user latent feature space with matrix size $m \times d$, and V denotes the low-dimensional topic latent feature space with matrix size $n \times d$, in which d represents the community dimensionality. By introducing two Gamma parameters α and β, we constrain the elements in the extracted latent feature space non-negative, which will make our community mining results well-grounded, then we turn the probabilistic factor model into the optimization problem of maximizing the following generalized objective function:

$$L = p(C|X)p(T|Y)p(U|\alpha, \beta)p(V|\alpha, \beta)p(Z|\alpha, \beta) \quad (1)$$

where $X = UZ^T$, $Y = UV^T$, in this way, we can find the best low-dimensional matrices U, V and Z with the partial derivative of the objective function (1).

Meaningful User Communities Discovery After we get the extraction results, it's easy to mine meaningful user communities: Each element u_{ik} ($k = 1, ..., d$) in U encodes the preference of user i to latent community k, and each v_{jk} in V can be interpreted as the affinity of topic j to the latent community k. The advantages of our model lie in: (i) It unifies the user-following relationship and user-content information simultaneously, which can help us find the meaningful communities effectively. (ii) We can regulate the weight of matrices C and T conveniently, thus helping us balance the biased impact of noises on the latent community discovery. (iii) By leading in the Gamma distribution parameters α and β, we make elements in the user community matrices non-negative, which will make our experimental results more explainable and meaningful.

3. EXPERIMENTS

3.1 Dataset and Parameter Settings

To evaluate the performance of our community mining model, we build dataset including user-following relationship and user-content information with time interval of 16 days from October 29th, 2011 to November 13th, 2011. After removal of users who post less than one tweet per day, we get 1879 users and 1640 topics. The influence parameter c was empirically set to be 0.5 to evenly weight the matrix C and matrix T. We also tune the parameter d which denotes the latent community dimension size from 4 to 30, and find the best performance at d=12 eventually. As to the Gamma distribution parameters α and β, we set them as the best performance value of 10, 0.02 respectively.

3.2 Results and Discussions

To demonstrate the validity of the proposed model, non-negative matrix factorization (NMF) method is conducted as the baseline, which only considers the user-content information. Looking into the community mining results shown in Table 1, we delightedly find that our model reveals more interesting phenomena of people's concerns in micro-blogging: First of all, our method obtains more personalized clustering results, as shown in part I, both models mine the user community which concerns the topic of South Korea Street Shot very much, besides, our model successfully takes in people who focus on South Korean clothing that often appears on the South Korea Street Shot. Secondly, our method merges user communities whose topics of concern are very close, as shown in part II, two separate user communities concerning similar leisure topics show up through NMF approach, however, considering users' close friendship, our method reasonably merges them. Thirdly, our method mines the meaningful community which can't be mined with NMF relying on user-content information, seeing Part III, our method mines people who care the Chinese drama *The Rhino in Love* very much, which is very popular but doesn't show up through NMF approach. Thus, by integrating users' interest of different aspects, we find more meaningful user communities, which will help us catch people's concerns better in micro-blogging.

Table 1. Community mining results in micro-blogging

Interesting Phenomena	People's Concerns	
	NMF Based Model	Our Integrating Model
Part I	*South Korea Street Shot*	*South Korea Street Shot & South Korean Clothing*
Part II	*Hey Gossip | Today's Fun*	*Hey Gossip & Today's Fun*
Part III	Null	*The Rhino in Love*

Besides, we use the mean value of soft modularity metric Q_s and users' cosine similarity based on content information to evaluate our community mining results. The higher mean value $\bar{\mu}$ means the closer friendship and closer topic interest of people falling into the same community. Table 2 shows the performance of our method compared to the NMF based model.

Table 2. Performance evaluation on community mining results

Evaluation Metric	NMF Based Model	Our Unified Model
$\bar{\mu}$	0.1211	**0.2718**

From Table 2, we can see that our unified model outperforms the baseline NMF-based model, which only considers the user-content information. The results reveal the validity of catching people's concerns from their multi-interest distribution. It is due to our model could contain more valuable latent clustering information than the baseline.

4. CONCLUSION

In this paper, we propose a PFM-based community discovery method by mining people's interest from their friendship network and content information. Preliminary experiments have proved the validity and effectiveness of our approach. The interesting clustering results will help us understand people's concerns and feel the pulse of our society easily. In the future, we would like to make advertising recommendations based on our existing work in micro-blogging, which will be very promising and interesting. This research is supported by the NNSFC project 61172106 and the BJNSF project 4112062.

5. REFERENCES

[1] Singh, V.K., Gao, M., Jain, R. 2010. Situation Detection and Control Using Spatiotemporal Analysis of Microblogs. In WWW'10, pp. 1181-1182.

[2] Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J. 2011. Who Says What to Whom on Twitter. In WWW'11, pp. 705-714.

[3] Lin, Y., Sun, J., Castro, P., Konuru, R., Sundaram, H., Kelliher, A. 2009. MetaFac: Community Discovery via Relational Hypergraph Factorization. In SIGKDD '09, pp. 527-536.

[4] Psorakis, I., Roberts, S., Ebden, M. 2011. Overlapping Community Detection using Bayesian Nonnegative Matrix Factorization. Phys. Rev. E 83, 066114 (2011).

[5] Ma, H., Liu, C., King, I., Lyu, M.R. 2011. Probabilistic Factor Models for Web Site Recommendation. In SIGIR'11, pp.265-274.