# Adaptive Combination of Tag and Link-based User Similarity in Flickr

Nhat Hai Phan
Advanced Technology Fusion Dept
School of Internet & Multimedia Eng.
Konkuk University, Seoul, Korea
+82-10-4326-3217
pnhai@konkuk.ac.kr

Van Duc Thong Hoang
Advanced Technology Fusion Dept
School of Internet & Multimedia Eng.
Konkuk University, Seoul, Korea
+82-10-3097-1512
hvdthong@konkuk.ac.kr

Hyoseop Shin
Advanced Technology Fusion Dept
School of Internet & Multimedia Eng.
Konkuk University, Seoul, Korea
+82-2-2049-6117
hsshin@konkuk.ac.kr

## ABSTRACT
Finding similar users is one of the probable applications in social media. The similarity between users can be measured in two different approaches: the semantic similarity and the similarity in terms of social relations. These two approaches can be combined with different weight factors. However, the conventional combination scheme has a critical drawback that the weight factors are fixed for every user and thus it is not optimized at those users that are using rare terms or do not have sufficient relations with other users. To address this problem, in this paper, we propose an adaptive combination scheme of tag-based similarity and link-based similarity in which the weight factors are dynamically determined for each user by evaluating each user's characteristics such as *tag commonness* and *link strength*. The experimental results with a Flickr data set show that the proposed scheme consistently outperforms the previous work by about 20%.

## Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models*, *information filtering, search process*

## General Terms
Algorithms, Measurement, Experimentation, Human Factors

## Keywords
User similarity, Tag commonness, Link strength

## 1. INTRODUCTION
Online photo services such as Flickr and Picasa have become one of the major types of social media on the web. Flickr allows users to share their photos with friends, family, and other members of the online community.

In this paper, we address the problem of finding similar users in photo sharing services. Finding similar users is one of the probable applications in social media. When a visiting user finds the photos of a user interesting, the visitor may want to find more unknown photo owners whose photos are *similar* to those of the given user. Here, the similarity between users can be measured in two different approaches: the semantic similarity and the similarity in terms of social relations. The semantic similarity can be captured in textual or image content-based approach. Though the state of the art of the content-based image retrieval is progressing, the textual annotations such as tags can be more effectively used for capturing the semantics of a photo. Flickr has been proved successful for searching purposes in letting users provide the semantic context of their photos through the manual annotations (i.e., tags). In this approach, tags are considered to describe content of photos posted by users. The advantage of this approach is that we can directly discover the interested topics of users, and thus find the similar users whose interested topics are similar as those of the given users. Other aspect of the similarity between users can be drawn from the social relations between users. The notion is here that if a visiting user has expressed an interest (i.e., established a link) on both user A's photos and user B's ones, then A and B are probably similar. These links can conveniently be captured in Flickr because each photo is linked to a set of users who pick the photo as a favorite. The link structure can provide additional insight about the relationships among users (e.g., Even among the photos of a same topic, a user can represent an interest only to a specific photo).


*MM'10*, October 25–29, 2010, Firenze, Italy.


The semantic similarity and the link-based similarity have been widely explored in the literature [1,2,3,4,5,6,7,8]. Those methods have been developed for comparing between generic web pages as well as between the documents of a specific type such as blogs. The two approaches have mostly been combined with different weight factors which could be determined by heuristics and machine learning, to improve the performance [3,5,6,8]. If we assume that tags are used for measuring semantic similarity, then the conventional combination can be formulated as follows:

$$similarity(a,b) = \omega_{tag} \times tag\_similarity(a,b) + \omega_{link} \times link\_similarity(a,b)$$

, where $\omega_{tag} + \omega_{link} = 1$.
This scheme, however, has a critical drawback; it is based on the assumption that each user uses *common tags* that other users may often use and each user has a *sufficient links* so that the links of the user can be compared with those of other users. In case a user is using only *rare* tags or has *insufficient* links, then the similarity of the user with other users cannot be fairly evaluated. Therefore, the conventional combination schemes based on the equation above will produce non-optimal results in the cases above. It was pointed out in Menczer's work [5] that any fixed-proportion combination of semantic and link-based similarity cannot produce optimal results.

