How Accurately Can One's Interests Be Inferred From Friends?

Zhen Wen
IBM T.J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532 USA
zhenwen@us.ibm.com

Ching-Yung Lin
IBM T.J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532 USA
chingyung@us.ibm.com

ABSTRACT

Search and recommendation systems must effectively model user interests in order to provide personalized results. The proliferation of social software makes social network an increasingly important source for user interest modeling, because of the social influence and correlation among friends. However, there are large variations in people's contribution of social content. Therefore, it is impractical to accurately model interests for all users. As a result, applications need to decide whether to utilize a user interest model based on its accuracy. To address this challenge, we present a study on the accuracy of user interests inferred from three types of social content: social bookmarking, file sharing, and electronic communication, in an organizational social network within a large-scale enterprise. First, we demonstrate that combining different types of social content to infer user interests outperforms methods that use only one type of social content. Second, we present a technique to predict the inference accuracy based on easily observed network characteristics, including user activeness, network in-degree, out-degree, and betweenness centrality.

Categories and Subject Descriptors: J.4 [Computer Applications]: Social and Behavioral Sciences

General Terms: Measurement, Experimentation

Keywords: User modeling, Social networks, Accuracy

1. INTRODUCTION

Modeling user interests to meet individual user needs is important for personalized search and recommender systems. Recently, the proliferation of online social networks spark an interests of leveraging social network to infer user interests, based on the existence of social influence and correlation among neighbors in social networks. For many applications, it is difficult to observe sufficient behavior of a large number of users. In such scenarios, inferring their interests from their friends can be the only viable solution. For example, for a new user in a social application, the application may only have information about his friends who are already using it. To motivate the new user to actively participate, the application may want to provide personalized recommendations of relevant content. To this end, the application has to infer his interests from friends.

However, there exists huge variation in the types and amount of information in social network. According to existing studies [1] on enterprise social networks, a small percentage of employees (e.g., $< 10\%$) may actively contribute social content using one or more social software (e.g., blogs and social bookmarking). But a large number of employees may seldom do so. That results in both a demand and a challenge for accurate user interest modeling, especially for *inactive users* that do not contribute social content. On one hand, search and recommendation systems need accurate user interest modeling to provide personalized results, and thus may help to increase the usage of social software. However, the available observations of users are sparse and exist in multiple types of social media. To address such a challenge, we first need to combine multiple types of social media to improve user interest modeling. Second, in order not to diminish the quality of personalized search and recommendation results by inaccurately inferred user interests, a computational method is demanded to measure inference accuracy based on observable features such as the user's social network characteristics.

This paper presents a study addressing this challenge. We focus on the accuracy of user interests inferred from neighbors in a large scale organizational social network. In addition, we examine a set of social network factors to predict the inference accuracy based on network characteristics that can be easily observed. Our work is based on a privacy-preserving organizational social network analysis system [3] that gathers, crawls and mines various types of data sources within an organization, including people's communication data such as email, instant message communications, and Web 2.0 social media such as blogs, wiki, social bookmarking and file sharing. The system is deployed within a large-scale corporation in more than 70 countries for over 3 years. After anonymizing the identity and the content of these data, we are able to quantitatively infer the social networks of 400K employees within the organization. In this organization, nearly 9K volunteers contributed their electronic communication records, 16K people used social bookmarking, 14K people shared files, and 5K employees blogged.

2. INFER USER INTERESTS

Because users' contributed content reveal their interests, we model user interests as a set of latent topics extracted from their communication data and contributed social content. We use LDA to extract latent topics and define a $U \times T$ matrix $\mathbf{S}$ to describe user interests using topics, where U is the total number of employees and T is the total number of topics ($T = 1200$). An element s_{ij} in $\mathbf{S}$ denotes the degree the i-th employee is interested in the j-th extracted topic.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

Condition	Max	Mean	Min	St. Deviation
1	59.4%	19.2%	5.1%	10.7%
2	44.9%	12.7%	3.0%	7.2%
3	62.1%	29.6%	3.8%	14.1%
4	100%	45.1%	4.2%	21.7%

Table 1: Accuracy of user interests inference.

Next, we use network autocorrelation model [2] to infer user interests from their neighbors in social networks. In our study, we estimate the degree the i-th user is interested in the j-th topic as $s_{ij} = \sum_{k=1}^{U}(w_{ki} \cdot s_{kj})$, where the weight w_{ki} as an exponential function of the social distance between user k and user i. It is defined by considering the degree of separation between them and the amount of their communication [4].

