A Comparative Study of Correlation Measurements for Searching Similar Tags[*]

Kaikuo Xu[1], Yu Chen[1], Yexi Jiang[1], Rong Tang[1,2],
Yintian Liu[2], and Jie Gong[1]

[1] School of Computer Science, SiChuan University, ChengDu, 610065, China
[2] Chengdu University of Information Technology, ChengDu, 610225, China

Abstract. In recent years, folksonomy becomes a hot topic in many research fields such as complex systems, information retrieval, and recommending systems. It is essential to study the semantic relationships among tags in folksonomy applications. The main contributions of this paper includes: (a) proposes a general framework for the analysis of the semantic relationships among tags based on their co-occurrence. (b)investigates eight correlation measurements from various fields; then appliying these measurements to searching similar tags for a given tag on datasets from del.icio.us. (c) conducts a comparative study on both accuracy and time performance of the eight measurements. From the comparison, a best overall correlation measurement is concluded for similar tags searching in the applications of folksonomy.

1 Introduction

Taxonomy, a traditional top-down classification method, is considered not sufficient to solve web classification problems [1]. When taxonomy is used for web classification, domain experts construct a hierarchical classification structure and the features of a certain class are normally identified. By this means, documents can be classified according to the expert-constructed hierarchy. However, there are three main problems to use this method: (1) The hierarchy and the features may not fully reflect the real classification of the documents since the experts' domain knowledge are limited; (2) The updates of the hierarchy may not describe the increasing on timely since the growth of web pages is too fast; (3) The classification may not stand for all web users' mind since it is just the opinion of the experts who are a small fraction of users. Folksonomy [2,3,4], also known as 'collaborative tagging', is introduced to alleviate all problems mentioned above. In collaborative tagging, web users are exposed to a web page and freely associate tags with it. Users are also exposed to tags previously entered by themselves and other users. The collective tagging activity creates a dynamic correspondence between a web page and a set of tags, i.e. an emergent categorization in terms of tags shared by a community of web users. Tags stand for the

[*] Supported by the 11th Five Years Key Programs for Sci. &Tech. Development of China under grant No. 2006BAI05A01, the National Science Foundation under grant No 60773169, the Software Innovation Project of Sichuan Youth under Grant No 2007AA0155 and the Development Foundation of Chengdu Univeristy of Information Technology(KYTZ200811).

C. Tang et al. (Eds.): ADMA 2008, LNAI 5139, pp.709–716, 2008.

users' true opinion to classify the web pages. From the description above, it is clear that tag is the core concept of folksonomy and the classification is conducted through tags of all users. Thus to study the semantic relationships among tags is the key for applications of folksonomy.

A general framework to analyze the semantic relationships among tags based on their co-occurrence is proposed in this paper. In the framework, the whole analysis process is partitioned into eight steps. The task of each step is identified and the difficulties in each step are discussed in detail. The task to search similar tags for a given tag is left for further research. The definition of 'similar tags' is given, and eight correlation measurements from applications of various fields are investigated. The comparisons among correlation measurements are conducted over the datasets from del.icio.us on both the accuracy and time performance. The experiment results are analyzed and a remark on all the eight correlation measurements is given.

The rest of the paper is organized as follows. Section 2 describes the general framework to analyze the semantic relationship among tags. Section 3 investigates the correlation measurements to search similar tags for a given tag, and section 4 demonstrates the experimental evaluation. At last, Section 5 concludes the paper.

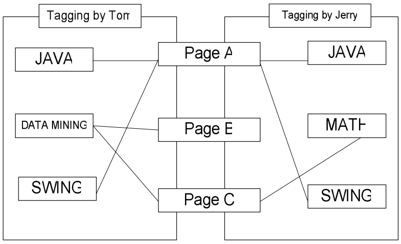
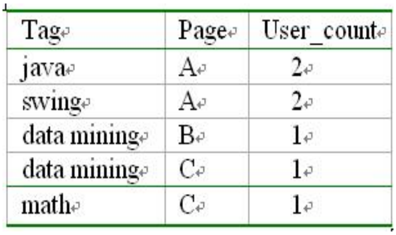


Fig. 1. An example of tagging in folksonomy



Tag	Page	User_count
java	A	2
swing	A	2
data mining	B	1
data mining	C	1
math	C	1

Fig. 2. An example of tagging in folksonomy

2 A General Framework for Tag Analysis

According to the observation, co-occurrence between tags is widespread in folksonomy systems. Figure 1 shows such an example. On the purpose of understanding user count, Figure 1 is transformed into Figure 2,which contains an attribute 'user_count'. Let's formulate the notions first. Let the tag set be T, the number of pages that tag A annotates be *La*, the number of pages tag B annotates be *Lb*, the number of common pages both tag A and tag B annotate be *Lab* and the number of web pages be *N*.

