Connecting Users with Similar Interests via Tag Network Inference

Xufei Wang and Huan Liu
Computer Science and Engineering
Arizona State University
Tempe, AZ 85281
{xufei.wang, huan.liu}@asu.edu

Wei Fan
IBM T.J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532
weifan@us.ibm.com

ABSTRACT

The popularity of social networking greatly increases interaction among people. However, one major challenge remains — how to connect people who share similar interests. In a social network, the majority of people who share similar interests with given a user are in the long tail that accounts for 80% of total population. Searching for similar users by following links in social network has two limitations: it is inefficient and incomplete. Thus, it is desirable to design new methods to find like-minded people. In this paper, we propose to use collective wisdom from the crowd or tag networks to solve the problem. In a tag network, each node represents a tag as described by some words, and the weight of an undirected edge represents the co-occurrence of two tags. As such, the tag network describes the semantic relationships among tags. In order to connect to other users of similar interests via a tag network, we use diffusion kernels on the tag network to measure the similarity between pairs of tags. The similarity of people's interests are measured on the basis of similar tags they share. To recommend people who are alike, we retrieve top k people sharing the most similar tags. Compared to two baseline methods triadic closure and LSI, the proposed tag network approach achieves 108% and 27% relative improvements on the BlogCatalog dataset, respectively.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Sociology; I.5 [**Pattern Recognition**]: Applications

General Terms

Algorithms,Experimentation

Keywords

Tag Network, Diffusion Kernel, Like-Minded Users

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

1. INTRODUCTION

Social networking is increasingly becoming an integral part of social life in which friend recommendation is an important feature. However, friend recommendation merely based on connectivity has two main limitations: it is inefficient and incomplete. For instance, the space complexity of an exhaustive search is exponential; an incomplete search risks of not being able to find anybody. Reaching the long tail users who accounts for 80% of the populatioin in a social network is not trivial, and in certain scenarios, some long tail users are not reachable at all. Thus search methods based on links could fail.

Nonetheless, recommending people of similar interests is an important task. Connecting to "people like you" also has a psychological component [6]: a sense of self-worth and fulfillment, being reassured of their worth and value, a sense of belonging to a community, the need to both seek help from and provide help to others, etc. Instead of recommending people of similar interests, following links alone can mostly recommend people you may (already) know. Thus, the proposed work of connecting like-minded people beyond using links is of practical significance.

- Connecting to people who are alike could accelerate information sharing and problem solving. For instance, we may want to find other people who have prior experience in using a system or a tool for advice when we have difficulty using them. We may want to find the right people for advice before we buy a product. We may want to find collaborators in a research field, etc.
- Connecting like-minded people in a social network is rewarding from the perspective of service providers who may be more interested in the behavior of groups of people than of individuals. For example, they may design more effective services, or place more relevant advertisements by understanding the common interests of a particular group.
- Connecting a small amount of people who are alike in the long tail is a challenging task which is not suitable for traditional data mining approaches. Searching for a few long tail users who are far apart from each other by leveraging connectivity between users is ineffective and inefficient.

The challenges of connecting like-minded users are summarized below. First, people only have an egocentric view of the social network, i.e., users only see their immediate

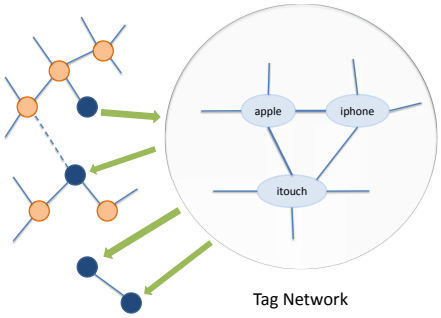


Figure 1: Connecting Like-Minded Users in a Tag Network Approach. The blue nodes represent Apple fans. The dashed line represents users who are connected via some intermediates. Given a seed user on the top left, the toy example demonstrates returning other Apple fans within the network.

