Deriving Ontological Structure from a Folksonomy

Saurabh Lalwani
University of South Carolina
Computer Science and Engineering Department
Columbia, SC, 29208
+1-803-777-2880
lalwanis@engr.sc.edu

Michael N. Huhns
University of South Carolina
Computer Science and Engineering Department
Columbia, SC, 29208
+1-803-777-5921
huhns@sc.edu

ABSTRACT
In this paper we describe our investigation of tagging systems and the derivation of ontological structure in the form of a folksonomy from the set of tags. Tagging systems are becoming popular, because the amount of information available on some websites is becoming too large for humans to browse manually and the types of information (multimedia data) is unsuitable for the indexers used by conventional search engines to organize. However, tag-based search is very inaccurate and incomplete (low precision and recall), because the semantics of the tags is both weak and ambiguous. The basic problem is that tags are treated like keywords by search engines, which consider individual tags in isolation. However, there is additional semantics implicit in a collection of tagged data. In this paper, we innovate and investigate techniques to make the implicit semantics explicit, so that search can be improved in both precision and recall and additional utility can be derived from the tags that people associate with multimedia items (pictures, blogs, videos, etc.). Our approach is to propose hypotheses about the ontological structure inherent in a collection of tags and then attempt to verify the hypotheses statistically. We conducted more than one hundred experimental searches on *Flickr* with different tags. By statistical analysis of the search results, we discovered information about how tags are assigned by users and what ontological knowledge is implicit in these tags that can be made explicit and, ultimately, exploited.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Abstracting methods, Dictionaries, Indexing methods, Linguistic processing, Thesauruses.*

General Terms
Experimentation, Human Factors

Keywords
Folksonomy, Tagging, Ontology Induction

1. INTRODUCTION
The structure of the World Wide Web has changed enormously in the past few years. Transitions from Akamai to BitTorrent, Britannica Online to Wikipedia, personal websites to personal blogs, publishing to participation, content management systems to Wikis, stickiness to syndication feeds, and directories to folksonomies are some of the indicators of the changes underway. The resultant Web now consists of Internet communities, social networking sites, data sharing sites, wikis, blogs, and tagging systems. Consequently, huge amounts of information of different types are now accessible to users.

One of the aspects of the vision for the Semantic Web is that Web pages will have metadata that helps to specify the semantics of the contents of the pages. The metadata will be machine-understandable and machine-processable, which are needed for computers to be able to assist humans with use and management of the massive amounts of data.

The concept of tagging items on the Web is not a new one. Originally, it was done only by domain and computer experts. The problem they encountered is that it is difficult for typical website developers to add the metadata precisely and formally, because they do not always know how the website and its metadata will be used. As a partial solution, some websites in Web 2.0 environments have achieved success by enabling users to add metadata in the form of natural language tags. The result is that increasing amounts of on-line information are being categorized by associating each piece of information with tags by the users themselves in ways that are comfortable and natural for them.

Users are allowed to classify the content according to their liking by adding freely-chosen keywords as tags with no restriction on the use of a controlled vocabulary. The users are not given any guidance about the form or structure of the tags, making the tagging process easy and straightforward for them.

Some common examples of websites used most frequently by users to tag content are:

- http://del.icio.us/ – for *bookmarks*
- http://www.flickr.com – for *photographs* and *videos*
- http://amazon.com – for a variety of things *to be bought and sold*
- http://www.librarything.com – for *books*
- http://www.gmail.com – for *e*-mails
- http://odeo.com – for *podcasts*

The tags can then be used for keyword-based searches. The tags ideally are used to filter out the enormous amount of data present on the Web and display only the data of interest. When a user runs a tag-based search on the Web, only the information that has been

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ACMSE '09 March 19-21, 2009, Clemson, SC, USA.
Copyright 2009 ACM 1-58113-000-0/00/0004…$5.00.

tagged the same by other users is shown and the user can then select the most appropriate content among the results displayed.

Although the tags are assigned easily in an unstructured manner, there is an implicit structure within a set of tags, and this implicit structure is called a *folksonomy*. It is a structure that emerges bottom-up as tags are added to a system of information items.

Intuitively, this tagging of data and then searching for tags of interest to get the appropriate content seems to be very easy and to work well superficially. However, although tag-based search filters out the irrelevant data, it also filters out a large amount of relevant information that is marked with similar, but not exactly the same tags as used in the search. This is because, in a tagging system, there is no information about the semantics of a tag. As a result, there is no means to zero-in on the information related to a particular tag and include it in the result of the search. For example, a search for {*utensils}* returns only the items that have been tagged with that exact keyword, ignoring all the other items that may be related to it. A user would like this search ideally to display all the items related to {*utensils}*, such as {*spoon}*, {*fork}*, and {*knife}*.

