Connecting Users with Similar Interests Across Multiple Web Services

Haewoon Kwak and **Hwa-Yong Shin** and **Jong-Il Yoon** and **Sue Moon**
Computer Science Department, KAIST
335 Gwahangno, Yuseong-gu, Daejeon, Korea
haewoon@an.kaist.ac.kr, shinhy83@gmail.com, cyril.yoon@gmail.com, sbmoon@kaist.edu

Abstract

Most online social networking services provide a feature for users to build interest groups. Based on the profiles and behavior data, web services can assist users to join groups by recommending relevant interest groups. In this paper, we propose a novel method to connect users across multiple services based on user-labeled tags. Tags represent interests of a user and have advantages in terms of privacy, up-to-datedness, and service coverage. We have collected tags from six popular web services, and analyzed usage patterns. We observe that the popularity of tags is highly skewed and dependent on the web services. We have also found that a set of tags of a single user frequently changes over time. Through user study, we demonstrate that the vector space model combined with intra-personomy normalization is good enough to recommend other users with similar interests.

Introduction

Most online social networking services provide a feature for users to build interest groups: 'Groups' in Facebook and MySpace, 'Community' in Orkut, and 'Cafe' in Daum and Naver, top two portal services in Korea, just to name a few. It connects users with similar interests. A user logs in to a web service, searches for a group whose members have similar interests, and joins the group. Based on the profiles and behavior data, web services can assist users to join groups by recommending relevant interest groups.

In this paper, we propose a method to connect users with similar interests across *multiple* services based on user-labeled tags. Our goal is to help users reach beyond the boundary of a single service and find others with similar interests. The key to our idea is the user-labeled tags. They represent interests of a user and have advantages in terms of privacy, up-to-datedness, and service coverage. First of all, tags are publicly accessible and we are free from privacy issues. Next, tags also eliminate the need to update one's interest constantly. Once a user attaches tags to contents as a way to summarize succinctly, the set of tags used by the user reflects one's current interests. For example, tags from photos uploaded, videos watched, and music listened represent one's interests in all those media in words. Finally, we can apply our approach to any online service that supports tags. Our approach becomes a convenient vehicle to add a social networking feature to any web service.

A tag set of a single user is called personomy (Hotho et al. 2006), for it typically serves as a personal taxonomy. We have collected tags from six popular web services, and analyzed tag usage patterns. We report that the popularity of tags is highly skewed and their vocabulary dependent on the web services. We have also found that a set of tags of a single user frequently changes over time. We propose a simple vector space model combined with intra-personomy normalization. Through user study, we demonstrate that the vector space model combined with intra-personomy normalization is good enough to recommend other users with similar interests. Additionally, we report that the top 30 most frequently used tags of a user are sufficient in similarity calculation.

We note that identifying the same user across multiple web services is out of the scope of this work. The main reasons are two-fold. First, we do not have any user profile to detect the same user represented by different online identity. Once we consider user profiles, our approach is not free from privacy concern. Second, our goal is to recommend other users with similar interests across many web services. A set of tags of a user is accessible, but one's profile is not in many web services. Thus, we require only personomy that we can easily get.

The remainder of the paper is organized as follows. We first introduce web services covered in this work. We analyze tag usage patterns to assess the feasibility of finding users of similar interests. Then we evaluate the calculated similarity by a user study. We add a discussion on our future directions and then conclude after a section on related work.

Dataset

We use tag data from six popular web services: a social bookmarking site, Del.icio.us, a photo sharing site, Flickr, the world's largest video sharing site, YouTube, a blog portal site, LiveJournal, a social music site, Last.FM, and a meta-blog service in Korea, AllBlog[1]. Our choices of web services cover all major media: bookmarking, photos, videos, music, and blogs.

No site other than AllBlog allows access to its user base in entirety. Typically only a small selection of users is dis-

[1] http://allblog.net

played on the site's main page. We utilize this selection and crawl the homepages of those users in the selection. We summarize our dataset in Table 1. We note that we refine all tags in this work by using API of WordNet (Fellbaum 1998). The total number of users varies from $6,363,000$ to $54,464$ and the number of tags from $71,724$ to slightly over a million. The average number of tags per user on Del.icio.us is far larger than on any other service. Many users of Del.icio.us are tech-savvy people working in IT industry; they tend to be more familiar with tags and use more than average users (Li et al. 2007).