Figure 2: A Snapshot of a Blog Description. Six semantically relevant tags are used to describe the purpose of the website: *apple*, *ipod*, *iphone*, *mac*, *apple iphone*, *iphone apps*. Mobile Tech and Gadgets are the two categories the blog is listed under.

contacts. Second, the scale of a social network website like Facebook, Twitter, or LinkedIn makes manual search unrealistic. Thus, more effective and efficient tools are necessary. Third, as shown earlier, recommendation based on links have limitations due to the long tail distribution.

1.1 Recommend via Tag Network Inference

We propose to connect like-minded users via tag network inference. The basic idea is illustrated in Figure 1. Nodes with different colors represent users in a social network. Some users are in the largest component of the network, whereas other users are disconnected, thus either isolated or in small groups. The four nodes highlighted in blue (dark) are fans of Apple products such as iPhone, iTouch, etc. Thus, the four users are deemed "like-minded" (A formal definition will be given in Section 3). The right part of the figure represents a tag network in which each node represents a tag, and the weight between two tags corresponds to users who use the two tags simultaneously.

Providing "wisdom of the crowd", the tag network describes the semantic relationships among tags. The similarity in the use of tags provides a way of measuring how similar two users are. For example, assume we want to connect other Apple fans to the upper left user in blue (dark). Instead of traversing links, we use the tag network, and return the other three Apple fans in the lower left.

The proposed approach requires no parameter tuning, thus is easy to use in practice. It is also efficient since the time complexity for recommendation is linear in the number of users and tags in a social network.

1.2 Summary

Connecting people of similar interests is not a link prediction problem [8]. The key difference between these two concepts is their objectives. The former attempts to find "familiar strangers" who have similar experiences, opinions, interests, etc but who do not know each other, whereas link prediction attempts to recommend people you may already know. There is no causal relationship between similarity and friendship. For instance, people who have similar interests may not be friends, and vice versa.

- We propose to use a tag network to represent the semantic relevance between tags, and show that the tag network is more powerful in capturing the semantic relevance than latent semantic indexing (LSI).
- We propose to connect like-minded users via tag network inference by leveraging the collective wisdom from the crowd. The tag network approach is effective compared to baseline methods.

2. TAG NETWORK CONSTRUCTION

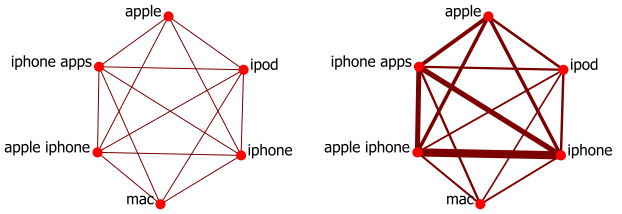
Tagging is an activity for organizing various objects like bookmarks and blogs for future browsing, management, and sharing using informal vocabularies. Tags can be words or phrases, and *informal* means they may not be found in any dictionary. Figure 2 is a snapshot of a description for a blog on BlogCatalog including tags[1]. As shown in the figure, the blog, which is a news and review website on iPhone and iPhone applications, was added on September 2008. It has a major category *Mobile Tech* and a sub-category *Gadgets*. Categories imply the owner's interests. Six tags that are semantically relevant are specified by the owner such that other readers can easily discover the topics of the blog without browsing hundreds of articles within it.

Tagging is a knowledge that reflects thoughts on various web resources [4]. Collective wisdom emerges by combining many people's tagging knowledge together. The hypothesis of the work is that collective tagging naturally brings semantically relevant tags closer. For example, if two tags (e.g., iPhone and Apple iPhone) are used simultaneously by many people, there could be a semantic relevance between them. We represent the connectivity of tags in a network format: *Tag Network*.

We illustrate the steps to construct a tag network on the BlogCatalog dataset (more about the data in Section 6).

- For each object (e.g. blog) and its descriptive tags, we connect the tags as a clique as shown in Figure 3 (a);
- For each person, we combine all cliques corresponding to the objects she owns and form one or more **unweighted** tag networks, since her tags may or may not be connected in a tag network;

[1]http://www.blogcatalog.com/blogs/apple-iphone-news-and-app-reviews-ifonescom#



(a) Tag Network of a Blog (b) Tag Network of a Site

Figure 3: Examples of Tag Networks: (a) an unweighted tag network corresponds to the blog shown in Figure 2, and (b) a snapshot of the tag network constructed on BlogCatalog (other tags not shown). The weight of a link represents the number of users who use the two tags simultaneously.

