Who Should I Interact With?

Ba Quan Truong, Xiao Chen, David Frankl
{bqtruong, markcx, dfrankl}@stanford.edu

December 10, 2014

1 Introduction

The growth of social networks such as Facebook, Google+, and Twitter has provided us a great medium to communicate with many people across the world. However, as it turns out, behind a keyboard and a screen, not everyone is nice and not everyone shares our interests. Wrongly choosing people to interact may not only waste our time but can also negatively affect our habits, our emotion, our health and our lifestyles [1,5]. Therefore, even in a social network, it's important to find good users and good friends. We call this *the user quality problem.*

The first challenge of this problem is: what are properties of a good user? In this project, we identify two aspects of a good user:

1. First, a good user in a social network can be good simply because he is just generally good at producing quality content. On a Community Question Answering(CQA) such as stackoverflow, a good user is typically an expert who can provide good answers. Thus, goodness here equates to expertise [14]. On Twitter, a good user is usually an authoritative twitterer who can set the trend and influence many people. On Flickr, a good user can be a user who posts many high-quality photos. Meanwhile, on a chatting site like Chatous, a good user can just be a good conversationalist. With such diverse definitions of "user quality", how can we quantify it? In this project, we shall use a simple, generic definition for "user quality", a good user should be *someone who are endorsed by many in the network.*

2. Second, while a user may be generally well-liked by the community, she may only produce quality content with respect to certain topics. For a user with different interests, her *perceived* user quality may not very high. For example, a user on stackoverflow may be an expert in Java, but she may have low value to a theoretical computer scientist. Similarly, while an American Chatous user may be a great conversationalist since she knows a lot about American pop culture, her conversation with a Chinese person may not turn out well. Thus, the *perceived user quality depends on the topic similarity between two users.*

In this project, we shall jointly address both aspects of user quality. In particular, first, we shall build a signed network in which each edge explicitly or implicitly represents an endorsement from a user to another user. Thus, without considering the user's topics, a high quality user is simply a person with many incoming edges from other high-quality users. Hence, a natural model to capture such measurement is using PageRank [11], a standard and popular model to measure node centrality.

Next, if we disregard the endorsement in the network, a standard method to extract a user's topics topics of interest (typically from their history) is Latent Dirichlet allocation (LDA). We shall explore Topic-Sensitive PageRank, a model which incorporates both PageRank and LDA to jointly address both aspects of user quality.

2 Related Work

Our topic-sensitive user quality problem has been explored separately in different applications. For instance, on Twitter, [7, 13] went beyond the number-of-follower measurement used in Twitter to quantify the authoritative/influential each Twitterer is. In particular, [13] followed a similar approach to ours by first distilling the topics of each user based on their tweet history and then introducing a new algorithm, TwitterRank, which is an extension of PageRank to compute user's authoritativeness. Similarly, on the StackOverflow network, [14] introduces a probabilistic generative model named Topic Expertise Model to predict the expertise of each user in each topic.

A key application of accurate user quality prediction is prediction/recommendation of links on a social network. Link prediction/recommendation on social networks is a well explored topic, starting from Liben-Nowell and Kleinberg [10] with the premise that people/entities are better linked if they share many existing friends/neighbors. Their results suggest that Adamic and Adar's node proximity measure [2] performs best. Meanwhile, Tong *et.al.* uses Random Walks to leverage information about nodes beyond the direct neighbors to estimate the node proximity in graphs [12]. The strength of these approaches is that they take advantage of the networks' features (e.g. the interactions among users) and admit that pure node characteristics (e.g. the user profile) are not good enough features to predict the links. Specifically, Backstrom *et.al.* pointed out that applying classification using user's profile features is challenging, because the network is very sparse, so the number of predicted edges compared to the number of possible edges is very small [4]. Classification generally performs purely on such imbalanced datasets. Then, they proposed a link prediction model using Supervised Random Walks, which combines both network features and node features, and produces significant improvement. While these results are very promising, one of the drawbacks of these approaches is that they only consider networks with positive edges (e.g. friendships) without accounting for the fact that a user may specifically refuse to communicate with another user. For example, on Facebook, a user may specifically block another user, and in our Chatous dataset, a user may report another user. It is probably not a good idea to recommend friendship between not only these people, but also their friends as well.