Table 1 summarizes the performance of the conventional combination scheme of tag-based and link-based similarity in different cases. The conventional schemes are just effective if both users to be compared use common tags and they have sufficient links with the other users. On the other hand, the schemes usually are not effective when either user uses rare tags or has insufficient links to other users.

**Table 1. Disadvantage of Tag-based similarity and Link-based similarity**

| | user 2 | | |
|---|---|---|---|
| user 1 | **Tag** | **Rare** | **Common** |
| | **Rare** | bad | bad |
| | **Common** | bad | good |

| | user 2 | | |
|---|---|---|---|
| user 1 | **Link** | **Insufficient** | **Sufficient** |
| | **Insufficient** | bad | bad |
| | **Sufficient** | bad | good |

To address this problem, in this paper, we propose an adaptive combination scheme of tag-based similarity and link-based similarity in which the weight factors, $\omega_{tag}$ and $\omega_{link}$, are dynamically determined for each user by evaluating each user's characteristics such as *tag commonness* and *link strength* in order to optimize the precision of the similarity between the users. For example, if a query user is using many common tags but has insufficient links, then the tag-based similarity will get a high weight value but the link-based similarity will get a low weight value by the proposed scheme.

This paper is organized as follows. Section 2 introduces and formalizes the concepts of *tag commonness* and *link strength* for each user. Section 3 presents the adaptive combination scheme in detail. Section 4 shows the experimental results and Section 5 presents conclusion and future work.

# 2. TAG COMMONNESS AND LINK STRENGTH

Topics, categories, and other related information of user-generated contents can be captured by user-annotated tags. Thus, tags can be useful in finding similar users for a given user. However, the performance of the tag-based similarity scheme may depend on the commonness of the tags that are used. The performance of similarity will be improved if users are using many common tags rather than rare tags, which is not always true; some users use common tags but others use rare tags. To address this issue, we present an algorithm to measure the commonness of the tags (tag commonness, TC) for each user.

In social media such as Flickr, users interact with each other by using online activities such as posting, commenting, giving feedback on posts. Users can establish links with other users by giving or receiving feedbacks. These links can be used for finding similar users because two users could be considered similar if they both receive links from a same group of users. The link-based similarity can be effective only when a query user receives a sufficient number of links from others. If a user has little connections with other users, the link-based analysis would be inadequate. Consequently, the link strength of a user is a crucial characteristic to identify whether the link-based similarity is effective or not. Hence, we propose an algorithm to measure the link strength (LS) of each user.

## 2.1 Tag Commonness (TC)

*TC definition:* The *proficiency* in tag usage of each user that is used to measure how well authors use tags to describe photo's content. If the proficiency in tag usage of the user A is larger than that of the user B, then the TC of the user A should be larger than the TC of the user B and vice versa.

We propose that the proficiency in tag usage of users depends on the number of common tags and the proportion between the number of common tags and the number of rare tags. Thus, our heuristic function to evaluate TC of users is designed as follow:

$$TC_{u,\mu} = \log(NumCT+1) * NumCT/(NumCT+NumRT)$$

Where u is user u, μ is the threshold to separate common tags and rare tags, and NumCT and NumRT denote the number of *common tags* and *rare tags*[1], respectively, that are used by the user u.

To evaluate TC of users, the important process is to separate common tags and rare tags. In previous work [7], a random threshold was chosen to identify common tags and prune rare tags as illogical tags. The method is not convincing because the boundary between common tags and rare tags cannot be decided by a simple threshold and the threshold can be dependent on each tag. In addressing this issue, we consider all the possible cases of thresholds. For each threshold $\mu_i$, the $TC_{u,\mu_i}$ is evaluated by following the equation. Then, $TC_u$ is computed by averaging of those TC values as below:

$$TC_u = \sum_i^k TC_{u,\mu_i} \Big/ k$$

, where $k$ is the number of thresholds: $\Delta=\{\mu_1, \mu_2,\ldots, \mu_k\}$. The following describes the algorithm.

