Exploring the Existing Category Hierarchy to Automatically Label the Newly-arising Topics in cQA*

Guangyou Zhou, Li Cai, Kang Liu, and Jun Zhao
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing 100190, China
{gyzhou, lcai, kliu, jzhao}@nlpr.ia.ac.cn

ABSTRACT

This work investigates selecting concise labels for the newly-arising topics in community question answer. Previous methods of generating labels do not take the information of the existing category hierarchy into consideration. The main motivation of our paper is to utilize this information into the label generation process. We propose a general framework to address this problem. Firstly, we map the questions into Wikipedia concept sets, which are more meaningful than terms. Secondly, important concepts are identified to represent the main focus of the newly-arising topics. Thirdly, candidate labels are extracted from Wikipedia category graph. Finally, candidate labels are filtered and reranked by combination of structure information of existing category hierarchy and Wikipedia category graph. The experiments show that in our test collections, about 80% "correct" labels appear in the top ten labels recommended by our system.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval - *information filtering , selection process*; H.3.5 [**Information Systems and Applications:**]: Web-based services

General Terms

Algorithms, Experimentation, Performance

Keywords

Category Hierarchy, Newly-arising Topics, Community Question Answering

1. INTRODUCTION

Community question answering (cQA) portals are a typical form of user-generated content that is gaining a large audience in recent years. Many cQA sites have emerged in the past few years as an enormous market to fulfill the information needs. For example, Yahoo! Answers[1], AnswerBag[2], WikiAnswers [3] are such cQA services.

In cQA, asker-posted questions are generally organized into a category hierarchy. Answerers navigate through this category hierarchy to search questions they are interested in. Therefore, the category hierarchy is of great importance for both askers and answers [1, 2, 3, 20]. This category hierarchy is usually maintained by human efforts, and its structures remain unchanged in a fairly long period [9]. Questions belonging to none of the existing categories would be assigned by users into the pseudo "Other" categories, e.g., "Other-Internet" in the "Internet" domain, as shown in Figure 1. Consequently, the current categories are definitely unable to capture newly-arising topics which are attracting intensive public attentions [9]. These accumulating "Other" questions bring difficulties and inconvenience to both users and cQA service providers [9].

Miao et al. [9] studied extensively the problem of new category identification in cQA, which aims to find potential categories not included currently. However, it is not enough only to detect and capture the new or emerging topics which are not included in the existing category hierarchy; it is equally important that the new hierarchical categories should have high-quality labels [4],which can improve users' ability to browse the collection [16].

In this paper, we address the problem of automatically labeling the newly-arising topics in cQA. The task is to take the question archives under the existing category hierarchy as input and generate labels for newly-arising topics from the pseudo "Other" categories. For example, Figure 2 shows the category hierarchy of *"Internet"* and its question archives. We can see that the input of our system is the category hierarchy of *"Internet"* and its question archives. And the output is the ranked labels for newly-arising topics in the *"Internet"* domain (e.g *"Twitter"*, etc), which are extracted from the category of *"Other-Internet"*.

Traditional approaches mainly aim to create and label totally new hierarchical clusters or flat clusters, not for automatically generate labels for newly-arising topics in consistency with the Existing Cat-

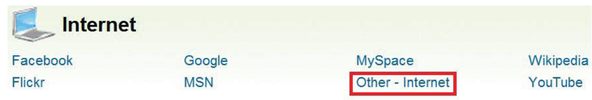

Figure 1: Category hierarchy of "Internet" domain of Yahoo! Answers.

*Guangyou Zhou and Li Cai have equal contribution to this work.

[1]http://answers.yahoo.com/
[2]http://www.answerbag.com/
[3]http://wiki.answers.com/

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

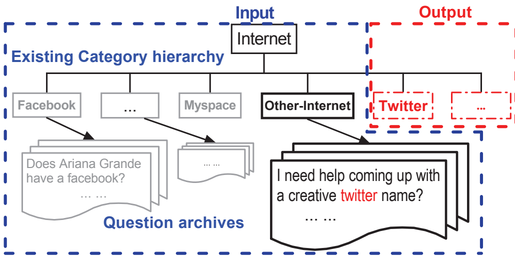

Figure 2: An illustration of our task.

