Duan JW, Chen YH, Liu T *et al.* Mining intention-related products on online Q&A community. JOURNAL OF COM-PUTER SCIENCE AND TECHNOLOGY 30(5): 1054–1062 Sept. 2015. DOI 10.1007/s11390-015-1581-7

Mining Intention-Related Products on Online Q&A Community

Jun-Wen Duan (段俊文), Yi-Heng Chen (陈毅恒), Member, CCF Ting Liu* (刘 挺), Senior Member, CCF, ACM, and Xiao Ding (丁 效), Member, CCF

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

E-mail: {jwduan, yhchen, tliu, xding}@ir.hit.edu.cn

Received November 17, 2014; revised July 9, 2015.

Abstract User generated content on social media has attracted much attention from service/product providers, as it contains plenty of potential commercial opportunities. However, previous work mainly focuses on user consumption intention (CI) identification, and little effort has been spent to mine intention-related products. In this paper, focusing on the Baby & Child Care domain, we propose a novel approach to mine intention-related products on online question and answer (Q&A) community. Making use of the question-answering pairs as data source, we first automatically extract candidate products based on dependency parser. And then by means of the collocation extraction model, we identify the real intention-related products from the candidate set. The experimental results on our carefully constructed evaluation dataset show that our approach achieves better performance than two natural baseline methods.

Keywords consumption intention, product extraction, Q&A community

1 Introduction

People are used to conveying their needs and desires on social media platform, which appeals to service/product providers. Previous work^[1-4] mainly focuses on the consumption intention (CI) identification from user generated content, and little effort has been spent to mine intention-related products, possibly for the following reasons.

• Constructing an intention-related product database requires a lot of domain-specific knowledge (e.g., medical care).

• It would be both time and labor consuming, and nearly impossible to cover all related products.

However, mining appropriate products to satisfy users' intention is an important and challenging problem for product providers^[3].

The development of online Q&A (question and answer) community offers new opportunities for solving this problem. Famous communities, such as Yahoo! Answers and Baidu Zhidao, have accumulated millions of user-generated question-answering pairs^[5], which can serve as a crowdsourcing knowledge database.

Example 1.

Q: My baby is calcium deficient. What can I do?

A: Try some calcium supplement.

Example 1 shows a real post on Q&A community. In this paper, our goal is to identify intention-related product "calcium supplement" to satisfy the user's intention "calcium deficient". Then when similar user intentions are detected, we can recommend "calcium supplement" to him/her immediately.

The problem is new and important, and there are two main challenging problems.

• Social media text is extremely noisy^[6], so how to extract products from the text accurately?

• How to rank candidate products and recommend the most intention-related products to users?

To address above challenges, in this paper, we propose a framework to automatically mine intentionrelated products. First, we assume that intention key-

Regular Paper

Special Section on Social Media Processing

The research is supported by the National Basic Research 973 Program of China under Grant No. 2014CB340503 and the National Natural Science Foundation of China under Grant Nos. 61133012, 61202277, and 61472107.

A preliminary version of the paper was published in the Proceedings of SMP 2014.

^{*}Corresponding Author

 $[\]textcircled{O}2015$ Springer Science + Business Media, LLC & Science Press, China

words are available because they can be extracted by previous work^[4]. Second, we collect question-answering pairs that contain intention keywords from online Q&A community. Third, we extract candidate intentionrelated products based on some heuristic rules. Fourth, we rank these candidate intention-related products based on a novel collocation extraction model. We carefully construct an evaluation dataset and present our observations. The experimental results show that our approach achieves better performance than two baseline methods, i.e., Co-occurrence and BM25.

The major contributions of our work are as follows.

• We make an attempt to mine knowledge from online Q&A community for commercial purpose, which will benefit service/product providers.

• We first propose the task of intention-related products identification. Our proposed approach achieves better performance than baseline methods.

• We carefully construct an intention-related product dataset in Baby & Child Care domain based on online Q&A community question-answering pairs, which can put forward related research.

This paper is organized as follows. We define the problem in Section 2. Our intention-related product mining approach is presented in Section 3. Data and our observations are in Section 4. Experimental setup and details are presented in Section 5. There, we also analyze the result. We then review the related approaches in Section 6. And finally in Section 7, we conclude our work and outlook possible future work.

