Mining Intention-Related Products on Online Q&A Community

Junwen Duan, Xiao Ding, and Ting Liu

Reseach Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
{jwduan,xding,tliu}@ir.hit.edu.cn

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, we propose a novel approach to mine intention-related products on online Question & Answer (Q&A) community. Making use of the question-answer 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. Our method is general enough for domain adaptation.

Keywords: Consumption Intention, Product Extraction, Q&A Community.

1 Introduction

People are used to conveying their needs and desires on social media platform. For example, "What toys are appropriate for children under three?" expresses the user's intention to buy toys. Previous works [1, 3, 5] mainly focus on the CI identification from user generated content, and little effort has been spent to mine intention-related products, possibly for the following reasons.

- Constructing such an intention-related product database requires a lot of domain-specific knowledge, especially domains like health care.
- It would be both time and labor consuming, and nearly impossible to construct a database containing all related products.

However, mining appropriate products to satisfy user's intention is important for product providers. If product providers could make immediate response to user intention, the sales volume would be greatly improved.

The development of online Q&A community provides new opportunities for solving this problem. Online Q&A community, as a product of the open and sharing Internet, is continuously gaining popularity. Famous online Q&A communities, such

H.-Y. Huang et al. (Eds.): SMP 2014, CCIS 489, pp. 13–24, 2014.

as Yahoo! Answers and Baidu Knows, have accumulated millions of user-generated question-answer pairs [10], which can serve as a natural knowledge database. Consider the following scenario: a user posted a question on online Q&A community, e.g. "*what toys are appropriate for child under three*", later someone else replied with a possible answer "*stuffed toys*". Thus we can make a connection between the intention keyword *toy* and candidate product "*stuffed toys*". After intention-related product identification process, we consider "*stuffed toys*" as intention-related product. Then whenever someone else presents similar intention, we could immediately recommend "*stuffed toys*" to him/her.

The problem is new and important, however, there are at least two main challenges.

- Social media text is notoriously noisy [6], how to extract products from the text accurately?
- Even with candidate products at hand, how can we identify the intention-related ones from the intention-unrelated ones without much human labor?

To address above challenges, in this paper, we propose a general framework to automatically mine intention-related products. Because CI identification is out of scope of this paper, we assume that the intention keyword set is available. We first make use of the online Q&A community question-answer pairs as data source and pick the pairs with intention keyword inside. We combine pattern-based method with dependency parser to automatically extract the candidate products from the answers. In order to identify the intention-related products, we further propose a novel approach based on 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 *Jaccard Coefficient.*

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 has open up a new way to make use of online Q&A user-generated content.
- We first propose the task of intention-related products identification, and innovatively adopt collocation extraction model. Our approach achieves a remarkable performance without preprocessing.
- We carefully construct intention-related product dataset based on online Q&A community question-answering pairs.

This paper is organized as follows. Data and our observations are in Section 3. We then introduce our intention-related product mining approach in Section 4. And experiment 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. Intention keyword is a single word or a phrase that can indicate user's CI most, and a user post with CI must at least contain one intention keyword. According to [5], CI could be further classified into explicit and implicit ones, thus a CI is the desire that explicitly or implicitly expressed by consumer to buy something.

Example 1: my baby is calcium deficient, what can I do?

Example 1 shows a possible post with intention keyword "*calcium deficient*", because it may give rise to the user's further action to buy calcium supplement products.

Intention-Related Products. Intention-related products are the products that consumers with such intention would like to buy. Here products that we extracted consist of the following two different levels.

- Full product name, which refers to a certain product e.g. JNJ baby soap (强生婴儿香皂);
- Product category, which refers to a category of products, e.g. milk powder (奶粉).

Our approach takes both levels into consideration. Because we notice that some consumers are quite sure what products could satisfy their intention, and we could recommend certain products to them. However, others may not have a clear product in mind, thus we could recommend a category of products.

In this paper, we only focus on intention-related products mining in the Baby & Child Care domain. If not specially stressed, all the running cases are in this domain. And we suggest that the domain picked should satisfy the following conditions:

- The intention in the domain must be general and urgent, thus they can arouse hot discussion;
- Do not require too much domain-specific knowledge, so it would be easier to evaluate.

