Finding Domain Experts in Microblogs

Shao Xianlei, Zhang Chunhong and Ji Yang
Mobile Life and New Media Laboratory, Beijing University of Posts and Telecommunications (BUPT), Beijing, China
shaoxianlei@163.com, {zhangch.bupt.001, ji.yang.0001}@gmail.com

Keywords: Domain Experts Finding System, Microblog Lda, GBDT, User Features.

Abstract: As users and contents of microblogging services gain a sharp increase, it presents the challenge of finding domain experts who are of high profession but generally don't have followers widely. To address this, we propose a domain experts finding system, which consists of three modules: data preprocessing module, user features extracting engine, experts identifying and ranking module. Firstly, we extract three kinds of features for characterizing social media authors, including user profile features, tweeting behavior features and linguistic content features which are generated by our Microblog Latent Dirichlet Allocation(Microblog Lda) model. Secondly, by casting the problem of finding domain experts as a 0-1 classification problem, we use the Gradient Boosted Decision Trees (GBDT) framework to do probabilistic classification over these features, execute a ranking procedure and yield a list of top N users for a given domain. Experimental results on actual datasets show our Microblog Lda outperforms LDA(Latent Dirichlet Allocation) and our system has a high accuracy in the task of finding domain experts in Microblogs.

1 INTRODUCTION

Millions of people turn to microblogging services such as twitter which is known to all and Sina Microblog which is the most influential microblogging services in China to gather real time news or opinions about people, things, or events of interest. Such services are not only used as social networking to stay in touch with friends and colleagues but also used as publishing platforms to create and consume content from sets of users with overlapping or disparate interests.

Through a survey on users' following decisions on Twitter (Ramage, 2010), we can know that the most two common reasons for users to make following decisions are "professional interest" and "technology". From this conclusion and our long-term observation of user behavior, it is not difficult to find that meeting users' demand to access domain expertise of users would make a great significance for both the advancing of microblogging services and the efficiency of using microblogging.

In order to meet users' demand to access expertise, finding the users that are recognized as sources of relevant and trustworthy information in specific domains is an important challenge. But currently, Twitter and Sina Microblog interface fails to support such kinds of services.

Despite the important role of domain expert users in microblogging, the challenge of identifying true experts is trickier than it appears at first blush. Content in microblogging systems is produced by tens to hundreds of millions of users. In microblogging contexts, for any given domain, the number of these content producers even in a single day can easily reach tens of thousands. While this large number can generate notable diversity, it also makes finding the true experts, those generally rated as learned and authoritative in a given domain, challenging.

Furthermore, most domain experts are not as well known as some celebrities known by many people, they are less discoverable due to low network metrics like follower count and the amount of content produced to date. Thus, we cannot use traditional graph-based methods of discrimination degree of user authority to find domain experts. Besides, graph based algorithms are computationally infeasible for near real time scenarios (Pal, 2011) and social graph information has a negligible impact on the overall performance of identifying a user (Pennacchiotti, 2011).

In this paper, we propose a new method for finding domain experts in microblogs. To sum up, the contributions of this paper are: (1) we propose a domain experts finding system which can identify

true experts in Microblogs with high accuracy. (2) A user feature engine is build to extract user features that are useful to identify one's authority. (3) Microblog Lda, which is based on Lda (Blei, 2003) but is more suitable for microblogging-style informal written genres, is proposed to extract users' linguistic content features.

The rest of the paper is organized as follows: Section 2 places our research in the context of previous work. Section 3 gives the framework of our domain experts finding system. Details of each module of our system are provided separately in Section 4 and Section 5. Results of experiments, which are provided in Section 6, show that the Microblog Lda can obtain significant performance gains and the system, as a whole, can achieve high accuracy in finding true experts in a given domain.

2 RELATED WORK

Within the microblogging research field, little work has explored the issue of domain expert identification. There have been several attempts to measure the influence of Twitter users and thereby identify influential users or experts (Bakshy, 2011; Cha, 2010; Romero, 2011). To our knowledge, there have been only two notable efforts that have approached the problem of identifying experts in specific topics (Weng, 2010; Pal, 2011). (Weng, 2010) proposed a Page-Rank like algorithm TwitterRank that uses both the Twitter graph and processed information from tweets to identify experts in particular topics. On the other hand, (Pal, 2011) used clustering and ranking on more than 15 features extracted from the Twitter graph and the tweets posted by users.

