A User-Oriented Special Topic Generation System for Digital Newspaper

Xi Xu[1,2,3(✉)], Mao Ye[2], Zhi Tang[1,2], Jian-Bo Xu[2], and Liang-Cai Gao[1]

[1] Institute of Computer Science and Technology, Peking University, Beijing, China
{xuxi}@founder.com.cn, 10548887@pku.edu.cn
[2] State Key Laboratory of Digital Publishing Technology,
Peking University Founder Group Co. LTD., Beijing, China
{yemao.apb,tommie,xu.jb}@founder.com.cn
[3] Postdoctoral Workstation of the Zhongguancun Haidian Science Park, Beijing, China

Abstract. With the coming of digital newspaper, user-oriented special topic generation becomes extremely urgent to satisfy the users' requirements both functionally and emotionally. We propose an applicable automatic special topic generation system for digital newspapers based on users' interests. Firstly, extract subject heading vector of the topic of interest by filtering out function words, localizing Latent Dirichlet Allocation (LDA) and training the LDA model. Secondly, remove semantically repetitive vector component by constructing a synonymy word map. Lastly, organize and refine the special topic according to the similarity between the candidate news and the topic, and the density of topic-related terms. The experimental results show that the system has both simple operation and high accuracy, and it is stable enough to be applied for user-oriented special topic generation in practical applications.

Keywords: Digital newspaper · User-oriented special topic · Latent Dirichlet allocation · Synonymy word map

1 Introduction

In the day of information and network, digital reading is increasingly important and people have unprecedentedly strong demand for news. The demand mainly embodies the quick picking up of the latest, the deepest and the most comprehensive news. The future of the digital newspaper has arrived. News can be custom-made to suit your own requirements. All these will make news expression more interactive than ever. Thus, it is necessary to innovate and develop the reorganization of digital news.

Many researchers have made attempts to organize and generate special topic, while there are few achievements of user-oriented automatic special topic generation. In Ref. 1, a set of topic-related news is extracted by textual retrieval using query expansion algorithm. Ref. 2 gets the candidate set of subject headings by integrating the result of meaningful string recognition algorithm into the subject heading vocabulary compiled in advance, and then filter the candidate set and calculate the weight of each subject heading in accordance with certain heuristic rules. Incremental clustering

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 484–491, 2015.
DOI: 10.1007/978-3-319-25207-0_45

algorithm and named entity are introduced to detect and represent the topics in Ref. 3, and Support Vector Machine (SVM) [4] is also adopted to organize and track the topics. Ref. 5 represents each piece of news as a sequence of words by Vector Space Model (VSM) [6] and trains self-defined topic model on documents around the same topic to estimate whether a piece of news belongs to the topic.

All above-mentioned methods not only can not generate fine-grained special topic automatically according to users' interests but also have two major defects. One is that a subject heading vocabulary needs to be compiled manually in advance. The other is that the impacts of polysemant and synonyms are not taken into account and reduced. In application, as newspapers contain a large amount of information and update very fast, there is hardly any subject heading vocabulary and it is a fussy and high subjective task to add subject headings manually. The number of polysements is small, but the probability of their occurrence is very high. Synonyms are sometimes the relationship between meanings because of polysemants. Polysemants and synonyms increase the difficulty of semantic similarity computation of texts.

To overcome above weaknesses and improve interactivity, we designed and implemented a user-oriented special topic generation system for digital newspaper. We first filter out function words based on part of speech tagging, localize Latent Dirichlet Allocation (LDA) [7] and train the LDA model to extract subject headings of the topic which a user is interested in. The function word filter and localized LDA model prevents us from compiling subject heading vocabulary manually or recognizing named entity. Then, we construct a synonymy word map according to Chinese thesaurus [8] to filter synonymy words out of the subject headings. The removing of semantically repetitive subject headings effectively reduces the impacts of polysemant and synonyms. At last, we develop an algorithm to compute the similarity between a text and a topic and organize the special topic according to priority transformed from similarity. The similarity calculation algorithm can globally expresses similarity under localized LDA and subject heading vector.

