FolksoViz: A Subsumption-based Folksonomy Visualization Using Wikipedia Texts

Kangpyo Lee
School of Computer Science & Engineering
Seoul National University
Seoul, Korea
kplee@idb.snu.ac.kr

Hyunwoo Kim
School of Computer Science & Engineering
Seoul National University
Seoul, Korea
hwkim@idb.snu.ac.kr

Chungsu Jang
School of Computer Science & Engineering
Seoul National University
Seoul, Korea
cschang@idb.snu.ac.kr

Hyoung-Joo Kim
School of Computer Science & Engineering
Seoul National University
Seoul, Korea
hjk@snu.ac.kr

ABSTRACT
In this paper, targeting del.icio.us tag data, we propose a method, FolksoViz, for deriving subsumption relationships between tags by using Wikipedia texts, and visualizing a folksonomy. To fulfill this method, we propose a statistical model for deriving subsumption relationships based on the frequency of each tag on the Wikipedia texts, as well as the TSD (Tag Sense Disambiguation) method for mapping each tag to a corresponding Wikipedia text. The derived subsumption pairs are visualized effectively on the screen. The experiment shows that the FolksoViz manages to find the correct subsumption pairs with high accuracy.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Storage and Retrieval

General Terms
Management, Measurement, Design, Human Factors

Keywords
Folksonomy, Collaborative Tagging, Subsumption, Visualization, Wikipedia, Web 2.0

1. INTRODUCTION
Folksonomy is 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'. If we refer to tags as web metadata that describe a web document, collaborative tagging is achieved collaboratively by many taggers who assign a list of keywords, or tags, as 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 will be open to the public in a form of posting history. Unfortunately, there have been no adequate ways to visualize this folksonomy other than using tag clouds. However, a tag cloud is just a representation of the top-k tags according to their frequency, and this may not be useful to provide an intuitive summary of the whole folksonomy. Furthermore, it does not provide any information about the relationships between the tags. Under this situation, if we are able to find semantic relationships 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 method, named FolksoViz, for deriving subsumption relationships between tags and visualizing the derived subsumption pairs on the screen.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. DERIVING SUBSUMPTION PAIRS

2.1 Basic Idea and Assumptions
In this section, we propose a statistical model for deriving subsumption relationships based on the co-occurrence and the frequency of each tag on Wikipedia[2] texts (Figure 1). Each tag is mapped to one corresponding Wikipedia text that describes the sense of the tag. Then, our metric for retrieving subsumption relationships is applied to every tag pair. The pair that has the calculated value over a predefined threshold is chosen as the subsumption pair. The reason why we chose the Wikipedia texts was that Wikipedia can act as the best reference to the senses of each tag in del.icio.us (In fact, Wikipedia is known as the best reflection of 'the wisdom of crowds' or 'the collective intelligence'). For this purpose, we needed to make three assumptions. First, all the tags in del.icio.us are treated as nouns. Second, each tag in del.icio.us is mapped to at least one Wikipedia text. And lastly, the information from that Wikipedia text is good enough to fully describe the sense of the tag.

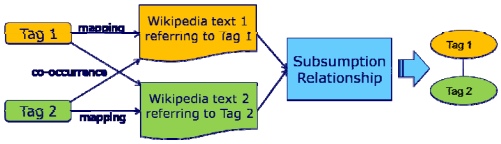

Figure 1. Deriving subsumption relationships.

2.2 Modeling for Deriving Subsumption Pairs
Our model adopted the basic idea from [3]. However, to reflect the characteristics of del.icio.us tags and Wikipedia, we made a slight modification to the original model. It is defined as follows, for two tags, x and y, x subsumes y if

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

where $Wiki(a)$ is the Wikipedia text that tag a is mapped to, $TF(b|Wiki(a))$ 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 if 1) x is more frequent on the Wikipedia text of y than y is on the Wikipedia text of x, and 2) x occurs on the Wikipedia text of y to some degree. And, empirically, the quality was best when μ was 0.01.

2.3 Tag Sense Disambiguation (TSD)

One of the important steps that need to be applied to the aforementioned model is to find a Wikipedia text that a tag will be mapped to. Considering that a word can have a number of distinct senses, it is essential to find the right Wikipedia text that best describes the sense of a tag. For example, a tag 'apple' may refer to either a sort of fruit or the Apple Incorporation. Therefore, we should determine which Wikipedia text will best refer and correspond to a tag. This kind of work is called word sense disambiguation (WSD). However, we renamed it as tag sense disambiguation (TSD). The basic idea of TSD is that a sense of a tag can be determined by the help of its neighboring tags (Figure 2). This is plausible because those tags that are attached to the same target are very likely to be related to each other in their meanings. We choose among the several senses the sense whose Wikipedia text records the largest sum of term frequencies of its neighboring tags. The detailed algorithm will not be introduced here due to the space limitation.