Definition 1 (Similarity between Tags). Let A and B be two Tags and Lab be the number of shared web pages of A and B. Let the threshold be δ, δ>= 1 .if Lab > δ the A and B is called similar tags. 'similarity' is used to describe the degree how they are similar to each other. 'similarity' between A and B is denoted as s(A,B).

Definition 2 (User Count). The number of users who use tag A to annotate Page P is defined as the **user count** of A to P.

A process for tags' analyzing based on the co-occurrence among tags is described below. The analysis process consists of eight steps and the output of each former step is the input of its latter step.

Step 1. Preprocessing. The main goals of this step are to: (a) Eliminate system tags. System tags are provided to users by folksonomy systems to complete some special tasks. For example, when users want to import bookmarks from other Social Bookmarking Service (SBS) providers to del.icio.us, the folksonomy system adds tag 'imported' to these bookmarks. These tags are collected to build a dictionary. Then tags matching any tag in the dictionary will be eliminated. (b) Eliminate meaningless tags. Users' mistakes or the system's mistakes produce meaningless tags. For example, a ' ' is taken as a meaningless tag. One solution to identify these meaningless tags can be the outlier checking algorithm.

Step 2. Classification. This step is to classify tags into three classes[5]: (a) Personal tags. These tags are used by only one user but assigned to more than one pages. The tags may be understood as personal vocabulary. Thus, they are useful for individual retrieval but useless for the rest users of the community. (b) Unpopular tags. These tags are assigned to different resources by different users but only once or several times. These tags can be treated as unpopular tags when they are only used by a small fraction of users quite occasionally. (c) Popular tags. These tags are assigned to various resources by different users frequently and they can be taken as global tags which are generally used by many users.

Step 3. Natural language processing. This step deals with tags under the class label 'Personal tags' and 'Unpopular tags'. Since these tags do not appear frequently, the traditional natural language processing methods are considered to be sufficient to solve the problem. The simplest way is to directly get insight into these tags by 'http://wordnet.princeton.edu/'.

Step 4. Frequent pattern mining. This step uses the mature frequent pattern mining techniques like *fpgrowth* to capture the co-occurrence between tags. The most difficult problem in this step is how to specify the threshold δ to judge whether a pattern is a frequent pattern. Obviously the deficiency of the support-confidence framework [6] still exists. Therefore, the result of this step needs further investigation in the future.

Step 5. Similar tag searching. This step searches the similar tags for a given tag. The correlation analysis is applied on the result of frequent pattern mining in our research. Since there are many correlation measurements on both statistics and data mining, the problem that needs to be solved is how to choose a 'better' measurement.

Step 6. Core-based clustering. Given a tag and its similar tags, the similar tags are clustered into groups in this step. Tags in each group should be the same 'close' to the given tag semantically. The given tag is taken as the 'core'. Both the hierarchical clustering and the density-based clustering are considered as candidate methods.

Step 7. Similar tag searching (multi-tag). This step searches the similar tags for a given tag group. The co-occurrence is used here and correlation analysis will takes effect in this step too. However, it is very difficult to reach a satisfy precision by existing methods because of two major reasons: (a) the occurrence of tag groups is rare; (b) it is not easy to judge what users mean only from a given tag group.

Step 8. Tag network building. The hierarchical structure implies a 'containing' relationship among class labels in taxonomy: the web pages that higher level class labels annotate contains the web pages that the lower level class labels annotate. However, the case is different for tags. Based on our observation, large number of tags normally annotates the same web pages. But these tags themselves usually annotate many unique web pages. The ultimate relationship among tags is likely to be a network, and each tag is a node in the network. It still needs further investigation whether this thought can be implemented.

The rest of the paper concentrates on step 5.

3 Similarity Measurements for Comparison

In this section, eight similarity measurements are discussed and compared in order to obtain the better measurements. These eight measurements are: Augmented Expected Mutual Information [7], Simrank[8], Pythagorean Theorem [9], Pessimistic Similarity [10],Cosine similarity [11], adjusted Cosine similarity [11], Pearson coefficient [11], and TF/IDF [12]. These measurements are divided into two groups: measurements with user count and without user count.

