# Refined Feature Extraction for Chinese Question Classification in CQA

Lei Su(✉), Bin Yang, Xiangxiang Qi, and Yantuan Xian

School of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650093, China
s28341@hotmail.com, yangbin0724@126.com, qixiangfighting@163.com,
yantuan.xian@gmail.com

**Abstract.** Community-based Question Answering (CQA) services, such as Baidu Zhidao, have attracted increasing attention over recent years, where the users can voluntarily post the questions and obtain the answers by the other users from the community. Question classification module of a CQA system plays a very important role in understanding the user intents, which could effectively enhance the CQA systems to identify the similar questions and retrieve the candidate answers. However, the poor semantic information could be obtained from the questions because of the short sentences. This paper proposes a refined feature extraction method for question classification. The method aims to use Wikipedia to expand the semantic knowledge of sentences, and extract the features step by step to overcome the shortness of semantic knowledge. Experimental results on 714,582 Chinese questions crawled from Baidu Knows show that the proposed method could effectively improve the performance of question classification in CQA.

**Keywords:** Community-based Question Answering · Wikipedia · Question Classification · Semantic Knowledge

## 1 Introduction

During the last few year, Community-based Question Answering (CQA) websites, such as Baidu Zhidao (zhidao.baidu.com), Sina iAsk (iask.sina.com.cn) and SOSO Ask (wenwen.soso.com), have emerged and become a popular form of online service. In these communities, web users can voluntarily ask and answer questions. Unlike the traditional search engines which retrieve a large number of candidate pages for users, CQA is an interactive platform where the posted questions could get a feedback by other volunteers. CQA provides a similar list of resolved history questions to the post items, where some good quality answers could be obtained by a small number of experts among the large population of users. These communities assure the quality of questions and answers through the mechanisms of voting, badges and reputation [1].

Understanding the user intent behind the questions would help a CQA system to find similar questions, recommend questions and obtain potential answers [2]. The goal of question classification is to accurately label the questions into predefined target categories. Question classification is an essential part of question answering systems, because it can not only impose constrains on the possible answers but also narrow the scope of finding answers. For example, if the question "how much to repair my

V.C.M. Leung et al. (Eds.): TridentCom 2014, LNICST 137, pp. 318–326, 2014.
DOI: 10.1007/978-3-319-13326-3_30

iphone5?" can be correctly classified into the category of maintenance in Consumer Electronics, the search scope for the answer will be significantly focused on the price instead of each word in the candidate documents.

Various machine learning algorithms have been proposed for question classification, which extract syntactic and semantic features from large quantities of training corpus to build the learning model. D. Zhang and W.S. Lee [3] used a special kernel function called tree kernel to enable the SVM to take advantage of the syntactic structures of questions. A. Moschitti et al. [4] defined tree structures based on shallow semantic encoded in predicate argument structures for question classification. A prominent achievement in Chinese question classification is the modified Bayes model proposed by Y. Zhang [5], where the accuracy rate reaches 72.4% on the 65 Chinese question classes. Compared with normal texts, questions in CQA are usually short and cannot provide sufficient syntactic and semantic features. To tackle the problem of data sparseness, Hotho et al. [6] used the synonym and hypernym included in WordNet to expend the text characteristics. Cai et al. [7] proposed a two-stage approach for question classification in CQA. The large-scale categories are pruned to a small subset, and then the questions are enriched by leveraging Wikipedia semantic knowledge (hypernym, synonym and associative concepts).

Unlike normal texts and documents, questions in CQA are usually short. Therefore, the traditional learning model based on bag-of-word in vector space model extracts a lot of feature value with zero due to the data sparseness. There is another difficulty in question classification in CQA. The traditional methods can classify the questions into several limited categories, while the number of categories in CQA is very large. For example, category level in Baidu Knows is roughly divided into three layers. From the top layer to the bottom layer in the taxonomy, the larger number of categories may cause a significant decline in classification accuracy.

In order to solve the above problems, we propose a refined feature extraction approach for question classification in CQA. First, the Wikipedia semantic library is constructed where the theme relationship can be used to extend the semantic knowledge of questions. Then, the proper nouns table, features extracting from categories and the refined feature extracting method could be employed based on bag-of-word. Experimental results on the 714,582 Chinese questions crawled from Baidu Knows show that the proposed method could significantly improve the classification accuracy in CQA.