If the implicit structure can be made explicit in a formally represented folksonomy, then searches can be improved. This is one of the major motivations for making a folksonomy explicit, as more and more tagged systems are appearing on the Web.

To make these tagging systems more efficient, a methodology needs to be devised so that all the data related to keywords of interest gets displayed rather than just the data containing those keywords. The objective of our research is to make folksonomic metadata explicit. Our approach is to formulate plausible hypotheses based on our previous work in the construction of ontologies [4,5,7,8], and then evaluate the hypotheses using data from existing on-line systems of tagged items. Hence, in this paper, we outline several hypotheses to derive additional utility from the tags that people are associating with items that are on the Web or, more precisely, *we are investigating how to derive ontological structure from a folksonomy*.

Specifically, we have formulated the following three hypotheses that we believe might describe the implicit structure in a folksonomy:

Hypothesis 1: For a group of items, if the number of occurrences of Tag1 is less than the number of occurrences of Tag2; and if there are items where Tag1 co-occurs with Tag2, then Tag2 is a subclass of Tag1. The general heuristic rule we hypothesize is that the more a tag is used, the higher the level (closer to a root) it will be in an ontology.

Hypothesis 2: Two tags can be claimed to be related if and only if the ratio of their co-occurrences to the subclass is greater than some threshold value, with confidence based on cardinality.

Hypothesis 3: For a group of items, if Tag1 co-occurs with Tag2 and Tag1 also co-occurs with Tag3, but Tag2 does not co-occur with Tag3, then Tag2 and Tag3 are subclasses of Tag1, and Tag2 is disjoint with Tag3 with a confidence based on the respective cardinalities.

To evaluate the veracity of these Hypotheses, we have conducted a number of searches on Flickr. Our first set of searches concentrated on evaluating the first Hypothesis. Each search contained three parameters: *{Tag1}* – displaying the number of items tagged with Tag1, *{Tag2}* – displaying the number of items tagged with Tag2, and *{Tag1, Tag2}* – displaying the number of items tagged with both Tag1 and Tag2 and conclusion was based on the number of occurrences.

For the second Hypothesis, each search contained two parameters *{Tag1, Tag2}* and *{Tag2}* and then a mathematical analysis was done to calculate the threshold ratio along with allowable variance.

For the third Hypothesis, the numbers of co-occurrences were compared to the threshold value and then analyzed to identify whether some relationship exists or the tags are disjoint. This hypothesis inherently considers Hypothesis 2 to be true.

2. BACKGROUND

As the World Wide Web is becoming enhanced by Web 2.0 techniques [11,14], there is increasing use being made of tagging and folksonomies. According to Wikipedia, a ***folksonomy (also known as collaborative tagging, social classification, social indexing, and social tagging) is the practice and method of collaboratively creating and managing tags to annotate and categorize content*** [12]**.** Folksonomies began gaining popularity in 2004 as a part of social software applications. The term was coined by Thomas Vander Wal [10] as a portmanteau of the words *"folks"* and *"taxonomy"* though it has little to do with either of those words. A *taxonomy*, in reference to the Web, refers to an ontological way of categorizing data; whereas a folksonomy categorizes content with *"tags,"* which do not have any implicit hierarchy defined and are all treated to be at the same level, i.e., they are theoretically "equal" to each other. Using these tags, a folksonomy is intended to make information retrieval extremely easy and fast. It can also be used, as demonstrated with the tags from *http://delicious.com* [9], to customize searches.

Tags themselves are keywords used to categorize the content of information items, typically on the Internet. Aggregating the tags of many users creates a *folksonomy*. Again according to Wikipedia**, *a tag is a non-hierarchical keyword or term assigned to a piece of information (such as an Internet bookmark, digital image, or computer file). This kind of metadata helps describe an item and allows it to be found again by browsing or searching*** [13].

The history of tags goes back to 2003 when Joshua Schachter, the founder of the most famous social bookmarking site *http://delicious.com*, pioneered the use of tags for the user's bookmarks. *Flickr* joined in with the same concept and allowed users to tag pictures and videos making them easy to search. With the success of *Flickr* and *Delicious*, collaborative tagging gained popularity and consequently many other websites, such as *YouTube*, *Picasa*, and *Technocrati*, started implementing tagging.

