FolksoViz: A Semantic Relation-Based Folksonomy Visualization Using the Wikipedia Corpus

Kangpyo Lee
Seoul National University
Gwanak-gu, Seoul, Korea
kplee@idb.snu.ac.kr

Hyunwoo Kim
Seoul National University
Gwanak-gu, Seoul, Korea
hwkim@idb.snu.ac.kr

Hyopil Shin
Seoul National University
Gwanak-gu, Seoul, Korea
hpshin@snu.ac.kr

Hyoung-Joo Kim
Seoul National University
Gwanak-gu, Seoul, Korea
hjk@snu.ac.kr

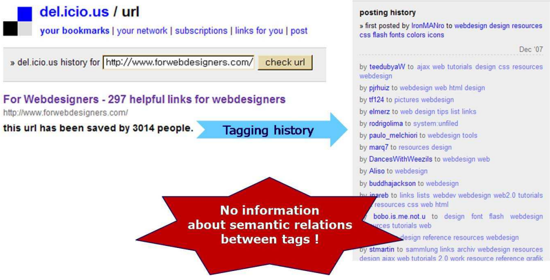
ABSTRACT

Tagging is one of the most popular services in Web 2.0 and folksonomy is a representation of collaborative tagging. Tag cloud has been the one and only visualization of the folksonomy. The tag cloud, however, provides no information about the relations between tags. In this paper, targeting del.icio.us tag data, we propose a technique, FolksoViz, for automatically deriving semantic relations between tags and for visualizing the tags and their relations. In order to find the equivalence, subsumption, and similarity relations, we apply various rules and models based on the Wikipedia corpus. The derived relations are visualized effectively on the screen. The experiment shows that the FolksoViz manages to find the correct semantic relations with high accuracy.

Keywords: Folksonomy, Collaborative Tagging, Semantic Relation, Visualization, Wikipedia, Web 2.0

1. INTRODUCTION

Tagging has become one of the most popular services in Web 2.0. A tag is a relevant keyword assigned to Web documents or resources. A noticeable role of tags is that they can act as good metadata that best describe the Web document or resource, because many of them are carefully chosen by taggers. Folksonomy is also one of the most noticeable features in the current Web 2.0, which originated from combining the words ‘folk’ and ‘taxonomy’. Folksonomy is also widely known as collaborative tagging. Collaborative tagging is achieved collaboratively by multiple taggers who assign a list of tags as the metadata. Del.icio.us [1] is said to be the true implementation of collaborative tagging. It provides an online social bookmarking service that enables users to register their own bookmarks and share them with others. Each user can assign several tags to a URL, and the whole set of tags created for that URL by many taggers are open to the public in the form of posting history. Figure 1 shows the collaborative tagging in del.icio.us. A URL regarding the web design was registered by the first poster, and he or she assigned several tags to the URL. After that, many other users also assigned their own tags to the URL. A long posting history is given at the right side of the screen. This process gradually constructs a folksonomy.



del.icio.us / url
your bookmarks | your network | subscriptions | links for you | post

» del.icio.us history for http://www.forwebdesigners.com/ check url

For Webdesigners - 297 helpful links for webdesigners
http://www.forwebdesigners.com/
this url has been saved by 3014 people.

posting history
» first posted by IronMANro to webdesign design resources css flash fonts colors icons
Dec '07
by teedubyaW to ajax web tutorials design css resources webdesign
by pjrhuiz to webdesign web html design
by tf124 to pictures webdesign
by elmerz to web design tips list links
by rodrigolima to system:unfiled
by paulo_melchiori to webdesign tools
by marq7 to resources design
by DancesWithWeezils to webdesign web
by Aliso to webdesign
by buddhajackson to webdesign
by inareb to links lists webdev webdesign web2.0 tutorials resources css web html
bobo.is.me.not.u to design font flash webdesign urces tutorials web
design reference resources webdesign
by stmartin to sammlung links archiv webdesign resources design ajax web tutorials 2.0 work resource reference grafik

Figure 1. Collaborative tagging in del.icio.us.