**Input**:U={$u_1$, $u_2$,…, $u_m$} is a set of users, with $u_i$ = (($t_1$,$w_{i,1}$), ($t_2$,$w_{i,2}$),…, ($t_n$,$w_{i,n}$)) τ={($t_1$,$w_1$), ($t_2$,$w_2$),… ($t_n$,$w_n$)} is set of tags; $t_i$ is a tag, $w_i$ is the number of the users who used $t_i$, and $w_{i,j}$ is the number of times the user j used the tag $t_i$.

1. For each μ in Δ
2. | For each $u_a$ in U
3. | | $TC_{u_a,\mu} = \log(NumCT+1) * NumCT/(NumCT+NumRT)$;
4. | End
5. End;
6. Result = {ϕ};
7. For each $u_a$ in U
8. | $TC_a = \sum_i^k TC_{u_a,i} \Big/ k$;
9. | Result = Result ∪{($u_a$,$TC_a$)};
10. End;
11. **Output**: Result

## 2.2 Link Strength

*LS definition:* the *noticeability* of a user in a social network that is used to measure the degree of links that are exchanged with other users. If a user has a high value of LS, she may have received a large number of feedbacks from other users; therefore, this user can easily be noticed by other users.

In designing LS for each user, we consider the properties of LS:

**property 1:** LS of a user depends on the number of links this user has received from other users. The more links a user receives, the more noticeable she is.

**property 2:** LS of a user is related to the weight of each link this user has received from each user. The more weight each link has, the more noticeable she is.

**property 3:** LS of a user is affected by the variation of the weights of the links this user has received from each user. If a user has a low variation, LS of the user would be high.

Based on the properties above, our heuristic LS is defined as follows:

$$\sigma_u = \sum_{i=1}^n \left| fb_{u,i} - \frac{\sum_{i=1}^n fb_{u,i}}{n} \right| \Big/ n ; LS_u = \begin{cases} n*(1/(\sigma_u+1))*\left(n \Big/ \sum_{i=1}^n \frac{1}{fb_{u,i}}\right) (n>1) \\ 0 (n=1) \end{cases}$$

, where $n$ is the number of links of a user $u$, and $fb_{u,i}$ is the weight of the $i$-th link of the user $u$.

# 3. ADAPTIVE COMBINATION OF USER SIMILARITY MEASURES

In our adaptive combination framework of user similarity, the tag commonness, TC, and the link strength, LS, are used in

[1] A common tag is a tag that has been used by at least as many as a given threshold, μ, users, otherwise it is a rare tag.

computing the weights of the tag-based similarity and the link-based similarity, respectively.

Suppose we have a query user p, and the normalized TC and LS values of hers are denoted as $TC_p$ and $LS_p$. Let $\Gamma_{p,q}$ denote the combination similarity of tag-based similarity and link-based similarity of the user p against a user q. $\Gamma_{p,q}$ is computed as follow:

$$\Gamma_{p,q} = \frac{TC_p}{TC_p + LS_q}\sigma_{tag,p,q} + \frac{LS_p}{TC_p + LS_q}\sigma_{link,p,q}$$

The idea of the suggested scheme is to flexibly evaluate the combination proportion of tag-based similarity approach and link-based similarity approach to obtain the optimal similarity search result for each user.

As for the similarity measures, we propose a variation of *tf*idf* cosine similarity as the tag-based similarity and a variation of Jaccard similarity as the link-based similarity. We describe these schemes in the following subsections.

## 3.1 Tag-based User Similarity

The tag-based similarity between users is computed by the consine similarity between the tf*iuf vectors. Against tag $t_i$ for a user $u_j$ , *tf* and *iuf* are computed as follows:

$$tf_{i,j} = n_{i,j} \Big/ \sum_k n_{k,j}\,;\; iuf_i = \log \frac{|U|}{\left(1 + |\{u : t_i \in d_u\}|\right)}\,;\; tf \times iuf_{i,j} = tf_{i,j} * iuf_i$$

where $n_{i,j}$ is the number of occurrences of the tag $t_i$ of the user $u_j$, $\Sigma_k n_{k,j}$ is the sum of the occurrences of all the tags of user $u_j$, $|U|$ is the total number of users, $|\{u:t_i \in d_u\}|$ is the number of users who use tag $t_i$.