Service	# of users	# of tags	# tags / user
Del.icio.us	40,072	1,092,534	227.2
Flickr	6,366	71,724	32.4
YouTube	9,481	171,990	56.5
LiveJournal	49,792	729,975	44.49
Last.FM	54,464	95,901	10.95
AllBlog	24,559	383,374	44.04

Table 1: Summary of tag data from 6 services

Analysis of Tag Usage Patterns

In this section, we analyze characteristics of a user's tag usage pattern. Three characteristics are our main foci: tag popularity, service dependence, and evolution. These characteristics shows what proportions of tags are personal, how many tags are used only in one service, and how frequently a set of tags of a user changes.

Tag Popularity

Previous work on Flickr reports that tag popularity exhibits a highly skewed pattern: a small number of tags are shared by most people and a larger number of tags are shared only by a small number of people (Xu et al. 2006). We have looked into the distribution of tag popularity of all six services and found the same breakdown. In the case of AllBlog, top 20% of tags are associated with about 80% of posts, while top 20% tags in other services are associated with over 90% of items.

Unpopular and esoteric tags, if shared by others, represent specific focus topics of users, but personally meaningful words such as 'ToDo-Until-3rd-Dec' cannot be used as is in grouping users of similar interests. We have found that 75.5% of all tags aggregated over the six services are used by only one user. We remove those 75.5% of tags as we deem them irrelevant in our work.

Service Dependence of Tag Vocabulary

Next we investigate if tags of one service are different from those of another. If tags turn out to be highly service dependent, tags from different services have very few in common and leave little room for our approach.

From each of the five services, we choose the top $15,000$ tags and count the union of the tags of matching ranks in an cumulative manner. We leave out tags from AllBlog in this analysis, since they are in Korean. We see that about 20% of tags belong to more than one service. Half of those 20% tags belong to two services and only 600 tags out of $75,000$ tags belong to all five services. This result shows a great disparity among popular tags from different services, and implies a difficulty in finding similar tags across services. One solution we are considering to close this disparity is to group tags of similar meanings into one. It is somewhat similar to using a thesaurus. A more advanced technique is to extract tag clusters based on tag co-occurrences (Begelman, Keller, and Smadja 2006; Schmitz 2006). We will consider these solutions in our future work.

Dynamically Changing Tags

We would like to know if a user's interests stay steady or change over time. We use AllBlog data to see how tags of a user change over time, for AllBlog is the only dataset that has time information associated with tags. In order to gain a long-term perspective on tag usage, we calculate the most popular tags per user and then compare them against the tags by the month in Figure 1. Interestingly the largest number of matches occurs at the beginning and the number of matches slowly decreases over time. We interpret this as a diversification of a user's interests over time. Initially, a user joins a service with a handful of topics in mind, but gradually explores other topics on the same service. Thus, both the large time window to capture diverse interests and the small time window to capture the latest interests are relevant to track user interests.

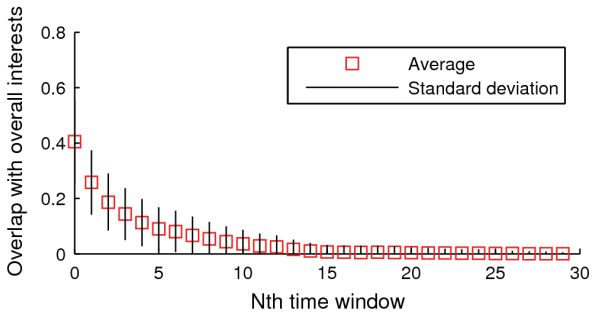

Figure 1: Changing interests over time

Weight Assignment Schemes and Algorithms for Similarity Calculation

In order to calculate similarities between users, we assign a weight to each tag. A weight of a tag reflects the importance of the tag. We normalize the number of times tags are used. That is, to normalize in each individual personomy. This represents the relative importance of the interest. It is basically the same as the term frequency in information retrieval and data mining. Then, we consider each tag as one dimension of a user vector, and calculate the cosine similarity between two users. This is the basic approach in information retrieval and document processing (Salton and McGill 1983).