- We construct a **weighted** tag network by aggregating all tag networks belonging to each person. In the weighted tag network, tags correspond to the union of all users' vocabularies, and the weight of each link represents the number of users who use both tags simultaneously.

A snapshot of the weighted tag network is demonstrated in Figure 3 (b). Note that other tags and the corresponding links are not shown. We count *the number of users* instead of *the number of times* two tags are used simultaneously as the weight of each link to avoid bias from spam users, i.e., those who may use automated tools to assign the same group of tags many times. However, it could be interesting to consider user influence in assigning link weights as future work. Tags are available in most social networking sites in different forms such as user interests, bookmarks, labels, etc. Thus, the construction process can be easily adapted.

3. PROBLEM STATEMENT

A social network $\mathcal{G} = (\mathcal{U}, \mathcal{E})$ is represented as an undirected graph, in which $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$ represents a set of n users and $\mathcal{E} = \{e_1, e_2, \ldots, e_\ell\}$ represents ℓ connections amongst the set of users. Each user subscribes to a certain number of tags. We denote the tag subscription relationship as a matrix $U \in R^{t \times n}$, in which each entry represents the number of times a tag is used by a given user. Let the number of unique tags associated to u_i be $\|u_i\|$. Denote $u_i^{interest}$ as a set of interests (e.g., categories specified by users on BlogCatalog) explicitly declared by the i-th user. Two users are said to be like-minded if they share some interests, e.g., both of them are fans of Apple iPhone.

$$u_i^{interest} \cap u_j^{interest} \neq \emptyset, 1 \leq i, j \leq n \tag{1}$$

However, two Apple fans may not necessarily use the same tags, e.g., one person likes to use *iPhone* as tags, the other person prefers to use *apple iPhone*.

A tag network $W \in R^{t \times t}$ is a symmetric graph in which each node represents a tag that could be a word or a phrase, a non-zero entry w_{ij} in W represents the number of users who use the two corresponding tags simultaneously. A diffusion kernel K_β defined on a tag network is utilized to measure the tag similarities, where β is the parameter which controls the speed of diffusion. Table 1 summarizes the notations. The problem is then defined as follows:

Table 1: Notations

Notations	Description
$\mathcal{G}$	Social Network
U	User Tag matrix
W	Tag Network
u_i	The i-th user in $\mathcal{U}$
$u_i^{interest}$	Interests of the u_i
$\|u_i\|$	Number of unique tags of u_i
S_i	The set of top k most similar users of u_i
k	Number of users to be selected
K_β	Diffusion kernel with parameter β
MSI_j	Mean Shared Interests between u_i and the j-th user in S_i averaged on all u_is in $\mathcal{G}$

- **Input:** Given a social network $\mathcal{G}$, a user u_i $(1 \leq i \leq n)$, a tag network $W \in R^{t \times t}$, and a scalar k.
- **Output:** top k most similar users from $\mathcal{G}$.

4. DIFFUSION ON TAG NETWORK

The tag network enables us to measure the similarities between any pair of tags within it. The simplest measure of similarity between two tags is the shortest path distance. However, the shortest path distance is susceptible to change in graph structure, i.e., newly added or removed tags and links might dramatically affect the distance between two nodes. Therefore, we prefer to average all path distances between two given tags for a more robust similarity measure, which leads to the idea of random walk with varying steps, equivalent to a diffusion kernel on a network [7, 9]. The concept of diffusion kernels is well established, thus readers who are familiar with it can simply skip.