2 Problem Statement

Intention Keyword. According to Hollerit *et al.*^[4], consumption intention(CI) is the intention explicitly or implicitly expressed by consumers to buy something. Hence, an intention keyword is a single word or a phrase that can indicate users' CI most. A user post with CI at least contains one intention keyword.

Example 2. Recommend me a feeding bottle?

Example 2 shows a real post with intention keyword "feeding bottle", because the user may move forward to buy a feeding bottle.

Intention-Related Products. Intention-related products are the products that can satisfy the user's specific intentions. For example, to intention keyword "calcium deficient", "calcium supplement" is an intentionrelated product. However, although "glove" is a product, it cannot satisfy the intention "calcium deficient". Hence, it is intention-unrelated. It is noteworthy that the product here can be either a specific product or a category of products.

In this paper, we only focus on intention-related products mining in a specific domain (Baby & Child Care domain) for the following reasons.

• The intention in this domain is common and urgent, so we can collect enough data.

• It does not require too much domain-specific knowledge, so it would be easy to evaluate.

3 Approach

As demonstrated in Fig.1, our approach consists of three steps. First, we collect user CI. Second, we extract candidate product names based on dependency parser. Third, we identify intention-related products. The details of each component are introduced as follows.



Fig.1. Process of our proposed approach.

3.1 Candidate Product Extraction

Given an intention keyword k, we collect the question-answering pairs with intention keyword k in the questions.

We adopt the pattern-based method introduced by Hollerit *et al.*^[4] to extract all possible products presented in the answers. Patterns are constructed based on our observations on how people recommend related products in the answers. Table 1 shows some patterns we use in our method. Via pattern-based method, we are able to take answers that may include candidate products.

 Table 1. Illustrative Patterns and Their Samples in

 Candidate Product Extraction

Pattern	Sample
试试(try)	为什么不试试贝新的奶瓶?
	Why not try milk bottle by Pigeon?
买(buy)	我买的是妈咪宝贝的纸尿裤.
	I bought diapers by Mamy Poko.
推荐(suggest)	我推荐来自德国的SI NA积木.
	I suggest puzzle blocks from German SINA.

We further apply semantic analysis to the answers based on LTP^[7] (Language Technology Platform). The

1056

dependency relations we make use of are VOB (verbobject), COO (coordinate), and SBV (subject-verb). By analyzing the features of each relation, we can easily obtain the products inside them. Table 2 demonstrates a sample for each dependency relation. Taking VOB (verb-object) relation for example, there is a VOB arc between the pattern "use" and the noun "pacifier". Hence, we extract "pacifier" as a candidate product.

Table 2. Dependency Relation and Sample



3.2 Intention-Related Product Identification

We have obtained a dataset containing the intention keywords and their candidate products. We view the intention-related product identification task as a collocation evaluation problem. An intention keyword and a product forming a collocation means that the intention keyword has once appeared in the question and the product has been extracted from its corresponding answer. Note that we may extract the same product from an answer for multi times; however, it would be only taken into account for once. An intention keyword may collocate with many products, and a product may collocate with many intention keywords as well. A product with higher collocation probability to the intention keyword is more likely to be the intention-related product. Thus we can make use of the method introduced by Liu et al.^[8] to deal with intention-related product identification. Based on the extraction result, we

J. Comput. Sci. & Technol., Sept. 2015, Vol.30, No.5

can easily figure out the frequency that intention keyword k_i collocates with candidate product p_j , denoted as $freq(k_i, p_j)$. Similarly, we estimate the probability that k_i collocates with p_j using (1). The probability that p_j collocates with k_i is calculated using (2). The average collocation probability of product p_j to intention keyword k_i is calculated using (3).

$$p(p_j|k_i) = \frac{freq(k_i, p_j)}{freq(k_i)},\tag{1}$$

$$p(k_i|p_j) = \frac{freq(k_i, p_j)}{freq(p_j)},$$
(2)

$$\bar{p}(k_i, p_j) = \frac{p(k_i|p_j) + p(p_j|k_i)}{2},$$
(3)

$$\bar{p}(k_i, p_j) = \frac{p(k_i|p_j) + p(p_j|k_i)}{2} \times e^{\frac{-b}{\log(freq(k_i, p_j))+1}}.$$
 (4)

However, in the experiment, we notice that collocations with low frequency may achieve a high collocation probability under circumstance that product p_j is infrequent. In order to penalize the collocations with low frequency, we add a penalization factor. The final intention-relatedness score is calculated by (4), where b is a constant parameter. According to the central-limit theorem^[9], sample set size larger than 30 is sufficiently large; thus, we set b = 5.9.

