I.J. Wireless and Microwave Technologies **2012, 1, 10-16**
Published Online February 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijwmt.2012.01.02

Available online at http://www.mecs-press.net/ijwmt

Identifying Sentiment in Web Multi-topic Documents

Na Fan

School of Information Engineering Chang 'an University Xi' an, China

Abstract

Most of web documents coverage multiple topic. Identifying sentiment of multi-topic documents is a challenge task. In this paper, we proposed a new method to solve this problem. The method firstly reveals the latent topical facets in documents by Parametric Mixture Model. By focusing on modeling the generation process of a document with multiple topics, we can extract specific properties of documents with multiple topics. PMM models documents with multiple topics by mixing model parameters of each single topic. In order to analyze sentiment of each topic, conditional random fields techniques is used to identify sentiment. Empirical experiments on test datasets show that this approach is effective for extracting subtopics and revealing sentiments of each topic. Moreover, this method is quite general and can be applied to any kinds of text collections.

Index Terms: Analyzing Sentiment; Multi-topic Text; Parametric Mixture Model

1. Introduction

With the rapid growth of semantic web page (such as movie reviews, product reviews) in Internet, the analysis and detection of attitudes, feelings, or opinions expressed in a text has attracted more and more attention in the community of natural language processing and information retrieval. A challenging problem in this area is identification of sentiment. There is a wide range task in extracting sentiment form text. Previous work is focused on polarity classification [1, 2], opinion extraction [3] and opinion source assignment [4, 5]. However, a common deficiency of all this work is that the proposed approaches extract only the overall sentiment of a text or query, but can neither distinguish different subtopics within a text, nor analyze the sentiment of a subtopic. Since a text often covers a mixture of subtopics and may hold different opinions for different subtopics, it would be more useful to analyze sentiments at the level of subtopics. For example, a user may like the price and color of a new Nokia mobile telephone, but dislike its appearance. Indeed, people tend to have different opinions about different features of a product [6, 7]. In reality, a general statement of good or bad about a query is not so informative to the user, who usually wants to drill down in different facets and explore more detailed information. In all these scenarios, a more in-depth analysis of sentiments in specific aspects of a topic would be much more useful than the analysis of the overall sentiment of a text.

How to extract subtopics in web texts and reveal sentiment of these subtopics is a key problem. To the be of

Corresponding author:
E-mail address: fnsea@chd.edu.cn

our knowledge, no existing work has been conducted on exactly this kind of problem.

In this paper, we propose a novel probabilistic generative model of document with multiple topics and a method of identifying sentiment of each topic based on CRF. The proposed model is parametric mixture model (PMM). PMM models documents with multiple topics by mixing model parameters of each single topic. Moreover we identify sentiment of each topic and further model the dynamics of each topic and its associated sentiments.

The proposed approach is quite general and has many potential applications. The mining results are quite useful for predicting user behaviors, monitoring public opinions, and making business decisions.

The rest of the paper is organized as follows. In section 2, we propose the PMM model to extract subtopics from texts and analyzing sentiment of each subtopic. In section 3, we identify sentiment of each subtopic by conditional random fields model and reveal sentiment dynamics of each topic. In section 4, we present our experiment results. And we conclude in section 5.

2. Method of Extracting Subtopics

Most documents which appear on Internet have tended to be assigned with multiple topics in this situation; it is more and more important to analyze a relationship between a document and topics assigned to the document.

From a viewpoint of document classification, it can be said that documents have tended to be classified into multiple topics. In multiple-topic classification, a probabilistic generative model approach has gotten a lot of attention. A probabilistic generative model for documents with multiple topics is a probability model of the process of generating documents with multiple topics. By focusing on modeling the generation process, we can extract specific properties of documents with multiple topics. That is, we can analyze a relationship between a document and topics assigned to the document.

In this section, a probabilistic generative model approach is used to extract subtopics in multi-topic documents.

2.1. Parametric Mixture Model

Parametric mixture model is a typical probabilistic generative model. PMM models documents with multiple topics by mixing model parameters of each single topic.

Firstly, terminological words used in this section are explained. K is the number of explicit topics. V is the number of words in the vocabulary. d ={d1,…,dN} is a sequence of words where dN denotes the nth word in the sequence. d denotes a document itself and is called words vector. x is a word-frequency vector and xv denotes the frequency of word v. y={y1,y2,…yK} is a topic vector into which a document d is categorized, where yi takes a value of 1 or 0 when the ith topic is or not assigned with a document d. A probabilistic generative model for documents with multiple topics models a generation probability of a document d in multiple topics y using model parameter θ, i.e., models $P(d|y,\theta)$. The model parameters are learned by documents $D = \{(d_j, y_j)\}_{j=1}^{M}$, where M is the number of documents.