Given a tag network $W \in R^{t \times t}$, where t represents the number of unique tags in a social network, we define a matrix L, whose negation is called Laplacian matrix, as follows,

$$L = W - D, \tag{2}$$

where D is a diagonal matrix in which the i-th diagonal entry corresponds to the summation of the entries in i-th column of matrix W. The diffusion kernel K_β of a tag network is defined as follows,

$$e^{\beta L} = \lim_{s \to \infty} (I + \frac{\beta L}{s})^s, \tag{3}$$

where $\beta \geq 0$ is a user specified parameter which controls the speed of diffusion. A larger β value means a faster information diffusion speed on the network; and there is no diffusion when β is set to 0. The diffusion kernel is positive semi-definite, thus is a valid kernel for measuring similarity between any pair of two tags [9].

The computation of a diffusion kernel requires an eigen-decomposition of L such that $L = V \Sigma V^\top$,

$$\begin{aligned} K_\beta &= e^{\beta L} \\ &= I + \beta L + \frac{(\beta L)^2}{2!} + \frac{(\beta L)^3}{3!} + \ldots \\ &= V\big(I + \beta\Sigma + \frac{\beta^2}{2!}\Sigma^2 + \frac{\beta^3}{3!}\Sigma^3 + \ldots\big)V^\top \\ &= V e^{\beta\Sigma} V^\top \end{aligned} \tag{4}$$

where the columns of V are the eigenvectors, Σ is a diagonal matrix whose diagonal entries are eigenvalues, and $(e^{\beta\Sigma})_{ii} = e^{\beta\Sigma_{ii}}$, other non-diagonal elements are all zeros.

5. RECOMMEND LIKE-MINDED USERS

Let u_i be a seed user, K_β be the kernel, the goal is to select the top k most relevant users in terms of similarity from the social network. The similarity between two users is aggregated on the pair-wise tag similarity given below,

$$sim(u_i, u_j) = \sum_{t \in u_i, t' \in u_j} \frac{u_i(t)}{\sqrt{\|u_i\|}} \cdot K_\beta(t, t') \cdot \frac{u_j(t')}{\sqrt{\|u_j\|}}, \quad (5)$$

where $u_i(t)$ represents the number of times the tag t is used by the i-th user and two normalization terms $\sqrt{\|u_i\|}$ and $\sqrt{\|u_j\|}$ are applied to the two users, respectively. The normalization is necessary because it prevents selecting spammers who use a large number of tags. But users who share more semantically relevant tags are credited thus we use the square root for both normalization terms. The intuition of Equation (5) is that two users are more *like-minded* if they share more semantically relevant tags.

Denote Z as a diagonal matrix whose diagonal entries are $Z_{ii} = \frac{1}{\sqrt{\|u_i\|}}$. We rewrite the similarity between u_i to other users in the social network as follows,

$$sim(u_i, \cdot) = u_i^\top \cdot K_\beta \cdot U \cdot Z \quad (6)$$

We discard the normalization term $\|u_i\|$ since it does not affect the final ranking. Without prior knowledge, determining parameter β is difficult in practice. However, tag network does provide heuristics for β selection. Tags that are frequently used simultaneously are semantically relevant, which is also the basic idea behind Latent Semantic Indexing (LSI) which leverages term co-occurrence in articles [2]. In a tag network, many semantically relevant tags are close or even immediate neighbors, thus it is desirable to select *small* values of βs.

6. DATASET AND EXPERIMENTAL SETUP

BlogCatalog[2] is an online blog service which enables bloggers to register, manage, share, and connect blogs. A blog in BlogCatalog is associated with various pieces of information such as the categories that the blog is listed under, blog level tags, blog statistics such as the average rating and recent viewers, posts within the blog, and reviews from peer bloggers. A blogger also connects to other bloggers to form her social circle on BlogCatalog. A blogger's interests could be gauged by the categories (e.g. arts, business, education, etc) she publishes her blogs in. We obtained in total 60 categories in the processed BlogCatalog dataset. We notice that a blogger can specify more than one category for each blog. On average each blogger lists their blog under 1.69 categories. In the rest of the paper, categories are treated as bloggers' interests. Bloggers in this social network form the largest component, thus any blogger can be connected to any other blogger through some intermediate bloggers. The social network is undirected. After post processing, we obtain a dataset with 88,784 bloggers, 5,713 unique tags[3], and 60 categories. The BlogCatalog dataset is published at Social Computing Data Repository at Arizona State University[4].

