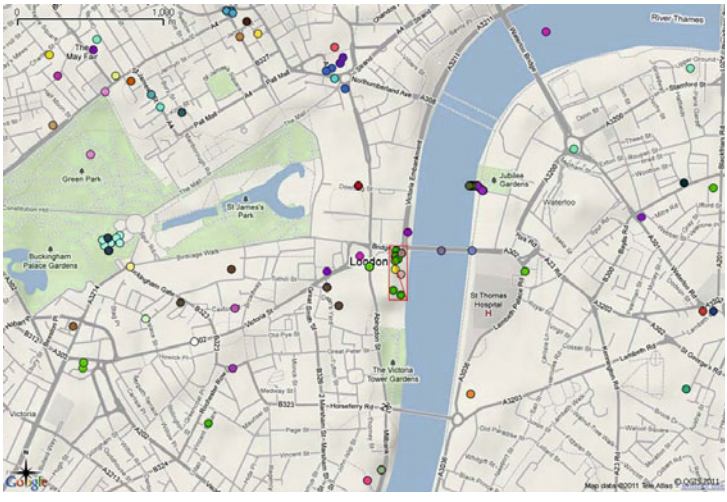

Fig. 3. place clusters after applying spatial and textual clustering

5.2 Measuring the Uncertainty

In order to measure the uncertainty of the Folksonomy Shannon's information gain [16] is used as follows:

$$I(t) = -\sum_{i=1}^{m} \log_2 p(x_i) \tag{4}$$

Where t is any given tag. m is the number of places annotated by the tag t and $p(x_i)$ defined by:

$$P(x) = \frac{w_{t,x}}{\sum_{j=1}^{m} w_{t,x_j}} \tag{5}$$

Where w is equal to the weight of the link between t and place x. The value of $p(x)$ will increase if the number of user votes increases and vice versa, high values of $p(x)$ indicates a high degree of certainty (lower information gain) of using tag t with place x.

5.3 Evaluation Results

To understand the density of the spatial groups (considering WOEID as group) it is worth seeing how the place instances are distributed over the WOEIDs. Figure 4 shows the histogram of the number of place instances over WOEIDs; the WOEIDs that group only 2 place instance are 1653 groups, this number drops to 627 (less than half) for the WOEIDs that group only 3 place instance. Again, this number drops to 350 (around half) for the WOEIDs that group only 4 places and so on.

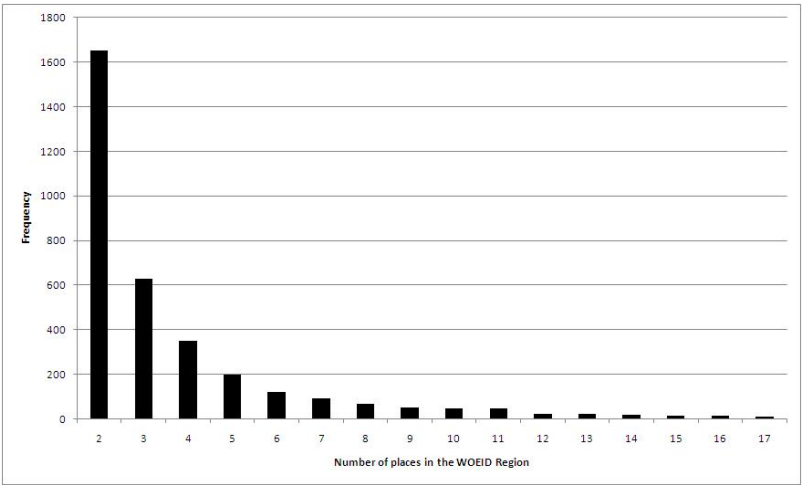

Fig. 4. Histogram of the number of places groupd by WOEIDs

To evaluate the effect of identifying the place instances of the same place concept and build a richer geo-folksonomy that includes user, the information gain is calculated for the Geo-Folksonomy before and after using the proposed method. The results show that the information gain before is 4011.54 and after is 3442.716 which is around 14% reduction in the uncertainty.

The uncertainty reduction is caused by the regions that have increased place annotation activities, in which it is likely to have multiple users annotating the same place using similar names. Table 2 shows a sample of WOEID regions, the number of places in each region and the information content before and after using the proposed method.

