2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Topic Detection for Discussion Threads with Domain Knowledge

Mingliang Zhu Weiming Hu Ou Wu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
{mlzhu, wmhu, wuou}@nlpr.ia.ac.cn

Abstract

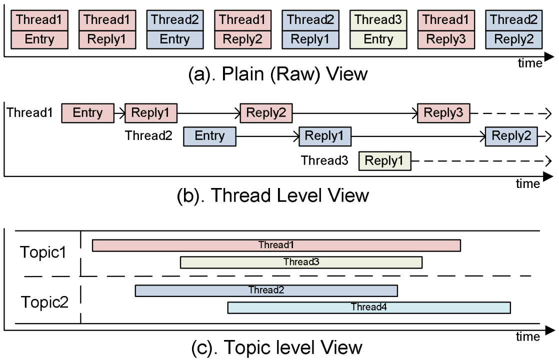
The online communities are becoming so popular along with the development of the web but indexing and searching for the discussion data are big challenges to current applications. Topic detection was proposed to solve the problem but the accuracy is still not satisfactory, mainly because key elements are usually implicit or ambiguous which literal content comparison cannot handle. In this paper, we propose to improve the basic topic detection model by combining domain knowledge. The domain knowledge can be automatically extracted from a collection of external knowledge sources and applied to the content analysis of the threads. Two approaches, i.e. the LDA and the Concept Mapping, are proposed to implement the knowledge extraction and integration. Experimental results show that both approaches make the detection accuracy outperform the previous model. The LDA approach achieves better overall performance while the Concept Mapping is more suitable for dynamic knowledge sources.

1. Introduction

Online discussion communities are becoming more attractive and important. Many communities organize the discussions into *threads*, based on replying relationships among discussion posts. Figure 1 (a) and (b) illustrates the post and thread level structures of the communities. Discussion data are difficult to summarize and index because of their very nature. They are much looser and more distributed in structure. Furthermore, the languages used are very casual, which brings a lot of noises. As a result, the discussion communities are great challenge to current data indexing and searching models.

One way to address these problems is to analyze *topics* among the threads [14]. Topics reflect the discussion trends, and provide a higher view above the thread level (see Figure 1 (c)). Threads in the same topic are relatively highly connected to each other. Topics often relate to real event in people's lives.

The original Topic detection and tracking (TDT) [1] are designed for news stories, and do not work well on discussion data. In [14] a TDT model was introduced for discussion threads but the result was still far from satisfactory. One reason is that certain key elements are often implicit or ambiguous in real-life discussion but human users are able to infer those with prior knowledge about the topic. For example, in a basketball game topic, a certain player may be referred as different nick names, or there may be other threads talking about the game statistics. Literal content comparison is likely to fail on such conditions.



(a). Plain (Raw) View

(b). Thread Level View

(c). Topic level View

Figure 1. Levels of views on discussion community

In this paper, we propose to make use of domain knowledge to improve the topic detection accuracy for discussion data. The domain knowledge can be obtained from some type of sources, such as encyclopedia articles or news stories. Two models, i.e. the *LDA* (Latent Dirichlet Allocation [2]) and the *Concept Mapping*, are introduced to extract knowledge and combine it into the basic framework. The LDA approach works for relatively stable sources while Concept Mapping is designed for dynamic sources.

2. Related work

Topic detection and tracking (TDT) for news stories has been widely studied for years [1][4][12][13], but because of the very differences between discussion data and news stories, these models do not work very well directly on the discussion data [14].

Several methods have been developed to improve text related tasks by combing certain forms of prior knowledge, such as WordNet [7] and ontologies. These knowledge bases usually require manual creation so that their utilities are somewhat limited. Another approach is to use statistical methods, such as latent topic analysis models including pLSA [9] and LDA [2], and word co-occurrence based models [11]; they can

978-0-7695-4191-4/10 $26.00 © 2010 IEEE
DOI 10.1109/WI-IAT.2010.68

automatically extract the semantic connections of terms. There are also search engine based approaches, for example, the semantic similarities of term P and Q can be estimated based on the search results for the queries P, Q and $P\ AND\ Q$ [3].