[2]www.blogcatalog.com

[3]Tags that are used by less than 10 users are removed. This process helps to reduce noisy tags or typos in tags.

[4]http://socialcomputing.asu.edu/

Table 2: Statistics on BlogCatalog

Measure	BlogCatalog
Nodes	88,784
Edges	1,409,112
Average Contacts	49
Unique Tags	5,713
Average Tags	4.0

We first introduce baseline approaches and evaluation metrics for verification in the BlogCatalog social network. Then, we present and discuss detailed experimental results.

6.1 Baseline Methods

Practical friend recommendation systems are often application-oriented and domain-dependent, and it is very challenging to implement such systems for experimental evaluation. Thus, we illustrate two baseline methods that are generally applicable to any friend recommendation task.

Triadic Closure (Transitivity Principle) seeks to find similar users in terms of the number of mutual friends, and is solely based on links. This approach returns the top k people who are two hops away (friends of friends) in a social network. Note that it may return potential friends, but not necessarily return the most similar users.

Latent Semantic Indexing (LSI) is used to capture semantic correlation by applying Singular Value Decomposition (SVD). This approach computes the cosine similarity between an arbitrary pair of users in the latent space and can connect like-minded users who are far apart in a social network.

6.2 Evaluation Metrics

The quality is evaluated by the number of shared interests between the seed user and the selected users. More specifically, if the users selected by approach A share more interests with the seed user than those by approach B, intuitively, we say approach A is better.

On BlogCatalog dataset, each individual has explicit categories (or interests) which serve as the ground truth for evaluation purposes. The metric, Mean Shared Interests (MSI), is formally defined in Equation (7),

$$MSI(j) = \frac{1}{n} \sum_{i=1}^{n} \|u_i^{interest} \cap S_i(j)^{interest}\|, \ 1 \le j \le k, \quad (7)$$

where u_i represents the seed user, $S_i(j)$ $(1 \le j \le k)$ represents the j-th recommended user for u_i, noting each user set S_i (ranked in descending order) depends on u_i. We average the shared interests over all users in a social network.

7. EMPIRICAL FINDINGS

In this section, we first demonstrate the effectiveness of the tag network approach with properly selected parameters. Then we show the advantages of the proposed approach compared with other baseline methods. We also study the top k selected users in depth.

7.1 Comparative Study

The diffusion parameter β is sensitive to the outcomes. In Section 2, we suggest small values of β be selected. Figure 4 shows the MSI values with respect to different β values range from 10^{-1} to 10^{-5}. The performance stabilizes when

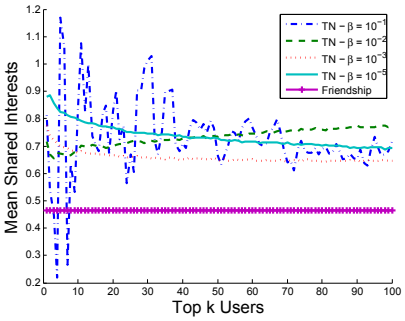


Figure 4: The Shared Interests Correspond to Different Selection of β: Large value of β yields large variation on the user's shared interests. Users selected by tag network approaches share more interests with the seed user than her immediate friends in the social network.

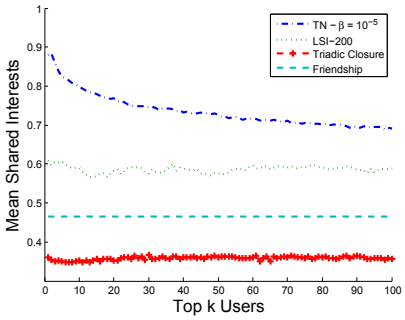


Figure 5: The Shared Interests w.r.t Different Approaches: the tag network approach outperforms the other approaches.