The task of discovering semantic relations between concepts (e.g., subsumption, disjointness, or named relations) is core to productive use of the Semantic Web. As such, Scarlet [6] harvests the Semantic Web by automatically finding and exploring multiple and heterogeneous online knowledge sources. The result is discovered relations, which can be used for tasks such as ontology matching, ontology learning, word sense disambiguation, and ontology enrichment. The sources for this effort are not tag sets, but partial ontologies of a domain.

In a similar vein, Angeletou [1,2] developed FLOR, a tool that performs semantic enrichment of folksonomy tagspaces by

exploiting online ontologies, thesauri, and other knowledge sources. The result is improved semantics for tags, but not relationships among tags.

Tags are chosen freely, generally without a controlled vocabulary, by the item's creator or viewer, depending on the system, with no limit placed on the number of tags an item can have. This flexibility provided by tagging systems allows people to classify data in a manner they find useful, and associating a larger number of tags with an item facilitates in finding more relevant information (i.e., greater recall).

Because of their demonstrated utility, there have been several attempts to improve upon the semantics of the tags to increase their utility even more. In one of these [3], a domain was seeded with an existing taxonomy. User-generated tags were then added to the taxonomy, so that the resultant set of tags would have a structure. In contrast, our approach is to induce a structure from an existing unstructured set of tags.

Though very useful, tags have quite a few disadvantages as well. First, since tags are freely chosen, synonyms, homonymy and polysemy are very likely to arise, thereby degrading the efficiency of searching. For example, a user could tag an item as *{Sport}* or *{Sports}* and if we try searching for items having these tags separately, the result for each search will be different.

Second, tags are used as just keywords, which do not convey information about their semantics. Consequently, when an item is tagged with a word that can represent more than one meaning, the search results are bound to display some results that might be irrelevant to the user. For example, a user can tag an item as *{Orange}* which can refer either to the color orange or the fruit orange.

Third, items on the Web are tagged by many different people and everyone has a personalized way of tagging, which may or may not match with the ways of others. As a result, people may have to search quite a few times before they find appropriate information.

Last, tags are non-hierarchical structures, so any tag-based search returns only the content containing the exact same keyword. For example, consider three pictures where the first is tagged as *{Sports, Soccer}*, the second is tagged as *{Sports}*, and the third is tagged as *{Soccer}*. A search for the tag *{Sports}* fetches the first and second pictures as the result, but not the third picture, even though *{Soccer}* is a subclass for *{Sports};* as there is no means of knowing this relation implicitly, we are left with just the first two pictures (i.e., low recall).

3. ANALYSIS

To understand the ontological knowledge among tags used by people in different domains, we have performed a mathematical analysis based on the results obtained from the experimental searches done on *http://www.flickr.com* (only on pictures that were available publicly). Specifically, this analysis provides us with the additional knowledge implicit in the tags that can be used to derive a hierarchical structure (along with a few non-statistically significant exceptions). The obvious cause of these exceptions is the use of tags freely chosen by users.

3.1 Hypothesis 1

For a group of items, if *{Tag1}* occurs less frequently than *{Tag2}* and if we have a reasonable count for items where *{Tag1}* co-occurs with *{Tag2}*; then *{Tag2}* can be termed a subclass of *{Tag1}*.

Note, we consider here the co-occurrences of tags only if their number is at least 5% of the number of less occurring tags, to cover the margin for errors that can occur due to the use of an uncontrolled vocabulary. This assumption is based purely on observation.

This hypothesis was an intuitive guess taken initially while investigating how users have tagged content on *Flickr*. The first thing we observed was that for any given hierarchy of classes, users tend to use the leaf nodes as tags more often than they use superclass terms for tags. For example, the tag *{Spoon}* is used more frequently than the tag for its superclass, *{Utensil}*. The results returned from the experimental searches were evaluated by counting the occurrences of *{Tag1}* and *{Tag2}* individually and then counting the occurrences of both tags together. The cardinality ratio of *{Tag2} / {Tag1}* for all the individual observations was calculated and then its arithmetic mean was taken. The average of the cardinality ratio was calculated to be **3.86,** which might seem to conclude that most of these experimental searches abide by this hypothesis. But the reality is just the opposite. The average came out to be on the higher end because a few search results gave a high ratio of up to 32. A large number of these searches defied the hypothesis, which is clearly shown in the histogram below.