The rest of this paper is organized as follows. Section 2 introduces the method of constructing the Wikipedia knowledge library. Section 3 describes the refined feature extraction for question classification. Section 4 reports and analysis the experimental study on the Chinese question classification in CQA. Finally, section 5 summarizes this paper and introduces the future work.

# 2 Wikipedia Knowledge Library Construction

## 2.1 Wikipedia Semantic Knowledge

Wikipedia is a free, open-content online collaborative encyclopedia, which provides link designed to guide the user to related pages with additional information. It can be an effective knowledge base resource because of the rich semantic knowledge. In particular, research has been done to exploit Wikipedia for document categorization [8–10] and text cluster [11–13]·

Each article in Wikipedia describes a topic or a concept, and it has a short title, which is a well-formed phrase like a term in a conventional thesaurus. Each article belongs to at least one category, and hyperlinks between articles capture their semantic relations. Specifically, the represented semantic relations are: equivalence (synonymy), hierarchical (hyponymy), and associative [9]. Articles in Wikipedia form a heavily interlinked knowledge base, enriched with a category system emerging from collaborative tagging, which constitutes a thesaurus [14]. Thus, Wikipedia contains a rich body of lexical semantic information, which includes knowledge about named entities, domain specific terms or domain specific word. To use Wikipedia semantic knowledge, we preprocess the Wikipedia articles to construct the topic library and the category library. The topics in Wikipedia are organized as a theme tree, where the topic pages are equivalent to the top nodes and linked to the relational nodes. According to the degree of relationship, each category with the different level can be organized as leaf in the Wiki tree.

Following [5], the semantic relations, such as synonym, polysemy, hypernym and associative relation, can be extracted from the article pages. The synonym relations mainly come from the redirect hyperlinks whose means are usually similar. Wikipedia provides disambiguation for a polysemy concept. Wikipedia categories contain the hypernym relations by hierarchical relations, including relations between categories and links. The associative relation of each hyperlink between Wikipedia articles could be measured by three kinds of method: content-based, out-link category-based and distance-based [10].

(1) Content-based measurement ($S_{tfidf}$) is based on vector space model. The relatedness of two articles is evaluated by the extent to which they share terms using *tf-idf* scheme.

(2) Out-link category-based measurement ($S_{olc}$) could be defined as the out-link category similarity. If most of the out-linked categories of two articles focus on several same ones, the concepts described in these two articles are most likely strongly related.

(3) Distance-based method ($D_{cat}$) measures semantic distance as the number of nodes in the category taxonomy along the shortest path between two conceptual nodes. This measurement is normalized by taking into account the depth of the taxonomy.

The overall relatedness evaluation is defined as:

$$S_{overall}=\lambda_1 S_{tfidf}+\lambda_2 S_{olc}+(1-\lambda_1-\lambda_2)(1-D_{cat}) \tag{1}$$

where $\lambda_1$ and $\lambda_2$ are the weight parameters.

## 2.2 Semantic Knowledge Library from Wikipedia

Java-based Wikipedia Library (JWPL) is already freely available for research purpose and is used to construct semantic knowledge library [14]. The category and topic databases from Wikipedia are established respectively.

Categories can reflect semantic relations in Wikipedia, so question expansion could use category information. The category database includes four tables:

(1) Category table: the patent and child categories are stored from Wikipedia, where two fields *CategoryID* (ID of category) and *Name* (name of category) are included.

(2) Category_inlinks table: the relations between the categories and their parent categories are stored, where two fields CategoryID (ID of category) and Inlinks (ID of parent category) are included.

(3) Category_outlinks table: the relations between the categories and their child categories are stored, where two fields CategoryID (ID of category) and Inlinks (ID of child category) are included.

(4) Category_pages table: the relations between the categories and the topic articles which belong to the categories are stored, where two fields CategoryID (ID of category) and Inlinks (ID of topic article) are included. Therefore, the categories and the topics can be connected by this table.