3. The basic model

In this section we briefly describe the common model used for topic detection, as shown in the "Basic Model" dashed line box in Figure 2.

The algorithm runs periodically and each time newly created posts go into the model. The raw post data are first pre-processed to extract the posts properties and perform typical lexical regulations. Then, during the content analysis process, for each new or updated thread, its content is compared with each earlier thread. Usually the TF-IDF strategy is used for content term weighting. For content similarity comparison, the Hellinger distance is reported to perform better than cosine metric in TDT tasks [4][14]. The content similarity values are then compared with a pre-defined threshold θ_c. If the new thread is similar enough to some old thread, it is treated as in some known topic, or else treated to be in a new topic.

In [14] a few improvements were made, such as filtering out uninformative posts and user activity analysis. Some other factors, such as temporal information [5], may also help. Our domain knowledge extension is designed to work with others transparently so that it can be applied with almost no modifications to existing frameworks.

4. Incorporating the domain knowledge

As discussed in section 1, topic detection in the online discussion environment is still far from good. One critical problem with discussion data is that many key elements may be implicit or ambiguous but can be inferred from context. One idea to solve that is to find semantic connections among terms in addition to literal comparison. For example, if the terms "Kobe" (an NBA player) and "rebound" can be connected with "basketball", it is possible to tell that some thread talking about Kobe or rebound statistics is in the topic of a basketball game. We use a collection of external documents as knowledge source. Then two methods, i.e. LDA and Concept Mapping, are used to extract the domain knowledge and combine it for content analysis of discussion threads. Online encyclopedia documents and news stories are examples of knowledge sources: they are easy to collect and are naturally divide into different domains.

4.1. The LDA approach

LDA (Latent Dirichlet Allocation) [2] is a generative probabilistic model, which models a text document as a mixture over an underlying set of hidden topics. Figure 3 demonstrates the graphical model representation of LDA. Note that the concept *hidden topic* is not equivalent to the *discussion topics* we are trying to detect. The hidden topic is a finite set and is a virtual layer upon which LDA assumes that a document is generated. In the rest of this paper, we will use the term *h-topic* to avoid confusion.

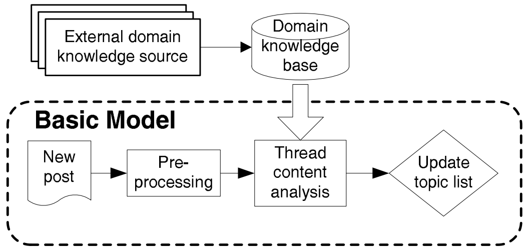


Figure 2. Outline of the framework

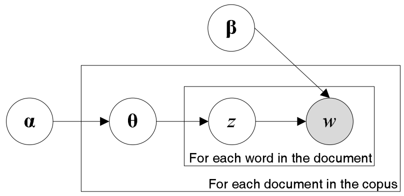


Figure 3. Graphical model representation of LDA

The details of the LDA model estimation and probability inference algorithms can be found in [2]. During the training process, the LDA parameters can be estimated upon the external knowledge source. And then in actual topic analysis of discussion data, the h-topic distribution, i.e., the $\boldsymbol{\theta}$ parameter in LDA model can be inferred on each thread, noted by $\underline{\boldsymbol{\theta}}$.