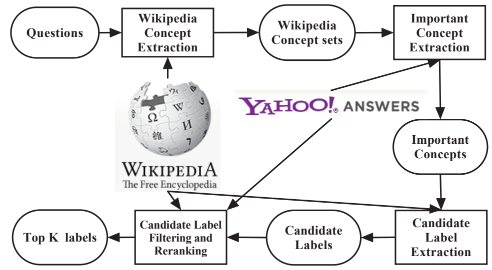

Figure 3: A general framework of our system.

Figure 4: An example of the identified Wikipedia concepts for question Q_1

egory Hierarchy[4](ECH). Therefore, the criteria for judging a high-quality label cannot be totally the same with those of the previous methods. Using the example shown in Figure 2, we illustrate the criteria for judging a high-quality label as follows.

- The generated labels should cover the main newly-arising topics that web users are interested in. This criterion is the same with that of previous methods.
- The generated labels should differentiate the newly-arising topics from their existing sibling and parent categories. For example, the top two candidate labels for the newly-arising topics in the existing *"Internet"* domain are *"Twitter"* and *"Social networking services"*. Intuitively, compared with *"Social networking services"*, *"Twitter"* is a better label for a new sub-domain in *"Internet"* domain. The reason is that some existing sibling categories(e.g *"Facebook"*, *"Myspace"*, etc) belong to *"Social networking services"*. That is to say, *"Social networking services"* overlaps some its existing sibling categories.
- The generated labels should be consistent with the ECH. Just as the example shown in Figure 2, *"Twitter"* is of the same granularity and more like a sibling node of the existing category hierachy (e.g *"Facebook"*, *"Myspace"* and etc).

The rest of this paper is organized as follows. Section 2 describes our proposed approach. Section 3 presents our experimental results. Section 4 concludes with ideas for future work.

2. GENERAL FRAMEWORK OF OUR PROPOSED METHOD

We propose a general framework for selecting concise labels for the newly-arising topics in consistency with the ECH in cQA using external resources. The reasons that we select Wikipedia for this task from many other external resources are as follows [4, 11]. The aim of our system is to select concise new category labels that are similar to what a person might create manually. A good new category label should not only indicate the main newly-arising concept, but also differentiate the new category from its siblings and its parent node.

The framework of our proposed method is shown in Figure 3, which includes four main components: Wikipedia concept extraction, important concept extraction, candidate label extraction and candidate label filtering and reranking. The general flow of the system can be summarized as follows. The system receives question collection as input. Initially, the questions are mapped into Wikipedia concept sets. The system extracts a set of important concepts that are estimated to best represent the topics of questions. Then, candidate labels can be selected by leveraging the Wikipedia. Finally, the set of candidate labels are reranked and a list of top recommended labels is returned by the system. In the rest of this section, we describe each of the system components in detail.

[4]The existing category hierarchy in our paper refers to the category hierarchy in cQA(e.g Yahoo! Answers)

2.1 Wikipedia Concept Extraction

To use the Wikipedia thesaurus, one of the key issues is how to map words in questions to Wikipedia concepts. Considering frequently occurring synonym, polysemy and hypernym in questions, accurate allocation of words in Wikipedia is really critical in the whole classification process.