4 Data and Observations

As the task is first proposed by us, there is no dataset available for evaluation. In this section, we first introduce how we construct our experiment and evaluation dataset. We then present our observations over the dataset.

4.1 Data Collection

We carefully pick three most famous Baby & Child Care websites in China, namely Taobao Wenda⁽¹⁾, BabyTree⁽²⁾ and Sina Baby & Child Care Q&A⁽³⁾. We crawl the question-answering pairs from above three websites, and obtain more than 700 thousand pairs, which almost cover all aspects of Baby & Child Care.

Intention Keyword Set. We first randomly choose 4 000 questions from the question-answering dataset. We manually annotate the intention keywords and obtain 1 380 questions with intention keywords. There remain 245 after removing the duplicated ones. We then

⁽¹⁾http://baobao.wenda.taobao.com, July 2015.

⁽²⁾http://www.babytree.com, July 2015.

⁽³⁾http://ask.baby.sina.com.cn, July 2015.

randomly pick 30 intention keywords and they make up the intention keyword set. Following lists eight intention keywords in our set: 磨牙 (molar), 冲奶 (mix milk powder), 消毒 (disinfect), 吃手 (eat hands), 便秘 (constipation), 缺钙 (calcium deficient), 学步 (learn to walk), and 枕禿 (pillow baldness).

Intention-Related Product Standard Set. We ask two annotators to annotate each candidate product as related and unrelated to a given intention keyword. The agreement between annotators is measured using Cohen's Kappa Coefficient^[10].

We only keep the candidate products that both annotators judged as intention-related. Due to space limit, we do not present the annotation guidelines here. We obtain the highest Kappa = 0.86 and the lowest Kappa = 0.78, which is substantial. As a result, we construct an intention-related product gold set. Table 3 lists part of intention-related products to intention keyword "disinfection".

 Table 3. Part of Intention-Related Products to Intention

 Keyword "Disinfection"

Intention Keyword	Related Product
Disinfection (消毒剂)	Pressure cooker (高压锅)
	Alcohol (酒精)
	Disinfectant (消毒剂)
	Disinfection cabinet (消毒柜)
	Disinfectant soap (消毒皂)

4.2 Observations

Table 4 shows the statistics of our constructed dataset. From the table, we find that intention-related products only account for less than 10 percent in all candidate products extracted.

 Table 4. Statistics of Constructed Dataset

Intention keywords	30.0
Average candidate products per intention keyword	345.7
Average intention-related products per intention keyword	33.3

The remaining 90 percent are made up of extraction errors and intention-unrelated products. Hence, we further apply the intention-related product identification algorithm.

We further characterize the distribution of intention keywords and intention-related products. Fig.2 presents the distribution of intention keywords on our dataset.



Fig.2. Distribution of intention keywords on our dataset.

We find that a majority of intention keywords have a high frequency (occurred more than 500 times in questions); this is the foundation of our candidate product extraction. Fig.3 shows the distribution of intentionrelated products. As shown in Fig.3, we notice most intention keywords have more than 20 related products, which ensures the diversity of products could be recommended.



Fig.3. Distribution of intention-related products on our dataset. 56% intention keywords have more than 20 related products.

Does more frequent occurrence of an intention keyword in questions lead to more related products? Fig.4 shows the relation between the occurrence frequency of an intention keyword and the amount of its related products.

The number of intention-related products varies among intention keywords. But the general trend is that the number of intention-related products correlates with the number of candidate products. Some intention keywords have more intention-related products. That could explain why we observe several peaks in Fig.4. In summary, with more candidate products at hand, we can cover more related products.



Fig.4. Relation between the occurrence frequency of an intention keyword and the amount of its related products.

5 Experiments and Analysis

5.1 Baseline Methods

To evaluate the effectiveness of our proposed approach, we compare it with the following two baseline methods.

