Mining Semantically Similar Tags from Delicious

딜리셔스에서 유사태그 추출에 관한 연구

Kwan Yi *

ABSTRACT

The synonym issue is an inherent barrier in human-computer communication, and it is more challenging in a Web 2.0 application, especially in social tagging applications. In an effort to resolve the issue, the goal of this study is to test the feasibility of a Web 2.0 application as a potential source for synonyms. This study investigates a way of identifying similar tags from a popular collaborative tagging application, Delicious. Specifically, we propose an algorithm (FolkSim) for measuring the similarity of social tags from Delicious. We compared FolkSim to a cosine-based similarity method and observed that the top-ranked tags on the similar list generated by FolkSim tend to be among the best possible similar tags in given choices. Also, the lists appear to be relatively better than the ones created by CosSim. We also observed that tag folksonomy and similar list resemble each other to a certain degree so that it possibly serves as an alternative outcome, especially in case the FolkSim-based list is unavailable or infeasible.

초 록

자연언어에서 유사어의 처리는 사람과 컴퓨터간의 의사소통에 적지않은 장애가 되어왔고, 이는 사용자의 임의적 단어사용에 기반을 두고 있는 웹 2.0 애플리케이션, 특히 소셜태깅분야에 있어서 그 장애의 정도가 더 심각해질 수 있다. 본 연구는 한 대표적인 웹 2.0 애플리케이션에서 자동 유사어 추출에 관한 문제를 다루고 있다. 더 구체적으로, 가장 널리 사용되는 소셜북마킹 애플리케이션인 딜리셔스를 기반으로, 유사태그를 추출하는 방법(FolkSim)을 제시하고자 한다. 제시한 방법의 평가를 위하여, 문서유사도의 측정을 위해서 쓰여진 고전적 벡터모델에 의거한 유사태그를 추출하는 방법(CosSim)과 그 결과들을 서로 비교분석하여 보았다. 몇 가지 면에서 FolkSim가 더 나은 결과 산출해내는 증거들이 관찰되어졌다. 또한, FolkSim 방법에 의한 유사태그가 만들어지지 않는 경우에 대비하여, 그 대안 또한 제시하고 있다.

Keywords: tag similarity, Delicious, synonym extraction, Folksonomy, web mining
유사태그, 딜리셔스, 유사어 추출, 폭소노미, 웹 마이닝

---

* Assistant Professor. School of Library and Information Science, University of Kentucky, USA (kwan.yi@uky.edu)

▪ Received : 16 May 2009 ▪ Revised : 21 May 2009 ▪ Accepted : 2 June 2009
▪ Journal of the Korean Society for Information Management, 26(2): 127-147, 2009.
〔DOI:10.3743/KOSIM. 2009.26.2.127〕

1. Introduction

Different terms are often used to describe the same concept or object (synonym issue), and also the same term is often used for different concepts or objects (polysemy issue). The vocabulary issues have been a barrier in human-computer communication and natural language understanding (Furnas et al. 1987). The issue has more significantly rekindled with the recent popularity of Web 2.0 applications due to the creation and circulation of vocabularies by general users.

As a type of Web 2.0 application, collaborative tagging applications (or social bookmarking applications) have been rapidly growing with large numbers of users. A collaborative tagging service allows information users to create or introduce online resources, and to organize and share them with others through the Web. Examples of collaborative tagging services include: Delicious (http://delicious.com) for Web resources, Flickr (http://www.flickr.com) for pictures, YouTube (http://youtube.com) for video clips, to name only a few.

Social bookmarking applications support a new way of sharing resources with the public. However, they walk in the opposite direction from the traditional way of indexing and organizing information using controlled vocabularies. In such an application, its participants or registered users introduce and annotate resources into the application and label them with any freely chosen arbitrary words, called tags. The applications may help assign labels by providing some suggestive tags. However, there are no formal specific guidelines or rules to be applied in tagging. Instead, a collection of tags is a user-centered, user-created, and user-oriented vocabulary - selected by information users, not information professionals. As a consequence of the diversity of users' choices, the vocabulary issues are greatly magnified in social tags.

To deal with the vocabulary issue, especially synonym, in Web 2.0 context, this study investigates a way of collecting similar tags from a popular collaborative tagging application, Delicious. Specifically, we propose an algorithm for measuring the similarity between social tags from Delicious and evaluate the quality of the proposed method in comparison to a popular similarity method. The goal of this study is to test the feasibility of a Web 2.0 application as a potential source for synonyms.

2. Related Work

2.1 Semantically Similar Terms from Domain Vocabularies

The conventional approach for alleviating the vocabulary issue is to automatically identify terms that are semantically similar or related to each other on the basis of corpora. Three different types of corpora have been utilized: newspapers(Lin 1998), web documents(Turney 2002), and dictionaries(Wu & Zhou 2003). The underlying assumption of this approach is that similar terms co-occur

in similar documents. Traditional information retrieval (IR) techniques, such as cosine similarity measure, and term discrimination values, such as mutual information, were applied in this context(Chen & Lynch 1992, Crouch 1990). In the web-based approach, binary or weighting term frequency, co-occurrence between terms, and distance measure among terms are among popular methods(Lin et al. 2003, Turney 2001). The dictionary-based approach assumes that synonyms or near-synonyms of a term co-occur in its definition in dictionaries. The relationships between terms and definitions are represented by a graph (dictionary graph), and web graph methods such as Kleinberg's hubs and authorities is employed in this approach(Jannink & Wiederhold 1999). The limitation of this approach for social tagging is that user-centered vocabulary may be not listed in conventional domain vocabularies.

2.2 Semantically Similar Terms from Social Tags

To examine the extent of tag inconsistencies, Choy and Lui (2006) conducted a study of the similarity of pairs of tags. In the study, more than 300,000 tag tuples, each of which consists of a web resource, a user, and a tag, were collected from Delicious. To assess the tag similarity, the Latent Semantic Analysis was applied to a tag-url matrix constructed with the collected tuples. For the result of the proposed algorithm, only a few selected cases were reported and no further assessment was conducted. Hotho el at. (2006) proposed a ranking method of shared resources by using an adapted version of the Google's PageRank. It does not directly measure the similar tags, but it can be done indirectly by comparing query tags submitted for ranking and tags assigned to some top-ranked resources. Tag suggestion is an application for similar tags. A recent study by Garg and Weber (2008) presented a method of tag suggestion for Flickr that aimed at drawing some related tags. In the study, a best result was obtained with a method in which previously assigned tags by others were highly weighted. Yi (2008) proposed co-occurrence-based and correlation-based methods to discover similar terms for terms that are not listed on or not available in a conventional dictionary, and suggested a correlation-based method with the co-occurrence frequency of 0.7 to obtain a large pool of similar terms.