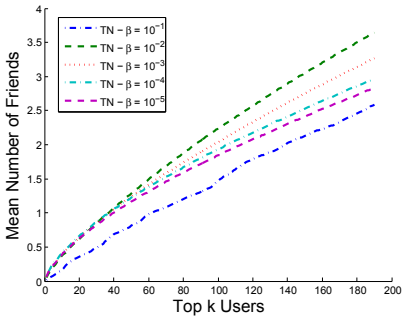


Figure 6: The Number of Friends in the Top k Most Similar Users: immediate friends account for around 2% of the people who are deemed as like-minded.

βs are set to smaller than or equal to 10^{-5}. The x-axis represents the top 100 users sorted in descending order in terms of similarity with the seed user. The y-axis denotes the MSI values between the j-th selected user (excluding the seed user's immediate contacts) with the seed user. The plots suggest that the best performance is achieved when β is set to 10^{-5}, since we often recommend few users as candidates, e.g., 10 or 15. We also notice that large β values cause large variations. For instance, when β is set to 0.1, the performance is not stable. As a baseline measure, we compute the average shared interests between the user and her immediate neighbors, denoted by the lower solid line in Figure 4. The higher MSI values of the proposed approach suggest that more like-minded users could be returned.

Theoretically, in a connected network, there is a path from any user to any other user. Thus, it is possible to connect all like-minded users by following links. However, exhaustive search is expensive and inefficient for a contemporary social network which can have hundreds of millions of nodes. As an alternative, applying triadic closure only searches for candidates up to two hops away. Therefore, the search by triadic closure principle is incomplete.

For comparison, we include all three approaches: triadic closure, LSI, and tag network with a specified parameter. The results are plotted in Figure 5. The LSI approach does provide improvement to some extent compared to the baseline measure as indicated by Friendship. It should be noted that the best performance for LSI is obtained when the latent dimension is set to 200 for the studied dataset. The proposed method outperforms the LSI approach significantly under t-test ($p < 0.001$). In computing the MSI values for above two approaches, the seed user's immediate contacts are excluded. The approach based on triadic closure is not as effective as the other two approaches, as indicated by the bottom curve in Figure 5. Comparing to the baseline methods (or measures), on average, the relative improvements of the tag network approach are 27%, 60%, and 108% for LSI, Friendship, and Triadic Closure, respectively.

Further Discussions Tag network and Latent Semantic Indexing are both capable of capturing the semantic correlation between tags, but diffusion on tag network appears to be more capable than LSI. The probable reasons for this are (1) the collective wisdom from the crowd brings the semantically relevant tags close to each other in terms of the number of hops; (2) although LSI also leverages the tag co-occurrence for dimension reduction, the diffusion kernel is more capable of measuring the similarity between any pair of two tags. We interpret the difference between LSI and diffusion on tag networks: LSI uses one path (i.e. the co-occurrence of two tags), whereas diffusion kernel combines all paths between any two tags (i.e. random walks with different number of steps on the tag network).

7.2 Correlation Analysis

In this section, we demonstrate the overlap between the true friends of the seed user and the top k most similar users. We find that a small set of selected users are actually the user's friends. The correlation between the friends and the returned top k users are presented in Figure 6. The x-axis represents top k most similar users sorted in descending order; y-axis represents the number of users who are actually friends, noting that y-axis values are averaged over all users in the social network. Consistent with the conclusion we draw in Section 7.1, most similar users (around 98%), thoese who share interests with the seed users, are not her immediate friends. We evalute different kernels but they all show very similar performance.

7.3 Hop Distance from the Seed Users

We observe on the BlogCatalog dataset that users that are multiple hops away could be like-minded. Thus, we compute the number of hops between the seed users and their top k

Table 3: Distance Distribution of Top k Candidates

# of Hops	1	2	3	4	5	6
Top 10(%)	1.555	20.197	**55.825**	12.160	0.263	0.002
Top 50(%)	1.062	29.035	**57.406**	12.229	0.264	0.003
Top 100(%)	0.875	28.528	**58.072**	12.260	0.261	0.003