Unfortunately, there has been no adequate way to visualize this folksonomy other than tag clouds. A tag cloud, however, is just a representation of listing the top-k popular tags according to their frequencies, and this may not be useful to provide an intuitive summary of the whole folksonomy. Furthermore, it does not provide any information about the semantic relations between tags. Under this situation, if we are able to find the semantic relations between tags created through collaborative tagging and visualize them, it can help users understand the web metadata more intuitively. In this paper, we propose a technique, called FolksoViz, for automatically deriving the semantic relations between tags and for visualizing the derived relations on the screen.

The remainder of the paper is organized as follows. In section 2, we discuss the previous work related to the semantic relation extraction between terms. We then describe the proposed technique, FolksoViz, for deriving semantic relations between tags based on the Wikipedia corpus in section 3. Section 4 demonstrates the analysis and evaluation of the FolksoViz. Finally, in section 5, we conclude this paper.

2. RELATED WORK

A variety of approaches in computational linguistics and information retrieval communities have been proposed to automatically extract the semantic relations between terms. It, however, still remains as a challenging task because it is not so easy for machines to understand the semantics in human language.

Researches on the similarity between terms have been made most widely. Among them, Lin's similarity measure [2] is accepted as a good indicator of how similar two terms are to each other. According to his similarity theorem, the similarity between two objects A and B can be measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are:

$$sim(A,B) = \frac{\log P(common(A,B)}{\log P(description(A,B)}$$

In his following work [3], he developed his idea based on the distributional pattern of words. He used the dependency triples, which consist of two words and the grammatical relationship between them.

Researches on deriving a hierarchical organization of concepts have also been made. Sanderson and Croft proposed a statistical model [4] to derive subsumption term pairs from a set of documents based on a type of term cooccurrence. Term *x* subsumes term *y* if most of the documents which *y* occurs in are a subset of the documents which *x* occurs in:

$$P(x|y) >= 0.8,\ P(y|x) < 1$$

3. DERIVING AND VISUALIZING SEMANTIC RELATIONS BETWEEN TAGS

This section describes the proposed technique for deriving semantic relations between tags. To apply our rules and models and to derive the semantic relations between del.icio.us tags, we use Wikipedia [5] as a corpus. Wikipedia is an online encyclopedia which gets the most popularity among internet users. Wikipedia is known as the best reflection of 'the wisdom of the crowds' or 'the collective intelligence' in Web 2.0 because anyone can be an author of any pages on Wikipedia. And, at the same time, it provides the high-quality information. Furthermore, it is currently known to be the largest knowledge repository on the Web. It contains much information about the words that are not defined in a dictionary, e.g. technical terms or newly created words on the Web. We can be sure that it covers almost all concepts that exist in the world. These interesting features of Wikipedia satisfy the qualifications of a good balanced corpus.

In this context, we need two assumptions:

Assumption 1. Wikipedia contains the information full enough to describe all del.icio.us tags.

Assumption 2. A Wikipedia page is a basic unit of context for describing about a topic.

Details of Assumption 2 will be covered in subsection 3.3.

Figure 2 shows the whole process of folksonomy visualization. The semantic relations between del.icio.us tags are derived using the Wikipedia corpus, and the tags and the derived relations are visualized on the screen.