Following Hu et al.[7], we build a phrase index which includes the phrases of Wikipedia concepts, their synonyms, and polysems in Wikipedia thesaurus. Based on the generated Wikipedia phrases index, all candidate phrases in the question can be recognized. We use the Forward Maximum Matching algorithm [18] to search candidate phrases, which is a dictionary-based word segmentation approach. In this process, it is necessary to do word sense disambiguation to find its most proper meaning mentioned in questions if a candidate concept is a polysemous one. Wang et al.[17] proposed a disambiguation method based on document similarity and context information, and the implemented method show high disambiguation accuracy. We adopt Wang et al.[17]'s method to do word sense disambiguation for the polysemous concepts in the question. Figure 4 shows an example of the identified Wikipedia concepts for question Q_1 using the above method. The phrase "social network" in Q_1 is mapped into Wikipedia concept *"Social Network"*, "twitter" in Q_1 is mapped into Wikipedia concept *"Twitter"*.

2.2 Important Concept Extraction

Given a set of Wikipedia Concepts $C \in \mathcal{C}$ as input, we now wish to find a list of concepts $\mathcal{T}(C) = (t_1, t_2, ..., t_k)$, ordered by their important weight. Existing term weight approaches extract term weight by comparing term distributions of a cluster to a reference collection and taking the statistically most discriminative terms[16, 4, 15, 6, 5]. However, we cannot fix the number of clusters and assign initial clusters for dynamic Web information services. Our approach eliminates the problem of determining the number of clusters and assigning initial clusters. Inspired by previous work [10], We capture the difference of concept distribution with regard to a sub-domain, and employ *Jensen Shannon Diver-*

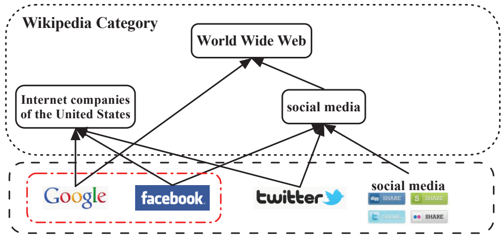

Figure 5: An example of the parent and child category nodes of *"Facebook"*

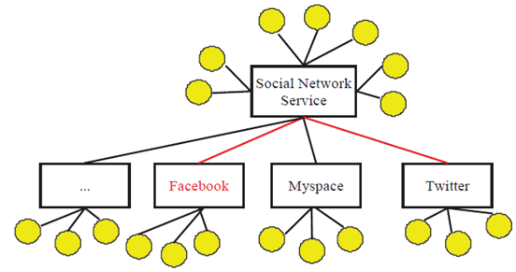

Figure 6: An example of the distance in WCG

gence(JSD). We examine the point-wise function for each concept as follows:

$$score_{jsd}(c, S||D) = \frac{p_s(c)\log\frac{2p_s(c)}{p_s(c)+p_d(c)} + p_d(c)\log\frac{2p_d(c)}{p_s(c)+p_d(c)}}{2} \quad (1)$$

where S and D denote the sub-domain-specific and the domain-specific concept set, with their probability distribution $p_s(c_i)$ and $p_d(c_i)$ obtained respectively by the Maximum Likelihood Estimator on the sub-domain-specific and the domain-specific concept set. We look for a set of concepts that maximize the *JSD* distance between the sub-domain-specific S and the domain-specific D. Each concept is scored according to Equation (1). The top scored concepts are then selected as the important concepts. We will experimentally show the superiority of this set of terms over the top weighted sets of terms in traditional term weight schemes.

2.3 Candidate Label Extraction

In Wikipedia, each concept belongs to one or more categories. Among them, there are many administrative categories. We use the following methods to filter out these meaningless labels. First, we utilize the methods proposed in[13] to derive generic hierarchical relation from category links. Second, we keep the category labels which have their corresponding equivalent concepts as candidate labels. After these processs, we can get meaningful categories from Wikipedia category graph.