While somewhat similar to paper (Pal, 2011), our method differs in several important ways. Firstly, in paper (Pal, 2011), authors only emphasized users' tweeting behavior features but ignored the precise linguistic content features which can make great significant to domain experts finding task. In our paper, we choose several features used in (Pal, 2011) which are suitable for our target users – Sina Microblog users but also add some more features. Secondly, apart from users' tweeting behavior, we also make use of users' profile features and linguistic content features and use a new method to build the features of users. Finally , our approach offers the potential advantage over network-based calculations in that it is less likely to interface by a few users with high popularity(i.e., celebritieds).

Outside microblogging, finding authoritative users generally has been widely studied. Authority finding has been explored extensively on the World Wide Web. Amongst the most popular graph based algorithms towards this goal are PageRank, HITS and their variations (Page, 1998; Kleinberg, 1998; Farahat, 2002). Also predating microblogging, several efforts have attempted to surface authoritative bloggers. (Java, 2006) model the spread of influence on the Blogosphere in order to select an influential set of bloggers which maximize the spread of information on the blogosphere.

Authority finding has also been explored extensively in the domain of Community question answering(CQA). Among most of the models proposed, some authors used network modeling approach(i.e., Agichtein, 2008). Others modeled CQA as a graph induced as a result of a users' interactions with other community members (Jurczyk, 2007; Zhang, 2007). Still other approaches used characteristics of users' interactions (Bouguessa, 2008, Pal, 2010).

In the domain of academic search, authority identification also has been studied extensively. (Tang, 2008) studied the problom of expertise search in their academic search system-ArnetMiner. (Kempe, 2003) modeled the spread of influence in co-authorshipnetworks.

Summarizing related work, the problom of finding authority has been explored extensively in other domains. Among these work, some used network analysis approaches which is computationally expensive, some used structed information (i.e., users' interaction behaviors) and some used both appraoches in an integrated way. Our domain of interest, microblogging, has seen far less attention. As mentioned above, we feel our approach extends research in the following points:apart from users' interaction behaviors, we also use users' linguitic content features which carry rich information about users; without using graph-based approach, we use a classification approach which is computationally tractable.

3 DOMAIN EXPERTS FINDING SYSTEM

Our domain experts finding system mainly consists of three parts: data preprocessing module, user features extracting enginee, experts identifying and ranking module. The framework of our systen is shwon in the following Figure1.

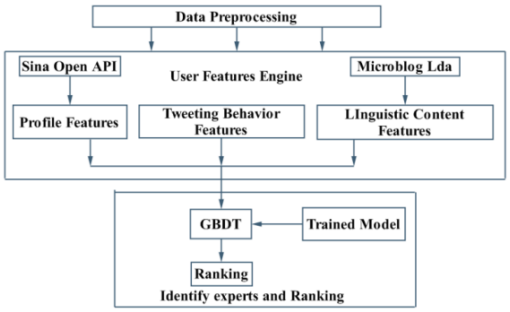

Figure 1: Framework of Domain Experts Finding System.

The work of data preprocessing module is to prepare cleaned source data for features engine and experts identifying and ranking module. Details of this module's workflow are described in Section 6.

In our proposed system, user features extracting engine can automatically construct user features and extract numerous features that are useful in domain expert authentication. In Section 4, we will describe the details of user features extracting engine and give a comprehensive analysis to the features we choose.

In Section 5, we would describe how we use the features extracted in Section 4 in our classification model to identify experts. The module will eventually generate the experts list and give the top N experts.

4 USER FEATURES EXTRACTING ENGINE

To learn the classification model, we use a large set of features that can reflect the impact of users in the system and their expertise. According to the nature they aim to capture, the features can fall into three main categories: profile features, tweeting behavior features and linguistic content features.

The rest of this section will further describe in depth these main categories of user features.

4.1 Profile Features

To start we present the list of valuable profile features in Table 1.