Table 2. Information content (Uncertainty) sample

WOEID	Instances	(I) Before	(I) After	Reduction %
2441564	106	126	115	8.7%
2491521	86	11.7	6.9	41%
2352127	83	129	119	7.8%
2377112	80	23.6	18.8	20.3%
2480201	68	24.6	21.6	12.2%

6 Discussion and Future Work

The geo-folksonomy generated in Web 2.0 mapping-based social bookmarking application has introduced a different type of resource on the Web, namely, geographic places. However, these resources cannot be uniquely identified even within the same social bookmarking application. Technically, the cause of this problem is the user interface used to annotate the places. Exposing existing

places resources already annotated by users to new users might address this problem. However, this is controversial and is not adopted by current applications, as this may influence the tagging behaviour of those new users.

An alternative solution is to create a centralized Web service that is responsible for creating and maintaining unique identifiers for place entities. Whenever any social bookmarking application needs to create a new place instance, it can query the centralized service with attributes such as name and location and get a unique identifier for this place. Yahoo's WOEID Web service is an example of this centralized service. However, Yahoo's WOEID Web service generates unique IDs for collection of places up to street level and not to the level of individual places.

Despite creating an overhead, where social bookmarking applications need to integrate with the centralised service to maintain the unique IDs, this solution will support standardised reference to place instances across different applications and therefore can allow the linking and integration of multiple resources.

The methods used for identification and clustering place instances in this work were shown to be successful in removing a significant percentage of redundant place instances. Moreover, the number of links between tags and place resources was dropped from 65,893 to 62,759 where each link is weighted by the number of users who agreed to use the tag-resource pair it connects.

References

1. Almeida, A., Sotomayor, B., Abaitua, J., López-de-Ipiña, D.: folk2onto: Bridging the gap between social tags and ontologies. In: 1st International Workshop on Knowledge Reuse and Reengineering Over the Semantic Web (2008)
2. Auer, S., Lehmann, J., Hellmann, S.: LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 731–746. Springer, Heidelberg (2009)
3. Hao Chen, W., Cai, Y., Fung Leung, H., Li, Q.: Generating ontologies with basic level concepts from folksonomies. ICCS 2010 1(1), 573–581 (2010)
4. Farooq, U., Kannampallil, T., Song, Y., Ganoe, C., Carroll, J., Giles, L.: Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In: Proceedings of the 2007 International ACM Conference on Supporting Group Work, pp. 351–360. ACM (2007)
5. Twaroch, F.A., Jones, C., Abdelmoty, A.: Acquisition of Vernacular Place Footprints from Web Sources. In: Baeza-Yates, R., King, I. (eds.) Weaving Services and People on the World Wide Web, pp. 195–214. Springer, Heidelberg (2009)
6. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (2006)
7. Heyer, L., Kruglyak, S., Yooseph, S.: Exploring expression data: identification and analysis of coexpressed genes. Genome Research 9(11), 1106 (1999)
8. Heymann, P., Garcia-Molina, H.: Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford InfoLab (April 2006)

9. Lee, S., Won, D., McLeod, D.: Tag-geotag correlation in social networks. In: Proceeding of the 2008 ACM Workshop on Search in Social Media, pp. 59–66. ACM (2008)
10. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady 10, 707–710 (1966)
11. Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., Stumme, G.: Evaluating similarity measures for emergent semantics of social tagging. In: Proceedings of the 18th International Conference on World Wide Web, pp. 641–650. ACM, New York (2009)
12. Mathes, A.: Folksonomies-cooperative classification and communication through shared metadata. In: Computer Mediated Communication, LIS590CMC (Doctoral Seminar), Graduate School of Library and Information Science. University of Illinois Urbana-Champaign (2004)
13. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In: Web Semantics: Science, Services and Agents on the World Wide Web, vol. 5 (2007)
14. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 103–110. ACM, New York (2007)
15. Schmitz, P.: Inducing ontology from flickr tags. In: Collaborative Web Tagging Workshop at World Wide Web, Edinburgh, Scotland (2006)
16. Shannon, C.: A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review 5(1), 55 (2001)
17. Tsui, E., Wang, W.M., Cheung, C.F., Lau, A.S.M.: A concept-relationship acquisition and inference approach for hierarchical taxonomy construction from tags. Inf. Process. Manage. 46(1), 44–57 (2010)
18. Van Damme, C., Hepp, M., Siorpaes, K.: Folksontology: An integrated approach for turning folksonomies into ontologies. Bridging the Gap between Semantic Web and Web 2, 57–70 (2007)
19. Wal, T.V.: Folksonomy (2007), `http://www.vanderwal.net/folksonomy.html`
20. Wu, H., Zubair, M., Maly, K.: Harvesting social knowledge from folksonomies. In: Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, p. 114. ACM (2006)