2.3 Clustering Semantically Similar Terms

Begelman, Keller, and Smadja (2006) proposed a tag-clustering algorithm to identify semantically related tags. An undirected weighted graph of a tag is built in which the frequency of co-occurrences of a pair of tags is assigned to an edge associated with the tags. The graph is partitioned into clusters recursively on the basis of the spectral bisection algorithm(White & Smyth 2005). The top N similar tags in each cluster are selected for the semantically related tags for a given tag. However, these studies

do not describe how similar tags from different clusters are compared, as all clusters are associated with a same tag. Also, except for reporting some cases, these studies do not analyze the proposed algorithm in a systematic way.

The study of tag-clustering can be directly applicable to the study of mining similar tags. In the co-occurrence-based graph, similarity between tags is based solely on the context-free tag occurrence, whereas in our approach, tag similarity is based on context-bounded tag relationship via resource folksonomy (see Section 4).

3. Folksonomies in Delicious

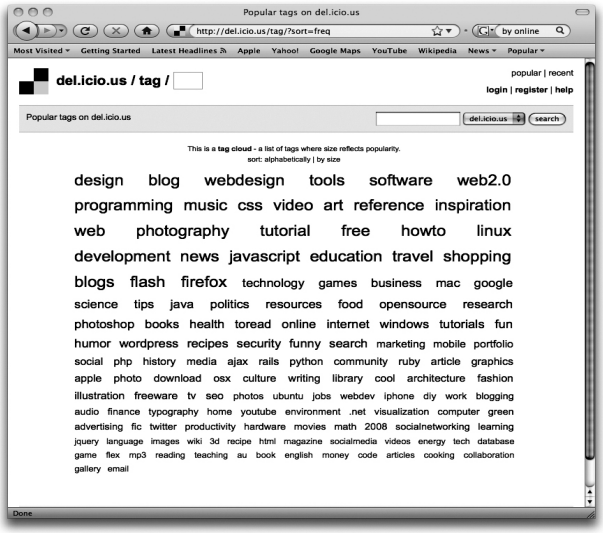
A tag can be considered a string of any characters such as alphanumerics and special symbols. In Delicious, a space is not allowed as a part of a tag. In this study, the term *folksonomy,* coined by Thomas Vander Wal (2004), refers to the collection of all tags created and assigned. A tagging event or activity in Delicious can be characterized by a tuple of (r_i, p_j, t_k), $r_i \in R$, $p_j \in P$, $t_k \in T$ where R, P, and T are a set of resources, participants, and tags, respectively. A tag event (r_i, p_j, t_k) symbolizes a resource (r_i) annotated by a participant (p_j) with a tag (t_k).

Two different levels of folksonomies are available in Delicious[1]: *global-level* and *resource-level.* A global-level folksonomy refers to an entire set of tags used in Delicious. The top popular tags in the global-level of folksonomy are displayed in <Figure 1>, where the more popular a tag is in Delicious, the larger the size of the tag. The “design, “blog, and “webdesign tags were the three most popular tags at the time that the screenshot was taken. A resource-level folksonomy refers to a complete set of all tags assigned to a single resource. <Figure 2> illustrates a part of the resource-level folksonomy in Delicious for the resource titled “Folksonomy - Wikipedia, the free encyclopedia. The example indicates that 1741 people participated in tagging. As shown in the “common tags section, the word “Folksonomy was the most common tag selected by 291 different people, and then “tagging and “web2.0 by 181 and 175 people, respectively. All the tags used for the resource are available from the “posting history section located right below the “common tag section. Note that tagging activity is conducted in real time so that the results shown in the figures reflect only the snapshot at the time taken.

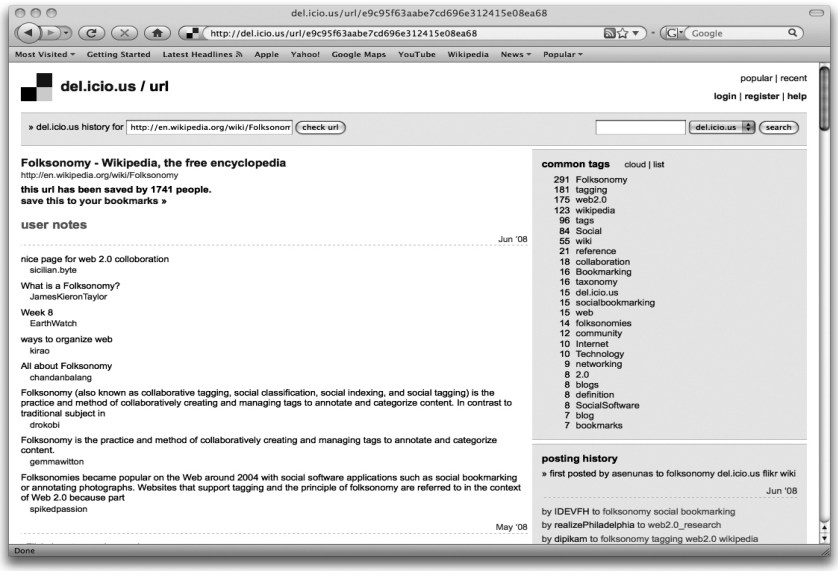
The following notations will be used throughout this paper: (1) Let $F_r = \{(t, p) \mid t$ is a tag and p is the number of participants using the tag $t\}$ be a resource-level folksonomy for a resource r. That is, a resource-level folksonomy can be represented by a collection of pairs of tags and their corresponding tag frequency; (2) Let p_{max} in F_r be a value p of the greatest value in F_r, and let

1) Delicious website had a change not only in the official URL from http://del.icio.us/ to http://delicious.com/ but also in interface as of 31 July 2008 (http://blog.delicious.com/blog/2008/07). Since we collected data from Delicious on 15 May 2008, the screenshots shown in 〈Figure 1 & 2〉 were copied from the old version of the interface from http://del.icio.us.

t_{max} in F_r be a tag t paired with the p_{max} in F_r; (3) The t_{max} in the folksonomy shown in <Figure 2> is “folksonomy and the corresponding p_{max} is 1,741; and (4) Let P_r in F_r be the number of all different participants who have been involved in creating the folksonomy F_r.



〈Figure 1〉 An example of a global-level folksonomy in Delicious



〈Figure 2〉 An example of a resource-level folksonomy in Delicious for the resource titled “Folksonomy - Wikipedia, the free encyclopedia”

4. Proposed Approach for Deriving Synonyms

The proposed approach is composed of three steps: first, collect relevant folksonomies for a target term; second, merge folksonomies collected in the previous step into a single folksonomy; third, measure the similarity of terms. To obtain tags, we used the popular social tagging application, Delicious. Delicious is one of the very popularly used social bookmarking applications and it has been adapted for the tag research in many previous research. However, the real reason for the choice lies in that Delicious is the almost unique social tagging application that covers any generic web resources rather than being limited to any specific type of resource.