Having registered the service, users would have several profile features such as PF1-5 which are maintained by the microblogging service system automatically. Through the open API (application program interface) service of microblogging, we can get these profile features of users.

Experimental, a domain expert is more likely to have higher PF1, PF4 and PF6 because of his identity of information provider. PF2 is a service provided by microblogging system. If a user is authenticated, his identity is more likely to be true.

Table 1: Profile Features.

Name	Feature
PF1	Followers Count
PF2	Verified
PF3	Friends Count
PF4	Statuses Count
PF5	Favorites Count
PF6	Followers per Friend
PF7	Description Score
PF8	Tags Score

In self-descriptions and tags, users would like to use some words or sentences to describe themselves and choose tags provided by microblogging system to stand for them. Hence, from users' descriptions and tags we can partially know their interests and domains. In this paper, we convert user's description and tags to two features, PF7 and PF8. By counting words used in description of training users in the domain we care, we get top N words in all users' descriptions according to their word frequency, which is expressed as D_{domain}. PF7 is calculated using formula (1).

$$PF7=\frac{|D_i \cap D_{domain}|}{|D_{domain}|} \quad (1)$$

Where D_i is the words in ith user's descriptions.

Similarly, PF8 is calculated using the following formula (2).

$$PF8=\frac{|T_i \cap T_{domain}|}{|T_{domain}|} \quad (2)$$

Where T_{domain} is top N tags in all users' tags with high frequency and T_i is tags of the ith user.

4.2 Tweeting Behavior Features

Tweeting behavior is characterized by a set of statistics capturing the way the user interacts with the microblogging service. In paper (Pal, 2011), the authors listed several tweeting behavior features that reflect the impact of users in microblogging system. In our paper, we use some of features that listed in paper (Pal, 2011), and add more features that can be extracted from Sina Microblogging service. The valuable tweeting behavior features we used are listed in Table2.

In paper (Java, 2007), the authors suggested

that users who often post URLs in their tweets are most likely information providers. Giving an URL in microblogs is an efficient way to supply information in depth. In our work, we use feature TBF1 to record number of links user shared.

Hashtag keywords (TBF2) are words starting with the # symbol and are often used to denote topical keywords in microblogs. These keywords can clearly reflect the topic of microblog.

Table 2: Tweeting Behavior Features.

Name	Feature
TBF1	Number of links shared
TBF2	Number of keyword hashtags(#) used
TBF3	Number of conversation microblogs
TBF4	Number of retweeted microblogs
TBF5	Number of mentions (@) of other users by author
TBF6	Number of unique users mentioned by the author
TBF7	Number of users mentioned by the author
TBF8	Average number of messages per day
TBF9	Average comments per microblog
TBF10	Average reports per microblog

In paper (Boyd, 2010), retweeting or reposting someone's post were discussed. A user can mention other users using the "@user" tag. In paper (Honeycutt, 2009), authors discussed @user. And in papers (Naaman, 2010) and (Ritter, 2010), authors modeled the conversations. It's not difficult to know that features TBF2-7 can make a big difference in identifying domain experts. As an information provider, a domain expert tends to tweet several or even dozens of messages a day. TBF 8 can measure the impact of this behavior. Because the content of microblogs tweeted by domain experts is of high value, follows of experts would comment or even repost it. Statistics show that the higher the features TBF9 and TBF10 are, the higher user's authority is.

4.3 Linguistic Content Features

According the results in paper (Pennacchiotti, 2011), user's microblogs content makes most of the contribution in user features extraction. Making a good use of microblogs content would determine the performance of our system in a large extent.

Linguistic content information encapsulates the user's behavior of lexical usage and the main topics the user is interested in. Several studies, e.g. (Rao, 2010), have shown that bag-of-words models usually outperform more advanced linguistic ones.

Different from other primarily spoken genres previously studied in the user-property classification literature, microblogging-style informal written genres has its own characteristic.

The content of microblog can fall into three categories: original microblog, which is produced by the author; conversation microblog, which is replied by the author; reposted microblog, which is produced by someone else and forwarded by the author with some additional comments. In Sina Microblog service, the format of conversation microblog and reposted microblog is shown as follow:

Conversation microblog:

回复(reply)@user: content of reply//@user: source content.