3 Data and Observations

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

3.1 Data Collection

We carefully pick three most famous Baby & Child Care websites in Chinese, namely Taobao Wenda[1], BabyTree[2] and Sina Baby & Child Care Q&A[3].We start our web

[1] Taobao Wenda: `http://baobao.wenda.taobao.com`

[2] BabyTree: `http://www.babytree.com`

[3] Sina Baby & Child Care Q&A: `http://ask.baby.sina.com.cn/`

spider to crawl the question-answer pairs from the three websites, and obtain more than 700 thousand pairs, which nearly cover all aspects of Baby & Child Care.

Intention keyword set. We first randomly choose 4000 questions from the question-answer dataset, and manually annotate the intention keywords inside them. After that, we have obtain1380 questions with intention keywords. There remain 245 after removing the duplicated keywords. We then randomly pick 30 intention keywords and they make up the intention keyword set. Table 1 lists nine intention keywords in our set.

Table 1. Intention keyword samples

便秘	(constipation)	磨牙	(molar)	消毒	(disinfect)
吃手	(eat hands)	冲奶	(mix milk powder)	缺钙	(calcium deficient)
学步	(learn to walk)	枕秃	(pillow baldness)	拉稀	(diarrhea)

Intention-related product standard set. For each keyword in the intention keyword set, we look it up in the question-answer pairs. After that we obtain question-answer pairs with intention keyword in the question. Via method that will be mentioned in 4.1, we obtain the candidate products for each intention keyword. We hire two annotators to annotate the candidate products as related and unrelated to the intention keyword, and the agreement between annotators is measured using Cohen's Kappa Coefficient [2]. We only keep the candidate products that both annotators judged as intention-related. Due to space limit, we don't present the annotation guidelines here. We obtain highest Kappa = 0.86 and lowest Kappa = 0.78, which is substantial. As a result, we construct an intention-related product standard.

3.2 Observations

Table 1 shows the statistics of our constructed dataset, from which we can find that intention-related products only account for less than 10% in all candidate products extracted.

Table 2. Statistics of constructed dataset

Intention keywords	30
Average candidate products per intention keyword	345.7
Average intention-related products per intention keyword	33.3

We review the candidate products and discover that products judged as intention-unrelated could be further divided into two categories.

- *Not a real product*, this is due to limitation of pattern-based method and poor performance of dependency parser. Text matches the pattern will be extracted as a product and precision of dependency parser fall sharply in noisy social media text.
- *Intention-unrelated*, although the product has been mentioned in the answer, however, it has nothing to do with the intention.

This accounts for why we further carry out the intention-related product identification process.

We further make statistics of distribution of intention keyword and intention-related products. Figure 1 presents the distributions of intention keyword on our dataset. We can find that a majority of intention keyword has a high frequency (occurred more than 500 times in questions); this is the foundation of our candidate product extraction. Figure 2 shows the distribution of intention-related products. According to Table 1, on average per intention keyword relates to 33.3 products, from Figure 2 we can further discover that most intention keywords have more than 20 related products, which ensures the diversity of products could be recommended.

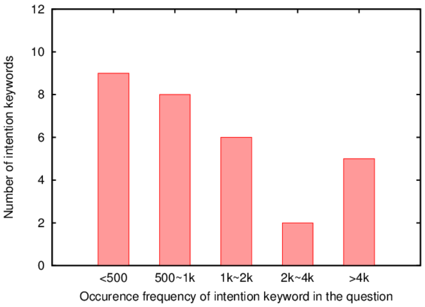


Fig. 1. Distribution of intention keywords on our dataset

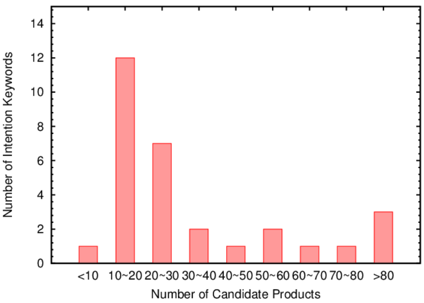


Fig. 2. Distribution of intention-related products on our dataset

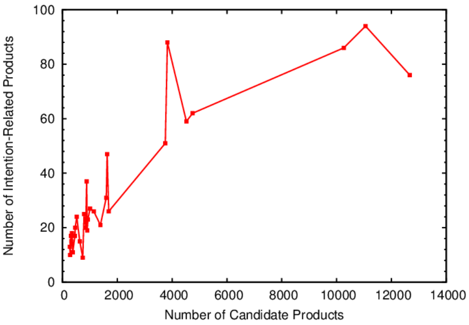


Fig. 3. Relation between occurrence frequency of an intention keyword and amount of its related products

Does more frequent occurrence of an intention keyword in questions leads to more related products? Our answer is “most times it does, but not always”, which accords with our common knowledge. Figure 3 shows the relation between occurrence frequency of an intention keyword and amount of its related products. Frequent occurrence means more candidate products extracted, thus more related products identified. However, some intention keyword has a quite limited related product set, which means increase in occurrence frequency does not lead to more related products. However, with more candidate products at hand, we can cover more related products.