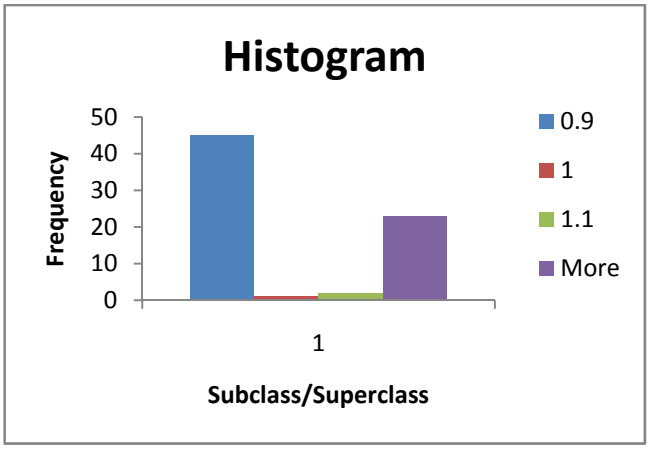


Figure 1. A histogram of the experimentally found subclass-cardinality / superclass-cardinality ratios

According to the hypothesis, the frequency for *Subclass / Superclass ratio > 1* should have been more than the frequency for *Subclass / Superclass ratio < 1*, but the histogram reveals just the opposite. The main cause of this inconsistency is the use of an uncontrolled vocabulary by the users, which prevents us to get the actual count of the tag and consequently the failure of this hypothesis. For instance, consider a picture of a "human being". Different users can tag this picture with {Human, Humans, Homo sapiens, Man, Woman, etc.}, which, although they mean the same thing, are different when counting tags.

Another reason that can be observed here is that users tag content differently for different domains. For some categories, users tend to use the superclass rather than the leaf nodes and vice versa for other categories. For example, the cardinality ratio for pictures tagged as *{Fork}* vs. the pictures tagged as *{Utensil}* is 32

claiming the hypothesis to be true, whereas the cardinality ratio for pictures tagged as *{Eagle}* vs. the pictures tagged as *{Bird}* is 0.1 which causes the hypothesis to be considered false.

There are many similar examples and hence no certain conclusions can be drawn about the subclass-superclass relationships of tags from this hypothesis. Hence, it is not verified.

3.2 Hypothesis 2

Two tags can be declared to be co-related if and only if the ratio of their co-occurrences to the subclass is greater than some threshold value, with confidence based on cardinality.

In the first hypothesis, we made a vague assumption about tags being co-related. So, in this hypothesis, we claim that for two tags to be co-related, the ratio of the cardinality of their co-occurrence to the cardinality of the individual tags must be greater than a specific plausible value (determined experimentally). In order to calculate this value, three consecutive searches were done – first having both the tags and then a search for obtaining each individual tag. The result of the first search is then divided by the result of both the other searches separately to calculate two ratios for cardinality. Here, two ratios of cardinality have been calculated separately because, currently, we do not know whether any relationship between the tags exists and, therefore, we need to consider both cases. Similar calculations have been made for approximately two-hundred search results on Flickr. The mean of the ratios calculated, thus, gives us a good estimate of the final cardinality ratio to decide whether a subclass/superclass relationship exists or not; or more precisely, it gives us the threshold value.

For example, the cardinality of items with tag *{Animal, Dog}* is divided both by the cardinality of items with tag *{Animal}* and the tag *{Dog}*. The cardinality ratios are calculated to be 0.12 and 0.06. So, the final cardinality is the mean of these two, which is 0.09.

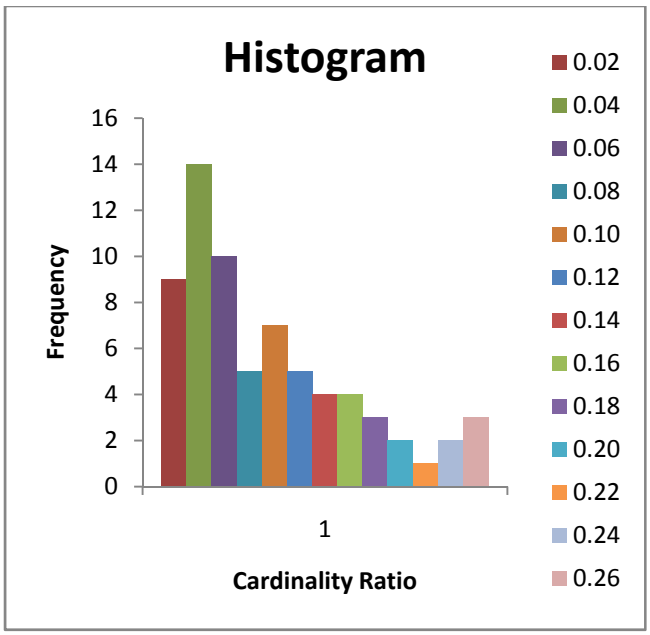


Figure 2. A histogram of the cardinality ratios used to determine a threshold for deciding whether two tags are related or not

The histogram in Figure 2 shows the results of our experiment. The tags for this histogram are deliberately chosen to have an inherent subclass-superclass relationship, which will then help us in determining the threshold value.