Suppose we have a set of tags, $t = \{t_1, t_2, \ldots, t_n\}$, then the tag-based cosine similarity ($\sigma_{tag,pq}$) for the pair of users p, q is measured as below:

$$p\left((t_1, tf \times iuf_1), (t_2, tf \times iuf_2), \ldots; (t_n, tf \times iuf_n)\right)$$

$$q\left((t_1, tf \times iuf_1), (t_2, tf \times iuf_2), \ldots; (t_n, tf \times iuf_n)\right)$$

$$\sigma_{tag,pq} = \left(\sum_{i=1}^{n} p_i * q_i\right) \Big/ \left(|\,p\,| * |\,q\,|\right)$$

## 3.2 Link-based User Similarity

The link-based similarity between users is computed by the Jaccard similarity between the link weight vectors. Suppose $w_{p,i}$ denotes the link weight from a user *i* to user *p*. Then, the link-based Jaccard similarity ($\sigma_{link,pq}$) between the user *p* and the user *q* is computed as below:

$$p(w_{p,1}, w_{p,2}, \ldots, w_{p,n});\; q(w_{q,1}, w_{q,2}, \ldots, w_{q,n})$$

$$\sigma_{link,pq} = \sum_{i=1}^{n} \min(w_{p,i}, w_{q,i}) \Big/ \sum_{i=1}^{n} \max(w_{p,i}, w_{q,i})$$

# 4. EXPERIMENTAL RESULTS

## 4.1 Experimental Data Set

We collected as many Flickr posts as possible that have been posted between January and March of 2009, by using "the interesting photos of the day" (500 photos a day) as the seed photos. In Flickr, each photo is related to a set of users who pick the photo as a favorite. In our experiments, we interpret this favorite action as the user feedback onto a post. Table 2 summarizes the data set we used in the experiments.

**Table 2. Data Set Description**

| no.posts | no.feedbacks | no.users | no.posters | no.feedback givers |
|---|---|---|---|---|
| 51,742,309 | 24,991,762 | 1,454,042 | 756,064 | 800,393 |

## 4.2 User Studies

### *4.2.1 TC and LS user study*

To estimate the efficiencies of the proposed TC and LS schemes, we performed two-step user studies. 1) *Generic user study*: we chose 50 random users and made $C_2^{50}$ user pairs for the user study. Then, the tag and link information of the users in each pair of users is presented to testers and each tester picks the better user in TC and the better user in LS. 2) *Narrow user study*: to more delicately evaluate the efficiency of the proposed methods, we tried to choose more competing users. We first sorted all the users in descending order of the TC score and the LS score that are computed from our schemes. Then, we chose the 50$^{th}$–ranked user as the first user and picked every 100$^{th}$ user to make a selection of 50 users. The following steps are the same as the generic user study.

Table 3 shows the TC and LS user study results. In generic user study, both TC and LS achieved quite high ratios of correctness (about 80%). In narrow user study, LS was relatively higher (76%) than TC (61.4%). Overall results, however, appear to be acceptable.

**Table 3. TC and LS user study results**

| | | Correct | Incorrect | Unidentified | %Correct |
|---|---|---|---|---|---|
| **Generic User Study** | **TC** | 918 | 232 | 75 | 79.8 |
| | **LS** | 905 | 220 | 100 | 80.4 |
| **Narrow User Study** | **TC** | 685 | 431 | 109 | 61.4 |
| | **LS** | 846 | 266 | 113 | 76 |

We computed the Kappa statistic [9] that is a common measure for agreement between judges (i.e., test users). The average Kappa statistic of the user study in this paper is computed as follows:

$$K = 1/C_2^N * \sum_{x \in U} \sum_{y \in U, y \neq x} (P_A(x,y) - P_E(x,y)) / (1 - P_E(x,y))$$

, where N is the number of test users, $P_A(x,y)$ denotes the proportion of the times that two test users, x and y, agreed and $P_E(x,y)$ denotes the proportion of the times that two test users, x and y, would agree by chance. As table 4 shows, the kappa statistics of our user study appear to be at quite high level.