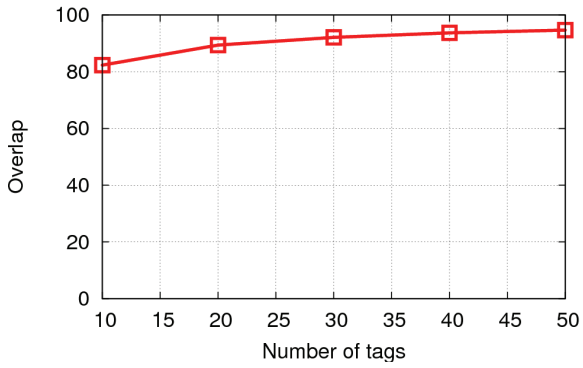

Figure 2: User match percentage vs the number of most popular tags used

We need to address one last detail in our approach. We have not put any limit on the number of tags used in similarity calculation, although the number of tags differs greatly from one user to another. A fixed number of tags in similarity calculation would simplify the calculation as well as data structures of the system. We extract 10, 20, 30, 40, and 50 most frequently used tags of a user and obtain the 30 most similar users based on cosine similarities. If the 30 users from the limited tags match those from all tags, then we could safely use a fixed number of tags in our similarity calculation.

Figure 2 plots the average of the tag match percentage against the number of the most popular tags used. It shows higher than 92% beyond 30 tags. In other words, even with the 30 most popular tags from a user we can find a decent match of similar users.

User Study for Evaluation

In this section we set out to answer the following two questions through user study. What is the acceptable level of similarity that humans perceive? From which pair of services are users most likely to connect? The first question helps map the quantitative measure of similarity to human perception and guides in algorithm design. The second question addresses which service is more amenable to cross-domain user match.

To answer the first question, we develop our user study as a web-based application. We cut the screen horizontally into half and display the home pages of two users. The question "Do these users have similar interests?" is placed in the middle. The respondent can choose one out of "yes", "no", and "I don't know". We put a screen shot of our user study page in Figure 3

In our user study we recruited 20 respondents and presented them with 3 questions from each level of similarity and a total of 12 questions. The gap between each level of similarity is 0.25. Figure 4 summarizes the outcome of our user study. The number of "yes" decreases as the level of similarity decreases. Thus our measure of similarity does correspond to human perception. However, our solution is not perfect. Note that even in the "very similar" category,

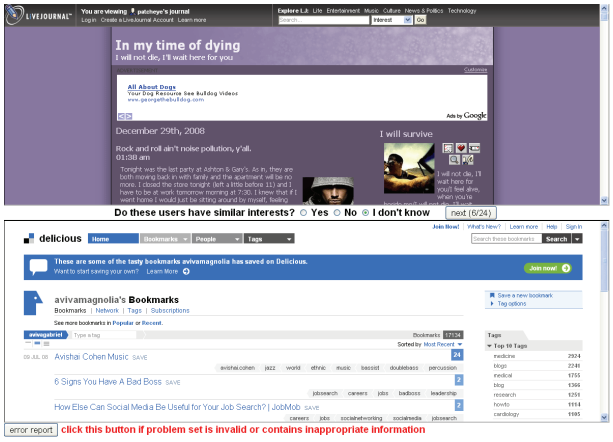

Figure 3: A screen shot of our user study page

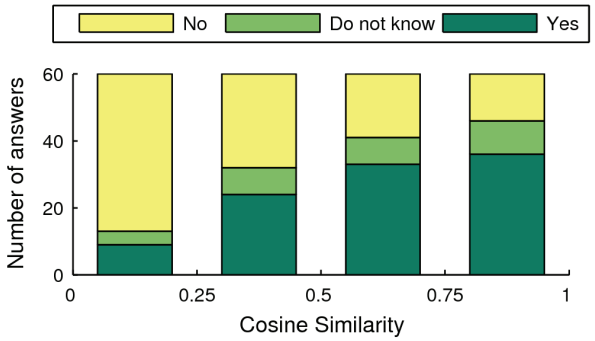

Figure 4: Quantitative measure of similarity against human perception

about 40% responded lukewarmly or negatively, while in the "very dissimilar" category, some respondents found a good match of similarity. The overall feedback from the user study is that due to the disparate interface design of each service respondents found the user study time consuming and demanding. At this point we do not know the partial disagreement between respondents' reaction and our quantitative measure of similarity is due to the interface design or points at room for improvement in our algorithm design. We leave this for future work.