Reposted microblog:

Additional comments //@user: source content.

4.3.1 Microblog Latent Dirichlet Allocation

Reply and repost characterize the relation between microblogs. In general, content of reply in conversation microblog and additional comments in reposted microblogs shares related topics with source content of microblog. In this paper, we take into account the above two relationships, extend the original Lda (Blei, 2003), and propose our Microblog Lda.

Microblog Lda adopts the basic idea of topic model, namely each microblogging exhibits multiple topics which are represented by probability distributions over words, denoted as $P(z|w)$ respectively. The Bayesian network of Microblog Lda is shown as follow in Figure 2.

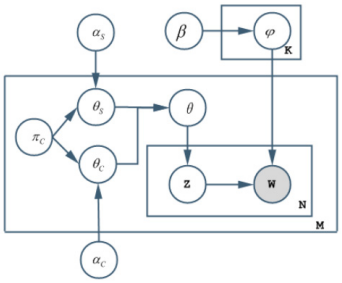

Figure 2: Bayesian network of Microblog Lda.

Apart from special instructions, symbols in Microblog Lda follow the definitions in (Blei, 2003).

Microblog Lda generates microblogging in the following process:

1. Random choose a topic distribution over words.
2. Judge whether a microblogging is retweeted or replied. If so, mark π_c as 1, random choose a contactor-topic distribution θ_c, which is sampled from a Dirichlet distribution with

hyperparameter α_c, then assign the value of θ_C to θ_S; if not, random choose a document-topic distribution θ_S, whose id sampled from a Dirichlet distribution with hyperparameter α_S.The probability distribution of θ is shown as follows:

$$\begin{aligned} & P(\theta;\alpha) \\ & = P(\theta;\alpha,c) \\ & = P(\theta_c;\alpha_c)^{\pi_c} P(\theta_S;\alpha_S)^{1-\pi_c} \end{aligned} \tag{3}$$

3 . Draw the specific word w_{dn} from the Multinomial distribution with parameter $\varphi_{z_{dn}}$.

For a microblogging, the joint probability is :

$$\begin{aligned} & P(W,Z,\theta,\varphi;\alpha,\beta) = \\ & \prod_{i=1}^{K} P(\varphi_i;\beta)\times \\ & \prod_{j=1}^{M} P(\theta_j;\alpha_C)^{\pi_C} P(\theta_j;\alpha_S)^{1-\pi_C}\times \\ & \prod_{t=1}^{N} P(W_{j,t}\mid\varphi_{Z_{j,t}})P(Z_{j,t}\mid\theta_j) \end{aligned} \tag{4}$$

Generative process is shown as follows:

Algorithm 1: Microblog Lda.

For each topic $k \in \{1,2,...T\}$ do
  Draw $\varphi_k \sim Dir(\beta)$
End for
For each microblog d do
  Judge whether d is conversation or reposted microblog
  If true
    Draw $\theta_S = \theta_c \sim Dir(\alpha_c)$
  Else
    Draw $\theta_S \sim Dir(\alpha_S)$
  For each word w_{dn} do
    Draw $z_{dn} \sim Multi(\theta_S)$
  End for
End for

4.3.2 Topic Features

Our Microblog Lda model is an adaptation of the original Lda proposed in paper (Blei, 2003), where documents are replaced by user's stream. Our hypothesis is that a user can be represented as a multinomial distribution over topics. While (Blei, 2003) represents documents by their corresponding bag of words, we represent users in microblogging service by the words of their tweets.

Results from (Pennacchiotti, 2011) shown that Lda system outperforms the tf-idf baseline with statistical significance. These prove our claim that topic models are good representations of user-level interests.

User's multinomial distribution over topics can clearly reflect his interest. Therefore domain experts' multinomial distribution over topics would be distinct. In our paper, we used results of Microblog LDA as linguistic content features of user and modeled each user by a topic-vector, where the weights are the probabilities to emit the topic.

5 EXPERTS IDENTIFYING AND RANKING

In Section 4, we generated user features, including profile features, tweeting behavior features and linguistic features, using our user features engine. In this section, we would use features generated above to identify domain experts and rank the result list.