The h-topic distribution $\underline{\boldsymbol{\theta}} = \{\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_K\}$ are integrated into the original TF vector using the strategy similar to the one introduced in [10], except for that decimal fraction TF values are allowed in our case: For each of the K h-topics in LDA, append an extra element to the original vocabulary, so that each topic becomes a virtual term, denoted by "{HT$_1$}", …, "{HT$_K$}". The values for the virtual terms in the TF vector of a thread d are transformation of the $\underline{\boldsymbol{\theta}}$ distribution (denoted $\{\underline{\theta}_1{}^{(d)}, \underline{\theta}_2{}^{(d)}, \ldots, \underline{\theta}_K{}^{(d)}\}$) of the thread:

$$tf(d, \{HT_k\}) = \begin{cases} 0, & \text{if } \underline{\boldsymbol{\theta}}_k^{(d)} \leq \dfrac{1}{K} \\ a \cdot \log_K(K\underline{\boldsymbol{\theta}}_k^{(d)}) + b, & \text{otherwise} \end{cases} \quad (1)$$

The above equation maps h-topic distribution value from $(1/K, 1]$ into term frequency range $(b, (a{+}b)]$. $a > 0$ and $\mathrm{b} \geq 0$ are parameters controlling the impact of the virtual terms. $a = 8$ and $b = 2$ are appropriate for discussion data since the transcripts are usually to be short. The impact of virtual terms may be weakened in long thread cases, but it is likely that terms directly from these threads contain enough information. Also, the global DF vector is appended with the K virtual terms with $df(\{\mathrm{HT}_k\})$ being the number of seen threads whose $tf(d, \{\mathrm{HT}_k\})$ is not 0.

The described process transforms the h-topic distribution values into virtual terms, so that the domain knowledge analysis by LDA is integrated into the basic model. Although the original TF vector consists of only integers, the content similarity calculation does not require this, so the original similarity calculation does not need any modification.

4.2. The Concept Mapping Approach

The LDA parameter estimation is usually time consuming and if the knowledge source updates, there is no easy way to update the model but perform the estimation process over again. As a result, LDA is not very suitable for dynamic sources (such as a news collection). We propose *Concept Mapping* approach for this situation. The Concept Mapping is based on the term co-occurrence frequencies. In [6] Deerwester gave an empirical discussion that the co-occurrence can be used to represent the latent topic structure that is quite like the pLSA and LDA do.

From the external knowledge base, a term co-occurrence matrix $\mathcal{C}$ is created, with $\mathcal{C}(w_i, w_j)$ be the number of documents that both term i and term j appear. Also, a term occurrence vector $\mathcal{O}$ is created and $\mathcal{O}(w_i)$ is the number of document term i appears.

The Concept Mapping is then taken during the content analysis for discussion threads. Before calculating the similarity between thread d_0 and a previous thread d_i, we try to map each term that only appears in d_0 into another term that appears in thread d_i (but not required to be still in d_0). For such a term w_m in d_0, search among terms in d_i for the one with maximum Jaccard term co-occurrence similarity:

$$\begin{aligned} w^{(i)}_{m'} &= \arg\max_{w\in d_i} \mathrm{Jaccard}(w_m, w) \\ &= \arg\max_{w\in d_i} \frac{\mathcal{C}(w_m, w)}{\mathcal{O}(w_m)+\mathcal{O}(w)-\mathcal{C}(w_m, w)} \end{aligned} \tag{2}$$

If a $w^{(i)}_{m'}$ with positive Jaccard similarity is found, term w_m in the TF vector of thread d_0 is mapped into term $w^{(i)}_{m'}$ by multiplying the found maximum Jaccard similarity as a mapping ratio:

$$\begin{aligned} &tf'(d_0, w^{(i)}_{m'}) = tf(d_0, w^{(i)}_{m'}) + \mathrm{Jaccard}(w_m, w^{(i)}_{m'}) \cdot f(d_0, w_m) \\ &tf'(d_0, w_m) = 0 \end{aligned} \tag{3}$$

Similar to the LDA approach, the term frequencies may no longer be integer values after mapping. Then the TF-IDF weighting and similarity comparison is taken upon the tf' vector instead of the original tf vector. The similarity calculation remains unchanged.