Meanwhile, there is another research trend which explores graphs with both positive and negative links and attempts to predict the sign of an edge link. Lescovec *et.al.* consider the voting graphs in several datasets such as Wikipedia, Stackoverflow and Epinions as directed graphs, in which a vote is a positive edge expressing positive attitude from a user to another user and, similarly, a down-vote corresponds to negative edge/attitude [8,9]. Using that graph modeling, [8] predicted that the sign of the edges follows the balance and status models. The balance model follows the premise that "a friend of my friend is my friend" and "a friend of my enemy is my enemy". Using these intuitions, the balance model considers each edge in terms of its triads, and uses the sign of the other edges in the triads to predict the sign of the current edge. Meanwhile, the status model argues that positive/negative edges indicate higher/lower status of the recipient compared to the voter and then uses the deduced status to predict the sign. Such models are very intuitive and the results in [8] suggested that these properties appear in many datasets. Meanwhile, [9] confirmed these properties on a specific case study, Wikipedia voting, and also explored the effect of the sequential nature of voting where each voter's decision is also affected the dynamic of voting played out over time. Anderson *et.al.* expanded it further by considering the effect of the similarity between the voter and the recipient on the probability of receiving positive votes [3]. They discovered that highly similar voters tend to vote more positively and be less vulnerable to the status difference with the candidate. They also discovered the "selection effect" where high-status voters tend to participate in elections which impact areas closest to their interests.

While these results are interesting and intuitive, we notice all of these papers only predict the sign of the edges without quantify their weight. This is important in our case, since our task is to recommend a small set of positive edges to users.

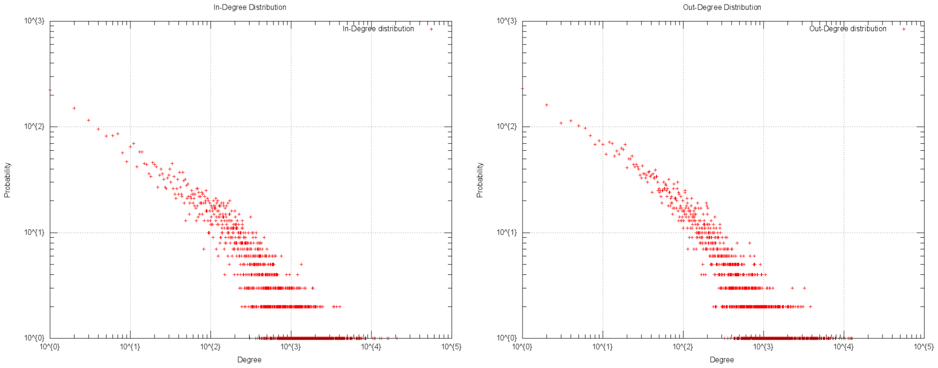
3 Datasets

3.1 Flickr

Our Flickr dataset is downloaded from SNAP repository[1]. The dataset contains images from 4 different sources: PASCAL, CLEF, MIR and NUS. For our analysis, we use the images from NUS, since this is the largest source among the four. The Flickr-NUS dataset contains 244,762 images uploaded by 48,870 unique users from Feb 24, 1989 to Sep 06, 2011. Among them, there are 10,915 users that have uploaded at least 5 images. Our analysis shall focus on these 10,915 users.

[1] http://snap.stanford.edu/data/web-flickr.html

Next, we use the Flickr API to crawl the favorite images of these users by July 1, 2008, and then extract the owner of each of these photos. If a user A favorites an image of user B, we make a directional edge from A to B in our network. Thus, our network is similar to a network of positive votes in which each edge from A to B represents an endorsement of A to B. Using this method, we form a network of 3,128,832 edges and 8,585 nodes. Self-edges are discarded (*i.e.* a user can't like his/her own images). There are two important notices here. First, the network is a multigraph since a user may choose multiple photos of the same user as favorites. Second, the number of nodes in our network is smaller than the original number of users we consider, because there are some users without any favorite activities and there are some users who are no longer accessible. Figure 1 shows the degree distribution of our Flickr favorite network. While our network is very dense (since most users in our network are very active), the degree distribution still roughly follows a Power Law.