4.1 Collecting Relevant Folksonomies for a Target Term

Assumption 1. Given a resource-level folksonomy F_r, the tag t_{max} in F_r is considered the user-assigned index term that best represents the content of the resource r, under the condition with the current participants P_r in F_r. Since a folksonomy changes dynamically over time, assumption 1 might hold only for a certain number of participants. That is, tag A becomes the value of t_{max} in F_r at time T_A, but tag B can be the value of tag t_{max} at time T_B.

Does any threshold value of participants exist, beyond which a tag is expected to stay as the t_{max} (the most frequently occurring tag) in a folksonomy? What would be the value for P_r in a folksonomy F_r? In relation to this question, Golder and Huberman (2006) claimed, "Empirically, we found that, usually after the first 100 or so bookmarks [tags], each tag's frequency is a nearly fixed proportion of the total frequency of all tags used. In addition, we conducted a pilot study with Delicious concluding that a participant assigns approximately two tags on average. Based on these two empirical facts, we set 50 to be the threshold value for P_r, and it gives rise to the following definition:

Definition 1. Let a set of resource-level folksonomies $\Phi_w = \{F_1, F_2, \cdots, F_n\}$ be defined as a *relevant* folksonomy set for a target term w when the following two conditions are satisfied: (1) the value of tag frequency P_i of F_i, $1 \leq i \leq n$, is equal to or greater than 50, and (2) the tag T_{max} of F_i must be w.

4.2 Merging folksonomies

Once some resource-level folksonomies for a single term w are collected, we intend to build a single folksonomy Ψ_w for the tag w from the collected folksonomies Φ_w in the following steps: 1) normalizing each folksonomy in the set $\Phi_w = \{F_1, F_2, \cdots, F_n\}$, and 2) combining the normalized folksonomies into one single folksonomy.

The first step (normalization): In this step, tag frequencies in a folksonomy need to be normalized. This step is necessary because non-normalized

same tag frequencies appearing in different folksonomies must contain different implications. That is, the frequency value five in a folksonomy with only five people involved should be quite differently interpreted from the one in another folksonomy with 500 people involved. Given a folksonomy F_i in the set Φ_w, each tag frequency, say p_j, in the folksonomy F_i is normalized based on all n tag frequencies, as shown in the Formula (1) below. This is simply equal to the relative frequency for a term given a resource:

$$Normalization(p_j/F_i)=\left(p_j/\sum_{k=1}^{n}(p_k\in F_i)\right) \quad (1)$$

The second step (combination): In this step, multiple appearances of the same tag across different folksonomies are gathered and conflated to a single appearance. This step is conducted to create a single folksonomy associated with a term, w. Given the set $\Phi_w = \{F_1, F_2, \cdots, F_n\}$, the tag frequency of a tag t is calculated by Formula (2). The summation of normalized tag frequencies is divided by the total number of folksonomies in the set Φw, no matter whether the tag under consideration occurs in some or all of the folksonomies in the Φ_w. It is divided by the total number (n in this case), in order to give a penalty to a tag appearing in some rather than all folksonomies. As a result of this combination step, the multiple folksonomies $\Phi_w = \{F_1, F_2, \cdots, F_n\}$ are conflated to a single folksonomy Ψ_w:

$$Combination(t/\Phi_w)=\sum_{k=1}^{n}Normalization(t/F_k)/n \quad (2)$$

The following is an example of the merging process. Let $\Phi_A = \{F_1, F_2, F_3\}$ be a set of resource-level folksonomies for the tag A as stated in the Definition 1, with F_1 = {(A, 50), (B, 35), (C, 10), (D, 5)}, F_2 = {(A, 40), (C, 30), (E, 7), (F, 3)}, and F_3 = {(A, 30), (B, 12), (D, 8), (F, 6), (G, 4)}. A corresponding folksonomy Ψ_A is obtained by Formula (2) above as follows:

$$Combination(A/\Phi_A)=\frac{\frac{50}{50+35+10+5}+\frac{40}{40+30+7+3}+\frac{30}{30+12+8+6+4}}{3}=\frac{1}{2}$$

Similarly, $Combination(B/\Phi_A) \approx 0.18$, $Combination(C/\Phi_A) \approx 0.16$, $Combination(D/\Phi_A) \approx 0.06$, $Combination(E/\Phi_A) \approx 0.03$, $Combination(F/\Phi A) \approx 0.05$, $Combination(G/\Phi_A) \approx 0.02$. Therefore, we finally have the merged folksonomy Ψ_A = {(A, 0.5), (B, 0.18), (C, 0.16), (D, 0.06), (E, 0.03), (F, 0.05), (G, 0.02)} for the tag A.

4.3 Measuring the term similarity

As a consequence of the previous sections, we obtain a collection of folksonomies $\Psi = \{(\Psi_k) \mid \Psi_k$ is derived from Φ_k for tag k$\}$. In this section, we will measure the similarity between tags on the basis of the folksonomy set Ψ.

A way of measuring the document similarity is by measuring the conceptual distances between documents in IR (Baeza-Yates & Ribeiro-Neto 1999). Here we apply the similar approach taken

in IR for measuring the similarity between terms: the similarity of folksonomies associated with the terms is used for the similarity of terms. The metric of the folksonomy similarity between two tags, *A* and *B*, is as follows: Given two corresponding folksonomies, $\Psi_A = \{(t_{iA}, p_{iA}), 1 \leq i \leq k\}$ and $\Psi_B = \{(t_{iB}, p_{iB}), 1 \leq i \leq j\}$, with only two common tags across the two folksonomies (t_{1A} in Ψ_A is equivalent to t_{1B} in Ψ_B and t_{2A} in Ψ_A is equivalent to t_{2B} in Ψ_B), the similarity between the two folksonomies is calculated by Formula (3):

$$Similarity(\Psi_A, \Psi_B) = \frac{(p_{1A} + p_{2A}) + (p_{1B} + p_{2B})}{\sum_{i=1}^{k} p_{iA} + \sum_{i=1}^{j} p_{iB}} \quad (3)$$

Formula (3) states that the summation of the frequencies associated with common tags is divided by the summation of all the frequencies in the two folksonomy sets Ψ_A and Ψ_B. It is postulated in the formula that the more tags with higher frequencies co-appear in both sets, the more similar (close or related) the two associated tags are. The value of "similarity (the outcome of the formula) indicates the degree of the similarity with 1 for the most similarity and 0 for the least similarity. The similarity will be equal to 1 only if the two folksonomy sets are identical and there are only two common tags included in each set, whereas it will be equal to 0 only if there is no single tag in common between the two sets. Therefore, the calculated similarity value can be utilized as a degree of the tag similarity.