In this paper, we cast the problem of identifying domain expert as a problem of 0-1 classification. As a classification algorithm, we use the Gradient Boosted Decision Trees – GBDT framework (Friedman, 2001). (Friedman, 2001) shows that by drastically easing the problem of over-fitting on training data (which is common in boosting algorithms). GBDT outperforms the state-of-the-art machine learning algorithms such as SVM with much smaller resulting models and faster decoding time (Friedman, 2006).

We use the features listed in section 4 to learn the classification model. After learning the GBDT model, we will use it to classify the large set of Sina Microblogging users and give the probability of a user judged as a domain expert.

In GBDT framework, results are shown in the format of probability of a user classified into classes. Having generated the probability of a user seen as a domain expert, we can ranking the probability and give the top N most liked experts of the domain we care.

6 EXPERIMENTAL EVALUATION

6.1 Data Preprocessing

Different from English, there are no spaces in words

interval of Chinese sentences. In order to process Chinese data, we should firstly segment sentences into words. In this paper, we use the ICTCLAS Chinese word segmentation system which has a high accuracy in Chinese word segmentation.

After word segmentation, we would discard all words that appear in a stop-word dictionary.

During July 1-15, we invited a pool of experts and seniors in the field of open source hardware. Through collecting their opinions extensively, we choose 200 users to train and validate our domain experts finding system, among them 92 are experts in open source hardware domain and 108 are not experts in open source hardware domain.

To train Microblog Lda model, we crawled all microblogs of these 200 users on Sina Microblog which is a microblogging service in China like twitter. There are 428 thousand microblogs totally.

6.2 Effectiveness Experiment

6.2.1 Performance of Microblog Lda

We conducted the comparative experiment between Microblog-lda and Lda using perplexity , measure of performance for statistical models which indicates the uncertainty in predicting a single word.

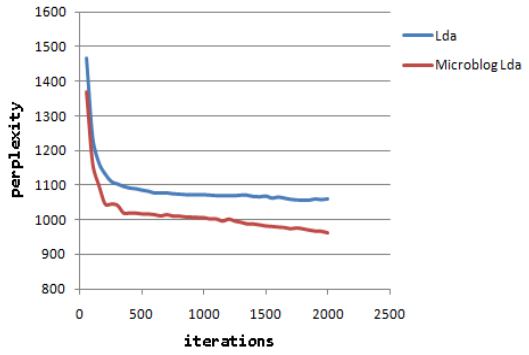

Figure 3: Perplexity of Lda and Microblog Lda.

Perplexity is used to measure the performance of LDA and Microblog-lda under the same hyperparameters setup, and the result is shown in Figure 3. From the result in Figure 3, we can see that Microblog Lda has plenty of performance gains compared with Lda.

6.2.2 Performance of Domain Experts Finding System

We compared our model with two baseline models as described below.

Baseline1: In this model, we used features listed in (Pal, 2011) only. Then, these features is used in our domain experts finding system and to give results on our dart base.

Baseline2: In this model, we used users' linguistic content features only.

Our: we used all kinds of features as mentioned above, including profile features, tweeting behavior features, linguistic content features.

After data processing and feature extraction, classification approaches are employed based on GBDT framework. The result is obtained with 10-fold cross validation in Figure 4. In this paper, we use ROC Area which refers to the area under ROC curve to measure the quality of our classifier and F-measure to measure the accuracy of our classifier comprehensively. We also give the results of Precision and Recall.

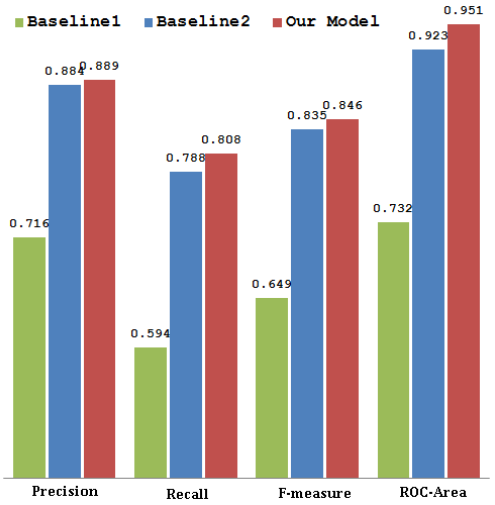

Figure 4: Classification results of training dataset.