2 Subject Heading Extraction

To avoid compiling a subject heading vocabulary beforehand, key words of the samples are extracted based on LDA as subject headings. In practice, the samples are selected by the user and are on the same subject that the user is interested in.

2.1 Text Preprocessing

Although there are many elements of news, such as the "press", "title" and "content", only the "content" is used to extract subject headings in our research. First, the "content" has to be segment by word segmenter according to Chinese grammatical and lexical analysis. Considering that not all parts of speech are qualified for subject headings, content word is much better than function word, the word segmenter must tag part of speech. Thus, the processing of "content" includes Chinese word segmentation and part of speech tagging.

There are many Chinese word segmenters, such as IKAnalyzer, Ansj and so on. Ansj is selected because of its high accuracy and function of part of speech tagging.

It is a Java implementation of the ictclas [9]. As the candidate terms of subject headings should keep in full meaning as far as possible and new terms always appear in news, coarse-grained method in Ansj, NlpAnalysis, is used in our research. User-defined dictionary is also added to improve the performance of new term discovery and the reliability of word segmentation.

Not all terms obtained from word segmenting are qualified for subject headings. Firstly, filter the function words from the terms within parts of speech. Specifically, stop preposition, conjunction, auxiliary, punctuation and so on. Secondly, filter the insignificant words from the rest of the terms within a stop-word list. Stop-words are words which are filtered out before or after processing of natural language text [10]. In our research, the stop-word is compiled based on the feature of the style of writing in news, such as "本报记者" (our reporter) and importing Chinese stop-word list. It covers a wide number of stop-words without getting too aggressive and including too many words which might point to the subject of news.

Eventually, the "content" of each piece of news is condensed into a sequence of content words and the candidate word set of subject headings are obtained.

2.2 Key Word Extracting Based on LDA

Topic model is introduced to our research to semantically extract the subject information according to the demand of special topic generation for digital newspaper. In natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents [11].

One of the most classic topic models is LDA. It was first presented as a probabilistic graphical model for topic discovery by David Blei et al in 2003. The basic idea and more details of LDA are described in Ref.12.

In digital newspapers, neologisms and new topics are constantly emerging. Thus, it is quite difficult to provide adequate corpus for LDA estimator. Furthermore, the topics attracting users are always fine-grained and the performance of LDA inferencer is not in sufficient reliability to satisfy demand. All these led us to apply LDA in an extremely novel way that is reducing the dimensionality of topic and using just LDA estimator. The basic idea of LDA can be simply represent by the following formula:

$$p(word \mid doc) = \sum_{topic} p(word \mid topic) \times p(topic \mid doc) \quad . \tag{1}$$

If set the number of topic the minimum, one, all documents belong to the same topic. Here one piece of news, which is composed of the sequence of content words obtained in Section 2.1, is assumed as a document and all documents around one topic are chosen by the user in real time. This method can be extended to two or more topics, but the size of corpus and the number of topics should keep small. The Dirichlet and conditional multinomial parameters for the LDA model that is α and β are iteratively computed with initial value using Expectation Maximization (EM) until convergence. Then the probability of word w_n under topic z_n can be calculated by multinomial distributions $p(w_n \mid z_n)$ with known α and β.

2.3 Remove Semantically Repetitive Subject Headings

As mentioned in Section 1, polysemant and synonym are common phenomenon, and therefore two different subject headings obtained above may have the same meaning especially when they are in the same text. Furthermore, the significance of the meaning, which more than one subject heading have, should be enhanced in the topic.

We construct a synonymy word map according to Chinese thesaurus to filter synonymy words out of the subject headings. The thesaurus was firstly compiled by Mei J.-J. et al. in 1983 [13] and was expanded by Information Retrieval Laboratory of Harbin Institute of Technology. Pertinent experiment has shown that although the expanded Chinese thesaurus is already covering a large number of Chinese words, there are still many words, especially proper nouns such as organization name and so on, appeared in the digital newspaper but not in the thesaurus. Thus, we once again supplemented the Chinese thesaurus based on neologism and abbreviations (e.g. "马航" and "马来西亚航空公司" (Malaysia Airlines)) in popular usage in newspaper.