The topic database includes six tables:

(1) Page table: the detailed topics are stored in this table, which is the most important table in the knowledge library. The four fields, *PageID* (ID of page), *Name* (name of page), Text (detailed topics) and *IsDisambiguation* (whether it is a disambiguation page or not), are included in this table.

(2) Page_category table: the relations between the categories and the topic pages are stored, where two fields PageID (ID of page) and CategoryID (ID of category that the page belongs to) are included.

(3) Page_inlinks table: the relations between the pages and the child pages are stored in this table, where two fields PageID (ID of page) and Inlinks (ID of page which links to the page) are included.

(4) Page_outlinks table: the relations between the pages and the parent pages are stored in this table, where two fields PageID (ID of page) and Outlinks (ID of page which the page links to) are included. The associate rules are extracted from this table.

(5) Page_redirects table: the relations between the pages and the disambiguation pages, where two fields PageID (ID of page) and Redirects (ID of page which the page redirects to) are included.

(6) Page_mapline table: the allover information about the redirection pages is stored in this table, where three fields, PageID (ID of page), Name (name of the topic) and Redirects (ID of page which the page redirects to) are included.

The topic pages provide the associative relations through a large number of links. The rich semantic knowledge, such as synonym, polysemy, hypernym and associative relation, could be used to construct the library from the categories in Wikipedia. Therefore, the semantic knowledge library could be applied into question expansion.

## 3 Refined Feature Extraction

Compared to text documents, questions contain fewer words in each sentence. Therefore, the sparse data is inevitable and maybe degrade the performance of classifier. To tackle this problem of data sparseness for Chinese question classification in CQA, we expand the questions by using Wikipedia semantic knowledge library. The nearest mirror of Wikipedia is used to construct the library, which contains almost 8 hundred thousand topics. The index of topic is constructed to improve the processing speed. Therefore, the corresponding synonyms, hypernym about the topics could be obtained.

### 3.1 Question Expansion with Wikipedia

ICTCLAS platform (http://ictclas.nlpir.org/) is used to do word segmentation for Chinese questions and stop words are removed. LTP platform [15] is employed to words or phrases disambiguation.

According to Wikipedia knowledge library, the questions could be expanded by using their synonyms. For example, the question “移动定制版三星 Note2 是否可以在美国国际漫游?” is processed into the word list “移动 定 制版 三星 Note2 是否 可以 在 美国 国际漫游 ?” after word segmentation. In this sentence, the term “移动” has three synonym words “中国移动通信”，

“运动(物理学)” and “移动电话”. Obviously, the first word “中国移动通信” is the most similar to the term and could be added to the sentence according Wikipedia knowledge. The question after expansion is shown in the following figure 1.

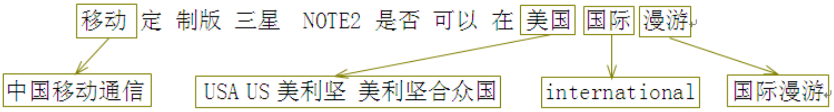

**Fig. 1.** Question expansion with Wikipedia

After question expansion, the word list has become

“移动 中国移动通信 定 制版 三星 Note2 是否 可以 在 美国 USA US 美利坚 美利坚合众国 国际 international 漫游 国际漫游”.

## 3.2 Refined Feature Extraction Methods

### 3.2.1 Domain Proper Nouns Table

In Community-based Question Answering system, there are some domain proper nouns which can contribute to category identification. For example, Dungeon-Fighter is a popular game and becomes a hot topic in the game community of Baidu Zhidao. The domain term “地下城” cannot be spitted into “地下 城”. Therefore, the domain proper nouns are collected and adopted as a dictionary for word segmentation tools. The dictionary contains 427 domain words and phrases with 13 top categories.

### 3.2.2 Feature Table upon Category Tree

The categories are organized as a tree from top level to bottom level in CQA. We collect labeled question in CQA belong to the category tree and extract the word feature. On top level, the features with the coarse categories are extracted, and the data sparseness is very obvious because of the huge amount of questions. Therefore, the fine features are extracted according to the bottom level.

We set a predefined dimension $D$, and extract the high-frequency terms from labeled questions. First, the number of terms belong to the category $i$ with high frequency is $Count_i$, and the number of overall terms is counted as $Sum$. Then, the proportion of the number of term with high frequency is set to $P_i= Count_i/Sum$. So, the number of features extracted from the labeled questions upon the bottom categories is $N_i= P_i*D$.