(a) In-Degree Distribution

(b) Out-Degree Distribution

Figure 1: Flickr's Degree Distribution

To experiment on our models, we also built a test dataset. The test dataset consists of 200 images in Flickr-NUS dataset that were uploaded after July 1, 2008 (thus, the test set and the training set do not overlap). Each of these 200 images is selected in a way such that it is uploaded by a user among our 10,915 test users and is favorited by at least one user among our 10,915 test users. Furthermore, the users favoriting these images are all different.

3.2 Chatous

Our Chatous dataset contains information about 332887 unique profiles, and 9050712 chats between randomly matched users. The Chatous raw dataset is a 2 week snapshot of user activity. User profile data follows the schema:

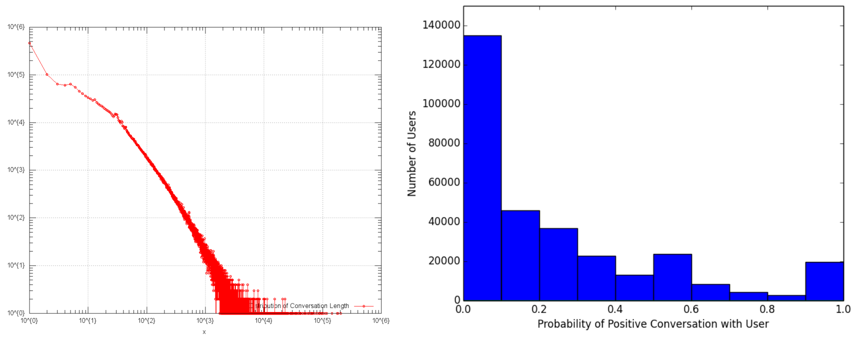
- Profile(Profile ID, User ID, Time Created, Age, Gender, Location, Location Flag);

and chat data follows the schema:

- Chat(Chat ID, Fuser ID, Suser ID, Friendship Status, Chat Created Date, Chat Finished Date, Length of Chat, Disconnector, Reported User ID, Reason for Reporting, Word Histograms);

The word histograms contain 925240 unique words, with actual words replaced by unique word identifiers, to protect the privacy of the users. The large majority of conversations consist of just 0 or 1 words. Figure 2a shows conversation length plotted on a log-log scale.

This issue of recommending conversation partners in the Chatous network is a difficult one, because most users are frequently involved in low-quality conversations. Figure 2b shows the lack of users that consistently have high-quality conversations, even with the relatively low standard we devise below for classification as a high-quality conversation.



(a) Conversation Length

(b) User Quality

3.3 Stack Overflow

Stackoverflow has a vast amount of data. We pick the math.stackexchange.com forum as our experiment dataset. There are 786162 posts and over 87811 users in the forum. For each post, we store the creator, post body, post tags, comment count, and post score.

4 Approaches

4.1 Similarity Measures

Similarity measures provide a good baseline approach to predicting positive interactions. The general idea is that the more similar two users are, the more likely they are to have a positive interaction. To measure the similarity between user word histograms in the Chatous data, we use Cosine Similarity:

$$\text{SIM}(a,b) = \frac{a \cdot b}{||a|| \, ||b||}$$

To measure the similarity of two users in terms of the distribution over their topic interests, we use Jensen-Shannon Divergence:

$$\begin{aligned} \text{SIM}(P,Q) &= 1 - \text{JSD}(P,Q) \\ &= 1 - \left(\frac{1}{2} D(P||M) + \frac{1}{2} D(Q||M) \right) \end{aligned}$$

where $M = \frac{1}{2}(P+Q)$, and $D(A||B)$ denotes Kullback-Leibler divergence of B from A.