In the synonymy word map, each meaning is identified by a character string composed of numbers and letters, named the meaning code. Thus, every monosemy is mapped to one character string and every polysemant is mapped to two or more meaning codes. We iteratively remove semantically repetitive subject headings using the synonymy word map. When two different subject headings have the same meaning code, the one with lower probability is removed and its probability is added to the other one with higher probability. The experiments show that the removing can reduce the subject headings by 10 to 15 percent. 20 percent of the highest probability words (terms in Section 2.1) of a topic are retained as subject headings considering the features of natural language and writing style in news. After that the probability of each subject heading is normalized within the topic. Until now, the condensed subject headings altogether construct a subject heading vector denoted by *topicwords* = ($tterm_1$, $tterm_2$, ⋯, $tterm_M$), where $tterm_m$ is the mth dimension of the vector. Moreover, the probability of $tterm_m$, denoted by p_m, is also obtained and

$$\sum_{i=1}^{M} p_i = 1 \quad . \tag{2}$$

As the probability of the subject heading, expressed by different words with the same meaning, is increased, the subject heading vector represents the topic more accurately.

3 Special Topic Generation

To meet the demand of user-oriented special topic generation, we use LDA estimator in a small corpus to extract subject headings and their corresponding probability. Obviously, the LDA model here is just local model because the global one is not suitable for our application, so LDA inferencer is not applicable any more. A new method is proposed to compute similarity between a new "content" and the topic to infer that whether a piece of news belongs to a topic. The similarity is actually between the key words of the new "content" and the subject headings of the topic.

3.1 Compute Similarity Between News and a Topic

In the practical application, there is a huge amount of news and the news is updated at very high speed. Moreover, many neologisms always appear in news. Therefore, it is quite difficult to train word vector under the corpus of news and we have to calculate the similarity in another way. For a piece of news to be observed, firstly, preprocess its "content" using the method described in Section 2.1 to obtain the sequence of content words, denoted by $seqterms = (term_1, term_2, \cdots, term_L)$, where $term_l$ is the lth word in the sequence. And compute the probabilities of the words in *seqterms* using the method "computeArticleTfidf", which is based on TF-IDF (term frequency-inverse document frequency) and supplied by Ansj. Secondly, remove semantically repetitive words by the method proposed in Section 2.3 and update the corresponding probabilities. Thirdly, sort the words in descending order by probability and the result is denoted by $keywords = (kterm_1, kterm_2, \cdots, kterm_N)$, where $kterm_n$ is the nth word in *keywords*. Finally, set the weight of $kterm_n$, denoted by w_n,

$$w_n = 1 - \frac{n-1}{2N} \quad , \tag{3}$$

where N is the total number of terms in *keywords* and n is the sequence number of $kterm_n$. Obviously, $1 \le n \le N$ and $0.5 < w_n \le 1$.

Now, we can compute the similarity between a piece of news to be observed and a certain topic. As semantically repetitive terms have already been removed in the *topicwords* using the synonymy word map (detailed in Section 2.3), the algorithm starts iteration by *topicwords* to calculate the similarity:

- Step 1: Choose a term in *topicwords*, denoted by $tterm_i$. Search *keywords* in descending order by probability using the synonymy word map until a term with the same meaning as $tterm_i$ is found or all terms are compared.
- Step 2: Mark $tterm_i$ processed. If no term is found, repeat step 1 until all terms in *topicwords* are processed. Else, denote the term found in Step 1 by $kterm_j$ and compute the $tterm_i$ component of similarity S_i, using the following equation:

$$S_i = p_i w_j \quad , \tag{4}$$

  where p_i is the probability of $tterm_i$ and w_j is the weight of $kterm_j$.
- Step 3: Remove $kterm_j$ from *keywords*. If *keywords* is not empty, repeat step 1 until all terms in *topicwords* are processed.
- Step 4: Define the similarity between a piece of news and a certain topic as

$$S = \sum_{i=1}^{M} S_i \quad , \tag{5}$$

  where M is the total number of terms in *topicwords*., We can deduce $0 \le S < 1$ according to Eqs. (2), (3) and (4).