	Del.icio.us	Flickr	YouTube	LiveJournal	Last.FM
Del.icio.us	0.820	0.005	0.006	0.142	0.027
Flickr	0.088	0.292	0.016	0.536	0.069
YouTube	0.154	0.024	0.416	0.289	0.117
LiveJournal	0.072	0.015	0.006	0.852	0.055
Last.FM	0.006	0.001	0.001	0.024	0.969

Table 2: The conditional probability that a user of a service in a row connects to users of the other service in each column

Previously we have shown that the vocabulary of tags is

not completely independent of the service. Does it mean that users of the same service are more likely to have similar interests than from another service? We analyze the conditional probability of similarity based on the service. We choose those pairs from the user study that produce the cosine similarity value higher than 0.75. Table 2 shows the conditional probability that a user of one service connects to users in other services in a row. The sum of a row is equal to 1. The first observation is that the probability within the same service is high. It means that recommendation within a service should work well. We note that Flickr is an exception and their users connect to LiveJournal users better. The conditional probability that Flickr and YouTube users connect to those in the same service is relatively low and their users find good matches in other services. LiveJournal attracts users from other services better than any other service. We conjecture that writing a blog posting is easier than creating multimedia contents and LiveJournal users cover a wide spectrum of topics.

Related Work

Our idea of finding users of interests across multiple, heterogeneous services bears similarity to the problem of finding similar documents in a large collection of documents (Resnick and Varian 1997). A tag set of a single user could be interpreted as a list of important keywords extracted from a document.

Several projects have focused on generating user profiles from tags in a single service (Diederich and Iofciu 2006; Firan, Nejdl, and Paiu 2007; Michlmayr and Cayzer 2007; Yeung, Gibbins, and Shadbolt 2008). We expand the scope of the problem to multiple domains and assess a variety of weight assignment schemes and similarity calculation algorithms.

In the past decade or so, we have witnessed great advancement in recommendation system design. For this work we have taken the two fundamental ideas of recommendation system based on the simple summation and cosine similarity. For future work we would like to add another dimension to the problem formulation. Relations between users and semantic relations between tags are two additions we consider pursuing.

Conclusion

In this paper, we have proposed a novel method to connect users across multiple services based on user-labeled tags. Tags represent interests of a user and have advantages in terms of the privacy, up-to-dateness, and service coverage. We have collected tags from six popular web services and analyzed tag usage patterns. We have observed that the popularity of tags is highly skewed and its vocabulary dependent on the web service. We have also found that frequently used tags of a single user change over time. We have demonstrated that the most frequently used 30 tags of a user are sufficient in similarity calculation. We have shown that the vector space model combined with intra-personomy normalization is promising approach to connect users with similar interests.

Acknowledgement

This work was supported by the IT R&D program of MKE/IITA [A1100-0801-2758, "CASFI : High-Precision Measurement and Analysis Research"].

References

Begelman, G.; Keller, P.; and Smadja, F. 2006. Automated tag clustering: Improving search and exploration in the tag space. In *Proceedings of the Workshop on Collaborative Tagging at WWW2006*.

Diederich, J., and Iofciu, T. 2006. Finding communities of practice from user profiles based on folksonomies. In *Proceedings of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice (TEL-CoPs'06), co-located with the First European Conference on Technology-Enhanced Learning*.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Firan, C. S.; Nejdl, W.; and Paiu, R. 2007. The benefit of using tag-based profiles. In *LA-WEB '07: Proceedings of the 2007 Latin American Web Conference*, 32–41. Washington, DC, USA: IEEE Computer Society.

Hotho, A.; Jäschke, R.; Schmitz, C.; and Stumme, G. 2006. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference*.

Li, R.; Bao, S.; Yu, Y.; Fei, B.; and Su, Z. 2007. Towards effective browsing of large scale social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, 943–952. New York, NY, USA: ACM.

Michlmayr, E., and Cayzer, S. 2007. Learning user profiles from tagging data and leveraging them for personal (ized) information access. In *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference (WWW2007)*.

Resnick, P., and Varian, H. R. 1997. Recommender systems. *Communications of the ACM* 40(3):56–58.

Salton, G., and McGill, M. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Schmitz, P. 2006. Inducing ontology from flickr tags. In *Proceedings of the Workshop on Collaborative Tagging at WWW2006*.

Xu, Z.; Fu, Y.; Mao, J.; and Su, D. 2006. Towards the semantic web: Collaborative tag suggestions. In *WWW2006: Proceedings of the Collaborative Web Tagging Workshop*.

Yeung, C. M. A.; Gibbins, N.; and Shadbolt, N. 2008. A study of user profile generation from folksonomies. In *Social Web and Knowledge Management, Social Web 2008 Workshop at WWW2008*.