As noted by Hu et al.[7], the higher level categories have less influence than those lower level categories since the lower level categories are more specific and therefore can depict the articles more accurate. In this paper, we collect only the first level categories of concepts as candidate labels. For the weight of candidate labels cl, we use Equation (2).

$$score(cl) = \sum_{i=1}^{n}(\alpha_i * score_{jsd}(c_i, S||D)) \quad (2)$$

where α_i controls the influence of concept c_i on candidate label cl. We perform an exhaustive grid search of step size 0.1 on [0,1] to find the parameter on a small development set of the question collection *"Other-Internet"*. We set α_i empirically as shown in Equation (3), as this setting achieves the best performance in the experiments.

$$\alpha_i = \begin{cases} 1 & c_i \text{ is equivalent concept as candidate label } cl \\ 0.1 & c_i \text{ is not equivalent concept as candidate label } cl \end{cases} \quad (3)$$

2.4 Candidate Label Filtering and Reranking

Given the input of the set of candidate labels and the corresponding part of Yahoo! Answers category hierarchy, we now wish to get the set of top-K recommended candidate labels for the new-arising topic. From above steps, we can get a list of candidate labels. However, many of them are not consistent with the ECH. We propose to filter and rerank the candidate category labels by utilizing the structure of ECH and WCG.

To make our generated labels differentiate the newly-arising topics from the existing sibling and parent categories in the ECH, we filter out the candidate labels which are the parent and child category of nodes of ECH in WCG. Figure 5 shows an example of the parent and child category nodes of *"Facebook"*. From Figure 5, *"Social media"* is the parent category of *"Facebook"* in WCG. Therefore, we filter out *"Social media"* from the candidate label list.

Wikipedia category graph has proved to be a successful knowledge source for semantical relatedness measures[19]. Therefore, we propose to utilize the structure information of hierarchy of Yahoo! Answers and Wikipedia category to rerank the candidate labels.

To make our generated labels consist with the ECH, we rerank the candidate labels by utilizing Zesch and Gurevych[19]'s method of computing semantic relatedness. Zesch and Gurevych[19]'s method of computing semantic relatedness essentially measures the distance of the two concepts in a taxonomy-like structure. Following Zesch and Gurevych[19], path length(PL) based and intrinsic information content(IIC) based measures are the two main ways of computing semantic relatedness(SR) in WCG. We define C_1 and C_2 as the set of categories assigned to Wikipedia concept a_i and a_j , respectively. We then determine the semantical relatedness value for each category pair (c_k, c_l) with $c_k \in C_1$ and $c_j \in C_2$. We choose the best value among all pairs (c_k, c_l), i.e. the minimum for PL based and the maximum for IIC based measures, which are shown in Equation (4).

$$SR(a_i, a_j) = \begin{cases} \min\limits_{c_k \in C_1, c_l \in C_2}(sr(c_k, c_l)) & \text{PL based} \\ \max\limits_{c_k \in C_1, c_l \in C_2}(sr(c_k, c_l)) & \text{IIC based} \end{cases} \quad (4)$$

Zesch and Gurevych[19] proved experimentally the two methods achieved the best result. Therefore, we adopt the above two methods to measure $sr(c_k, c_l)$, which is shown in Equation (5).

$$sr(c_k, c_l) = \begin{cases} 1/dist_{PL}(c_k, c_l) & \text{PL based[14]} \\ 2 * \frac{IIC(c_k \cap c_l)}{IIC(c_k)+IIC(c_l)} & \text{IIC based[8]} \end{cases} \quad (5)$$

To measure the semantic relatedness between the candidate labels and the ECH, we use the average score of the semantic relatedness between the candidate labels and sibling labels in ECH. The candidate label cl is scored using Equation (6):

$$score(cl, H) = \frac{\sum_{k=1}^{n} SR(cl, k)}{n} \quad (6)$$

where H denotes the domain-specific category hierarchy and $SR(cl, k)$ denotes the semantic relatedness between the candidate label cl and the existing label k , n denotes the total number of sibling labels in H.