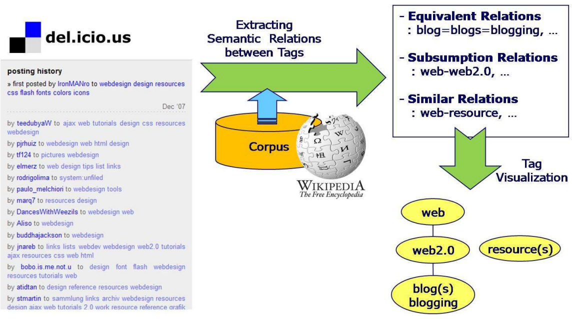

del.icio.us
posting history
» first posted by IronMANro to webdesign design resources css flash fonts colors icons
Dec '07
by teedubyaW to ajax web tutorials design css resources webdesign
by pjrhuiz to webdesign web html design
by tf124 to pictures webdesign
by elmerz to web design tips list links
by rodrigolima to system:unfiled
by paulo_melchiori to webdesign tools
by marq7 to resources design
by DancesWithWeezils to webdesign web
by Aliso to webdesign
by buddhajackson to webdesign
by jnareb to links lists webdev webdesign web2.0 tutorials ajax resources css web html
by bobo.is.me.not.u to design font flash webdesign resources tutorials web
by atidtan to design reference resources webdesign
by stmartin to sammlung links archiv webdesign resources design ajax web tutorials 2.0 work resource reference grafik
Extracting Semantic Relations between Tags
Corpus
WIKIPEDIA
The Free Encyclopedia
- Equivalent Relations
: blog=blogs=blogging, ...
- Subsumption Relations
: web-web2.0, ...
- Similar Relations
: web-resource, ...
Tag Visualization
web
web2.0
resource(s)
blog(s)
blogging

Figure 2. The process of folksonomy visualization

We introduce a 4-step process for our folksonomy visualization. Step 1 is finding equivalence relations between tags. For example, we find out that *blog*, *blogs*, and *blogging* are the equivalent tags. Step 2 is deriving subsumption relations between tags. For instance, we find out that *web* subsumes *web2.0* and *html* subsumes *css*. Step 3 is clustering similar tags, e.g. *apple*, *mac*, *leopard*, and *osx* are treated as the similar tags, and clustered in a same cluster. Finally, step 4 is visualizing all of the relations we have found. Each step will be covered in detail in the following subsections.

3.1 Finding Equivalent Relations

Before we proceed to find the equivalent relations between tags, we need to define the *equivalence*. It is given by the following:

Definition 1. Two tags are said to be *equivalent* iff their meanings are exactly the same, or they refer to exactly the same target. And then the two tags form an *equivalence relation*.

After making a careful examination of the del.icio.us tag data, we found out that there exist four types of equivalent relations between tags. The first and the most common type is the equivalence of singulars and plurals. For example, *computer* - *computers*, *utility* - *utilities*, and *woman* - *women* are the singular-plural pairs, each of which refers to exactly the same target. The second type is the equivalence of verbs or adjectives and their nominalized nouns. For instance, *blog - blogging* and *virtual - virtualization* are the pairs which have exactly the same meaning but different parts-of-speech. The third type is the equivalence of nouns and their abbreviations. For example, *newyorkcity* is equivalent to *nyc*, and *administrator* is equivalent to *admin*. Abbreviations of words are very common on the Web. And the last type is - (hyphen) or _ (underscore) embedded nouns. Example pairs of this type are *e-learning - elearning* and *social_networking - socialnetwroking*.

In order to find these equivalence relations, we apply simple rules and heuristics according to their types. The first equivalence type of singulars and plurals can be found by checking whether adding *–s*, *-es*, or *–ies* to the end of one tag makes the other tag. Of course, we have irregular plurals in our tag data set. In this case, consulting a predefined dictionary for irregular plurals can be one possible solution. But our observation reveals that irregular plurals are very rare when used as tags. The second equivalence type of nominalization can be found by checking whether adding *–ing*, *-ation*, or *–y* to the end of one tag makes the other tag. Of course, again, this rule may have exceptions, but they also turned out to be rare. The third equivalence type of abbreviation can be found by checking whether each character of one tag is located in the other tag in a same order. For example, *nyc* contains the characters *n*, *y*, and *c*, and they are located in *newyorkcity* in a same order. Unfortunately, this rule is not perfect when we are to find the abbreviation. And the last equivalence type of – or _ embedding can be easily found by checking whether removing – or _ from one tag makes the other tag. As illustrated, all of these rules and heuristics are not perfect to find the equivalence relations. Our examination, however, shows that these simple rules manage to find almost all of the equivalence relations in the del.icio.us tag data without the help of more powerful linguistic approaches.

3.2 Deriving Subsumption Relations

Again, we need to define what a *subsumption* is.

Definition 2. Tag *x* is said to *subsume* tag *y* iff tag *x* refers to a more general concept than tag *y*. And then the two tags form a *subsumption relation*.