• Co-Occurrence. To evaluate the intentionrelatedness, the first method we may come up with is the co-occurrence based one. The more often a candidate product co-occurs with the intention keyword, the more likely it is an intention-related product. Here cooccurrence means the intention keyword appears in the question, while the candidate product is in its answer. The co-occurrence score is calculated using (5).

$$SCORE(k_i, p_j) = freq(k_i, p_j).$$
 (5)

• BM25. BM25^[11] is a classical approach in information retrieval. The model leverages term frequency to measure the relevance between a given query and a document. We slightly change the model to fit our problem. We use the changed model to calculate the relevance between an intention keyword and a candidate product. The BM25 is calculated using (6). IDF is short for inverse document frequency, which is calculated using (7). In (7), N is the number of Q&A pairs and N_{p_j} is the occurrence frequency of product p_j . $f(k_i, p_j)$ is calculated using (8), where $freq(k_i, p_j)$ is the co-occurrence frequency of intention keyword k_i and candidate product p_j . $b_1 \in [1.2, 2.0]$ is a free parameter. We have tuned the parameter and set it to 1.25.

$$SCORE(k_i, p_j) = IDF(p_i) \times \frac{f(k_i, p_j) \times (b_1 + 1)}{b_1}, \quad (6)$$

$$IDF(p_j) = \log \frac{N - N_{p_j}}{N_{p_j}},\tag{7}$$

$$f(k_i, p_j) = \frac{freq(k_i, p_j)}{freq(k_i)}.$$
(8)

5.2 Evaluation Metrics

We adopt two evaluation metrics, namely R-precision ((9)) and WARP (weighted average R-precision) ((10)), to measure the performance.

R-precision measures the precision in top R ranked candidate products, or in another word, the number of candidate products that are related to the intention keyword k_i in top R. Because our data is highly skewed, some of the intention keywords are much more frequent than the others. As a result, these intention keywords will have more candidate products. Therefore, we use WARP, which takes frequency information into account, to calculate average R-precision.

$$= \frac{\text{number of related products to } k_i \text{ in top } R}{R}, \quad (9)$$
$$= \sum \frac{freq(k_i) \times R\text{-}precision(k_i)}{\sum freq(k_i)}. \quad (10)$$

5.3 Comparison of Results

Fig.5 shows the WARP of our method and the baseline methods. We can find that our method achieves more than 20% improvement compared with the better baseline method. With a maximum WARP of 63.3%, our method could meet the requirement of application.

Table 5 demonstrates top R precision of MWA and baseline methods on two randomly picked intention keywords. We can find that MWA generally outperforms the two baselines, and the superiority becomes even more obvious as R grows. We also notice that P@5 of MWA on "diarrhea" is zero, even worse than BM25. We review the dataset and find that the low frequency of candidate products of "diarrhea" may contribute to the worse performance of MWA at top 5. Jun-Wen Duan et al.: Mining Intention-Related Products on Online Q&A Community

Keyword	Method	P@5	P@10	P@15	P@20	P@25	P@30
缺钙 (calcium deficiency)	MWA	0.80	0.80	0.73	0.70	0.68	0.63
	BM25	0.60	0.70	0.53	0.50	0.40	0.37
	Co-occurrence	0.60	0.60	0.47	0.40	0.40	0.33
拉稀 (diarrhea)	MWA	0.00	0.30	0.27	0.25	0.24	0.30
	BM25	0.20	0.20	0.20	0.15	0.16	0.20
	Co-occurrence	0.00	0.10	0.13	0.20	0.20	0.20

Table 5. Performance of MWA and Baseline Methods on Intention Keywords "Calcium Deficient" and "Diarrhea"



Fig.5. Overall performance of MWA and baseline methods on our dataset.

We notice that the extracted candidate products contain a lot of noise. Simply considering co-occurrence information could not get rid of the noise. Thus instead of focusing on the frequency, we take both the specificity a product to an intention keyword and the specificity an intention keyword to a product into consideration.

We also notice that the BM25-based method is not much better than the occurrence-based one. That is because candidate products in top R have similar frequency so that it is difficult to tell intention-related ones from intention-unrelated ones.

5.4 Case Study

We study how our method and baseline methods perform and present a case study here. We select "calcium deficient" as our intention keyword. Table 6 lists the top 5 products retrieved by each method, and intention-related ones are in bold. In this case, our method obtains more intention-related products with less noise than the other two.

5.5 Error Analysis

To have a clear understanding of the shortcomings of our proposed method, we manually inspect the identification errors in our results. We group the errors into two categories.

• Extraction Error. This happens in the candidate products extraction process. It is due to the limitation of the pattern-based method and the poor performance of the dependency parser. Text matching the patterns will be further fed to dependency parser. However, the precision of dependency parser falls sharply on noisy social media text. As a result, numbers, URLs, and incomplete product names are often extracted as candidate products. For example, as listed in Table 6, "molecule (分子)" and "effect (效果)" are extracted as candidate products.