The following is an example of the similarity process. Let Ψ_X, Ψ_Y, and Ψ_Z be three folksonomies for the tags X, Y, and Z, respectively, where Ψ_X = {(X, 0.5), (Y, 0.4), (Z, 0.1)}, Ψ_Y = {(X, 0.4), (Y, 0.5), (Z, 0.1)}, and Ψ_Z = {(X, 0.1), (Y, 0.1), (Z, 0.8)}. The similarities between two out of the three folksonomies are calculated using Formula (3) above: Similarity(Ψ_X, Ψ_Y) = ((0.5 + 0.4) + (0.4 + 0.5)) / (1.0 + 1.0) = 0.9; Similarity(Ψ_X, Ψ_Z) = ((0.5 + 0.1) + (0.1 + 0.8)) / (1.0 + 1.0) = 0.75; Similarity(Ψ_Y, Ψ_Z) = ((0.5 + 0.1) + (0.1 + 0.8)) / (1.0 + 1.0) = 0.75. Among the three tags, the X and Y tags turn out to be more similar (the closest in distance) than any other pair of tags in terms of FolkSim.

5. Experiments

5.1 Data collection and process

We used the 140 most frequently used tags on Delicious posted on the site of http://del.icio.us/tag/, as of May 2008 (see <Appendix> for the full list of the tags shown in the decreasing order of popularity). We built associated folksonomy sets for the 140 most frequent Delicious tags as described in the Section of 4.1 through 4.3.

In collecting resource-level folksonomies for each tag, say X, we limited the collection to the first 5,000 folksonomies available from the site of http://del.icio.us/popular/X. The practical limitation is primarily due to the limitation of data available in Delicious. The size of resource-level

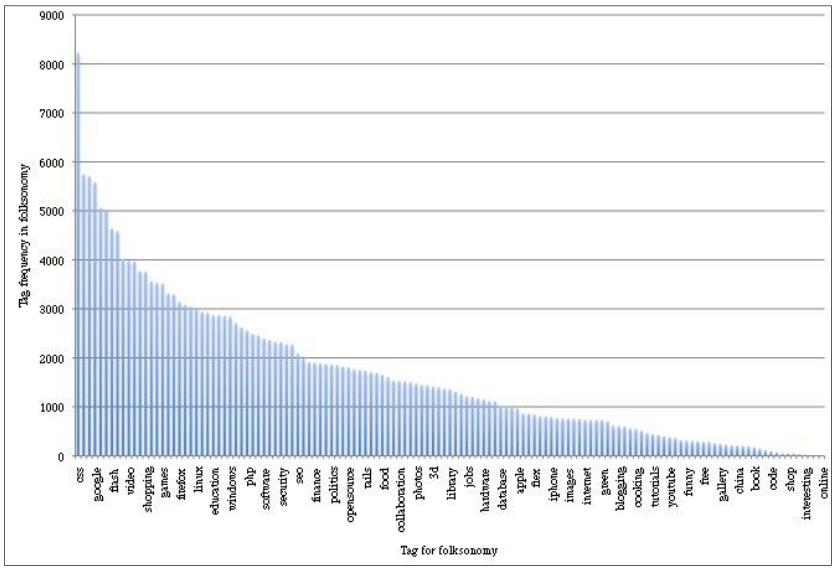
folksonomies (Φ) varies over tags with an average of 165.5 (738 for the tag “css that is the largest, only 1 for the tag “2008, and 0 for each of the following six tags - “article, “cool, “fic, “plugin, “toread, “webdev). There are seven tags with one or zero resource-level folksonomies, which are eliminated for the further process.

Also, it is important that we eliminate tags with very low tag frequency from resource-level folksonomies. This elimination process is necessary because they are apparently highly idiosyncratic annotation tags. This is somewhat similar to the process of eliminating low document frequency terms in IR. In the process of IR, a number between 2 and 5 is generally set as a threshold value for the lower boundary of document frequency, depending on different experimental environments. In this case, we set 2 as a threshold value for tag frequency after we conducted a series of experiments. That is, tags with less than or equal to 2 tag frequency are eliminated from the folksonomy set Φ. The elimination significantly reduces the average number of tags per folksonomy from 10,327 to 1,656 (approximately 16% of all tags).

5.2 Experimental results and discussion

5.2.1 Tag Folksonomies

<Figure 3> shows the distribution of the number of different tags over the folksonomies Ψ. The x-axis indicates a list of tags of the corresponding folksonomies, but the full list is not shown in the figure due to the limited space. The y-axis indicates the number of different tags in a folksonomy. The folksonomy (Ψ_{css}) for the tag “css contains 8,217

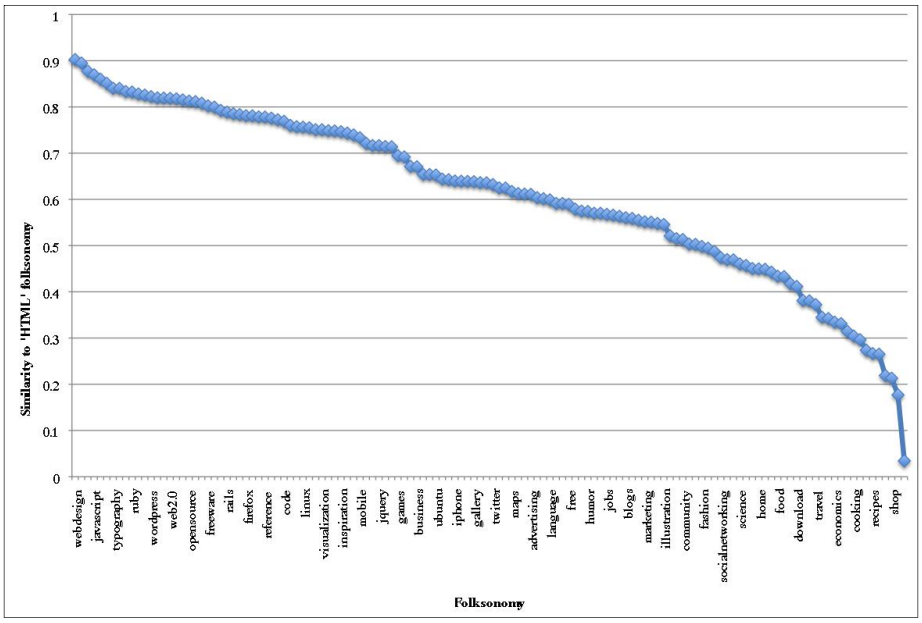


〈Figure 3〉 Distribution of the number of different tags appearing in tag folksonomies (Ψ)