4 Approach

Our approach consists of three steps. First, we extract user CI. In this paper, we suppose user CI keyword set is available. Second, we identify candidate product names based on dependency parser. Third, we extract intention-related products. The details of each component are introduced as follows.

4.1 Candidate Product Extraction

Given an intention keyword k , we have to search from the question-answer pairs for questions that contain the intention keyword k and their corresponding answers. Table 2 shows an illustrative sample for intention keyword *toy*.

Table 3. An illustrative sample for intention keyword *toy*

Intention Keyword: *toy*	
Question:	什么玩具适合三岁以下的儿童呢？
	(What toys are appropriate for children under three?)
Answer:	我推荐毛绒玩具。
	(I recommend stuffed toys.)

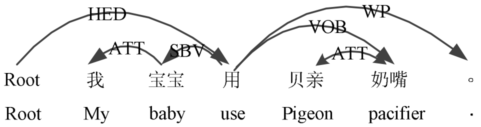
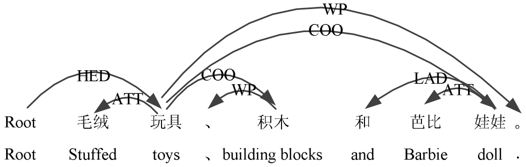
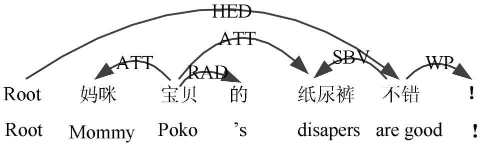
We adopt pattern-based method introduced in [5] to extract all possible products occurred in answers. Patterns are constructed based on our observations on how people recommend related products in the answers. Table 3 shows some patterns we use in our method. Our work is based on Chinese corpus, and we further apply semantic analysis to the answers apart from the part-of-speech information. We use the LTP[4] (Language Technology Platform) to carry out the language analysis. LTP [2] is a platform that has integrated Chinese word segmentation, part-of-speech tagger and semantic dependency parser. The dependency relations we make use of are VOB (verb-object), COO (coordinate) and SBV (subject-verb), by analyzing the features of each relation, we can easily obtain the products inside it. Table 4 demonstrates sample for each dependency relation.

Table 4. Pattern and Sample

Pattern	Sample
试试(try)	为什么不试试贝亲的奶瓶？ 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.
建议(suggest)	我推荐来自德国的SINA积木。 I suggest puzzle blocks from German SINA.

[4] LTP: `http://www.ltp-cloud.com`

Table 5. Dependency Relation and Sample

Dependency Relation	Sample
VOB(verb-object)	 HED WP VOB ATT SBV ATT Root 我 宝宝 用 贝亲 奶嘴 。 Root My baby use Pigeon pacifier .
COO(coordinate)	 WP COO HED COO WP ATT LAD ATT Root 毛绒 玩具 、 积木 和 芭比 娃娃 。 Root Stuffed toys 、 building blocks and Barbie doll .
SBV(subject-verb)	 HED ATT ATT RAD SBV WP Root 妈咪 宝贝 的 纸尿裤 不错 ! Root Mommy Poko 's disapers are good !

4.2 Intention-Related Product Identification

Via the product extraction step, we have obtained a dataset containing the intention keywords and their corresponding candidate products. We view the intention-related product identification problem as a collocation evaluation problem. An intention keyword and a product form a collocation means that the intention keyword has once appeared in the question and the product have 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 as well may collocate with many intention keywords. 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 in [7] to deal with intention-related product identification. Based on the extraction result, we can figure out the frequency that intention keyword k_i collocates with candidate product p_j, denoted as $freq(k_i, p_j)$, and then we estimate the probability that k_i collocates with p_j using Eq. (1), and the probability that p_j collocates with k_i using Eq. (2). The average collocation probability of product p_j to intention keyword k_i is calculated using Eq. (3).