In the results of our experiments, we give the performance comparisons of our domain experts finding system with baseline1 and baseline2. Compred with baseline1, both baseline2 and our model gain a great increase in preformance.In Figure 4, we can know that linguistic content features are highly valuable and contribute most of the classification confidence. From the index of ROC Area, we can know that our domain experts finding system is of high quality. Form the index of Precision and F-measure, we can know that our domain experts finding system has the ability to find experts in a particular domain with high accuracy.

6.2.3 Experts Indentifying and Ranking

In order to test performance of our system in real

production environment, we searched microblogs using keywords –"open source hardware" in search engine of Sina Microblog. The search engine would return the microblogs which content our search keywords. All microblogs were published recently. After parsing the returned microblogs and extracting the user id in the microblogs, we obtained initial users list which contents users who are likely to be expert in open source hardware domain. In our experiments, there were 3934 users in the users list.

Next, we used our domain experts finding system to analysis these users and identified 46 users who can be recognized as experts. In table 3, we give top 10 users in the domain of open source hardware.In order to compare preformance of our domain experts finding system with existing system, in table 4 we give top10 users returned by People Search System of Sina Microblog using keyword "open source hardware".

Table 3: Top 10 users returned by domain experts finding system.

Id	Screen Name
2171581500	SeeedStudio
2305930102	柴火创客空间(Chai huo chuang ke kong jian)
2524468112	Arduinos
3160959662	KnewOne
2055985387	王盛林 Justin(Wang sheng lin Justin)
3657027664	开放制造空间(Kai fang zhi zao kong jian)
1683765255	导通不能(Dao tong bu neng)
1906419177	新车间(Xin che jian)
1497878075	老黄(Lao huang)
1518434112	李大维(Li da wei)

In top 10 users returned by People Search System of Sina Microblog, the former six users's name have the search keyword "open source hardware". This means that People Search System of Sina System currently can not search out experts accurately, such as, a common user has screen_name containing the keywords, his is more likely to be returned.

In the users returned by our domain experts finding system, their have real people and organization farily. Specially, in order to evaluate the performance of our system, we made a questionnaire survey on 20 members of a club which focuses on open source hardware. From the feedback of these interviewees, we can get that 91.5% of users returned by our domain experts finding system can be recognized as experts in the particular domain.

Table 4: Top 10 users returned by People Search System of Sina Microblog.

Id	Screen Name
1750097377	开源硬件的星星之火(Kai yuan ying jian de xing xing zhi huo)
2334652932	赛灵思开源硬件社区(Sai ling si kai yuan ying jian she qu)
2497494380	开源硬件(Kai yuan ying jian)
3561629704	小米开源硬件俱乐部(Xiao mi kai yuan ying jian ju le bu)
2356441795	开源硬件平台
1906419177	新车间(Xin che jian)
2284986847	北京创客空间(Bei jing chuang ke kong jian)
2305930102	柴火创客空间(Chai huo chuang ke kong jian)
1715452481	54chen
1518434112	李大维(Li da wei)

7 CONCLUSIONS

In this paper, we proposed a domain expert finding system that could be used to produce a list of top N domain experts in Microblogs. We showed that: the thought of casting the problem of finding domain experts to a problem of 0-1 classification is feasible and of high accuracy in practice. From our experimental results, we can know that our domain experts finding system achieves good performance. In this paper, we use three kinds of user features, including profile features, tweeting behavior features and linguistic content features. Among them, linguistic content features show especially robust performance across tasks.

For further work, we wish to explore in detail running our system in parallel computing platform, like Hadoop. In addition, we wish to explore in detail how different features affect the final ranking and eliminate the influence of negative features.

ACKNOWLEDGEMENTS

The State Key Program of China- project on the Architecture, Key technology research and Demonstration of Web-based wireless ubiquitous business environment (2012ZX03005008).

REFERENCES

A. Java, P. Kolari, T. Finin, and T. Oates.2006. *Modeling the spread of influence on the blogosphere. I*n WWW (Special interest tracks and posters).