In order to derive subsumption relations between tags, we used the model proposed in our previous work [6]. The model adopted the basic idea from Sanderson and Croft [4]. To reflect the characteristics of del.icio.us tags and Wikipedia, however, we made a slight modification to the original model. It is defined as follows, for two tags *x* and *y*, *x* subsumes *y* iff

$$TF(y|Wiki(x)) < TF(x|Wiki(y)),\ \mu < TF(x|Wiki(y))$$

where *Wiki(a)* is the Wikipedia texts where tag *a* appears, *TF(b|Wiki(a))* is the term frequency of tag *b* on the *Wiki(a)*, and μ is the threshold value that is determined empirically. In other words, tag *x* subsumes tag *y* iff 1) *x* is more frequent on the Wikipedia texts of *y* than *y* is on the Wikipedia texts of *x*, and 2) *x* occurs on the Wikipedia texts of *y* to some degree.

3.3 Clustering Similar Tags

As a last step to derive the semantic relations between tags, we find the similarity relations and cluster the similar tags. The definition of *similarity* is given by the following:

Definition 3. Two tags are said to be *similar* to each other iff they share a certain degree of common characteristics. And then the two tags form a *similarity relation*.

Note that the definition of similarity is slightly different from that of computational linguistics, in which the similarity is usually defined as how similar the direct meanings of two words are. Our definition of similarity, however, is how much they share the common characteristics. For instance, *apple*, *mac*, *leopard*, and *osx* look no similar to one another when only their direct meanings are taken into account. But, in fact, they share a considerable amount of common concepts, i.e. they are related to the operation system of Macintosh. Our definition of tag similarity relation is quite reasonable in that the tags attached to a web document are not likely to be directly similar to one another, but many of them share common characteristics.

In this context, we propose a new measure for tag similarity, adopted from Lin's original idea of similarity measure. Our similarity measure is, again, based on the Wikipedia corpus. The Wikipedia similarity between two tags, t_1 and t_2, is the following:

$$sim_{wiki}(t_1,t_2) = \frac{IC_{wiki}(t_1,t_2)}{IC_{wiki}(t_1) + IC_{wiki}(t_2)}$$

where *ICwiki(t)* is the information content of tag *t* in Wikipedia, i.e. the logarithm of the number of Wikipedia pages which include the word *t*, and *ICwiki(t_1, t_2)* is the information content of tag t_1 and tag t_2 in Wikipedia, i.e. the logarithm of the number of Wikipedia pages which include both the word t_1 and the word t_2. Here, it is necessary to recall the Assumption 2. A Wikipedia page is assumed to be a basic unit of context for describing about a topic. This is why we define the information content of a tag as the number of Wikipedia pages which include the tag.

Let us calculate how similar the two tags *design* and *diagram* are. The numbers of Wikipedia pages which include the word *design* and *diagram* are 230100 and 13616 respectively. And the number of Wikipedia pages which include both *design* and *diagram* is 3693. Therefore, *ICwiki(design) = log(230100) = 12.34626, ICwiki(diagram) = log(13616) = 9.51900,* and *ICwiki(design, diagram) = log(3693) = 8.21419*. According to the similarity measure defined above, this leads to the following:

$$sim_{wiki}(design, diagram) = \frac{8.21419}{12.34626 + 9.51900} = 0.37567$$

So, we can conclude that *design* and *diagram* are similar.

After calculating the similarities of the possible pairs of all tags, we cluster the tags according to their similarities. In order to cluster the similar tags, we use the MCL (Markov Clustering Algorithm) [7]. The MCL is an unsupervised clustering algorithm for graphs based on the Markov assumption. When our model is applied to the MCL, a tag is represented as a node, a relation is represented as an edge, and a similarity is represented as an edge weight. After applying the MCL to our model, we get several clusters of similar tags.

3.4 Visualization

The final step is to visualize all of the relations (i.e. equivalence, subsumption, and similarity relations) that were derived through the previous steps. Our goal is to effectively and intuitively visualize the semantic relations between tags. This requires the following seven principles for a successful visualization.