Based on the $score(cl, H)$, we are able to rerank the candidate label cl to improve accuracy for labeling using Equation (7):

$$Score_{label}(cl, H) = (1 - \lambda) * score(cl) + \lambda * score(cl, H) \quad (7)$$

where $\lambda \in [0, 1]$ controls the relative importance of $score(cl, H)$. We set $\lambda = 0.2$ empirically as this setting achieves the best performance in the experiments. The final candidate labels are ranked based on the value of $Score_{label}(cl, H)$. The set of top-k scored candidate labels are recommended for the final labeling. Our assumption is that the semantic relatedness between candidate labels and sibling labels in ECH can contribute to select concise labels for the newly-arising topics in cQA. We will experimentally show the effectiveness of candidate label filtering and reranking.

3. EXPERIMENTS

3.1 Data Collection

We collect questions from all categories at Yahoo! Answers[5]. These questions have been issued over a period from March to November, 2011. We only focus on the resolved, meaningful questions that have been given their best answers. The resulting question repository contains 1659 categories. There are 26 categories at the first level, 309 categories at the second level, 1324 categories at leaf level. For preprocessing, we filter out the questions of less than three words. And we perform document frequency feature selection on the vocabulary: those words which appear in less than three questions are removed. For term weight labeling baseline, all the questions are converted into lower case. Each question is tokenized with a stop-word remover[6] and Porter stemming.[7]

We select 5 domains from Yahoo! Answers to evaluate our proposed method, two from second level, three from third level. The selected domains vary both in depth(between level two and level three with respect to the Yahoo! Answers Root) and in topics(from *internet* to *pets*). The selected domains are as follows: *Pets*, *Consumer Electronics*, *Business & Finance*, *Internet*, *Car Makes*.

Wikipedia data can be obtained easily from the website[8] for free research use. It is available in the form of database dumps that are released periodically. The version we used in our experiments was released on July. 22, 2011. The category names are processed with POS tagger developed by the Stanford NLP group[9].

3.2 Evaluation

The category hierarchy of cQA is always maintained by human editors. We propose a method by utilizing the existing labels assigned by human editors in Yahoo! Answer as ground-truth dataset to evaluate our proposed method. The process can be summarized as follows: For a specific domain, we put the questions from the "Other" category and one of its sibling categories together to form a new question collection. And the ground-truth "correct" label for this new question collection is the label of the above sibling category. Fox example, the questions under *"Other-Internet"* and *"Facebook"* are combined into a question collection. *"Facebook"* is used as the ground-truth "correct" label for this question collection. Then, every category besides "Other" category under the same domain is respectively used to form a new collection with its ground-truth "correct" label. For example, in domain "Internet", which can be seen in Figure1, seven question collections can be respectively formed with their corresponding ground-truth "correct" labels(e.g *"Facebook"*, *"Google"*, *"MySpace"*, *"Wikipedia"*, *"Flickr"*, *"MSN"* and *"Youbute"*). The goal of the experiments is to re-find the label X for the category "X+Other"(with X= *"Facebook"*, *"Google"*, etc.). We use these testing question collections with ground-truth "correct" labels to evaluate our proposed method.

For each configuration, we evaluate the system's performance using the following measure:

Match@K:The relative number of candidate labels for which at least one of the top-k labels is correct. The higher this measure is and the lower k is, the better the system's effectiveness.

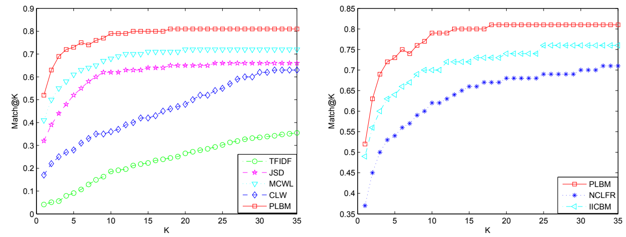

Figure 7: Left: The overall performance of all the methods. Right: The performance of the methods of NCLFR and IICBM

3.3 Experimental Setting

To evaluate the performance of our proposed method, we compare the following systems:

(1)**Maximum Term Weight Labeling(TFIDF)**: A naive method to extract term weighting is to use term weight scheme in traditional information retrieval model. Therefore, the TFIDF term weight method can be used as a classic baseline.