3.1 Measurements without User Count

Augmented Expected Mutual Information (AEMI). The concept of mutual information [13] is from the information theory. Given two objects, the mutual information measures how much the knowledge to one object reduces the uncertainty about the other. In the context of correlation analysis, it can be a common correlation metric. The augmented expected mutual information [7] is adopted in our study, as shown in formula 1.

$$\text{AEMI(A,B)} = \text{MI(A,B)} + \text{MI}(\overline{A}, \overline{B}) - \text{MI}(\text{A}, \overline{B}) - \text{MI}(\overline{A}, \text{B}) \tag{1}$$

Simrank. The formula proposed in [8] is an intuitive formula, and also the simplest one among the four, as shown in formula 2. Here, *C* is a constant between 0 and 1 and it shows less confidence on the similarity between *A* and *B* than the confidence between *A* and itself. In this paper, *C* = 0.8.

$$\begin{cases} \text{s(A,A)} = 1 / \text{La} \\ \text{s(A,B)} = \text{C*Lab/(La*Lb) if } \ \text{A != B} \end{cases} \tag{2}$$

Pythagorean Theorem (PT). *Pythagorean Theorem* is a classic theorem in Euclidean geometry. It describes the relationship among the three sides of a right triangle, as "The square on the hypotenuse is equal to the sum of the squares on the other two sides". Let *Lab* be the length of one leg, and (La - Lab) + (Lb - Lab) be the length of the other legs respectively; then the length of the hypotenuse is $\sqrt{Lab^2 + (La + Lb - 2 * Lab)^2}$. The formula to measure the similarity between two tags is shown in formula 3. According to formula 3, s(A,B) increases along the increasing of Lab and decreases along the increasing of La and/or Lb.

$$\text{s(A,B)} = \frac{Lab}{\sqrt{Lab^2 + (La + Lb - 2 * Lab)^2}} = \frac{1}{\sqrt{1+(\frac{La}{Lab}+\frac{Lb}{Lab}-2)^2}} \tag{3}$$

Pessimistic Similarity (PS). Pessimistic prune is a prune strategy adopted by C4.5 [10]. In our method, the pessimistic confidence on 'A is the same to B' is taken as the similarity between A and B. Let's consider a proposition P 'tag A is similar to tag B', for which La + Lb - Lab = K and the observed error rate is f. For this proposition there are K training instances to support. The observed error rate is simply the number of pages annotated by only one tag divided by K, i.e.(La + Lb -2 * Lab)/K. A random variable is also considered standing for the true error rate: a random variable X with zero mean lies within a range 2z with a confidence of Pr[-z <=X<=z] = c. According to Normal distribution, there is a corresponding z once c is specified. For example, Pr[-1.65 <= X <= 1.65] = 0.9. From the notation above, we can obtain X = $\frac{f - e}{\sqrt{e(1-e)/K}}$, where f is the error rate and e is the mean. According to the formula above, the range of true error rate e for rule R can be found based on the observed error rate f and the number of supporting instances K. Let the confidence range value be z, the confidence value corresponding to z be cf, the number of supporting instances count (La+Lb-Lab) be K, and the observed error rate be f. Then the upper bound on the estimated error e is U_{cf}(f,K)[10]. The pessimistic confidence value sim(A,B) for the rule 'A is the same to B' can be defined as formula 4. The pessimistic confidence value shows the confidence level on "tag A is similar to tag B". This can be also explained as "how much tag A is similar to tag B", i.e. the value of the similarity. In this paper, the confidence is set as c = 80%, then z = 1.28.

$$\text{s(A,B)} = 1 - U_{cf}(f,K) \tag{4}$$

3.2 Measurements with User Count

In this section, the set of pages that tag i annotates is denoted as I_i, and the number of users who have annotated page j by tag i is denote as $C_{i,j}$.

$$\text{s(A, B)} = \frac{\sum_{j \in I_A} C_{A,j} * C_{B,j}}{\sqrt{\sum_{k \in I_A} C_{A,k}^2}\sqrt{\sum_{k \in I_B} C_{B,k}^2}} \tag{5}$$

Cosine Similarity (CS). In information retrieval, the similarity between two documents is often measured by treating each document as a vector of word frequencies and computing the cosine of the angle formed by the two frequency vectors. This formalism can also be adopted to calculate the similarity between tags, where tags take the role of documents, pages take the role of words, and usercount take the role of word frequencies. Then the similarity is defined as formula 5.