The mean of the cardinality ratio of the tags used for the above histogram is 0.09 and the variance is 0.005. But as can be seen from the histogram, the frequency is highest between cardinality ratios of 0.04 and 0.08, so it can be concluded that the threshold value for two tags to be in a subclass-superclass relationship is 0.06, based on our statistical analysis.

To verify the above finding, we constructed another histogram (shown in Figure 3) for tags that are not related to each other in any way. The mean of the cardinality ratios for these tags was calculated to be 0.0008 and the variance was 1.5E-06. Hence, our experiment clearly shows that the threshold value of the cardinality ratio works well indeed and *Hypothesis 2* holds.

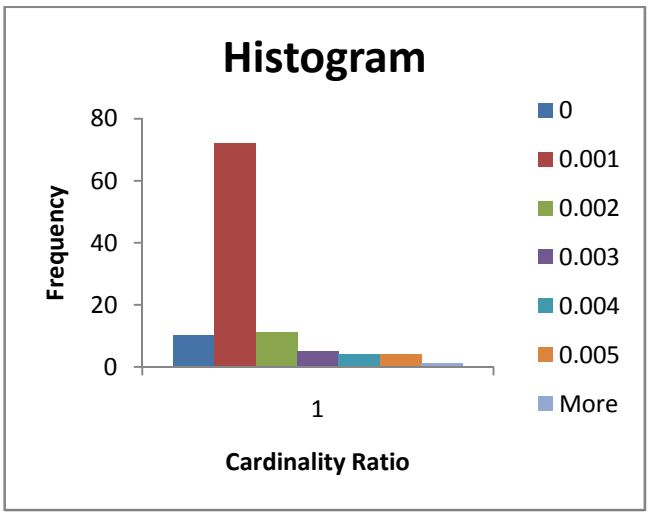


Figure 3. A histogram of the cardinality ratio among tags that are unrelated to each other, to provide a contrast with the histogram in Figure 2

There are certain cases where this hypothesis does not hold. For example, consider the tags *{Grass, Green}*. Obviously, *Green* is just the color of the grass and does not have any superclass-subclass relationship with *{Grass}*, but if this hypothesis were true, then it would result in *Green* being declared to be a superclass of *Grass*. This again shows that inconsistency in tagging content from the user's end can introduce inconsistencies in the behavior of any such ontological structure finding technique, and so errors are expected to occur. For the number of searches made, this hypothesis gave a success rate of more than 85%. Hence, this hypothesis is considered to be generally verified.

3.3 Hypothesis 3

For a group of items, if *{Tag1}* co-occurs with *{Tag2}* and *{Tag1}* also co-occurs with *{Tag3}*, but *{Tag2}* does not co-occur with *{Tag3}*, then *{Tag2}* and *{Tag3}* are subclasses of *{Tag1}*, and *{Tag2}* is disjoint with *{Tag3}* with confidence based on the measured cardinalities.

This hypothesis is a direct consequence of *Hypothesis 2*. In *Hypothesis 2*, we were able to derive a conclusion that some relation exists between two tags, but could not define the exact

relation. So, in this Hypothesis, we add one more tag to the observations and record the results as explained: the co-occurrence of {Tag1, Tag2} is recorded, then of {Tag1, Tag3} and finally of {Tag2, Tag3}. All these co-occurrences are then analyzed separately and cardinality ratios for each of them are calculated. The cardinality ratios for the first two searches are greater than the threshold and, hence, from *Hypothesis 2*, it can be concluded that there exists a relationship between them. No such relationship exists between *{Tag2}* and *{Tag3}*, as their cardinality ratio is far below the threshold. So, it can again be concluded that *{Tag2}* and *{Tag3}* are disjoint classes. Now, since *{Tag1}* co-occurs both with *{Tag2}* and *{Tag3}* with a high cardinality ratio, it becomes obvious that it is one level higher than the other two tags and hence becomes the superclass for the other two tags. The combined results shown in the histograms of Fig. 2 and Fig. 3 describes the statistical analysis of *Hypothesis 3*. The following example provides intuitive justification for the hypothesis.