The similarity calculation algorithm proposed in our research is aiming at a global similarity expression under localized LDA and subject heading vector.

3.2 Organize and Refine the Special Topic

To generate a special topic, it is essential to choose a candidate set from the database of news with restrictions such as the "date" and so on. As this operation is not the main point, here we just assume that the candidate set has already been offered.

For each piece of news in the candidate set, the similarity S between its "content" and the certain topic is calculated in by Eq. (5). A similarity threshold, denoted by θ, is introduced as a parameter between 0 and 1. Its value is determined by user and the default setting is 0.3. As only the piece of news which satisfied the inequality $S > \theta$ is collected to construct a rough special topic, the value of θ can adjust the grain size of the special topic. Finally, we obtain a rough special topic based on S and θ.

To increase the accuracy of the special topic, we filter each piece of news in the rough special topic by the density of topic-related terms. Firstly, construct the set of topic-related terms from all $kterm_j$, iteratively found in Step 2 of similarity calculation algorithm, and their synonyms in *keywords* (obtained in Section 3.1). Secondly, calculate the maximum distance d_{max} between any two topic-related terms in *seqterms* (obtained in Section 3.1). Lastly, if d_{max} is smaller than half of the total number of the terms in *seqterms*, remove this piece of news from the rough special topic.

After the filtering process, all pieces of news remained in the rough special topic are organized based on S to construct the final special topic. For any piece of news in the rough special topic, set its priority level *prior* the integer (100-100S). The smaller the *prior* is, the higher the priority level becomes. Then, sort all news remained in ascending order by *prior*. For all news with the same *prior*, first, choose a piece of news, denoted by *refnew*, and extract its subject heading vector using the same method proposed in Section 2; second, calculate the similarity S_{local} between *refnew* and any other piece of news with the same *prior*; last, set *refnew* is the standard news of this *prior* and group together *refnew* and all news satisfied $S_{local} > 0.8$. Repeat above steps until all pieces of news with the same *prior* are grouped.

Eventually, we generate the fine-grained special topic. All pieces of news are organized by the priority level and repetitive pieces under the same priority level are grouped together according to the similarity. As *prior* becomes bigger that is the priority level becomes lower, the grain size of the special topic gradually increases.

4 Experiments

There is no public database created for testing the performance of special topic generation methods, so it is difficult to do quantitative comparison experiments with other special topic generation methods and we evaluated the proposed system using real digital newspapers data. To verify our system (simply as UOSTG) has the better performance than muti-keyword retrieval (simply as MKR) when it is very difficult to accurately describe the subject by limited and certain keywords, we generated a special topic according to one piece of news focused on how foreign media report "2014 Kunming attack". Two thousand pieces of news was randomly chosen from twenty-one thousand pieces reported by People's Daily, Guangming Daily, Yunnan Daily and Global Times during two months, from Mar. 1st, 2014 to Apr. 30th, 2014. The

statistical result of the experiment is shown in Table 1. In muti-keyword retrieval, key terms are “昆明” (Kunming), “恐怖” (terror),“国外” (foreign) and “媒体” (press).

Table 1. Recall and accuracy of muti-keyword retrieval and user-oriented special topic generation.

Method	Recall (%)	Accuracy (%)
MKR	8.70	50.00
UOSTG	82.61	90.48

Table 1 shows that the user-oriented special topic generation has brought about huge improvement both in recall and accuracy. The recall of MKR is particularly lower because multi- keyword retrieval cannot accurately describe the topic.