(2)**Maximum Term Weight Labeling(JSD)**: This method exploits multiple aspects of the domain information to improve the performance of the existing retrieval models [10]. Therefore, the JSD term weight method can be used as a relatively strong baseline.

(3)**Carmel's Labeling using Wikipedia(CLW)**: Carmel et al.[4] obtained the state-of-the-art performance for cluster labeling by leveraging Wikipedia. In this experiment, we used top JSD weighting terms as query terms . Following the literature [4], we fixed the parameters to be 1) 20 important terms for querying Wikipedia, 2) 100 Wikipedia results for candidate extraction, and 3) the score propagation judge for candidate evaluation.

(4)**Maximum Concept Weight Labeling(MCWL)**: In this method, the weight scheme is the same with JSD. The only difference is that concepts replace terms, which are more meaningful.

(5)**We proposed path length based measure(PLBM)**: In this experiment, we use path length based measure to evaluate the candidate category: 1) 20 important terms for querying Wikipedia, 2) 100 Wikipedia results for candidate extraction, and 3) the score propagation judge for candidate evaluation.

3.4 Results and Discussion

The results are displayed in Figure 7(left). By comparing the results of different methods, we draw the following observations:

(1) JSD performs much better than TFIDF(up to 43.4% improvement with Match@10). The result is consistent with that reported in the experiment in [4], which JSD is measured between the cluster and the entire collection. We can further observe that MCWL pro-

[5]http://answers.yahoo.com/
[6]http://truereader.com/manuals/onix/stopwords1.html
[7]http://www.ling.gu.se/lager/mogul/porter-stemmer/index.html
[8]http://download.wikipedia.org
[9]http://nlp.stanford.edu/software/index.shtml

vides even better performance. We believe this is because Wikipedia concept representations contain much more information than the term representations do.

(2) CLW performs not that well as in [4]. Even the simplest JSD with terms achieves improvement with Match@10 26% over CLW. We believe this is because the top weight terms are not confined to a certain topic. Hence, Wikipedia articles retrieved using these terms as queries are irrelevant and introduce noise into the system's decision making scheme.

(3) Our proposed method achieves the best performance(up to 11% improvement with Match@10 over the best baseline MCWL).

It is also interesting to note that the method without using candidate labels evaluation require at least 30 candidate labels to achieve the precision of 70%, while the same effectiveness is achieved by a list of 3 candidate labels using candidate labels evaluation. To conclude, these result clearly show that the combination of structure information of ECH and Wikipedia category can contribute to select a concise label for the newly-arising topic in question archive. The combination of structure information of ECH and Wikipedia category yields significant performance improvement. This shows that candidate category filtering and reranking can contribute to select concise labels for the newly-arising topics in cQA.

3.5 The Effectiveness of Candidate Label Filtering and Reranking

We now explore the effectiveness and efficiency of candidate label filtering and reranking by utilizing structure information from ECH and WCG. For this evaluation, we carried out an experiment without using candidate label filtering and reranking(NCLFR). Moreover, an experiment using IIC based measure(IICBM) to evaluate the candidate category is also carried out for comparison. The results are displayed in Figure 7(right).

From Figure 7(right), we can see that using candidate labels filtering and reranking can significantly improve the Match@K value. It is interesting to note that NCLFR requires at least 30 candidate labels to achieve the precision of 70%, while the same effectiveness is achieved by a list of 3 candidate labels using path based measure to evaluate candidate labels. We conclude that the combination of structure information of ECH and WCG can contribute to select concise labels for the newly-arising topic in cQA.

PLBM achieves better performance than IICBM (up to 6% improvements with Match@10). The reason may be that semantically related terms are very likely to be categorized under the same category.