Example – For pictures having tags *{Soccer, Sports}*, the cardinality ratio was calculated to be 0.065 (greater than the threshold value, 0.06). For *{Basketball, Sports}* the cardinality ratio was 0.10 (> 0.06) and, finally, for *{Soccer, Basketball}*, the cardinality ratio was 0.005 (< 0.06). Indeed the hypothesis holds true, as Basketball and Soccer are subclasses of Sports.

Again, there are certain exceptions where this hypothesis fails. But considering the behavior of folksonomies and the reasons specified, exceptions of about 15% are assumed to be tolerable. Since the analysis for this hypothesis is dependent on *Hypothesis 2*, it primarily fails in cases when *Hypothesis 2* fails. Hence, this hypothesis can be declared as verified.

4. DISCUSSION AND CONCLUSION

In this paper, we have examined tagging and folksonomies: their emergence, importance, and future use. They began with one website pioneering this idea and now many developers are trying to use some variation of a tagging system for their websites. Considering the future of tagging systems and the way users are tagging items on the Web, it is going to become more difficult to find things using tags. Therefore, our research has proposed and investigated a few hypotheses that can facilitate tag-based search. The idea is to identify the network of related tags for a given tag and, for searches, retrieve all items within a short "tag distance." Based on the results of the experiments done, we can already conclude that it is feasible to derive an ontological structure from a given folksonomy and use it to retrieve additional relevant information.

Although our first hypothesis was not verified, a number of important inferences can be drawn from the results that we obtained. That is, users tend to tag data with different keywords that may or may not be the leaf nodes in a hierarchy depending on the domain to which the data belongs. For some domains, this hypothesis yields good results, whereas for other domains it fails. Consequently, work can be done to categorize data in different domains and hence a more domain specific hypothesis can be made that will yield more accurate results.

Our second hypothesis was a direct consequence of the first hypothesis. This hypothesis does not clearly tell us what relationship exists between two tags, but it does reveal whether a relationship exists or not. This hypothesis extracts a vague relationship, which is then used along with our third hypothesis to get the exact relationship. Our third hypothesis can be called an extension to the second one.

Our future work will be based on extending this research to either more specific domains or trying the same approach with different on-line tagging systems. Experiments on *Flickr* have produced good results, but they are insufficient to be generalized for other tagging systems. The primary reason is that the method and hypothesis given for tags used on *Flickr* may or may not be applicable to other systems, such as *del.icio.us*, since these are two completely distinct tagging systems. For instance, a superclass-subclass relationship derived in *Flickr* may turn out to be a subclass-superclass relationship in *del.icio.us* or the tags may not be related at all. Therefore, significantly more research needs to be done before we can generalize a statement applicable to all the tagging systems. The results of our experiments convey the same message.

Finally, to exploit the hypotheses we have formulated and verified, the discovered explicit structure can be captured in a formalism, such as OWL. It can then be mapped to an existing ontology expressed in the same formalism. It would then be possible to reason over the combination using a reasoner for that formalism (such as Pellet for OWL).

5. REFERENCES

[1] Sofia Angeletou, "Semantic Enrichment of Folksonomy Tagspaces," *Proc. International Semantic Web Conference*, Karlsruhe, Germany, 2008.

[2] Sofia Angeletou, Marta Sabou, Lucia Sepia, and Enrico Motta, "Bridging the gap between Folksonomies and Semantic Web: An Experience Report," *Proc. Workshop on Bridging the Gap between Semantic Web and Web 2.0, European Semantic Web Conference*, Innsbruck, Austria, 2007, pp. 30-43.

[3] Sarah Hayman, "Folksonomies and Tagging: New Developments in Social Bookmarking," in *Proc. Ark Group Conference, Developing and Improving Classification Schemes*, Sydney, Australia, June 2007.

[4] Jingshan Huang, Jiangbo Dang, and Michael N. Huhns, "Ontology Alignment as a Basis for Mobile Service Integration and Invocation," *Journal of Pervasive Computing and Communications*, vol. 3, no. 2, 2007, pp. 138-158.

[5] Michael N. Huhns and Larry M. Stephens, "Personal Ontologies," *IEEE Internet Computing*, vol. 3, no. 5, September/October 1999, pp. 85-87.