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

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

$$\overline{p}(k_i, p_j) = \frac{p(k_i \mid p_j) + p(p_j \mid k_i)}{2} \tag{3}$$

However, in the experiment we noticed 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 added a penalization factor, thus the intention-relatedness score is calculated by Eq. (4) .Here b is a constant parameter, according to central-limit theorem[4], sample set size larger than 30 is sufficiently large, thus, we set $b = 5.9$.

$$\overline{p}(k_i, p_j) = \frac{p(k_i \mid p_j) + p(p_j \mid k_i)}{2} \times e^{\frac{-b}{\log(freq(k_i, p_j)) + 1}} \tag{4}$$

5 Experiments

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 intention-relatedness, the first method we may come up with is the co-occurrence-based one for the simple idea that the more a candidate product co-occurs with the intention keyword, the more likely it is the intention-related product. Here co-occurrence means the intention keyword appears in the question and candidate product in its corresponding answer.The co-occurrence score is calculated using Eq. (5)

$$SCORE_{cooccurence}(k_i, p_j) = freq(k_i, p_j) \tag{5}$$

- Jaccard Coefficient. Jaccard Coefficient is generally used to evaluate the similarity of two sets. However, we adopt it here to calculate the intention-relatedness. Based on the idea that the more similar the distribution of a candidate product to the distribution of an intention keyword, or the extreme case that the candidate product only co-occurs with the intention keyword, the more likely the candidate product is an intention-related product. The Jaccard Coefficient is calculated using Eq. (6).

$$SCORE_{Jaccard}(k_i, p_j) = \frac{freq(k_i, p_j)}{freq(k_i) + freq(p_j) - freq(k_i, p_j)} \tag{6}$$

5.2 Evaluation Metrics

We adopt two assessment metrics, namely *R-Precision* (Eq.7) and *WARP* (*Weighted Average R-Precision*) (Eq.8), to measure the performance. Before we get evolution started, we calculate and rank the scores of candidate products using our method and baseline method.

R-precision measures the precision in top *R* ranked candidate products, or in another word, how many candidate products are related to the intention keyword k_i in Top *R*. Because our data is highly skewed, according to previous observations, some of the intention keyword is much more frequent than others; as a result, they will have more candidate products and thus are more likely to have related products. Therefore, we use *WARP*, which takes frequency information into account, to calculate average *R-Precision.*

$$R\text{-Precision}(k_i) = \frac{\#\text{Related products to } k_i \text{ in top } R}{R} \tag{7}$$

$$WARP(k_i) = \sum_i \frac{freq(k_i) \cdot R\text{-}precision(k_i)}{\sum_i freq(k_i)} \tag{8}$$

5.3 Results and Analysis

Table 6 demonstrates TOP *R* precision of MWA and baseline methods on two randomly picked intention keyword. We can find that MWA generally outperforms the two baselines, and the superiority become even more obvious as *R* grows. What's more, unlike *Jaccard* and *Co-occurrence* based methods, they drop sharply as R increases, while precision of MWA keeps even. Figure 4 shows the *WARP* of our method and the baseline methods, from which we can find that our method is remarkable better than the baseline methods, more than 20% higher *WARP* than the best baseline method. Note that our result is achieved automatically without preprocessing the candidate product sets. With a maximum *WARP* of 63.3%, our method could meet the requirement of application.

Our method is better because we innovatively regard it as a collocation identification problem. We notice that the candidate products extracted contains a lot of noise, including unrelated products noisy text. 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. What's more, we add a penalization factor to deal with infrequent cases.

Table 6. Performance of MWA and baseline method on intention keyword *calcium deficient* and *diarrhea*

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**
	Jaccard	0.60	0.70	0.53	0.45	0.40	0.33
	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**
	Jaccard	0.20	0.10	0.13	0.15	0.16	0.20
	Co-occurrence	0.00	0.10	0.13	0.20	0.20	0.20

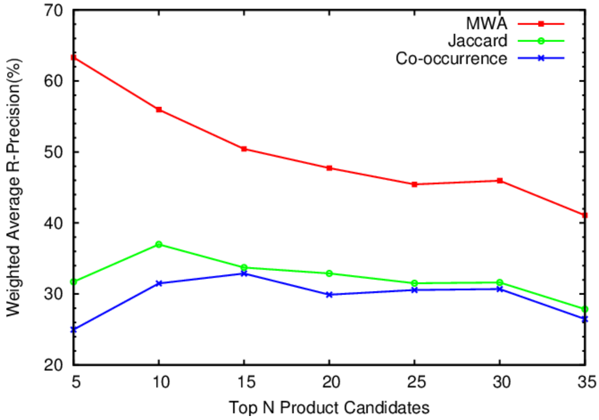


Fig. 4. Evaluation Result

We also notice that Jaccard-based method is not much better than occurrence-based one, because candidate products in TOP *R* have similar distribution so that it's difficult to tell them apart, or in another word, more similar distribution does not mean higher intention-relatedness.