4.2 PageRank

Informally, PageRank is an algorithm to capture the "node centrality" or "importance" of a network. It's based on the premise that important nodes receive many in-links and important nodes usually give their outlinks to other important nodes. To realize these two aspects, PageRank uses a random walk model in which the PageRank value of each node is the probability that the random walker is at that node. However, to avoid the sink nodes which have no out-links, PageRank's random walker has a probability of $1-\alpha$ to teleport to a random node in the graph. Typically, $\alpha = 0.85$.

Mathematically, the goal of PageRank is to compute a weight vector $\mathbf{r} \in \mathbb{R}^{n \times 1}$ where n is the number of nodes in the network. PageRank is an iterative algorithm where $\mathbf{r}$ is updated at each iteration. The output $\mathbf{r}$ is the converged vector after multiple iterations. At iteration k, the PageRank vector $\mathbf{r}^{(k)}$ is updated as follows

$$\mathbf{r}^{(k)} = \alpha \mathbf{M} \mathbf{r}^{(k-1)} + (1-\alpha)\mathbf{t}$$

where $\mathbf{M}$ and $\mathbf{t}$ are the transition matrix and the teleport vector, respectively. Typically, $\mathbf{M}_{ij} = \frac{1}{d_j}$ where d_j is the j-th node's out-degree and $\mathbf{t}$ is a stochastic vector in which all values are equal.

4.3 Topic-Sensitive PageRank

While PageRank is a simple but powerful tool to capture user quality, it does not take advantage of the topics of the user. Topic-Sensitive PageRank [6] resolves that problem by biasing the random walker towards a set of nodes belonging to a topic. Mathematically, Topic-Sensitive PageRank alters the teleport vector $\mathbf{t}$ such that the random walker is biased toward nodes of a particular topic (recalling that standard PageRank uses equally-valued stochastic $\mathbf{t}$ corresponding to a uniform distribution). Such biasing affects all iterations in the PageRank process and generally increases the PageRank values of nodes in the topics and their neighbors. Thus, it incorporates both the global endorsement of the network and the topics.

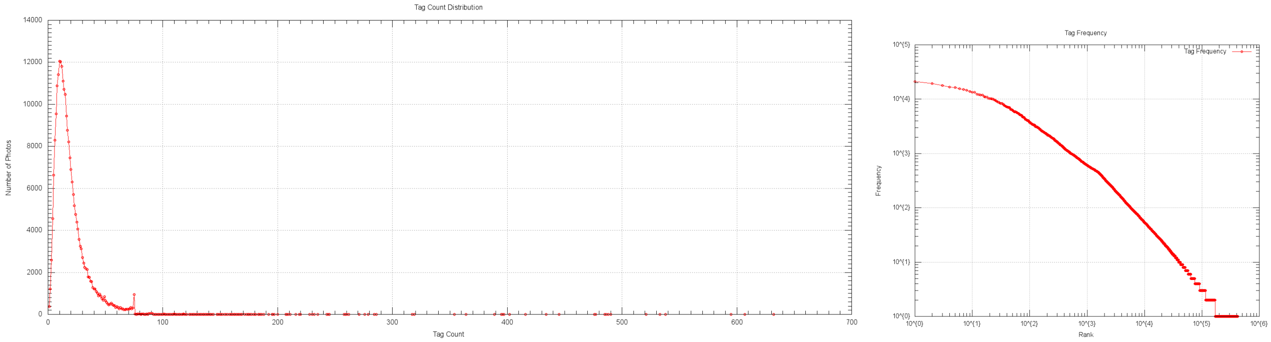
Using Topic-Sensitive PageRank, there are two options of altering $\mathbf{t}$. First, following [6], for each topic t, we could use a threshold τ on the probability p_{ut} that a user u is interested in a topic computed by LDA. Using τ, we divide the set of nodes V into two groups, S containing nodes belong to t and $V \backslash S$ containing nodes not belong to t. Then, $\mathbf{t}_i = \frac{1}{|S|}$ if the i-th node is in S and 0 otherwise. Naturally, the difficulty here is to find good τ. Alternatively, as suggested by [13], we could use normalize the p_{ut} across all users and use the normalized value of each user for the stochastic vector $\mathbf{t}$.