[6] Marta Sabou, Mathiue d-Aquin, and Enrico Motta, "SCARLET: SemantiC relAtion discoveRy by harvesting onLinE onTologies,"

[7] Munindar P. Singh and Michael N. Huhns, *Service-Oriented Computing: Semantics, Processes, Agents*, John Wiley & Sons, Ltd, West Sussex, England, 2005.

[8] Larry M. Stephens, Aurovinda K. Gangam, and Michael N. Huhns, "Constructing Consensus Ontologies for the Semantic Web: A Conceptual Approach," *World Wide Web Journal*, Kluwer Academic Publishers, vol. 7, no. 4, December 2004, pp. 421-442.

[9] Ching-man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt, "A Study of User Profile Generation from Folksonomies," *Proc. Social Web and Knowledge*

Management Workshop, April 2008, Beijing, China. (In press)

[10] Thomas Vander Wal, "Folksonomy coinage and definition," 2007. http://www.vanderwal.net/folksonomy.html

[11] Tim O'Reilly, "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," 2005. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[12] "Folksonomy," Wikipedia, 2008. http://en.wikipedia.org/wiki/Folksonomy

[13] "Tag (metadata)," Wikipedia, 2008. http://en.wikipedia.org/wiki/Tag_(metadata)

[14] "Web 2.0," Wikipedia, 2008. http://en.wikipedia.org/wiki/Web_2.0

Deriving Ontological Structure from a Folksonomy

ACM Southeast Conference Clemson, SC March 2009

Saurabh Lalwani and Michael N. Huhns
Department of Computer Science and Engineering
University of South Carolina

Research Objective

The objective of our research is to derive ontological structure in the form of a folksonomy from the set of tags. Tagging systems are becoming popular, because the amount of information available on some websites is becoming too large for humans to browse manually and the types of information (multimedia data) is unsuitable for the indexers used by conventional search engines to organize. However, tag-based search is very inaccurate and incomplete (low precision and recall), because the semantics of the tags is both weak and ambiguous. The basic problem is that tags are treated like keywords by search engines, which consider individual tags in isolation. However, there is additional semantics implicit in a collection of tagged data. We investigate techniques to make the implicit semantics be *explicit*, so that search can be improved in both precision and recall and additional utility can be derived from the tags that people associate with multimedia items (pictures, blogs, videos, etc.).

Research Methodology

To understand the ontological structure among tags used by people in different domains, we formulated hypotheses that might describe the implicit structure in a folksonomy, and then evaluated them statistically using the tags in Flickr, an on-line photo-sharing site.

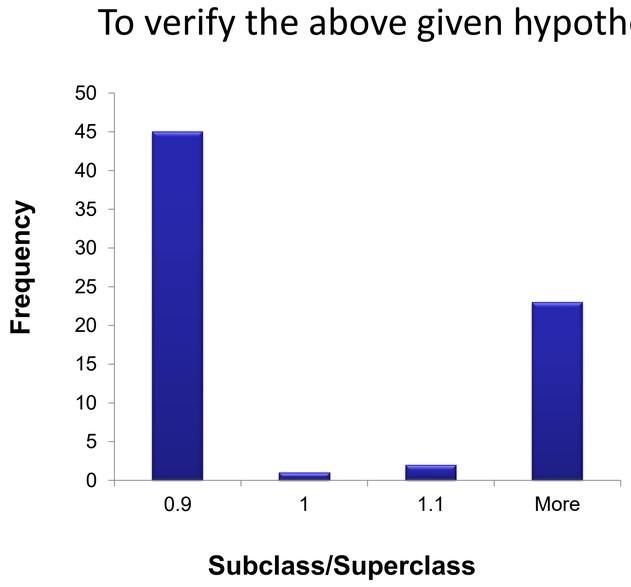
Hypothesis 1
For a group of items, if {Tag1} occurs less frequently than {Tag2} and if we have a reasonable count for items where {Tag1} co-occurs with {Tag2} (a threshold value is assumed), then {Tag2} can be termed a subclass of {Tag1}.