6 Related Work

Our work is related to the following:

Text Mining on Social Media. Social media as a real-time data source, a great deal of text mining work have been done on it. [13] proposes to summarize twitter content by extracting key phrases. [9] starts an experimental study to extract name entity from noisy tweets. [8] focuses on sentiment analysis and opinion mining based on twitter corpus. [11] makes use of real-time nature of twitter, and use each user as a sensor to extract earthquake event and report earthquake .

Online Consumption Intention Identification. With the increasing popularity of online communities, CI identification has long attracted attentions from researchers. [3] first gives a formal definition of online commercial intention(OCI), and proposed a supervised method to predict whether submitting a search query or visiting a webpage will lead to commercial activity. Later work [1] studied the relationship between query terms and ad click behavior. Making use of query log and ad click data, they apply Bayes Thorem to quantify how much a term in a query contributes to underlying commercial intention. [5] first starts the task of CI detection on social media, and classify CI into explicit and implicit ones. They learned a classification model using word and part-of speech n-grams.

The most related work to ours is [12]; they try to mine trend-driven CI. However, we have different focus; they focus on trend-related products mining on microblog platform while we focus on intention-related products mining on online Q&A community. What's more, trend-driven CI has a natural evaluation metric which is product sales.

7 Conclusion and Future Work

We make our own attempt to mine intention-related products on online Q&A community. Given a set of intention keywords, we automatically extract the candidate products from Q&A pairs. Our intention-related product identification method borrowed from collocation identification achieved a quite satisfying performance even on noisy data. The method describe in the paper could be further integrated into recommendation system. Our future work includes the following:

- Integrate with intention identification module, so that both intention identification and intention-related products recommendation could be done automatically;
- Add filter process after candidate product extraction, since many of the "products" extracted are not real products, by removing these "products" from the candidates, the precision of intention-related product identification would be greatly improved;
- Make use of the purchase data of intention-related products so that we can recommend more related products, even those not occurred in the candidate product set.

Acknowledgments. We thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of the National Basic Research Program (973 Program) of China via Grant 2014CB340503, the National Natural Science Foundation of China (NSFC) via Grant 61133012 and 61202277.

References

1. Ashkan, A., Clarke, C.L.: Term-based commercial intent analysis. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 800–801. ACM (July 2009)
2. Che, W., Li, Z., Liu, T.: Ltp: A Chinese language technology platform. In: Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, pp. 13–16. Association for Computational Linguistics (August 2010)

3. Dai, H.K., Zhao, L., Nie, Z., Wen, J.R., Wang, L., Li, Y.: Detecting online commercial intention (OCI). In: Proceedings of the 15th International Conference on World Wide Web, pp. 829–837. ACM (May 2006)
4. Grinstead, C.M., Snell, J.L.: Introduction to probability. American Mathematical Soc. (1998)
5. Hollerit, B., Kröll, M., Strohmaier, M.: Towards linking buyers and sellers: Detecting commercial Intent on twitter. In: Proceedings of the 22nd International Conference on World Wide Web Companion, pp. 629–632. International World Wide Web Conferences Steering Committee (May 2013)
6. Kaufmann, M., Kalita, J.: Syntactic normalization of twitter messages. In: International Conference on Natural Language Processing, Kharagpur, India (July 2010)
7. Liu, Z., Wang, H., Wu, H., Li, S.: Collocation extraction using monolingual word alignment method. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 2, pp. 487–495. Association for Computational Linguistics (August 2009)
8. Pak, A., Paroubek, P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: LREC (May 2010)
9. Ritter, A., Clark, S., Etzioni, O.: Named entity recognition in tweets: An experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534. Association for Computational Linguistics (July 2011)
10. Shah, C., Pomerantz, J.: Evaluating and predicting answer quality in community QA. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 411–418. ACM (July 2010)
11. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: Real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860. ACM (April 2010)
12. 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: EMNLP, pp. 1337–1347 (2013)
13. Zhao, W.X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E.P., Li, X.: Topical keyphrase extraction from twitter. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 379–388. Association for Computational Linguistics (June 2011)