### 3.2.3 Refined Feature Extraction

Considering the poor contributions to the classification by the single words in Chinese questions, these features with single words are removed from the feature table. Then, feature table upon categories tree contains only those features with two words or upon.

Because of the imbalance number of high-frequency words upon each category, the features extracted with types are different. With the increasing of the high-frequency terms on *i*-th category, the $P_i$ on the feature table is more than others. Because the total number of $D$ is fixed, the number of features extracted upon the other categories is decreased. Therefore, this imbalance maybe affects the classification accuracy. In order to solve this problem, those features where the threshold on frequency is below 10 are removed from the feature table.

# 4 Experiments

Chinese question data collected from Baidu Zhidao are used as training examples in the experiments. The types in the categories tree can be divided three layers from top to bottom. The top layer contains 13 categories, and the second layer contains 141 categories. In the second layer, there are 41 categories which can be divided into the third layers and then the third layer contains 289 categories.

We collect 714,582 questions from Baidu Zhidao and randomly select 87,149 questions for the training examples because the whole data set is too large. Each question belongs to one category at each category level. Ten times 10-fold cross validation is performed on the experimental data set. In detail, the data set is partitioned into ten subsets with similar sizes and distributions. Each fold is selected once as the test set with 10% examples of the whole set while the remaining nine folds are combined into the training set. The whole above process is repeated for ten times and the results are averaged.

Maxent Entropy Model is a general purpose machine learning framework that has proved to be highly expressive and powerful in NLP community. We employ the Maxent Entropy tool [16] as the classifier for questions classification in CQA.

## 4.1 Experiments on Top Layer

The top layer contains 13 categories. The high-frequency terms are extracted from the training dataset. According to the different thresholds, the dimensions of features are set to 1,500 and 2,000 respectively. The baseline method is the traditional TF-IDF. The second method uses the domain proper nouns table and the third method imports the feature table upon category tree. Here, the number of iteration in the Maxent Entropy is fixed to 20. The experimental results are as follows in table 1.

**Table 1.** Experiment Results on Top Layer

| The Method of Feature Extraction | Classification Accuracy Rate | |
|---|---|---|
| | Dimension | |
| | 1500 | 2000 |
| TF-IDF | 55.98% | 59.00% |
| Domain Proper Nouns Table | 60.58% | 63.60% |
| Feature Table upon Category Tree | 66.67% | 70.79% |

From the above experimental results, compared to the TF-IDF method, it can be seen that the accuracy rates could be effectively improved by the method of features extraction after importing both the domain proper nouns table and the feature table upon category tree. For instance, the accuracy rate by feature table upon category tree on 2000 dimension achieved 70.79% and increased clearly compared with TF-IDF.

## 4.2 Experiments on Middle Layer

Compared to the top layer, the second layer contains more categories. In this experiment on middle layer, the methods importing the domain proper nouns table and the feature table upon category tree are directly used. The dimensions of features are also set to 1,500 and 2,000 respectively. Furthermore, we adopt the refined method to extract features with removing the single words and remove those features whose frequency is bellow 10. Table 2 tabulates the detailed information of the experimental results.

**Table 2.** Experiment Results on Middle Layer

| The Method of Feature Extraction | Classification Accuracy Rate | |
|---|---|---|
| | Dimension | |
| | 1500 | 2000 |
| TF-IDF | 50.07% | 52.81% |
| Remove Features with Single Word | 57.45% | 58.89% |
| Remove Features Whose Frequency Is Below 10 | 61.23% | 62.23% |

From the experiments, it can be observed that the classification accuracy rates are increased with the refined methods. Furthermore, compared to the method of removing single words, the accuracy rates are improved after the low-frequency terms are removed from the feature table. For example, on 1500-dimension, the classification rate by the all above refined method achieved 61.23%.

## 4.3 Experiments on Bottom Layer

There are 289 categories on the third layer. First, both the domain proper table and the feature table upon categories are used to expand the questions. Then, the refined methods of feature extraction are employed by removing both the single words and the low-frequency words. Figure 3 is the comparison of classification accuracy rates by the different methods.