Adjusted Cosine Similarity (ACS). The adjusted cosine similarity is derived from cosine similarity, which is commonly used in collaborative filtering. However, this formula can not be applied directly to compute the similarity between tags since negative result may be obtained for Ci,j - $\overline{C_i}$. Therefore, data are preprocessed as follows. For a single tag, all the pages with user counts less than the average user count are eliminated. This maintains the pages that the tag primarily annotates.

$$\text{s(A, B)} = \frac{\sum_{j\in I_A}(C_{A,j}-\overline{C_A})*(C_{B,j}-\overline{C_B})}{\sqrt{\sum_{k\in I_A}(C_{A,k}-\overline{C_A})^2}\sqrt{\sum_{k\in I_B}(C_{B,k}-\overline{C_B})^2}} \tag{6}$$

Pearson Coefficient (PC): The Pearson's product-moment correlation coefficient is a measurement for the degrees of two objects linearly related. The correlation between tag A and tag B is shown in formula 7. Here the summations over i are over the urls to which both tag A and B are linked. For very small common annotations, this similarity metric returns inaccurate results. This needs to be solved by methods such as 'default voting' or 'minimum common votes'. The results of a given tag are pruned beforehand. Therefore, it is unnecessary to specify a minimum number of common annotations in order to calculate a valid similarity actually.

$$\text{s(A, B)} = \frac{\sum_{j\in I_A}(C_{A,j}-\overline{C_A})*(C_{B,j}-\overline{C_B})}{\sqrt{\sum_{i\in I_A}(C_{A,i}-\overline{C_A})^2}\sqrt{\sum_{i}(C_{B,i}-\overline{C_B})^2}} \tag{7}$$

IDF/TF: The IDF/TF method is a classic method in document classification. In this paper, it is applied to compute the similarity between Tags as in paper [10]. Let $\text{TF}_{\text{i,j}}$ be the ratio of tag i in all tags annotating page j, IDF_{i} be the rareness of tag i,

$$\text{TF}_{\text{i,j}} = \text{C}_{\text{i,j}} \;/\; \sum_i C_{i,j} \tag{8}$$

$$\text{IDF}_{\text{i}} = \log(\,\text{C}_{\text{i,j}} \;/\; \sum_i C_{i,j}\,) \tag{9}$$

Then the degree $\text{rel}_{\text{i,j}}$ of relation between tag i and page j is defined as formula 10.

$$\text{rel}_{\text{i,j}} = \sum_j \sum_i C_{i,j} \,/ \sum_j C_{i,j} \tag{10}$$

$$\text{s(A, B)} = \sum_j C_{A,j} * rel_{B,j} \tag{11}$$

At last, the similarity sim(A,B) of tag B from the view point of tag A is defined as formula 11. Here, sim(A,B) is not necessarily the same as sim(B,A) for formula 11. This requires more computations than other seven measurements.

4 Experiments

The real data sets from del.icio.us from Nov 30 to Dec 15 are collected. There are 234023 unique tags and 749971 unique web pages. Though web2.0 data provide valuable resource for data mining, there are no public benchmark data for research yet. Therefore, in our work, three human evaluators are asked to judge the performance of each formula. 30 tags are randomly selected as the given tags. Top-N similar tags can be obtained for each given tag A. For each tag T_i in the Top-N results, evaluators give a score to T_i as formula 12. Since our problem is actually a ranking problem, the classical evaluation method adopted in Information Retrieval is also used to solve our problem. Precision (P) at top N results is used as a measurement to evaluate the performance, as shown in formula 13.

$$score(T_i) = \begin{cases} 2 \text{ if } T_i \text{ is similar to } A \\ 1 \text{ if } T_i \text{ is somewhat similar to } A \\ 0 \text{ if } T_i \text{ is not similar to } A \end{cases} \tag{12}$$

$$\text{P@N} = \sum_{i=1}^{N} score(T_i) \ / \ (2*\text{N}) \tag{13}$$

All experiments are conducted on an INTEL core 2DuoProcessorE2160 with 2G memory, running UBUNTU OS. Table 1 shows the performance of eight measurements. There are two numbers in each cell. The first number is the average precision from the three evaluators and the second one in parentheses is the standard deviation. PC performs the worst among the eight measurements, which is the only one not in support of the counter evidence. This phenomenon indicates the necessity of the usage of counter evidence for similar tags' searching. CS is of the best performance among all eight measurements. Although ACS and IDF/TF are more complex than CS, their performance is worse than CS. The experiment result also shows that IDF/TF performs even worse than Simrank and PT although it takes user_count into account, while Simrank and PT are two measurements that do not use user_count. The performances of both AEMI and PS are out of our expectation: they are surpassed by Simrank and PT, the two simplest measurements in this paper. As we know, the performances of Simrank and PT are really good by considering their simplicity.