If new documents are added into the knowledge source, it is easy to update the knowledge base (i.e. matrix $\mathcal{C}$ and vector $\mathcal{O}$) incrementally by counting term frequencies in the newly added documents:

$$\begin{aligned} \mathcal{C}^{(new)}(w_i, w_j) &= \mathcal{C}(w_i, w_j) + \mathcal{C}^{(inc)}(w_i, w_j) \\ \mathcal{O}^{(new)}(w_i) &= \mathcal{O}(w_i) + \mathcal{O}^{(inc)}(w_i) \end{aligned} \tag{4}$$

where $\mathcal{C}^{(inc)}(w_i, w_j)$ denotes the number of newly added document in which both term w_i and term w_j appear, and $\mathcal{O}^{(inc)}(w_i)$ the number documents in which term w_i appears. This incremental strategy enables Concept Mapping approach for knowledge sources that updates frequently, such as news collections.

5. Experiments

Intensive experiments are conducted to evaluate the performance of our proposed method.

The experiments require two data sets: the online discussion data as well as a collection of documents served as domain knowledge source. The discussion data are from the "Olympic" board on the ShuiMu community[1], which discusses stuffs about the Olympic Games. All posts during Aug. 15, 2008 and Aug. 18, 2008, 4 of the days that the Beijing 2008 Games were held, are downloaded by our spider. The thread relations of posts are extracted, and community system posts are excluded from the data sets. totally there are totally 92605 posts of 10968 threads, averagely 23151 posts and 2742 threads every day. They are then divided into two subsets: (1) the base set $\mathcal{B}$ (posts during Aug. 15 and Aug. 16) for generating the initial DF vector; and (2) the testing set $\mathcal{T}$ (during Aug. 17 and Aug. 18) for testing. The threads in set $\mathcal{T}$ are manually clustered into topic collections, which are used as ground truth. Table 1 summarizes the online discussion data sets.

We use news stories from the official website of the Beijing 2008 Olympic Games [2] to serve as the knowledge source. All the news stories date from Aug. 17 to Aug. 18 were crawled and the main content of the stories were extracted. The two proposed approaches are implemented to extract the domain knowledge. The GibbsLDA++ library [10] is used for LDA calculations. Best performance is achieved when the LDA approach uses all the nouns while the Concept Mapping uses the named entities.

The C_{Det} evaluation metric is used to evaluate the performance of our methods by combining miss rate P_{Miss} (the probability that a new topic is not identified) and false alarm rate P_{FA} (an old topic is determined as new) of the algorithm:

$$C_{Det} = C_{Miss} \cdot P_{Miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{nontarget} \tag{5}$$

C_{Miss} and C_{FA} are costs for misses and false alarms. $C_{Miss} = 1$ and $C_{FA} = 0.2$ are used in the experiment. P_{target} is the probability of seeing a new topic while

[1] http://www.newsmth.net/

[2] http://www.beijing2008.cn/news/sports/headlines/

$P_{nontarget}$ is the probability of seeing an old one. The reported C_{Det} cost is normalized as:

$$(C_{Det})_{norm} = \frac{C_{Det}}{\min(C_{Miss} \cdot P_{target}, C_{FA} \cdot P_{nontarget})} \qquad (6)$$

We compare the results of the model to the previous proposed model in [14] (labeled by Zhu08). The normalized C_{Det} curve for the LDA approach, Concept Mapping approach as well as the previous model are shown in Figure 4 respectively. Table 2 shows the minimum $(C_{Det})_{norm}$ values on the curve as well as the θ_c values found to achieve the best performances for each approach. θ_c is used to threshold the content similarities and greater θ_c suggests that threads in the same topic have greater similarity values under the certain content analysis approach (refer to Section 3), so that whether or not a thread discusses a new topic is more distinguishable.