5 Experimental Results

5.1 Flickr

5.1.1 Topic Extraction

The first step of our approach is to extract the topics of our dataset. On Flickr, we crawled the tags of each image in Flickr-NUS and then use Latent Dirichlet allocation (LDA) on this image set (the number of topics is set at 50). Each image is considered a "document" and each tag is considered a "term". Since LDA assumes bag-of-word model, it fits our tag collection perfectly. Notice that we consider images instead of users as documents because each user may have a diverse set of interests while an image typically focuses on a single topic. Figure 3 displays the tag count (*i.e.* number of tags per photo) and tag frequency (*i.e.* number of photos per tag) distribution. Since most users typically provide some tags to their photos during uploading, the tag count peaks at 5. Meanwhile, we could see the tag frequency distribution follows power-law.



(a) Tag Count Distribution

(b) Tag Frequency Distribution

Figure 3: Flickr's Tag Distribution

The top-7 tags of 6 sample topics are shown below. We can see that each topic is very intuitive. For example, topic 1 is likely about Germany while topic 4 is about the UK. Topic 2 is about airplanes while topic 3 is likely about cats and zoos.

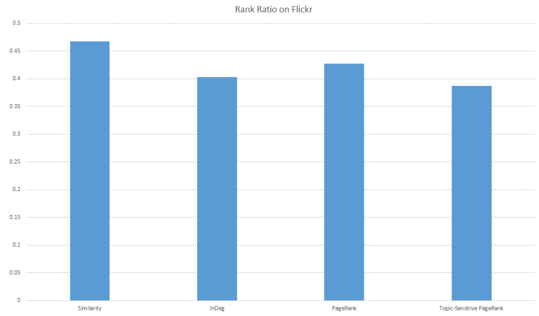


Figure 4: Flickr Results

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
germany	airport	animal	England	water	snow
deutschland	airplane	cat	church	Sky	mountains
berlin	plane	zoo	london	clouds	landscape
televisiontower	aircraft	specanimal	architecture	beach	ice
Europe	aviation	animals	UK	sunset	mountain
2006	flying	ImpressedBeauty	Britain	landscape	winter
bavaria	jet	wildlife	Europe	sea	nature

5.1.2 Results

Experimental Data. For Flickr, we shall evaluate of results using the recommendation task. Specifically, we select 200 images uploaded and favorited by 200 different test users among our 10,915 test users and were uploaded after July 1, 2008. Thus, the test set and the training set do not overlap.

Performance Metrics. For each test image, we compute the ground-truth score s_t of its uploader. Then, we random sampling 500 users from our test user and then find the rank of s_t among the 500 users' score and then divide it by 500 to get the rank ratio. Low rank ratio means our method successfully find the users of high quality who usually uploads high-quality images and who the user should favorite.

Results. Figure 4 displays the results of four approaches on Flickr. InDeg means that the score of each user is their in-degree in the network. Similarity means that the score of each user is the similarity score to the test photo. From the results, we could see that, first, all four approaches increase the performance compared to the random approach (which naturally has the rank ratio of 0.5). Second, Topic-Sensitive PageRank produces the best result as it incorporates both the centrality approach of PageRank and the topic distribution. Finally, we could see that the improvement of all approaches are minimal. There are multiple reasons for that. First, Flickr UI does not strongly emphasize the uploader of images. Thus, users have little incentive to like the images of users who they like a lot. Second, in photography, there is little correlation between the quality of past images to the current images. Amateur users could get good images if they capture good moments. Finally, to fully understand the quality of users and their images, the visual content must be analyzed which is totally neglected in our approach since we focus more on the network part.