To measure the quality of the inferred interests, we define the inference accuracy as: $C = \frac{1}{N}\sum_{j=1}^{N} \max_{t' \in \mathbb{T'}_N}[\cos(t_j, t')]$, where t_j is the j-th topic in the inferred top-N interests, $\mathbb{T'}_N$ is the ground-truth top-N interests. Intuitively, the equation calculates how many inferred top-N interests are similar to the top-N ground truth. In our study, we set $N = 10$ and use the interests extracted from a user's contributed content as his ground-truth interests.

Then, we perform 10-fold cross validation to compute the inference accuracy. In each round, we leave out the data of one-tenth of the users (testing set), and infer their interests using only the extracted interests for the other nine-tenths of the users (training set). To investigate the effectiveness of combining different types of social content, we conduct the experiment in four conditions: (1) using social bookmarking data only, (2) using file sharing data only, (3) using electronic communication data only, and (4) using all three types of data. Table 1 shows the inference accuracy in the four conditions. The comparison results demonstrate the significant advantage of combining multiple sources social information. We attribute the significant improvement to the much wider coverage of the combined social content. Table 1 also shows that, although the mean accuracy is reasonably good, the variance is huge. We hypothesize that the variance is due to the large variation in people's contributed content and their positions in the social network.

2.1 Predict Inference Accuracy

Because there is large variance in the accuracy of inferring user interests from social network, it is difficult for practical search and recommender applications to decide whether utilizing the inferred interests can improve results. Therefore, we examine a set of relevant network factors to predict the inference accuracy.

We hypothesize that a user's interests are impacted by the type and amount of content contributed in his ego network as well as the structural characteristics of the ego network. Specifically, we examine four factors including *user activeness* measured by the amount of contributed content, *network in-degree*, *network out-degree* and *between centrality*. We focus on studying the factors in the user's three-degree ego network. For each factor, we extract corresponding feature for three sub ego networks including one-degree neighbors, two-degree neighbors, and three-degree neighbors respectively. This allows us to assess the different influence from neighbors of different degrees of separation. In addition, the role of the user himself in the social network may influence interest inference. For example, a user that plays an important role in the information flow within the network may be more likely to share interests with his neighbors. Therefore, we also extract ego network feature of the user himself including in-degree, out-degree and betweenness.

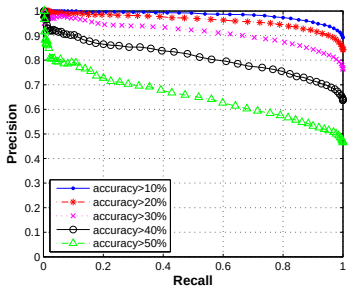

Figure 1: The precision-recall curves for inference accuracy classification.

Then, we choose to use support vector regression (SVR) determine the relationship between the set of features and the prediction accuracy. To evaluate the quality of the inferred top-N interests, we use the prediction to classify "accurate" inferred interests (e.g., accuracy $C > 50\%$). To classify "accurate" interest inference, we test whether the accuracy prediction is larger than a threshold TH. A precision-recall curve can be derived by varying the threshold TH. We use 10-fold cross validation to evaluate the classification performance based on support vector regressions. Five criteria are used to classify the inference accuracy. The resulting precision-recall curves are shown in Figure 1. Additional experiments also show that one- and two-degree neighbors are most useful in predicting a user's interests. Moreover, *user activeness* is the most important factors among the four factors studied.

3. CONCLUSION

In this paper, we present a study on the accuracy of inferring user interests from friends in one of the largest organizational social networks. We demonstrate that there exist large variance of the inference accuracy when user contributed content considerably vary and the content types are diverse. To allow search and recommendation applications make informed decisions on when to utilize inferred user interests, we further investigate relevant factors and present a method to predict inference accuracy based on easily observed network features including user activeness, network in-degree, out-degree and betweenness centrality. Our findings can be useful for social applications with widely varied participation rate so that the interests of many people can only be inferred from their friends.

4. REFERENCES

[1] M. Brzozowski, T. Sandholm, and T. Hogg. Effects of feedback and peer pressure on contributions to enterprise social media. In *CSCW*, pages 61–70, 2009.

[2] R. T. Leenders. Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, 24:21–47, 2002.

[3] C. Lin, K. Ehrlich, V. Griffiths-Fisher, and C. Desforges. Smallblue: People mining for expertise search,. *IEEE Multimedia Magazine*, 15(1):78–84, 2008.

[4] L. Wu, C. Lin, S. Aral, and E. Brynjolfsson. Value of social network – a large-scale analysis on network structure impact to financial revenue of information technology consultants. In *The Winter Conference on Business Intelligence*, 2009.