PMM employs BOW (Bag Of Words) representation and is formulated as follows.

$$p(d|y,\phi) = \prod_{w=1}^{V} (\phi(w,y,\phi))^{x_w} \quad (1)$$

ϕ is a K × V matrix whose element is $\phi_{iw} = P(w|y_i = 1)$. $\phi(w,y,\phi)$ is the probability that word w is generated from multiple topics y and is denoted as the linear sum of $h_i(y)$ and ϕ_{iw} as follows:

$$\phi(w,y,\phi) = \sum_{i=0}^{K} h_i(y)\phi_{iw} \quad (2)$$

Where $h_i(y)$ is a mixture ratio corresponding to topic i.

B. Model Parameter

In order to estimate parameter in PMM, a learning algorithm, which is an iteration method, is introduced. Model parameter θ is estimated by maximizing $\prod_{j=1}^{M} P(d_j | y_j, \theta)$ in learning documents $D = \{(d_j, y_j)\}_{j=1}^{M}$. Function f corresponding to a document j is introduced as follows:

$$f_{iw}^{j}(\theta) = \frac{h(y_j)\theta_{iw}}{\sum_{m=1}^{K} h_m(y_j)\theta_{mw}} \tag{3}$$

The parameters are updated along with the following formula (4).

$$\theta_{iw}^{(t+1)} = \frac{1}{C}\left(\sum_{j}^{M} x_{jw} f_{iw}^{j}(\theta^{(t)}) + \varepsilon - 1\right) \tag{4}$$

C is the normalization term for $\sum_{w=1}^{V} \theta_{iw} = 1$. x_{jw} is the frequency of word w in document j. ε is a smoothing parameter that is Laplace smoothing when ε is set to two. In this paper, ε is set to two as the original paper of PMM

According to PMM, subtopics can be extracted from documents.

3. Identifying Sentiment of Subtopic

In section 2, we extract subtopics of documents. In order to identify the polarity of each subtopic, sentiment of sentences in it should be taken into account. Sentiment of sentences in each subtopic influences each other and finally determines the sentiment of the subtopic. Because sentences in the subtopic are related in context, we treat the identification of sentiment of these sentences as a tagging task, and can use conditional random fields (CRFs) to address this problem.

CRFs are parametric families of conditional distributions that correspond to undirected graphical models and have been mostly applied to sequence annotation. The standard graphical structure is a chain structure on y with noisy observations x.

$$P(Y|X) = \frac{1}{Z_{\lambda}} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(Y, X)\right) \tag{5}$$

As mentioned before, we therefore define X as a sequence of sentences with Xi representing the i-th sentence in a subtopic. And the sequence y represents the sentiments of the sentences.

In sequence annotation a standard choice for feature is $\lambda_i f_i(Y, X)$ which is a binary function. We choose polarity words which may be adjectives or adverbs to represent feature of observation sequence X.

Firstly, we select seed words and label them with polarity values which may be –1or 1. The value of 1 means the positive polarity and the value of –1 means the negative polarity. Moreover, negation words have influence on sentiment words. If they modify sentiment words, the polarity of these words will be reversed. We build a list which includes familiar negation words. The effect of these words will not cross punctuation, such as commas, question marks, etc. Therefore they just will negate the polarity of the closest sentiment word.

Training CRFs model is typically estimated by maximum likelihood or MAP. In our work, maximum likelihood is adopted to train the model. Likelihood function is described in (6).

$$L(\lambda) = \sum_{i} \left\{ \lambda \cdot F(Y_k, X_k) - \log Z_{\lambda}(X_k) \right\} \tag{6}$$

According to CRFs model, the polarity of sentences in a subtopic is determined. We calculate the polarity of a subtopic as (7).

$$C = \sum_{j=1}^{l} S_j \tag{7}$$

Where C denotes the sentiment value of a subtopic, Sj denotes the sentiment value of the j-th sentence in the subtopic.

In order to explicitly reveal sentiment of every topic, we introduce the concept of sentiment dynamics. The sentiment dynamics for a topic is a time series representing the strength distribution of sentiment associated with the topic. The strength can indicate how much positive or negative opinion there is about the given topic in each time period and clearly reveal which sentiment dominates the opinions and the trend of this domination. This strength is reflected by the number of negative or positive documents.

4. Experiments

In this section, we describe the data and evaluations used in our experiments, and present our results.

4. 1 Data

To examine the ideas proposed in this paper, we need two types of data sets for evaluation. One is used to evaluate method of extracting topic. In order to evaluate efficiency of the method, we want the topics in this data set to be as diversified as possible. We collected a corpus, which includes 2000 on line reviews. Most of these reviews coverage multiple topics. The number of topics is 69. The size of vocabulary is 46,075. The proportion of documents with multiple topics on the whole dataset is 79.8%, i.e., that of documents with single topic is 20.2%. The average of the number of topics of a document is 3.2.

The other type of data is used to evaluate the method of analyzing sentiment and reveal sentiment dynamics. Such data need to have sentiment labels and time stamps. In particular, sentiment of every text was labeled manually and sentiment of all sentences was also annotated based on their context within every document. Sentences were annotated as neutral if they conveyed no emotion or had indeterminate sentiment from their context. The composition of the data set 2 is shown in table 1.