5.2 Chatous

While there are various methods for characterizing conversation quality, we opt for a simple metric that encompasses the notion of conversation quality with sufficient accuracy. First, if a participant in the conversation flags the conversation, it is automatically negatively characterized. Then, if the conversation lasts less than two words, it is negatively characterized. Otherwise, it is positively characterized.

We first implement a simple baseline approach: recommend a chat partner randomly. This approach leads to a lot of low-quality conversations. With random partners, the probability of high-quality conversation is very low:

$$P(\text{quality conversation with random partner}) = 0.187$$

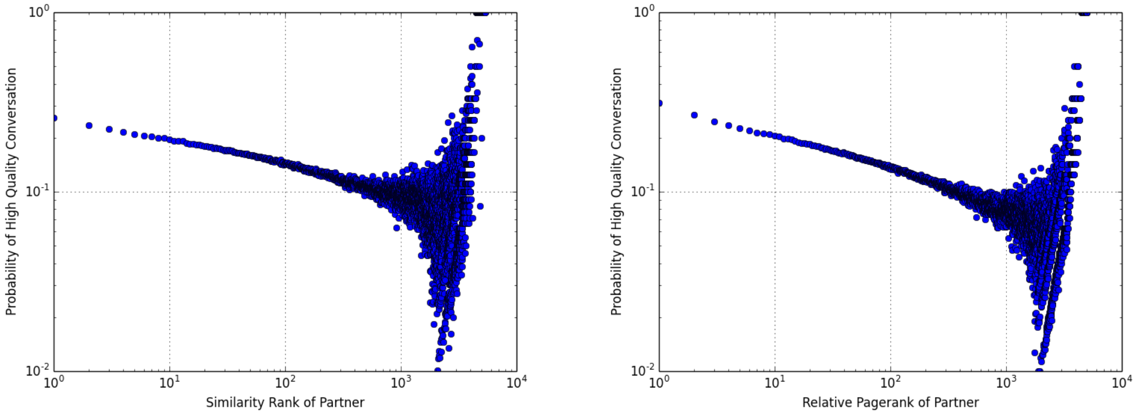
5.2.1 Similarity Measures

The challenge is to improve upon this figure, by matching users that are more likely to share conversational interests. As a first improvement, we use the Cosine Similarity to predict conversation quality. For each user, we aggregate the user's word counts over all conversations. Then, for each conversation partner v of a given user u, we calculate similarity score. We rank conversation partners, and match u with his or her most similar partner. By selecting the partner with highest similarity, we obtain a moderate improvement over the baseline approach:

$$P(\text{quality conversation with most similar partner}) = 0.259$$

Next, we more thoroughly explore the notion of similarity of interests. The fundamental idea behind Cosine Similarity is that each user's word histogram relates to their underlying interests, so user's with more similar word histograms should share common interests. We generalize this by using LDA to assign a distribution of interest topics to each user. These distributions allow us to use Jensen-Shannon Divergence to directly characterize the distance between the interests of two users. This approach again improves conversation quality with the most similar partner:

$$P(\text{Quality conversation with closest JSD partner}) = 0.354$$



(a) Relative Similarity

(b) Relative PageRank

Figure 5: Chatous Recommendation Results

5.2.2 PageRank

To convert the inherently undirected graph into a graph suitable for PageRank, we create a new, directed graph as follows. For each positive conversation between users u and v, add an edge from u to v and from v to u. For each flagged conversation, add an edge from the flagged user to the flagger. To test the efficacy of PageRank on predicting positive conversations, we assign to each user in our test set the conversation partner who has the highest PageRank score. This strategy yields a large improvement over the baseline:

$$P(\text{quality conversation with partner of highest PageRank}) = 0.458$$

It is important to note, however, that this approach scales poorly in reality. Since PageRank score is independent of conversation partner, we find that top users will be recommended too frequently to fulfill all conversations desired of them. The use of Topic-Sensitive PageRank partially addresses this issue, by increasing the specificity of the ranking, and giving users recognition for quality performance in specific topics.