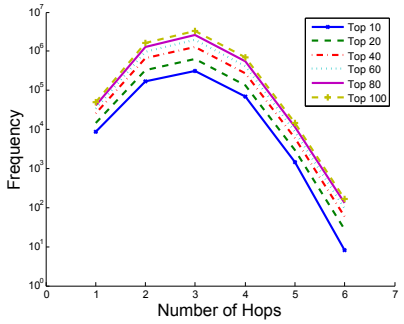


Figure 7: Hop Distance Distribution of the Top k Selected Users from Corresponding Seed Users: the majority of returned users are not immediate friends; small number of long-distanced users are also returned by the proposed approach.

most similar users. The computation is done by a breadth first search starting from a seed user, then each of the top k users is assigned the number of hops from the corresponding seed user. Finally we aggregate the number of users by hop distance from their corresponding seed users.

The distance distribution is presented in Figure 7, in which the curves from bottom to top represent top k (k = 10, 20, ..., 100) users who are considered. As shown in this figure, statistically, the majority of the most similar users are 2, 3, and 4 hops away. A small number of users who are 5 or 6 hops away from the seed users, (the diameter of the BlogCatalog social network is only 7) are also suggested as like-minded. The percentages of users with different hops from the seed users are summarized in Table 3. The immediate friends who are 1-hop away from a seed user account for less than 2%. The above results demonstrate that the tag network approach is capable of returning distant like-minded people for future interactions.

8. RELATED WORK

Recommendation Systems are growing in popularity in social media. Recommending items such as products, movies, services, or information entails trying to predict a user's preferences on things that may be attractive in the future. Collaborative Filtering (CF) is widely used in many applications. The key idea of collaborative filtering is that people who agreed in the past tend to agree in the future. Thus the collaborative filtering algorithm is to find the set of people of similar tastes and recommend items. A second type of recommendation system tries to recommend people instead of items. The Facebook recommendation system *People You May Know*[5] recommends potential friends by mutual friends or the triadic closure principle [10]: if two people have strong connections with a third person, it is more likely there is a strong or weak tie between them [11].

Link prediction is the task to infer future interactions between users in a social network with the knowledge at current time stamp. The key idea of this line of work is to recommend potential friends in terms of proximity with the seed user [8]. Many extensions have been studied recently by leveraging the user profile, activities and interactions, user-generated content [1, 5], network structure [11], etc.

[5]http://blog.facebook.com/blog.php?post=15610312130

9. CONCLUSION AND FUTURE WORK

In this paper, we propose to connect like-minded users via tag network inference. A tag network represents collective knowledge. We demonstrate that the tag network approach outperforms the two baseline methods based on triadic closure and latent semantic indexing. Experimental results also show the proposed approach is capable of recommending users with similar interests who are far apart when using links as the distance metric. The tag network approach can be used for on-line recommendation since the time complexity of recommending top k most alike users for a seed user is linear with respect to the number of users and the number of tags in a social network.

There are several topics that are worth further exploration. One direction is to apply the knowledge of tag networks to other applications such as tag recommendation, query expansion, or dimension reduction by utilizing the semantic relevance between tags, etc. Another line of work would be integrating the link information and tag network for improved friend recommendation. Another possible area is to study tag selection in tag network construction. For instance, some popular tags like news are not indicative; removing those tags may reduce noise in tag networks.

10. REFERENCES

[1] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. "make new friends, but keep the old" — recommending people on social networking sites. In CHI'09.

[2] S. T. Dumais. Enhancing performance in latent semantic indexing (lsi) retrieval. Unpublished manuscript, September 1992.

[3] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.

[4] P. Heymann, A. Paepcke, and H. Garcia-Molina. Tagging human knowledge. In WSDM'10.

[5] D. Horowitz and S. D. Kamvar. The anatomy of a large-scale social search engine. In WWW'10.

[6] M. Joel. *Six Pixels of Separation*. Business Plus, 2009.

[7] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In ICML'02.

[8] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In CIKM'03.

[9] B. Schölkopf, K. Tsuda, and J.-P. Vert, editors. *Kernel Methods in Computational Biology*, chapter Diffusion Kernels, pages 171 – 192. The MIT Press, 2004.

[10] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[11] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.