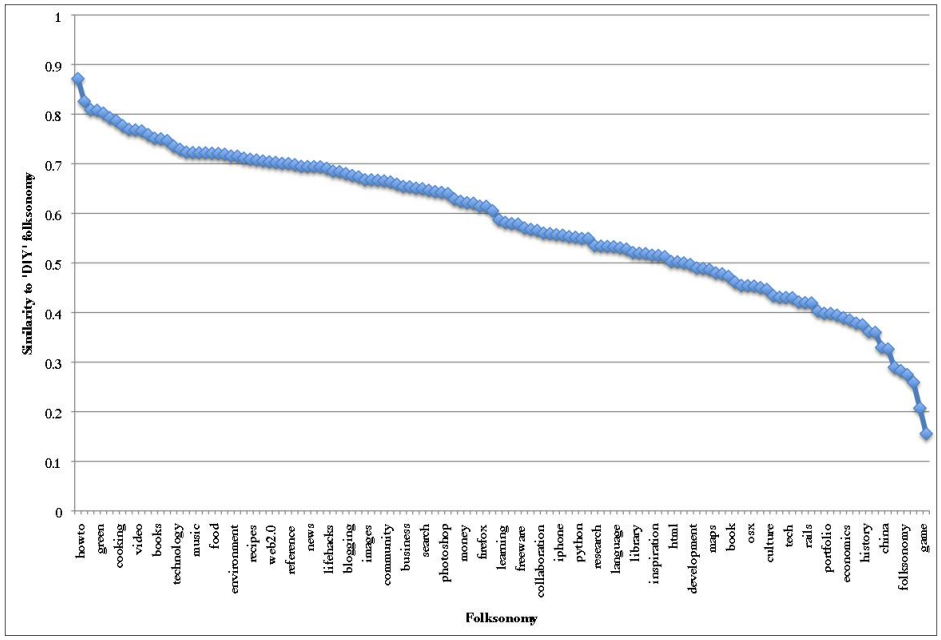
distinct tags, and the folksonomy (Ψ_{online}) of the tag "online has only 4 distinct tags, which are displayed at the extreme points on the x-axis of the figure. A folksonomy contains 1,655.8 tags on average. The "HTML, "DIY (Do It Yourself), and "Library folksonomies (Ψ_{HTML}, Ψ_{DIY}, $\Psi_{Library}$) were obtained by integrating 185, 101, and 108 resource-level folksonomies (Φ), respectively. The three distinct tags were intentionally chosen for the examination of different characteristics: HTML is selected as it is rather technology-related term, DIY is as it is not a full name term but an acronym, and Library as it is rather general term.

5.2.2 Analysis of the Results from the Proposed Approach

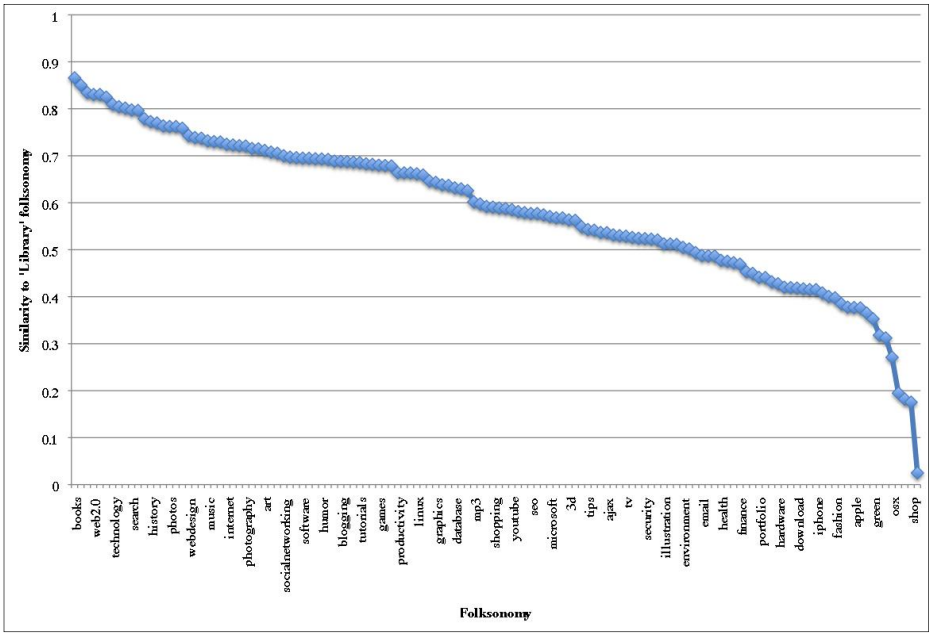
We used the method described in Section 4.3 to identify semantically similar terms. Note that the method yields a real number between 0 and 1 as a degree of the similarity between two associated terms - 1 for the maximum degree of similarity and 0 for the minimum degree of similarity. <Figures 4 through 6> illustrate the degrees of similarity of the 133 major tags against "HTML, "DIY, and "Library, respectively. In <Figure 4>, "webdesign, "css, "ajax, "javascript, and "php turn out to be the most similar terms for "HTML in decreasing order - all the tags are not shown in the figure, due to the limitation of the space in the x-axis. Except for "webdesign all the other four are web-related programming language names. On the opposite end, "game, "work, "shop, "online, and "recipe are the five least similar terms. <Figure 5> lists similar terms for "DIY, "Howto, "home,



〈Figure 4〉 Degree of similarity between "HTML" and the other main tags



〈Figure 5〉 Degree of similarity between "DIY" and the other main tags



〈Figure 6〉 Degree of similarity between "Library" and the other main tags

“shopping, “green, and “hardware are the five most similar ones, and “work, “jquery, “download, “game, and “online are the five least. For “Library as shown in <Figure 6>, “books, “education, “reference, “web2.0, and “research are among the most similar, and “osx, “game, “online, “shop, and “recipe are among the least similar.

An interesting result is that the three plots have a similar pattern, steadily decreasing to the middle range and sharply dropping at the end. However, they are different in percentage for a range of similarity level. For example, there are 5 (about 4%), 9 (about 7%), and 22 (about 17%) similar terms falling in the range of similarity from 1 to 0.8 for “DIY, “Library, and “HTML, respectively, and there are 33 (about 25%), 33 (about 25%), and 51 (about 39%) terms up to the range of 0.7 for the three cases. Does any similarity value exist that can be used as a threshold for an effective marginal line of similarity? Such a threshold appears to be somewhere above 0.8 for the three cases. In our demonstration, the identification of similar terms is limited only to a set of the 133 tags that are semantically quite apart from each other in many cases. Thus, any conclusive remark is too premature. Instead, we will see more examples of singular and plural cases, which are clearly very close in semantics: “blog and “blogs (0.79 in similarity), “book and “books (0.78), “game and “games (0.73), and “recipe and “recipes (0.76). All the similarity values from pairs of singular and plural terms are very close to 0.8, which is not very high. Nevertheless, the rankings of similarity values are very high in many cases: 1st with “games for “game 1st with “books for “book ; 2nd with “recipes for “recipe ; 3rd with “recipe for “recipes ; 7th with “blog for “blogs.