• *Identification Error.* It means that intentionunrelated products are identified as intention-related ones. Although the product has been mentioned in the

Table 6. Case Study of Our Method and Baseline Methods on Intention Keyword "Calcium Deficient"

MWA	BM25	Co-Occurence
White calcium (白钙)	Calcium (钙)	Calcium (钙)
Molecule (分子)	Molecule (分子)	Effect (效果)
Calcium (钙)	Effect (效果)	Molecule (分子)
Calcium (钙剂)	Calcium tablet (钙片)	Calcium tablet (钙片)
Bio-calcium (生物钙)	Cod-liver oil (鱼肝油)	Cod-liver oil (鱼肝油)

Note: the table only lists top 5 products identified by each method.

answer, it has nothing to do with the intention. Sometimes, the identification result has something to do with the intention; however, it is not sold in the market. For example, "human milk" is frequently mentioned to cure "calcium deficient", and we are not able to buy it though.

To solve above problems, we could make use of the online product search engines provided by E-Commerce companies. It will be able to filter a lot of noise. However, we leave it to future work.

5.6 Parameter Sensitivity

The only parameter in our model is b, which aims to penalize the collocations with low co-occurrence frequency. When b is set to higher values, frequent candidate products will rank top. On the contrary, candidate products with less co-occurrence frequency will benefit from a lower b. Fig.6 illustrates the overall performance of MWA on our dataset when we set b to different values. From the figure, we can see that b > 5 has significant increase compared with b < 5. And the performance reaches the peak when b = 15. And then continuing increasing b achieves no more promotion in performance. The performance gradually reaches steady after the top 20 candidates. Taking all into consideration, we set b = 5.9.



Fig.6. Overall performance of our model when setting constant b to different values.

6 Related Work

Our work is related to the following.

Text Mining on Social Media. Social media is a real-time data source, on which a great deal of text mining work has been done. Zhao *et al.*^[12] proposed to summarize Twitter content by extracting key

phrases. Through topic discovery, topic-related keywords ranking, and candidate keyphrases generation and re-ranking, they obtained a set of keyphrases that could summarize a topic. Ritter *et al.*^[13] started an experimental study to extract name entity from noisy tweets, in which they proposed a distantly supervised approach that makes use of LabeledLDA and constraints from open-domain database. Pak and Paroubek^[14] focused on sentiment analysis and opinion mining based on Twitter corpus. They extracted features from tweets and built a classifier based on multinomial naive Bayes. Sakaki et al.^[15] exploited the realtime nature of Twitter. In their work, they detected tweets that have mentioned earthquakes in real time. Location information was then extracted and location estimation techniques were applied. At last, they built an application that is able to report an earthquake in a high precision.

Online Consumption Intention Identification. With the increasing popularity of online communities, CI identification has long attracted attentions from researchers. Dai *et al.*^[2] were among the first to give</sup> a formal definition of online commercial intention (OCI). They proposed a supervised method to predict whether submitting a search query or visiting a webpage will lead to commercial activity. Later, Ashkan and Clarke^[1] studied the relationship between query terms and ad click behavior. Making use of query log and ad click data, Dai *et al.*^[2] applied Bayes Theorem to quantify how much a term in a query contributes to underlying commercial intention. Hollerit *et al.*^[4] first started the task of CI detections on social media, and classified CI into explicit and implicit ones. They learned a classification model using word and part-ofspeech *n*-grams.

The most related work to ours is that of Wang et $al.^{[16]}$, in which they tried to mine trend-driven CI. By exploiting the trend keywords provided by the microblog platform, they further extracted products that co-occur with the trend keywords in users' tweets and make use of the product search engine to retrieve more products. Wang *et al.* proposed a joint model to measure relevance and associativity of candidate products. However, we have a different focus. Wang *et al.* focused on trend-related products mining on microblog platform while we focus on intention-related products mining on online Q&A community. What is more, trend-driven CI has a natural evaluation metric, which is product sales.