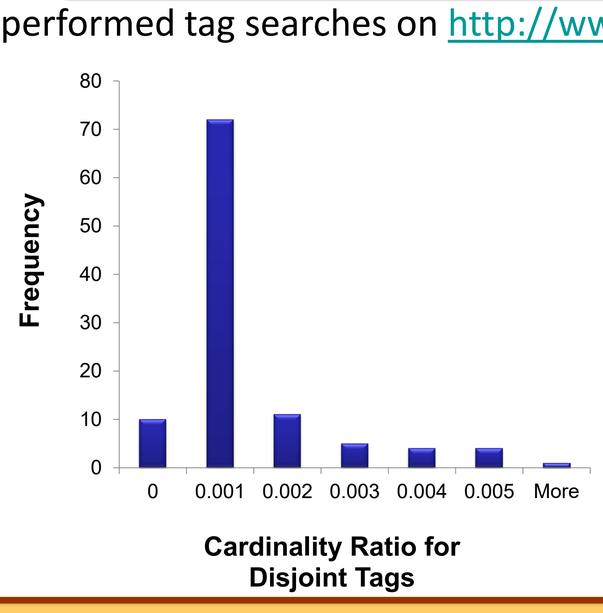
Hypothesis 2
Two tags are related if and only if the ratio of their co-occurrences to their occurrences is greater than some threshold value (calculated experimentally), with confidence based on cardinality.

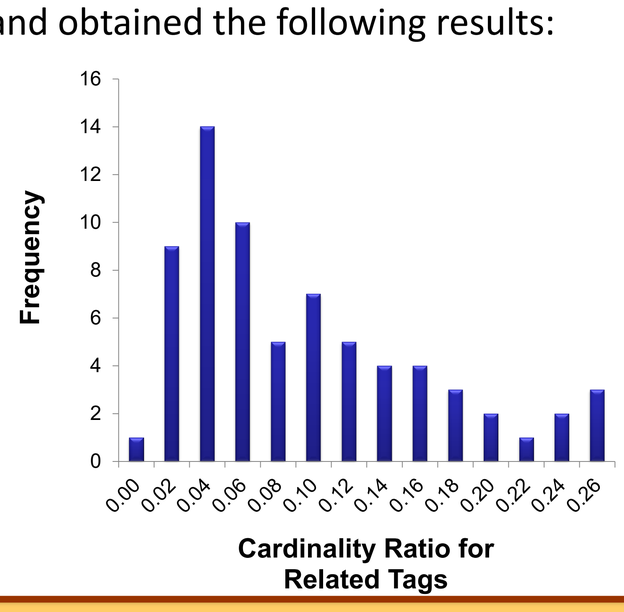
Hypothesis 3
For a group of items, if {Tag1} co-occurs with {Tag2} &{Tag3}, but {Tag2} does not co-occur with {Tag3}, then {Tag2} &{Tag3} are subclasses of {Tag1}; and {Tag2} is disjoint with {Tag3} with confidence based on the respective cardinalities.

Experimental Analysis

To verify the above given hypotheses, we performed tag searches on http://www.flickr.com/ and obtained the following results:







Hypothesis 1:
- Our heuristic is that the more a tag is used, the closer it will be to a root node in an ontology
- We tried to verify it by conducting ~100 experimental searches
- We found that more specific tags are used more frequently
- For example, *{Spoon}* is used more frequently than its superclass *{Utensil}*
- This hypothesis was thus not verified

Hypothesis 2:
- For each pair of tags, searches were made for *Tag1*, *Tag2*, and their co-occurrence *{Tag1, Tag2}*
- We calculated the ratios $|\{Tag1, Tag2\}|/(|Tag1|+|Tag2|)$, where $|\bullet|$ denotes cardinality
- The average ratio for related tags was found to be 0.06 and the average ratio for unrelated tags was 0.0008, thus providing a statistically significant difference to be used for a threshold
- This hypothesis gave a success rate of more than 85%, and hence, is deemed verfied

Hypothesis 3:
- This hypothesis assumes Hypothesis 2 to be true and extends it
- We compared the number of co-occurrences to a threshold value (evaluated in Hypotheses 2) and then manually identified whether some relationship exists or the tags are disjoint
- Our analysis showed that Hypothesis 3 holds ~85% of the time
- Again, considering the informal nature of folksonomies, we can consider this sufficient to deem this hypothesis verified

Conclusion And Future Work

•Considering the popularity of tagging systems, it is important to derive as much benefit from them as possible. Our research has shown that the implicit structure (folksonomy) of a set of tags can be made explicit
•We plan to investigate domain-specific hypotheses and other tagging systems. Experiments on *Flickr* have produced good results, so we will try to generalize them to other tagging systems, such as *del.icio.us*
•To exploit the hypotheses we verify, we will capture the discovered explicit structure in a formalism, such as OWL, and then map it to an existing ontology expressed in the same formalism. It would then be possible to reason over the combination using a reasoner for that formalism (such as Pellet for OWL)