Note that the similar tag lists generated in this experiment were bounded by the 133 tags only. Thus, not all similar terms are most likely considered in assessing similar tags or terms. For example, the *Merriam-Webster's Collegiate Thesaurus,* 1988 edition, defines “archive(s) and “athenaeum as synonyms of “library, but these terms were not included in the 133 candidate tags.

5.2.3 Comparing FolkSim with CosSim

To analyze the proposed folksonomy-based method (FolkSim), we generated list of similar terms for several words, and compared them with the ones obtained from a cosine similarity method (CosSim). Cosine similarity has been popularly used for measuring the similarity between clusters (Dhillon & Modha 2001). A simple cosine similarity will be used as a baseline method in this study. In cosine similarity, a tag folksonomy (Ψ) is represented by a corresponding vector (called a tag vector). CosSim(Ψ_A, Ψ_B) is equal to the cosine of the angle between V_A and V_B, where V_A and V_B are the corresponding vectors for the tag folksonomies Ψ_A and Ψ_B, respectively. A distinct tag in a tag folksonomy refers to a distinct dimension in a corresponding tag vector, and its associated tag frequency is used as a weight of the vector dimension. The cosine similarity of two vectors (V_A and V_B) is given by: dot(V_A, V_B) / $\| V_A \|*\|$

$V_B||$, where dot(V_A, V_B) is $V_A[0]*V_B[0] + V_A[1]*V_B[1] + ...$, and $\|V_A\| = sqrt(V_A[0]^2 + V_A[1]^2 ...)$ and $\|VB\| = sqrt(V_B[0]^2 + V_B[1]^2 ...)$.

<Table 1-3> contains two similar lists (one for FolkSim and the other for CosSim) and associated similarities. It is worthwhile to note that the direct comparison between two similarities is not meaningful because individual list of similarities is obtained in its own condition. Therefore, the use of similarities should be limited only to the ranking of tags in each list.

The <Table 1> shows the ranked list of similar tags from FolkSim and the cosine similarity method for the tag "HTML. The first column of the table contains the results of CosSim method and the second column shows corresponding cosine similarity values: the higher the value is, the more similar to the "HTML the corresponding tag is. The tag "php (a computer scripting language) holds the first position while tags like "javascript (a scripting language popular in web environment) or "webdesign which appear related are also under the top 20 list. The tags "free, "email, "mobile, however, are on positions 9, 10, and 16, but are not even related to "HTML. The third and fourth columns of the first table display the results of our FolkSim method for the tag "HTML. The tag "webdesign, which is probably the most related tag among the list, is ranked on the top. A closer look at the list reveals that this list also contains some tags such as "mac and "wordpress (an open-source blog publishing application) that are not obviously related to "HTML. Intuitively, however, the list from FolkSim looks better than the one from CosSim, as the most preferable tag is top ranked in the FolkSim list, and the entirely non-related tags are higher ranked in the CosSim list.

The <Table 2> lists the results for the tag "DIY. Among the terms on the two lists from the FolkSim and CosSim methods, the tag "howto is apparently the most similar tag, which is top ranked on the FolkSim list, but is second on the CosSim list with "photography top-ranked instead. In the CosSim list, the tags "culture, "environment, and "architecture are ranked 5^{th} to 7^{th}, all of which are little related to the concept of "DIY. The FolkSim list also contain some non-related tags as well like "games, "music, "apple, and "computer ranked on 17^{th} to 20^{th}.

The resultant lists for the tag "Library are shown in the <Table 3>. An online thesaurus source http://thesaurus.reference.com/ lists "books as a synonym of "library. We less arguably claim that the fact that the FolkSim list is better in ranking "books higher than the CosSim list. The tag "books is top ranked on the FolkSim list, but it is second ranked on the CosSim list. The same three tags co-appear within the top five tags on both lists: "education, "books, and "research. Two additional tags are "history and "health on the CosSim list, whereas these are "reference and "web2.0 on the FolkSim list. The tags "health and "web2.0 uniquely appear on each list. "Health appears to be a specific subject like "history or a specific type of library, such as a

health library. Also, considering that Delicious is a Web 2.0 service, and “Library is an area to which the 2.0 paradigm is actively being applied, its appearance is not surprising at all. As in the two previous lists, some much less or non-related tags unique to each list appear on lower ranking of both lists: “free, “media, and “recipes on the CosSim, and “language, “webdesign, and “jobs on the FolkSim.

The two sets of lists based on the two methods contain approximately 50% tags in common: 55% (11 out of 20) for “HTML, 40% (8 out of 20) for “DIY, and 45% (9 out of 20) for “Library. The rankings of common tags are inconsistent on both sets so that there is not any constant pattern to be found in the rankings. These three examples show that the FolkSim provides relatively good results compared to the results of the CosSim, in the experiment based on the 140 most popular Delicious tags. Overall, our experiments indicate that similar or related tags can be identified with FolkSim for any given set of tags.

5.2.4 Discussion of the Tag Folksonomy and the FolkSim Result

Given a tag A, is there any relationship between

〈Table 1〉 Similarity results for the tag “HTML”

	HTML			
Ranking	Tag	CosSim	Tag	FolkSim
1	php	0.97	webdesign	0.902
2	css	0.956	css	0.895
3	javascript	0.948	ajax	0.877
4	tutorials	0.945	javascript	0.869
5	flex	0.941	php	0.86
6	ajax	0.941	python	0.851
7	seo	0.932	typography	0.84
8	flash	0.925	tutorials	0.84
9	free	0.925	mac	0.833
10	email	0.919	ruby	0.832
11	database	0.914	programming	0.828
12	xquery	0.912	tools	0.826
13	tools	0.912	wordpress	0.822
14	webdesign	0.912	wiki	0.819
15	images	0.911	flash	0.819
16	mobile	0.911	web2.0	0.818
17	actionscript	0.908	seo	0.817
18	ruby	0.907	software	0.815
19	google	0.904	opensource	0.813
20	microsoft	0.903	google	0.811

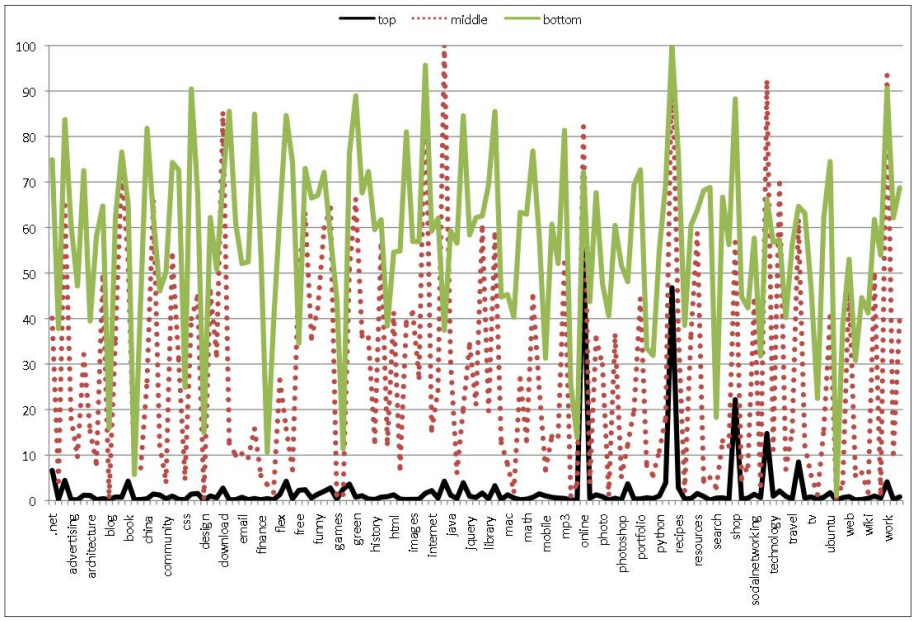
〈Table 2〉 Similarity results for the tag "DIY"