7 Conclusions

We made an attempt to mine intention-related products on online Q&A community. Focusing on the Baby & Child Care domain, given a set of intention keywords, we automatically extracted the candidate products from Q&A pairs. Our intention-related product identification method achieved better performance than two baseline methods. The method described in the paper could be further integrated into recommendation systems. Our future work includes the following:

• to integrate with intention identification module, so that both intention identification and intentionrelated products recommendation could be done automatically;

• to add filter process after candidate product extraction, since many of the "products" extracted are not real products, and by removing these "products" from the candidates, the precision of intention-related product identification would be greatly improved;

• to make use of the purchase data of intentionrelated products so that we can recommend more related products, and even those that did not occur in the candidate product set.

Acknowledgements We thank the anonymous reviewers for their constructive comments.

References

- Ashkan A, Clarke C L. Term-based commercial intent analysis. In Proc. the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, July 2009, pp.800-801.
- [2] Dai H K, Zhao L, Nie Z, Wen J R, Wang L, Li Y. Detecting online commercial intention (OCI). In Proc. the 15th International Conference on World Wide Web, May 2006, pp.829-837.
- [3] Ding X, Liu T, Duan J, Nie J Y. Mining user consumption intention from social media using domain adaptive convolutional neural network. In Proc. the 29th AAAI Conference on Artificial Intelligence, January 2015, pp.2395-2389.
- [4] Hollerit B, Kröll M, Strohmaier M. Towards linking buyers and sellers: Detecting commercial intent on Twitter. In Proc. the 22nd International Conference on World Wide Web Companion, May 2013, pp.629-632.
- [5] Shah C, Pomerantz J. Evaluating and predicting answer quality in community QA. In Proc. the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, July 2010, pp.411-418.
- [6] Kaufmann M. Syntactic normalization of Twitter messages. Studies, 2010, 2: 1-7.
- [7] Che W, Li Z, Liu T. LTP: A Chinese language technology platform. In Proc. the 23rd International Conference on Computational Linguistics: Demonstrations, August 2010, pp.13-16.

- [8] Liu Z, Wang H, Wu H, Li S. Collocation extraction using monolingual word alignment method. In Proc. the 2009 Conference on Empirical Methods in Natural Language Processing, August 2009, pp.487-495.
- [9] Grinstead C M, Snell J L. Introduction to Probability (2nd edition). American Mathematical Society, 1997.
- [10] Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, 70(4): 213-220.
- [11] Robertson S E, Walker S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Proc. the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 1994, pp.232-241.
- [12] Zhao W X, Jiang J, He J, Song Y, Achananuparp P, Lim E P, Li X. Topical keyphrase extraction from Twitter. In Proc. the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Volume 1, June 2011, pp.379-388.
- [13] Ritter A, Clark S, Mausam, Etzioni O. Named entity recognition in tweets: An experimental study. In Proc. the Conference on Empirical Methods in Natural Language Processing, July 2011, pp.1524-1534.
- [14] Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. In *Proc. LREC*, May 2010, pp.1320-1326.
- [15] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proc. the 19th International Conference on World Wide Web*, April 2010, pp.851-860.
- [16] Wang J, Zhao W X, Wei H, Yan H, Li X. Mining new business opportunities: Identifying trend related products by leveraging commercial intents from microblogs. In Proc. the Conference on Empirical Methods in Natural Language Processing, October 2013, pp.1337-1347.



Jun-Wen Duan received his B.E. degree in computer science and technology from Harbin Institute of Technology, Harbin, in 2013. Currently, he is a Ph.D. candidate in Harbin Institute of Technology. His current research interests include natural language processing, social computing, and

text mining.



Yi-Heng Chen received his B.E. degree in computer application from Northeast Forestry University, Harbin, in 2002, M.E. and Ph.D. degrees in computer science and technology, from Harbin Institute of Technology, Harbin, in 2004 and 2010, respectively. Cur-

rently, he is a lecture with Harbin Institute of Technology, Harbin. His current research interests include information retrieval, social computing, and text clustering.

J. Comput. Sci. & Technol., Sept. 2015, Vol.30, No.5



Ting Liu received his B.E., M.E. and Ph.D. degrees in computer science and technology from Harbin Institute of Technology, Harbin, in 1993, 1995 and 1998, respectively. Currently, he is a professor with Harbin Institute of Technology. His current research interests include natural language

processing, social computing, text mining, information retrieval, machine translation, etc.



Xiao Ding received his B.E. and M.E. degrees in computer science and technology from Harbin Institute of Technology, Harbin, in 2009 and 2011, respectively. Currently, he is a Ph.D. candidate in Harbin Institute of Technology. His current research interests include natural language processing,

social computing, and text mining.