4. CONCLUSION AND FUTURE WORKS

In this paper, we investigate to select concise labels for the newly-arising topics in consistency with the ECH in cQA by utilizing Wikipedia knowledge. We describe a general framework to address this task which extracts candidate labels, filters and reranks the candidate labels by utilizing the structure information of ECH and WCG. This process can be enhanced in several way. First, from the experimental results, we can see that the best Match@1 is 0.52, which is far from practical use. We attempt to improve the performance of Match@1 in future work. Second, other knowledge from WCG, such as siblings, can be used for selecting labels.

5. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 61070106), the National Basic Research Program of China (No. 2012CB316300), Tsinghua National Laboratory for Information Science and Technology (TNList), Cross-discipline Foundation and the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (No. 5026035403). We thank the anonymous reviewers for their insightful comments.

6. REFERENCES

[1] L. Cai, G. Zhou, K. Liu, and J. Zhao. 2011. Learning the Latent Topics for Question Retrieval in Community QA. In *IJCNLP*, pages 273-281.

[2] L. Cai, G. Zhou, K. Liu, and J. Zhao. 2011. Large-scale question classification in cQA by leveraging wikipedia semantic knowledge. In *CIKM*, pages 1321-1330.

[3] X. Cao, G. Cong, B. Cui, and C. S. Jensen. 2010. A generalized framework of exploring category information for question retrieval in community question answer archives. In *WWW*.

[4] D. Carmel, H. Roitman, and N. Zwerdling. 2009. Enhancing cluster labeling using Wikipedia. In *SIGIR*, pages 139-146.

[5] F. Geraci, M. Maggini, M. Pellegrini, and F. Sebastiani. 2007. Cluster generation and cluster labelling for web snippets: a fast and accurate hierarchical solution. Internet Mathematics, Vol. 3, No. 4: 413-444.

[6] E. Glover, D. Pennock, S. Lawrence, and R. Krovetz. 2002. Inferring hierarchical descriptions. In *CIKM*.

[7] J. Hu and L. Fang and Y. Cao and H. Zeng and H. Li and Q. Yang and Z. Chen. 2008. Enhancing text clustering by leveraging Wikipedia semantics. In *SIGIR*.

[8] D. Lin. 1998. An information-theoretic definition of similarity. In *ICML*.

[9] Y. Miao, C. Li, J. Tang, and L. Zhao. 2010. Identifying new categories in community question answering archives: a topic modeling approach. In *CIKM*.

[10] Z. Ming, T. Chua, and C. Gao. 2010. Exploring domain-specific term weight in archived question search. In *CIKM*.

[11] Z. Ming, T. Chua, and C. Gao. 2010. Prototype hierarchy based clustering for the categorization and navigation of web collections. In *SIGIR*.

[12] V. Nastase and M. Strube. 2008. Decoding Wikipedia categories for knowledge acquisition. In *AAAI*.

[13] S. Ponzetto and M. Strube. 2007. Deriving a large scale taxonomy from Wikipedia. In *AAAI*.

[14] R. Rada, H. Mili, E. Bicknell, and M. Blettner. 1089. Development and application of a metric on semantic nets. IEEE Trans. on Systems, Man, and Cybernetics, 19(1):17-30.

[15] H. Toda and R. Kataoka. 2005. A clustering method for news articles retrieval system. In *WWW*.

[16] P. Treerapituk, J. Callan. 2006. Automatically labeling hierarchical clusters. In *DGO*, pages 167-176.

[17] P. Wang, J. Hu, H. Zeng, L. Chen, and Z. Chen. 2007. Improving text classification by using encyclopedia knowledge. In *ICDM*, pages 332-341.

[18] P. Wong and C. Chan. 1996. Chinese word segmentation based on maximum matching and word binding force. In *COLING*.

[19] T. Zesch and I. Gurevych. 2007. Analysis of the Wikipedia category graph for NLP applications. In *NAACL-HLT*, pages 1-8.

[20] G. Zhou, L. Cai, J. Zhao, and K. Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *ACL*, pages 653-662.