**Table 4. Kappa statistics**

| General User Study | | Local User Study | |
|---|---|---|---|
| **TC** | **LS** | **TC** | **LS** |
| 0.8849 | 0.8608 | 0.8631 | 0.8337 |

### *4.2.2 User Similarity user study*

By this user study, we compare the performances of different user similarity schemes: our proposed adaptive combination, tag-only similarity, link-only similarity, combination schemes with different combination proportions. We picked 40 query users with different characteristics in tag and link: 10 users having *high* TC values and *high* LS values, 10 users having *high* TC values and *low* LS values, 10 users having *low* TC values and *high* LS values, and 10 users having *low* TC values and *low* LS values. Each algorithm generates top 10 similar users for each query user, and then those similar users are merged and presented to the testers. Then, each tester chooses at least 10 similar users among them.

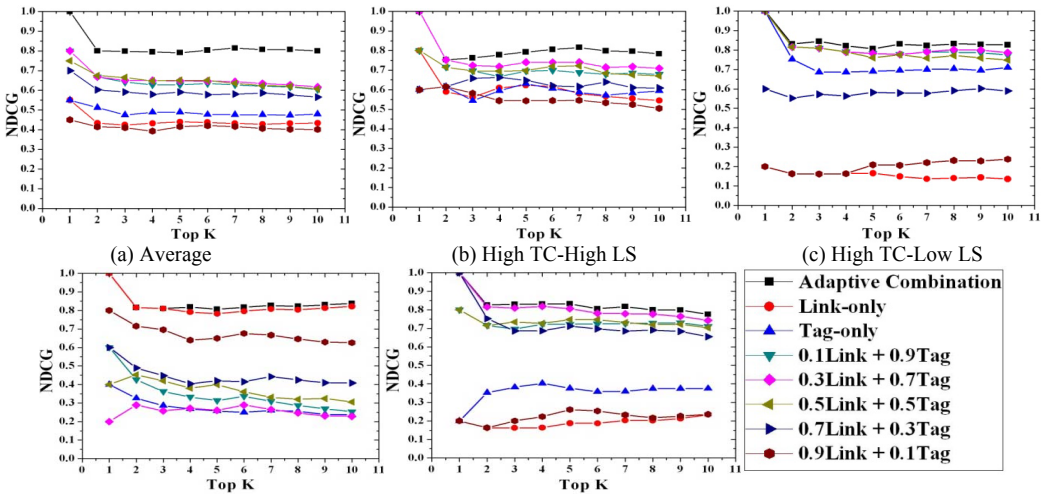
## 4.3 Performance Evaluation

NDCG [10] is used to consider the ranked position as well as the ratio of the relevant answers among top-k answers recommended by a ranking scheme.

Let $r_{Ri}$ denote the binary judgment (i.e., 1 for true and 0 for false) for the user ranked *i*-th by a ranking scheme R. Then, NDCG@k is defined as follows:

$$NDCG_R@k = \frac{DCG_R@k}{DCG_{ground-truth}@k}, DCG_R@k = \sum_{i=1}^{k} \frac{2^{r_{Ri}} - 1}{\log(i+1)}$$

Figure 1 shows the NDCG results for the different schemes. In average (figure 1-(a)), the proposed adaptive combination outperforms the other schemes. It consistently achieves over 80%




(a) Average (b) High TC-High LS (c) High TC-Low LS

(d) Low TC-High LS (e) Low TC-Low LS

**Figure 1. NDCG results**

which is at least 20% performance increase against others. Notice that the performances of the other schemes are not consistent against different cases. For example, Link-only scheme is quite good at (Low TC-High LS) (figure 1-(d)), but it shows the worst performance at (High TC-Low LS) users (figure 1-(c)). 0.3Link+0.7Tag combination is good at High TC-High LS, High TC-Low LS, and Low TC-Low LS, but it is the worst at Low TC-High LS. This performance fluctuation applies to all the fixed-proportion combination schemes. Meanwhile, our proposed adaptive combination scheme shows consistent performance for all the cases.