1. All tags and their relations should be displayed on one screen, while the displayed tags are the ones of interest (i.e. tag frequency > 10).
2. A node represents a tag and an edge between two tags represents a relation.
3. A font size is assigned to each node according to its tag frequency.
4. The equivalent tags are treated as one single tag and, thus, contained in one single node. Their tag frequencies are also summed to one value of tag frequency.
5. Tags that belong to the same cluster have the same color.
6. In handling transitivity, we maintain every edge of subsumption relations no matter when they are transitive or not. This is because some subsumption pairs are not transitive, e.g. *mac* < *apple*, *apple* < *corporation*, but, *mac* !< *corporation*.
7. Each node has a hyperlink for a tag search

According to these principles, all of the tags of interest and their relations are displayed as a directed graph on the screen by using the JGraph [8]. We named it FolksoViz, which means the folksonomy visualization.

4. ANALYSIS AND EVALUATION

4.1 Analysis

Figure 3 illustrates an example output of a FolksoViz cluster that visualized the del.icio.us tags assigned to a URL (http://www.exploratree.org.uk/). We present only one cluster from the whole picture due to space limitation. As shown in the figure, all derived relations were well displayed according to the seven principles in the previous subsection. The equivalent tags, such as *visual*, *visualisation*, and *visualization*, were gathered in one node. Subsumption relations also look good, e.g. *web* subsumes *web2.0* and *design* subsumes *graphic* (or *graphics*). And all nodes of similar tags were colored yellow because they formed a cluster. It is easy to notice that all tags in this cluster are related to the concept of web and design.

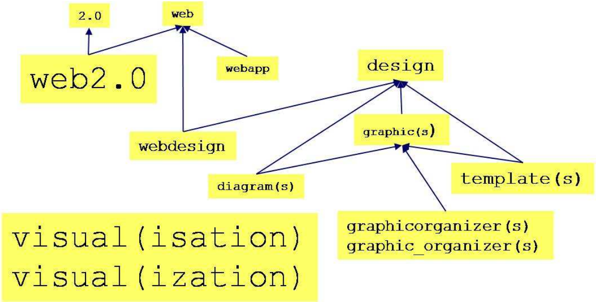


Figure 3. A cluster from a FolksoViz output.

4.2 Evaluation

The goal of our experiment is to figure out how correct the automatically derived semantic relations are. Unfortunately, we do not have any answer set that provides the correct relations between tags. This is mainly because tags on the Web are the special keywords, many of which are not defined in a dictionary or a thesaurus. One way to address this problem is a manual evaluation. For the first experiment, a group of 15 Ph.D. students were chosen as subjects, who were majoring in computer science and well-aware of a wide variety of technical terminologies. In other words, they were assumed to be the domain experts. The top-10 popular URLs and their tags were chosen from del.icio.us. Table 1 shows the basic information about those 10 URLs. From each of the URLs, 50 relations were chosen by random, i.e. total of 500 relations were chosen. For each relation, the subjects were asked to judge that the given relation of two tags looked a) Correct, b) Not correct, c) Inverted (in case of subsumption relation), or d) I don't know.

Table 1. del.icio.us Target URLs.

#	URL	Title	# of tags	# of taggers
1	http://fluidapp.com/	Fluid - Free Site Specific Browser for Mac OS X Leopard	8014	1902
2	http://www.sysresccd.org/	Main Page – SystemRescueCd	8355	1909
3	http://synergy2.sourceforge.net/	Synergy	7914	1873
4	http://www.exploratree.org.uk/	Exploratree - Exploratree by FutureLab	7810	1890
5	http://code.google.com/edu/	Google Code University - Google Code	7773	1900
6	http://www.shozu.com/portal/index.do	ShoZu	8030	1877
7	http://elgg.org/	Elgg.org	7737	1886
8	http://www.freenas.org/	FreeNAS: The Free NAS Server – Home	7830	1889
9	http://musicbrainz.org/	Welcome to MusicBrainz! - MusicBrainz	7829	1903
10	http://ccmixter.org/	ccMixter - Welcome to ccMixter	8142	1894

Table 2 shows the results. The high proportion of "Correct" (88.03%) is promising and means that the FolksoViz has a high precision. The proportions of "Not correct" (7.94%) and "Inverted" (1.03%) are fairly low. Some subjects answered with "I don't know" (3%). This may be because some relations were unobvious in judging from the tags alone.