	DIY			
Ranking	Tag	CosSim	Tag	FolkSim
1	photography	0.945	howto	0.871
2	howto	0.923	home	0.826
3	art	0.908	shopping	0.809
4	home	0.897	green	0.808
5	culture	0.894	hardware	0.802
6	environment	0.893	blog	0.792
7	architecture	0.891	cooking	0.787
8	design	0.891	photography	0.777
9	food	0.889	design	0.769
10	tips	0.889	video	0.768
11	money	0.882	tips	0.766
12	programming	0.881	productivity	0.759
13	photoshop	0.88	books	0.751
14	green	0.878	art	0.75
15	science	0.877	technology	0.735
16	humor	0.875	science	0.729
17	typography	0.875	games	0.723
18	visualization	0.87	music	0.722
19	3d	0.868	apple	0.722
20	mac	0.866	computer	0.722

〈Table 3〉 Similarity results for the tag "Library"

	Library			
Ranking	Tag	CosSim	Tag	FolkSim
1	education	0.963	books	0.866
2	books	0.96	education	0.85
3	research	0.953	reference	0.834
4	history	0.952	web2.0	0.83
5	health	0.949	research	0.83
6	travel	0.942	wiki	0.825
7	science	0.935	technology	0.81
8	writing	0.929	images	0.804
9	reference	0.928	opensource	0.801
10	maps	0.928	search	0.797
11	math	0.927	blogs	0.796
12	business	0.926	blog	0.779
13	environment	0.921	history	0.773
14	finance	0.919	news	0.77

15	free	0.919	google	0.763
16	media	0.917	photos	0.762
17	recipes	0.913	science	0.762
18	news	0.911	language	0.758
19	images	0.907	webdesign	0.743
20	wiki	0.906	jobs	0.739



〈Figure 7〉 Comparison of similar tags in terms of rankings within tag folksonomies

the folksonomy (Ψ_A) for tag A and a list of FolkSim-based similar terms for the tag? With the tag "HTML, the term "webdesign is calculated as the most similar term (see the "similarities between tag folksonomies section). Also note that the term "webdesign is ranked at the top of the tag folksonomy (Ψ_{HTML}) associated with "HTML (see the "Tag folksonomies section). Interestingly, the same holds with "DIY. For the case of "Library, "books is the most similar term. In the associated tag folksonomy ($\Psi_{Library}$), "libraries and "books are ranked at the first and second positions. In this case, the most similar tag is second ranked, not first. However, the tag "libraries is not on the list of the 133 tags that are used in assessing similar terms. In all the three cases, there seems to be a significant correlation between the most similar terms by FolkSim and the top ranked tags in tag folksonomies.

Does this hold for the other cases as well, i.e., between the degree of similarity of similar terms and their rankings in tag folksonomies? To answer

the question, three groups of similar terms were selected for each of the 133 main tags, each group consisting of ten most similar terms, ten least similar terms, and ten in the middle range of similarity. Then, the ranking of each of the thirty similar terms in the corresponding tag folksonomy was recorded. As different tag folksonomies hold different numbers of tags (resulting in varying ranges of rankings), we normalized all different ranges of rankings to a single scale of 100. That is, if a similar term A is ranked 500th out of a tag folksonomy containing 1000 tags, the ranking will be normalized to 50 (calculated from ((500/1000) x 100 = 50). After that, all normalized rankings were averaged, yielding a single averaged ranking for each similar group. For example, the averaged rankings of the top, middle, and bottom 10 ".net similar terms based on the ".net folksonomy are 6.62, 40.08, and 74.91, respectively (i.e., each of the ranking was averaged for a group of 10 similar terms and normalized in a scale of 100). It is highly possible that similar terms do not appear in the corresponding tag folksonomy. In such a case, the lowest normalized ranking, which is 100, was assigned to the tag as its ranking. In the ".net case, the resulting rankings seem to comply with our expectation that the more similar terms are, the higher their rankings are. However, according to our experimental results below, that does not always occur.

<Figure 7> plots the results with the 133 main tags. The main tags appear in the x-axis and the three calculated rankings for each tag are indicated in the y-axis (lower numbers in y-axis means higher rankings). The thick black line in the plot indicates the rankings for the group of top ten similar terms. The rankings stay under ten overall with the exception of four tags. The thick lighter line indicates the rankings for the group of the bottom ten. The rankings maintained the lowest (meaning the largest numbers) in most cases. The dotted line indicates the average rankings for the group of ten in the middle range. The rankings fluctuate between the ones for the top group and bottom group. As shown in the plot, sometimes the rankings are close to the top group ranking, but not always. Although some exceptions occur, the averages of 133 top-, middle-, and bottom-group rankings are 2.1, 27.1, and 57.5, respectively, which clearly shows a positive indicator for the relationship.

So, what does that mean? Remember that the similarity between terms is assessed by the degree of similarity between corresponding tag folksonomies. That is, generating similar terms costs more than creating tag folksonomies, and furthermore, the presence of tag folksonomies is a prerequisite for the creation of similar terms. The positive relationship may imply that the tag folksonomies can be alternatively used as a set of similar terms, in which case the ranking in a tag folksonomy may serve as similarity ranking instead.

6. Conclusion

In this paper, we presented a formal model of assessing tag similarity, the FolkSim algorithm that

takes into account tag folksonomies, and evaluation results on the most popular Delicious tags in a comparison with a cosine similarity method.

The FolkSim algorithm has been used to generate tag folksonomies by merging resource-level folksonomies, to measure similarity distance between tag folksonomies, and to produce a ranked list of similar tags. In the three examples, we have seen that the top-ranked tags on the similar lists, which are generated by FolkSim, tend to be among the best possible similar tags in given choices. Also, the lists appear to be better than the ones created by CosSim, based on the observation of the three examples.