Table 1. P@N of eight measurements& CPU time(s)

Measurement	N = 10	N = 20	N = 30	Rank	CPU time
AEMI	0.508(0.046)	0.567(0.042)	0.570(0.037)	7	10.31
Simrank	0.679(0.148)	0.643(0.122)	0.615(0.106)	4	0.68
PT	0.688(0.054)	0.609(0.035)	0.581(0.039)	3	1.84
PS	0.555(0.119)	0.538(0.104)	0.534(0.085)	6	4.01
CS	0.749(0.068)	0.662(0.071)	0.619(0.072)	1	9.52
ACS	0.710(0.075)	0.641(0.074)	0.603(0.076)	2	15.84
PC	0.526(0.130)	0.576(0.106)	0.561(0.093)	8	11.34
IDF/TF	0.665(0.045)	0.608(0.044)	0.577(0.045)	5	713.31

Table 1 also shows CPU time consumed by eight measurements, respectively. The CPU time for IDF/TF is one/two orders of magnitude longer than those of the other measurements. Due to its relatively bad performance, it is considered not suitable for similar tags' searching. Since the CPU time consumed by either Simarank or PT is extremely small, they are considered as the candidate measurements for the similar tags' searching. Although the CPU time consumed by CS is one order of magnitude longer than the former two, it is still considered as the candidate measurement for the searching as well. In real applications, many techniques, such as high performance clustering, parallel computing, etc, can be applied to improve the speed although it is very difficult to improve P@N. Considering these two factors, CS is considered the best overall measurement for similar tags' searching in this paper.

5 Conclusions

A framework to analyze the relationship among tags is proposed and discussed. The former five steps of the framework are under research and the latter three steps are in our vision for the whole project. To search similar tags for a given tag becomes the focus of this paper. The steps before this task in the framework are dealt with in a pessimistic way. Eight measurements are investigated in a whole. Experiments are conducted on datasets from del.icio.us. Both the accuracy and CPU time consumed by each measurement are compared from one another. Cosine Similarity is considered as the best overall measurement due to its high accuracy and relatively low CPU consumption. Simrank and Pythagorean Theorem are considered good as well because of their extremely low CPU consumption and relatively high accuracy.

References

[1] Shirky, C.: Ontology is overrated: Categories, links, and tags. Clay Shirky's Writings About the Internet Website (2005)

[2] del.icio.us, `http://del.icio.us/`

[3] Cattuto, C., Loreto, V., Pietronero, L.: Semiotic dynamics and collaborative tagging. Proceedings of the National Academy of Sciences United States of America 104, 1461 (2007)

[4] Cattuto, C., Loreto, V., Servedio, V.D.: A yule-simon process with memory. Europhysics Letters 76(2), 208–214 (2006)

[5] Lux, M., Granitzer, M., Kern, R.: Aspects of Broad Folksonomies. In: 18th International Conference on Database and Expert Systems Applications (2007)

[6] Brin, S., Motwani, R., Silverstein, C.: Beyond Market Baskets: Generalizing Association Rules to Correlations. In: Proceedings ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, USA, May 13-15 (1997)

[7] Chan, P.K.: A non-invasive learning approach to building web user profiles. In: KDD 1999 Workshop on Web Usage Analysis and User Profiling (1999)

[8] Jeh, G., Widom, J.: SimRank: A measure of structural-context similarity. In: Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (July 2002)

[9] Pythagorean Theorem, `http://mathworld.wolfram.com/PythagoreanTheorem.html`

[10] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)

[11] Deshpande, M., Karypis, G.: Item-based top-n recommendation algorithms. ACM Trans. Inf. Syst. 22(1), 143–177 (2004)

[12] Thorsten, J.: Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In: Proceedings of 14th International Conference on Machine Learning (1996)

[13] Rosenfeld, R.: A maximum entropy approach to adaptive statistical language modeling. Computer,speech, and language 10 (1996)