Table 2. Results for answering to the questions for the first experiment (%).

URL #	Correct	Not Correct	Inverted	Don't Know
1	89.5	8.4	0	2.1
2	84.3	12.4	3.3	0
3	86.7	8.8	2.8	1.7
4	90.5	6.4	0	3.1
5	86.4	6.2	0	7.4
6	87.4	10	0	2.6
7	92.5	7.5	0	0
8	83.3	4.6	4.2	7.9
9	88	10.1	0	1.9
10	91.7	5	0	3.3
Avg.	**88.03**	**7.94**	**1.03**	**3**

The second experiment is to test whether the FolksoViz managed to find all of the real relations between tags. The subjects were given a set of tag pairs that the FolksoViz regarded as the pairs whose tags had no relation at all. If the subject thinks that if it is right for the pair to have no relation at all, he or she chooses a) No relation. But if he or she thinks that the pair has any relation, he or she chooses b) equivalence, c) subsumption, or d) similarity. Here, multiple choices are allowed among b), c), and d). From each of the 10 URLs, 50 relations were chosen by random, i.e. total of 500 relations were chosen.

Table 3 shows the results. The high proportion of 'No Relation' (91.87%) shows that the FolksoViz managed to find almost all relations that really exist, and this means that the FolksoViz has a fairly high recall. Besides 'No Relation', 'Similarity' scores the second highest proportion (7.46%). This may be because the similarity defined in Definition 3 was somewhat ambiguous to the subjects.

Table 3. Results for answering to the questions for the second experiment (%).

URL #	No Relation	Equivalence	Subsumption	Similarity
1	92.3	1.4	0	6.3
2	91.5	0	0	8.5
3	89.7	1.3	0	9
4	90.7	0	0	9.3
5	92.4	1.5	1.5	6.5
6	90.1	0	4.4	8.5
7	93.5	0	0	6.5
8	91.4	0	4.8	8.5
9	95.2	2.4	0	3.4
10	91.9	0	0	8.1
Avg.	**91.87**	**0.66**	**1.07**	**7.46**

5. CONCLUSION

We proposed a technique that automatically derives the three semantic relations, i.e. equivalence, subsumption, and similarity, between del.icio.us tags based on the Wikipedia corpus. FolksoViz managed to display the semantic relations between tags in an effective and intuitive way to accomplish the folksonomy visualization. We fully exploited the characteristics of Web 2.0: the collaborative tagging in del.icio.us and the collective intelligence in the Wikipedia.

6. ACKNOWLEDGMENTS

This research was supported by the Brain Korea 21 Project, the Ministry of Knowledge Economy, Korea, under the Information Technology Research Center support program supervised by the Institute of Information Technology Advancement. (grant number IITA-2008-C1090-0801-0031), and a grant (07High Tech A01) from High tech Urban Development Program funded by Ministry of Land, Transportation and Maritime Affairs of Korean government.

REFERENCES

[1] del.icio.us, http://del.icio.us.

[2] Dekang Lin, "*An Information-Theoretic Definition of Similarity,"* in Proceedings of the 15th International Conference on Machine Learning, pp. 296-304, 1998.

[3] Dekang Lin, "*Automatic Retrieval and Clustering of Similar Words,"* in Proceedings of the 17th international conference on Computational linguistics, pp. 768-774, 1998.

[4] Mark Sanderson and Bruce Croft, "*Deriving Concept Hierarchies from Text,"* in Proceedings of the 22nd ACM Conference of the Special Interest Group in Information Retrieval, pp. 206-213, 1999.

[5] Wikipedia, http://wikipedia.org.

[6] Kangpyo Lee, et al., "*Folksoviz: A Subsumption-Based Folksonomy Visualization Using Wikipedia Texts,"* in Proceedings of the 17th International Conference on World Wide Web, pp. 1093-1094, 2008.

[7] Stijn van Dongen, "*A Cluster Algorithm for Graphs,*" Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science, Amsterdam, the Netherlands, 2000.

[8] JGraph, http://www.jgraph.com.