The experimental results show that both proposed domain knowledge approach improve the detection performance, having greater best θ_c thresholds. It can be inferred that by the help of domain knowledge, threads in the same topic will have greater content similarity measures. For comparison between the LDA and Concept Mapping approaches, LDA achieves better overall performance, especially on the side of the C_{Det} curve that favors low false alarm rate. If lower miss rate is favored, the performances of the two approaches are near. Also, as discussed before, the training step of LDA approach is much more time consuming, so if the knowledge source is quite dynamic, the Concept Mapping approach is preferred.

Table 1. summary of the discussion data sets

	Base set $\mathcal{B}$	Test set $\mathcal{T}$	Total
Posts	43237	49368	92605
Threads	4964	6004	10968
Topics	N.A.	1685	1685

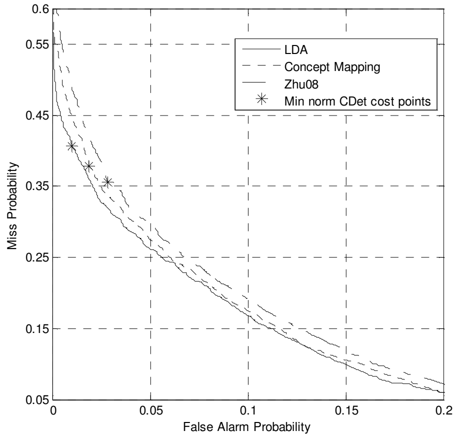


Figure 4. C_{Det} curve of the test results

Table 2. Minimum normalized C_{Det} Costs

	$\min(C_{Det})_{norm}$	θ_c
LDA	0.642	0.254
Concept Mapping	0.668	0.215
Zhu08	0.708	0.208

6. Conclusion

In this paper, we have proposed a framework to integrate domain knowledge into the topic detection model. We have designed two approaches to extract and integrate domain knowledge into the content analysis of the threads. Experimental results show that both approaches outperform the existing model. The LDA approach achieves better overall performance while the Concept Mapping is more suitable for dynamic knowledge sources.

7. Acknowledgement

This work was supported in part by the NSFC (Grant No. 60825204, 60935002 and 60903147).

8. References

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron and Y. Yang. “Topic detection and tracking pilot study: Final report”. *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[2] D.M. Blei, A.Y. Ng, M.I. Jordan. "Latent Dirichlet allocation". *Journal of Machine Learning Research*, 2003(3).

[3] D. Bollegala, M. Yutaka, and M. Ishizuka, "Measuring semantic similarity between words using web search engines," in *WWW '07*, 2007.

[4] T. Brants, F. Chen and A. Farahat. “A System for New Event Detection”. *SIGIR‘03*, 2003.

[5] C. C. Chen, M. C. Chen and M Chen. “An adaptive threshold framework for event detection using HMM-based life profiles”. *ACM Transactions on Information Systems*, 27(2), 2009.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman. “Indexing by latent semantic analysis”. *Journal of the American Society for Information Science*, 41(6), 1990.

[7] C. Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[8] T. L. Griffiths and M. Steyvers, "Finding scientific topics." *Proc. National Academy Science*, 101(1), 2004.

[9] T. Hofmann, "Probabilistic latent semantic analysis," in *UAI'99*, 1999.

[10] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. “Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections”, in *WWW2008.*

[11] M. Wettler, "Computation of word associations based on the co-occurrences of words in large corpora," *1st Workshop on Very Large Corpora: Academic and Industrial Perspectives*, 1993.

[12] Y. Yang, T. Pierce and J. Carbonell. “A Study of Retrospective and On-line Event Detection”, in *SIGIR98*.

[13] K. Zhang, J. Li and G. Wu. “New Event Detection Based on Indexing-tree and Named Entity”, in *SIGIR2007.*

[14] M. Zhu, W. Hu, O. Wu, “Topic Detection and Tracking for Threaded Discussion Communities”, in *WI08,* 2008.