To validate the advantage of removing semantically repetitive subject headings using synonymy word map, we generated another special topic focused on sea search for Malaysia Airlines incident. Three thousand pieces of news was randomly chosen from thirty thousand pieces of news reported by People’s Daily, Guangming Daily, Beijing Youth Daily and Beijing Morning Post, different presses from the above experiment to make result analysis easy, but during the same period. The subject heading vector without the synonymy word map (or simply as NoSWM) was {海域 (sea area) = 0.0432, 飞机 (airplane) = 0.0305, 客机 (airliner) = 0.0029, 马来西亚 (Malaysia) = 0.0208, 救援 (rescue) = 0.0203, ⋯}, while with the synonymy word map (also simply as UOSTG), {海域 (sea area) = 0.0468, 飞机 (airplane) = 0.0336, 舰 (warship) = 0.0318, 救援 (rescue) = 0.0289, 搜寻 (search) = 0.0275, ⋯}. Table 2 demonstrates the experimental result.

Table 2. Recall and accuracy with and without the synonymy word map.

Method	Recall (%)	Accuracy (%)
NoSWM	84.87	87.16
UOSTG	90.79	89.61

As shown in table 2, both recall and accuracy have risen because the subject heading vector refined by the synonymy word map can represent the topic more accurately. Most of the news, which appeared in the special topic but were regarded as extraneous pieces, related to Malaysia Airlines incident, but not to sea search.

Table 3 shows the experiment result of another five topics using user-oriented special topic generation system to further prove the universality of the system.

Table 3. Recall and accuracy of another five different topics.

Topic ID	Recall (%)	Accuracy (%)
1	85.96	92.45
2	93.90	79.38
3	84.73	88.80
4	76.47	92.85
5	95.52	87.67

5 Conclusion

To satisfy the users' requirements both functionally and emotionally, a user-oriented special topic generation system is designed, implemented and evaluated in our research. It has the following advantages. The function word filter based on part of speech tagging and localized LDA model prevents us from compiling subject heading vocabulary manually or recognizing named entity. The removing of semantically repetitive subject headings using the synonymy word map effectively reduces the impacts of polysemant and synonyms. Furthermore, the similarity calculation algorithm proposed in our research can globally expresses similarity under localized LDA and subject heading vector.

The experimental results show that the subject heading vector extracted by our methods can more accurately describes a topic than multi-keywords and the system has not only high recall but also high accuracy. The system is simple to use and stable enough to be applied for user-oriented special topic generation for digital newspaper in practical applications.

References

1. Fan, J.-R.: Research on Topic Generation and Retrieval of News Video Based on Text. Institute of Computing Technology, Chinese Academy of Science, Beijing (2008)
2. Li, H.-X., Zhang, H.-P.: Internet hot topic detection based on topic words. In: Proceedings of the 5th China Information Retrieval Conference, Shanghai (2009)
3. Wang, Z.-M.: Research on Web News Topic Organization and Acquisition System. College of Information Science & Engineering, Central South University (2008)
4. Cui, J.-M., Liu, J.-M., Liao, Z.-Y.: A Research of Text Categorization Based on Support Vector Machine. Computer Simulation **30**(2), 294–299 (2013)
5. Tan, H., Jia, Z.-Y., Shi, Z.-Z.: How to Organize and Generate News Topics with Great Efficiency. Science & Technology Review **7**, 48–51 (2004)
6. Erk, K., Padó, S.: A structured vector space model for word meaning in context. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2008)
7. Biggers, L.R., Bocovich, C., Capshaw, R., Eddy, B.P., Etzkorn, L.H., Kraft, N.A.: Configuring Latent Dirichlet Allocation Based Feature Location. Empirical Software Engineering **19**(3), 465–500 (2014)
8. He, D.: Retrospect of and Prospect for Chinese Thesaurus. Information Studies Theory & Application (2010)
9. Feng, G.-H., Zhen, Z.: Review of Chinese Automatic Word Segmentation. Library and Information Service **55**(2), 41–45 (2011)
10. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Pr., pp. 1–17 (2011)
11. David, M.B.: Probabilistic Topic Models. Communications of the ACM **55**(4), 77–84 (2012)
12. David, M.B., Andrew, Y.N., Michael, I.J.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
13. Mei, J.-J., Zhu, Y.-M., Gao, Y.-Q.: Cilin-thesaurus of Chinese words. Shanghai Lexicographic Publishing House, Shanghai (1983)