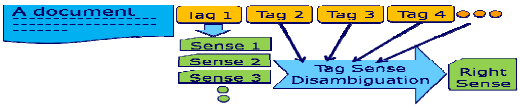

Figure 2. TSD using the neighbor tags.

3. FOLKSONOMY VISUALIZATION

3.1 Principles for Folksonomy Visualization

The retrieved subsumption pairs should be visualized so that the readers may understand it intuitively. Thus, we set five principles for effective folksonomy visualization. 1) All subsumption pairs should be displayed on one screen, and the displayed tags are the top-50 tags which users are most interested in. 2) The whole structure is a directed acyclic graph (DAG) since a tag can have more than one parent. 3) We assign a larger font size to a node whose tag has higher tag count (as in the case of tag clouds). 4) In handling transitivity, we maintain every edge of pairs regardless of when they are transitive or not since the subsumption relationships are not always transitive. 5) Each node has a hyperlink for a tag search (as in the case of tag clouds).

3.2 Analysis

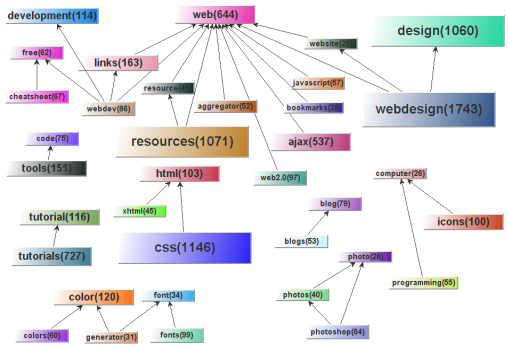

Figure 3. An example screenshot from the outputs.

According to the five principles, the subsumption pairs were visualized using JGraph[4]. Figure 3 shows an example of a screenshot that visualizes the del.icio.us tags attached to a URL regarding the web design. The figure shows that FolksoViz generally exhibits fairly good performance, e.g., 'web' subsumes 'web2.0', 'design' subsumes 'webdesign', and 'html' subsumes 'css'. However, some pairs may look a little awkward, e.g., 'code' subsumes 'tools' and 'free' subsumes 'webdev'. This may occur when the two tags are so closely interrelated that they often appear with each other. Moreover, FolksoViz is unable to handle singular-plural tags, e.g., 'resource' subsumes 'resources', 'color' subsumes 'colors', and 'blog' subsumes 'blogs'.

4. EVALUATION

The goal of our experiment is to figure out how correct the automatically derived subsumption pairs are. To achieve this goal, 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 (They were assumed to be the domain experts). The target 10 URLs were chosen from the top-10 popular topics of del.icio.us. From each of the URLs, 30 subsumption pairs were chosen by random (Total of 300 pairs were chosen). For each pair, the subject was asked to judge that the subsumption relationship of two tags looked a) Correct, b) Inverted, c) Synonymous, d) Not correct, but related, e) Neither correct nor related, or f) I don't know. Table 1 shows the results. The high proportion of "Correct" (58.4%) and the low proportion of "Inverted" (1.3%) and "Neither correct nor related" (7.8%) are promising. The proportions of "Synonymous" (4.1%) and "Not correct, but related" (14.8%) show the limitations in our method. And, many subjects answered with "I don't know" (13.8%), mainly because we could not handle the singular-plural tags, and partly because some relationships were unobvious in judging from the tags alone.

Table 1. Results for answering to the questions (%).

#	Topic	Correct	Inverted	Synonymous	Not correct, but related	Neither correct nor related	Don't know
1	mac	66.7	6.7	0.8	11.7	5.8	8.3
2	webdesign	54.2	0.8	0	13.3	10.8	20.8
3	music	63.3	0	4.2	16.7	3.3	12.5
4	web2.0	65.8	0	1.7	23.3	0.8	8.3
5	software	55.0	0.8	0	17.5	19.2	7.5
6	video	52.5	1.7	6.7	19.2	7.5	12.5
7	games	61.7	0	0.8	11.7	12.5	13.3
8	shopping	39.2	0.8	18.3	15.8	5.8	20.0
9	education	65.8	0	5.0	7.5	5.0	16.7
10	business	60.0	1.7	3.3	10.8	6.7	17.5
Avg.		**58.4**	**1.3**	**4.1**	**14.8**	**7.8**	**13.8**

5. CONCLUSION

FolksoViz managed to display the subsumption relationships between tags in an 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 in 2008 and the Ministry of Information and Communication, Korea, under the College Information Technology Research Center Support Program, grant number IITA-2007-C1090-0701-0031.

7. REFERENCES

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

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

[3] 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.

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