To apply Topic-Sensitive PageRank to the Chatous data, we again use LDA to group users into categories based on topical interests. We extract topic information based on each user's aggregated word histogram. This procedure gives a set of topics associated with a multinomial distribution over words. We choose to separate into 10 topics. For purposes of computational efficiency, we only take into account the 100 most used words, and apply LDA to a sampling of 10000 users.

Given topical groupings of words, we model the topical interests of a user based on the their word usage. We assign each word to the topic of greatest computed probability. Then, we assign to each user the topic which has the maximum count of words over the user's aggregated word histogram. To test our results, we rank each user's conversation partners according to their PageRank value on the assigned topic. We recommend the partner with highest ranking. This approach yields a slight improvement over general pagerank:

$$P(\text{quality conversation with partner of highest Topic Specifc PageRank}) = 0.495$$

Figure 5 gives two of the four plots for recommendation quality as a function of relative ranking among conversation partners. As similarity and PageRank rankings get worse (increase), the probability of quality conversation decreases as well. This holds consistent with our intuition that similarity and PageRank should be correlated with conversation quality. The plots for the omitted approaches as similar, with slight vertical shifts to reflect general difference in strength of recommendations.

5.3 Stackoverflow

As a first step, we parse the posts file using LDA, as above, to get the matrix of terms vs topics. We set topics number is 50, and we get following results. Noticing that terms × topics matrix is very large, we only illustrates top-4 frequent appeared terms from topic 0 to topic 4

Topic 0:	Topic 1:	Topic 2:	Topic 3:	Topic 4:
algebra-precalculus	calculus	reference-request	geometry	sequences-and-series
polynomials	multivariable-calculus	representation-theory	circle	real-analysis
roots	integration	lie-algebras	euclidean-geometry	power-series
functions	linear-algebra	lie-groups	triangle	convergence

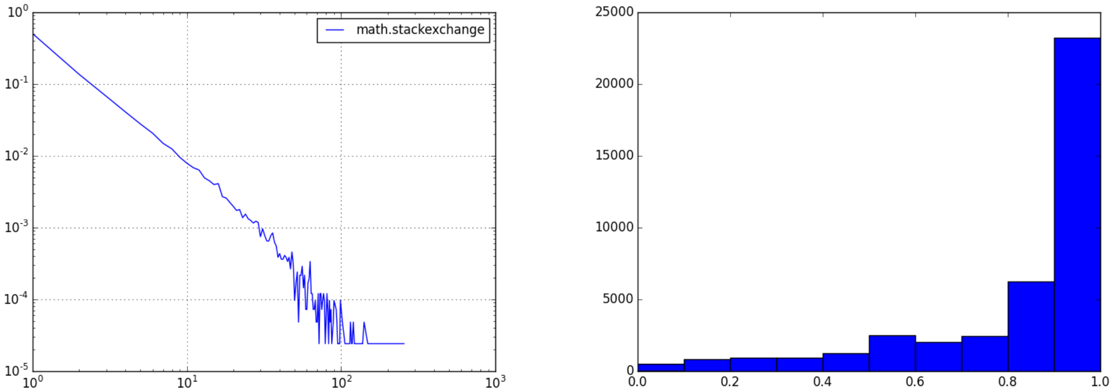
From the above table, we observe that topic 0 may related to algebra while topic 3 may focus on geometry. We run LDA to get each post's topic distribution. Then we aggregate multiple posts (if there are any) based on each user, because we are curious about the topics of interest for individual users. Thus we construct the user vs topics matrix. We manually inpect the results to discover that they are a very promising estimation. For example, we search argmax for topic 0 (treated as 'algebra') and found user '169311' has highest value. we then search the original post file to check if user '169311' is really interested in topic 0. It turns out that user '169311' is asking a question about prove an inequality by Taylor's formula. From here, we have the following issue to resolve: given a posted question, who is best qualified to answer it. We come up with the following underlying factors:

- If a user is interested in related topics, they are well qualified to answer the question.
- If a user has high reputation, they are well qualified to answer any question.

These assumptions lead us to construct a user endorsement based graph and incorprate the idea of topic specific pagerank.