Table 1 Basic statistics of data set 2

Data	positive	negative	Time period
fund	1355	1469	11/01/06-8/01/08
stock	965	855	11/01/06-8/01/08

4. 2 Results

Our first experiment is to evaluate the method of extracting subtopic. We use F-measure to evaluate the method of extracting subtopics. F=2PR/P+R, $P = |n_r \cap n_e| / |n_e|$ and $R = |n_r \cap n_e| / |n_r|$. n_r is a set of relevant topics and n_e is a set of estimated topics. A higher F-measure indicated a better ability to discriminate topics.

We respectively use 6 data sets, which include the 15%, 30%, 50%, 70%, 80%, and 95% of 2000 on line reviews to implement our method. Table 2 is the experiment result.

Table 2 result of extracting subtopic

Proportion of data set	P (%)	R (%)
15%	74	72
30%	75	74
50%	75	73
70%	77	75
80%	79	74
95%	79	75

Fig.1 shows the value of F in these six experiments.

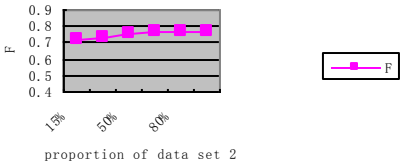

Figure 1. F changes with the size of test data

Fig.1 shows that the average of F arrives to about 75%. And with the increase of test data, the value of F also improves. The result demonstrates that the method has better robustness.

The second experiment is to evaluate the effectiveness of the method for analyzing sentiment. Firstly, we evaluate the efficiency of method in analyzing sentiment. Table 3 shows the result of identifying sentiment in data set 2. As with other information extraction task, we use precision (P), recall (R) and F to evaluate the result. F is the harmonic mean of precision and recall.

Table 3 Identifying sentiments in data set 2

Data	P (%)	R (%)	F (%)
fund	87.54	83.63	85.54
stock	87.36	86.75	87.05

Then we reveal sentiment dynamics of subtopic according to the change of time. In Fig.2 and Fig.3, we respectively present the sentiment dynamics of the subtopic “fund” and “stock”. Y axis denotes the number of negative or positive documents.

In Fig.2, we see that the positive sentiment about fund burst around July 2007; the negative sentiment, however, does not burst until several months later. This is reasonable, since investors need some time to determine whether fund marker changes badly.

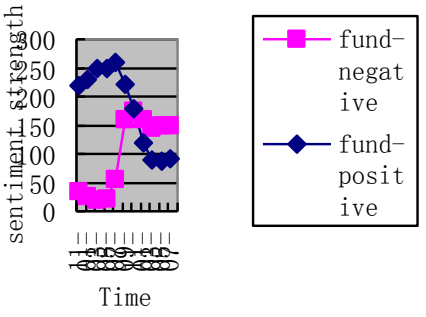

Figure 2. Sentiment dynamics of fund

In Fig.3, we see that in the beginning the positive sentiments dominate the opinions. However, in 11 2007, the negative sentiments shows a sudden increase of coverage.

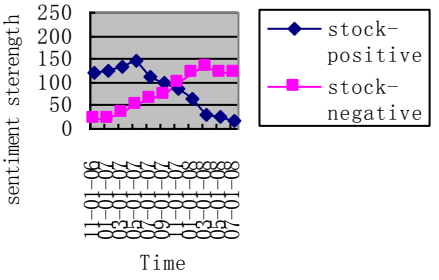

Figure 3. Sentiment dynamics of stock

Sentiment dynamics reveals the strength distribution of sentiment associated with the topic, and can provide more information to forecast trend of sentiment of the topic. So Sentiment dynamics can be used in many fields in reality, such as products survey.

5. Conclusion

In this paper, we propose a new method to extract subtopic and identify sentiment of subtopic in multi-topic web document. With this method, we could effectively extract topic and identify sentiments of every topic, especially represent the associated sentiment dynamics. We evaluate our method on different collections; the results show that the method is effective for web texts, which have multiple subtopics.

An interesting future direction is to further explore other applications of our method, such as user behavior prediction.

References

[1] Liu Jing, Zhong Wei-Cai, Liu Fang, Jiao Li-Cheng. Classification based on organization coevolutionary algorithm. Chinese Journal of Computers, 2003, 26(4): 446-453 (in Chinese).

[2] D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, Wiley Publishers, 2004.

[3] U. Fayyad, "Data mining and knowledge discovery in databases: implications for scientific database," Proc. of the 9th International Conference on Scientific and Statistical Database Management, pp. 2-11, 1997.
[4] C. Apte, B. Liu, E. P. D. Pednault, and P. Smyth,"Business applications of data mining," Communications of the ACM, vol. 45, no. 8, pp. 49-53, 2002.
[5] M. Garofalakis, R. Rastogi, and K. Shim, "Mining sequential patterns with regular expression constraints,"IEEE Trans. on Knowledge and Data Engineering, vol.14, no. 3, pp. 530-552, 2002.
[6] Zadeh, L.A. Fuzzy Sets. Information and Control 8:338:353 1965.