Figure 2 demonstrates the results of user similarity by the proposed adaptive combination. In High TC-High LS case (figure 2-(a)), some portion of the tags are overlapped between the similar users, and in High TC-Low LS case (figure 2-(b)), the tags between the users are quite similar. However, note that tags are quite different between the users in Low TC-High LS case (figure 2-(c)).

## 5. CONCLUSION AND FUTURE WORK

In combining tag-based and link-based user similarity measures, the conventional fixed-proportion combination schemes are not optimal in case users use rare tags and do not have sufficient relations with other users. To address this problem, this paper proposed an adaptive combination scheme in which weight factors are adaptively determined based on each user's tag commonness and link strength. The performance gain was quite impressive at about 20% and consistent for all the cases.

A future work would be the development of machine learning-based schemes for measuring TC and LS of users.

## 6. ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-A419-0070).

## 7. REFERENCES

[1] D. Fogaras and B. Racz. Scaling Link-Based Similarity Search. WWW 2005.

[2] A. A. Benczur, K. Csalogany, T. Sarlos. Link-Based Similarity Search to Fight Web Spam. AIRWEB'06.

[3] A.G. Maguitman, F. Menczer, H.Roinestad, A.Vespignani. Algorithmic Detection of Semantic Similarity. WWW 2005

[4] J. Lee, J.K. Min, C.W. Chung. An Effective Semantic Search Technique using Ontology. WWW 2009.

[5] F. Menczer. Combining Link and Content Analysis to Estimate Semantic Similarity. WWW 2004.

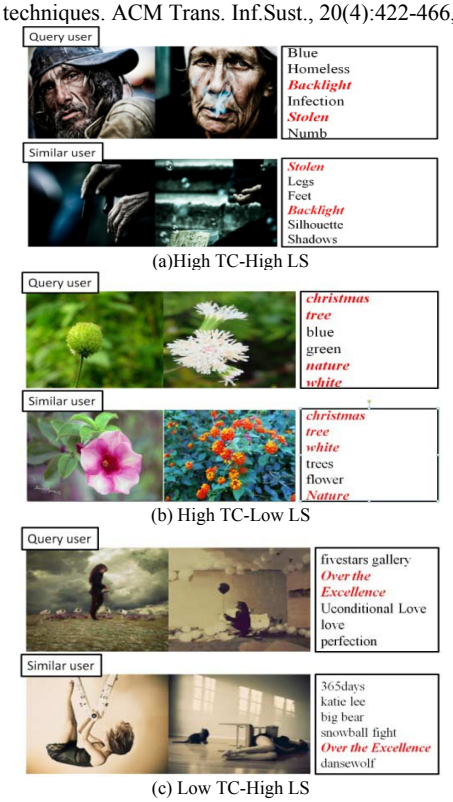
[6] S. Zhu, K. Yu, Y. Chi, Y. Gong. Combining Content and Link for Classification using Matric Factorization. SIGIR 2007.

[7] X. Li, L. Guo, Y. Zhao. Tag-based Social Interest Discovery. WWW 2008.

[8] K. Nikolaev, E. Zudina, and A. Gorshkov, Combining Anchor Text Categorization and Graph Analysis for Paid Link Detection, WWW 2009.

[9] J. Carletta, Assessing agreement on classification tasks: kappa statistic, Computational Linguistics 22:249-254, 1996.

[10] K. Jarvelin and J.Kekalainen. Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf.Sust., 20(4):422-466, 2002.




(a)High TC-High LS

(b) High TC-Low LS

(c) Low TC-High LS

**Figure 2. Similar user search demonstration**