A. Java, X. Song, T. Finin and B. Tseng. 2007. Why we

twitter: understanding microblogging usage and communities. *Joint 9th WEBKDD and 1st SNA-KDD Workshop* (WebKDD/SNA-KDD).

A. Pal and J. A. Konstan.2010.Expert Identification in Community Question Answering: *Exploring Question Selection Bias*. InCIKM.

A. Ritter, C. Cherry and B. Dolan. 2010. Unsupervised Modeling of Twitter Conversations. *In the 2010 Conference of the North American Chapter of the Association for Computational Linguistics*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research.*

C. Honeycutt, S. C.Herring. 2009. Beyond microblogging: Conversations and collaboration via Twitter. *In Hawaii International Conference on System Sciences* (HICSS).

D. Boyd, S. Golder, G. Lotan. 2010. Retweet: Conversational Aspects of Retweeting on Twitter. *In Hawaii International Conference on System Sciences* (HICSS).

D. Kempe.2003. Maximizing the spread of influence through a social network. *In KDD.*

D. M. Romero, W. Galuba, S. Asur, B. A. Huberman. 2011. Influence and passivity in social media. *In Proceedings of ACM Conference on World Wide Web* (WWW).

E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne.2008. Finding high-quality content in social media. *In WSDM.*

E. Bakshy, J. M. Hofman, W. A. Mason, D. J. Watts. 2011. Everyone's an influencer: quantifying influence on Twitter. *In Proceedings of ACM Conference on Web Search and Data Mining* (WSDM).

Farahat, A., Nunberg, G., & Chen, F. 2002. Augeas: authoritativeness grading, estimation, and sorting. *In Proceedings of the eleventh international conference on Information and knowledge management*.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*.

Friedman, J. H. 2006. Recent advances in predictive (machine) learning. *Journal of classification*.

Institute of Computing Technology, Chinese Lexical Analysis System, http://ictclas.org/.

J. M. Kleinberg.1998.Authoritative sources in a hyperlinked environment. *In SIAM symposium on Discrete algorithms* (SODA).

J. Tang, J. Zhang, L.Yao, J. Li, L. Zhang, and Z. Su. 2008.Arnetminer: Extraction and mining of academic social networks. *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (SIGKDD'08).

J. Weng, E. -P. Lim, J. Jiang, Q. He. 2010. TwitterRank: Finding Topic-sensitive Influential Twitterers. *In Proceedings of ACM Conference on Web Search and Data Mining*(WSDM).

J. Zhang, M. S. Ackerman, and L. Adamic. 2007. Expertise networks *in online communities: structure and algorithms.* In WWW.

L. Page, S. Brin, R. Motwani, and T. Winograd.1998. The pagerank citation ranking: Bringing order to the web. *Technical report, Stanford Digital Library Technologies Project*.

M. Bouguessa, B. Dumoulin, and S. Wang.2008. Identifying authoritative actors in question-answering forums: the case of yahoo! answers. *In KDD.*

M. Cha, H. Haddadi, F. Benevenuto, K. P. Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. *In Proceedings of AAAI Conference on Weblogs and Social Media* (ICWSM).

M. Naaman, J. Boase and C. H. Lai. 2010. Is it Really About Me? Message Content in Social Awareness Streams. *In Computer Supported Cooperative Work*.

Pal, A., & Counts, S. 2011. Identifying topical authorities in microblogs. *In Proceedings of the fourth ACM international conference on Web search and data mining* .

Pennacchiotti, M., & Gurumurthy, S. 2011. Investigating topic models for social media user recommendation. *In Proceedings of the 20th international conference companion on World wide web.*

Pennacchiotti, M., & Popescu, A. M. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* .

P. Jurczyk and E. Agichtein.2007. Discovering authorities in question answer communities by using link analysis. *In CIKM.*

Ramage, D., Dumais, S. T., & Liebling, D. J. 2010. Characterizing Microblogs with Topic Models. *In ICWSM.*

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. 2010. Classifying latent user attributes in twitter. *In Proceedings of the 2nd international workshop on Search and mining user-generated contents.*