Another valuable consequence can be found in tag folksonomy: it also holds meaningful and semantic information of the associated target tag in terms of tag and frequency. The successful operation of the FolkSim algorithm requires that the collective tagging for a target tag is mature enough to fully support the FolkSim algorithm. However, considered that the social tagging service is still young, there must be a considerable number of cases that the tagging for a target tag is not fully mature yet. That is, the FolkSim similar term list is unavailable or infeasible. For the pre-mature cases, tag folksonomy could serve as an alternative.

Both tag folksonomies and similarity lists reflect a blend of social tagging and service-dependence (Delicious in this case). Derived from a social tagging service, the outcome might keep being changed until it reaches the point where a social consensus is achieved. On the other hand, participants to Delicious are known to be interested in the subject of information technology. The fact greatly influences the choice of tags. For instance, tag “apple is much more likely to be used to refer to the well-known computer company, rather than the fruit, in Delicious. The Web 2.0 service is relatively young and still evolving; 100 million URLs were socially tagged through this specific service as of September 2007, compared to more than 8 billion websites indexed by Google at least two years ago. When folksonomy-based systems further grow, they will reach consensus (so-called collective intelligence in the sprit of Web 2.0) faster.

This study has contributed in proposing a method of mining similar tags from folksonomy and in examining the potential of folksonomy as a new potential source for the automatic discovery of similar terms. The contribution of this study can be further extended to the related research areas of tag-based information discovery, tag recommendation and clustering, and the understanding of social user vocabularies.

The majority of the study on tags has been centered on the nature of tags, tagging behaviors, etc. Few studies has been conducted on the problem of tag similarity. Our approach on the problem is distinct from a handful of previous studies that the context (as only specific resource folksonomies that are associated with the target tag were considered, not any general folksonomies) in which tags are used is counted for the tag similarity. The core of the proposed method is to properly extract context-bound tag only.

In this study, a number of the most popular Delicious tags have been considered as target tags. Future studies should further investigate the relationship between the degree of tag popularity and the feasibility of the corresponding tag folksonomies in order to examine the viable scale of the FolkSim algorithm. This study can serve as a pilot research on the study of similar user vocabularies in the context of social and collaborative collection. More extensive research is needed to understand the potential and limits to folksonomies as a new resource of social user vocabularies.

Acknowledgements

The author would like to thank anonymous reviewers for their valuable comments.

References

Baeza-Yates, R. and B. Ribeiro-Neto. 1999. *Modern Information Retrieval.* New York, NY USA: ACM Press.

Begelman, G., P. Keller, and F. Smadja. 2006. "Automated tag clustering: Improving search and exploration in the tag space. *Proceedings of the Tagging Workshop at the 15th International World Wide Web Conference: 22-26.*

Chen, Hsinchun and Kevin J. Lynch. 1992. "Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics,* 22(5): 885-902.

Choy, S. O. and A. K. Lui. 2006. "Web information retrieval in collaborative tagging systems. *Proceedings of the International Conference on Web Intelligence.* Hong Kong, 18-22 December 2006: 352-355.

Crouch, C. J. 1990. "An approach to the automatic construction of global thesauri. *Information Processing and Management,* 26: 629-640.

Dhillon, I. S. and D. S. Modha. 2001. "Concept decompositions for large sparse text data using clustering. *Machine learning,* 42(1): 143-175.

Furnas, G. W., T. K. Landauer, L. M. Gomez, and S. T. Dumais. 1987. "The vocabulary problem in human-system communication. *Communications of the ACM,* 30: 964-971.

Garg, Nikhil and Ingmar Weber. 2008. "Personalized Tag Suggestion for Flickr. *Proceedings of the World Wide Web conference,* Beijing, China, 21-25 April 2008: 1063-1064.

Golder, S. and B. A. Huberman. 2006. "Usage patterns of collaborative tagging systems. *Journal of Information Science,* 32(2): 198-208.

Hotho, Andreas, Robert Jäistoph Schmitz, and G.

Stumme. 2006. "Information retrieval in folksonomies: Search and ranking. *Proceedings of the 3rd European Semantic Web Conference,* Budva, Montenegro, 11-14 June 2006: 411-426.

Jannink, Jan and G. Wiederhold. 1999. "Thesaurus entry extraction from an on-line dictionary. *Proceedings of the Second International Conference on Information Fusion.* Sunnyvale CA.

Lin, D. 1998. "Automatic retrieval and clustering of similar words. *Proceedings of the 17th International Conference on Computational Linguistics.* Montreal, Quebec, Canada, 10-14 August 1998: 768-774.

Lin, Dekang, S. Zhao, L. Qin, and M. Zhou. 2003. "Identifying synonyms among distributionally similar words. *Proceedings of International Joint Conferences on Artificial Intelligence.* Acapulco, Mexico, 9-15 August 2003: 1492-1493.

Turney, Peter D. 2001. Mining the Web for synonyms: PMI_IR versus LSA on TOEFL. *Proceedings of the 12th European Conference on Machine Learning.* Freiburg, Germany, 3-7 September 2001: 491-502.

Turney, Peter D. 2002. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* Philadelphia, PA, 6-12 July 2002: 417-424.

Vander Wal, T. 2007. "Folksonomy coninage and definition. Retrieved on 15 June 2009: http://vanderwal.net/folksonomy.html

White, S. and P. Smyth. 2005. "A spectral clustering approach to finding communities in graphs. *Proceedings of the Fifth SIAM International Conference on Data Mining.* Newport Beach, CA, 21-23 April 2005: 274-285.

Wu, Hua and Ming Zhou. 2003. "Optimizing synonym extraction using monolingual and bilingual resources. *Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications.* Sapporo, Japan, July 11, 2003: 72-79.

Yi, Kwan. 2008. "Mining a Web2.0 service for the discovery of semantically similar terms: a case study with Del.icio.us. *Proceedings of the International Conference on Asia-Pacific Digital Libraries.* Bali, Indonesia, 02-05 December 2008: 321-326.

Appendix †

The full list of the tags shown in the decreasing order of popularity

design blog tools webdesign programming music software web2.0 art video reference linux inspiration web tutorial photography css free howto development education javascript flash travel shopping news blogs business google mac javafood tips politics technology games opensource research books health science security resources windows php toread recipes funny python online marketing twitter photoshop internet mobile search wordpress humor fun portfolio history social tutorials community graphics culture cool media photo ubuntu ruby ajax download rails illustration osx article finance library diy freeware audio fashion productivity architecture photos youtube hardware seo firefox visualization writing jobs advertising apple typography work socialnetworking webdev home wiki environment tv cooking images jquery database learning blogging .net mp3 recipe green fic money language microsoft china 3d lifehacks game computer math tech email html maps movies code 2008 interesting economics flex plugin collaboration gallery shop book actionscript iphone.

† Tags are listed in frequency order, from more frequently occurring tag to less frequently occurring tag. Actually, in the original list, the size of tags is proportionally associated with the frequency of the tags, as shown in the 〈Figure 1〉.