**Table 3.** Experiment Results on Bottom Layer

| The Method of Feature Extraction | Classification Accuracy Rate | |
|---|---|---|
| | Dimension | |
| | 1500 | 2000 |
| TF-IDF | 51.34% | 54.61% |
| Domain Proper Nouns Table and Feature Table upon Category Tree | 62.79% | 64.18% |
| Refined Feature Extraction | 69.20% | 71.21% |

From table 3, it can be seen that classification accuracy rates are obviously improved by the proposed refined method of feature extraction. For example, on 2000-dimension, the classification accuracy rate is 71.21% by the refined feature extraction. Note that the accuracy rate obtained on the third layer is even higher than on the second layer. Compared to the second layer, categories on the third layer become more finely. The categories on the third layer may be the most relevant to the features for question classification.

## 5 Conclusion

Community-based Question Answering has become a hot topic in recent years. Question classification is an important part for CQA systems. In this paper, the Wikipedia knowledge is used to expand the Chinese questions to enrich the features. Then, the refined method of feature extraction is proposed step by step. Experiments show the proposed method could significantly improve the classification accuracy rate in the Chinese question classification. Explore more powerful methods of feature extraction for the Chinese question classification in CQA is an interesting issue for future work.

**Acknowledgments.** This work was supported by the National Natural Science Foundation of China (No.61365010, 61363044), Yunnan Nature Science Foundation (2011FZ069), Yunnan Province Department of Education Foundation (2011Y387).

## References

1. Riahi, F., Zolaktaf, Z., Shafiei, M., Milios, E.: Finding Expert Users in Community Question Answering. In: Proceedings of the 21st International Conference Companion on World Wide Web, 791–798 (2012)
2. Chen, L., Zhang, D., Mark, L.: Understanding User Intent in Community Question Answering. In: Proceedings of the 21st International Conference Companion on World Wide Web, pp. 823–828 (2012)
3. Zhang, D., Lee, W.S.: Question Classification Using Support Vector Machines. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, pp. 26–32 (2003)
4. Moschitti, A., Quarteroni, S., Basili, R., et. al.: Exploiting syntactic and shallow semantic kernels for question answer classification. In: Proceedings of 45th Annual Meeting of the Association for Computational Linguistics: York, pp. 776–783 (2007)
5. Zhang, Y., Liu, T., Wen, X.: Modified bayesian model based question classification. Journal of Chinese Information Processing **19**(2), 100–105 (2005). (in Chinese)
6. Hotho, A., Staab, S., Stumme, G.: WordNet Improves Text Document Clustering. In: Proceedings of the Semantic Web Workshop of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto Canada, pp. 541–544 (2003)
7. Cai, L., Zhou, G., Liu, K., Zhao, J.: Large-Scale Question Classification in cQA by Leveraging Wikipedia Semantic Knowledge. In: Proceeding of the 20th ACM Conference on Information and Knowledge Management (2011)

8. Gebrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorication with encyclopedia knowledge. In: IJCAI, pp. 1301–1306 (2006)
9. Wang, P., Domeniconl, C.: Building semantic kernels for text classification using wikipedia. In: KDD (2008)
10. Wang, P., Hu, J., Zeng, H.-J., Chen, L., Chen, Z.: Improving text classification by using encyclopedia knowledge. In: ICDM, pp. 332–341 (2007)
11. Hu, J., Fang, L., Cao, Y., Zeng, H., Li, H., Yang, Q., Chen, Z.: Enhancing text clustering by leveraging Wikipedia semantics. In: SIGIR (2008)
12. Hu, X., Sun, N., Zhang, C., Chua, T.-S.: Exploting internal and external semantics for the clustering of short texts using world knowledge. In: CIKM (2009)
13. Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: KDD (2009)
14. Zesch, T., Müller, C., Gurevych, I.: Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: LREC (2008)
15. Che, W., Li, Z., Liu, T.: LTP: A Chinese Language Technology Platform. In: Proceedings of the Coling 2010:Demonstrations, Beijing, China, pp. 13–16 (August 2010)
16. Le, Z.: Maximum Entropy Modeling Toolkit for Python and C++. Software available at. http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolikt.html