We generate this graph based on acceptedAnwserId. If an user A's answer is accepted by user B, and we add a directed edge from user B to user A, which represents the vote/endorsement. The user endorse based graph has 41440 nodes and 144851 edges. The average clustering coefficient is 0.03739. Degree distribution is diplayed in Figure 6a

We run the basic pagerank as the baseline, and find the top 5 users. We also implement the topic specific pagerank to get different set of scores with top 5 users.



(a) stackoverflow degree distribution

(b) stackoverflow similarity count histogram

node	original pagerank	node	topic specific pagerank (topic 0)
246	0.0508387919318	246	0.000341706974449
128	0.0116853565044	128	7.03456621129e-05
3346	0.0113945691629	39	1.66142742961e-05
22	0.0110926747462	1856	7.2197020025e-06
651	0.0110023260158	224	5.24534635103e-06

We can see that the score from topic specific pagerank is lower than the original pagerank score. We take a further look at the data and find that it may be reasonable, because more stringent relations are focused on the network. And the original high score user who answer a lot of geometry (topic 3) may not be a high score user in algebra (topic 0). We also take 10% data out and estimate the user to user link pair base on pagerank and topic sensitive pagerank. The original pagerank give over 20.12% accuracy of link prediction, but the topic sensitive pagerank give over 27.13% accuracy (if the link is between the top 5 high score user we think it is correct)

For recommendation of answers, we randomly pick a user and calculate the similarity of all other user given such a user's topic distribution. We plot out the histogram and find that over 60% percent user have similar interested topics with the picked user (Figure 6b).

6 Difficulties

With respect to the Chatous data, we are limited by the fact that word usage data is scrubbed of real words, and replaced with word IDs. This limits our ability to analyze the content of conversations, and ultimately reduces the power of our evaluation metrics, because we cannot be sure that a conversation is high-quality unless we manually examine its content.

With the Flickr data, difficulties arise in generating an effective Topic-Sensitive PageRank. A couple of underlying factors contribute to this difficulty. For one, users commonly favorite a wide range of photos, with little correlation between photographers or topics. In addition, the quality of each photographer's photos may vary widely, which makes it difficult to predict another user's response.

7 Future Work

As discussed above, we are limited in analyzing the content of a conversation due to the anonymized nature of the dataset. Even with a lossy representation of conversation quality, we were able to produce a relatively effective recommendation engine using Topic-Sensitive PageRank. Should actual word data become available, for Chatous or another chat network, a more complicated analysis including NLP techniques such as Sentiment Analysis can be performed. A direction for future work is to perform Topic-Sensitive PageRank on a graph generated using more complicated parameters. By incorporating more data in the recommendations, it is likely that we can further improve recommendation accuracy.

References

[1] The positives and negatives of using social networking sites. `http://www.bbc.co.uk/schoolreport/22065333`.

[2] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[3] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Effects of user similarity in social media. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 703–712. ACM, 2012.

[4] Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 635–644. ACM, 2011.

[5] Nick Bilton. Reclaiming our (real) lives from social media. `http://www.nytimes.com/2014/07/17/fashion/reclaiming-our-real-lives-from-social-media.html?_r=0`.

[6] Taher H Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM, 2002.

[7] Alex Leavitt, Evan Burchard, David Fisher, and Sam Gilbert. The influentials: New approaches for analyzing influence on twitter. *Web Ecology Project*, 4(2):1–18, 2009.

[8] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650. ACM, 2010.

[9] Jure Leskovec, Daniel P Huttenlocher, and Jon M Kleinberg. Governance in social media: A case study of the wikipedia promotion process. In *ICWSM*, 2010.

[10] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[11] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[12] Hanghang Tong, Christos Faloutsos, and Yehuda Koren. Fast direction-aware proximity for graph mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 747–756. ACM, 2007.

[13] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[14] Liu Yang, Minghui Qiu, Swapna Gottipati, Feida Zhu, Jing Jiang, Huiping Sun, and Zhong Chen. Cqarank: jointly model topics and expertise in community question answering. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 99–108. ACM, 2013.