Word Polarity Analysis Method Based on Topic Model

Xiao-Nan FAN[1,a*], Shi-Min WANG[2,b]

[1]Haidian District Fucheng Road No. 11 Gemini building block A room1409, Beijing, China

[2]Haidian District Fucheng Road No. 11 Hard floor building room 806, Beijing, China

[a]xiaonan_smile@126.com, [b]minshiw@vip.sina.com

*Corresponding author

Key Words: Word polarity, LDA model, Random Walk, Positive and negative orientation.

Abstract. Along with the proliferation of new media, the user generated content becomes irreplaceable and providing main channel of daily information for people. By get rid of the shackle of the poor information, information technology has entered a big data era. Faced with the data overload, words polarity analysis research appeals the attention of numerous scholars and becomes the important role in national security and information filtering for Internet users, enterprises, and governments. However, due to the rapid change of internet words, the lexicon based sentiment analysis method shows its drawback. Because the traditional method cannot get the polarity of internet words to make the ideal corpus, they usually generate the bad results. This paper presented a topic-based word polarity analysis method which utilizes the LDA topic model and random walk method to get the polarity of a new word. Experiments show that our method achieves the proper accuracy and reasonable results.

Introduction

Sentiment analysis and opinion mining is an important part of Natural Language Processing. Meanwhile, word sentiment mining is a branch of sentiment analysis problem. Word sentiment mining aims at to infer the polarity of given word in specific context and contributes to information filter, dictionary expansion and mining sensitive information. People can make a buying decision through user’s product opinion; enterprise can learn the sense of identity of customer in terms of corporate culture in order to establish the enterprise brand image; the government can understand the public opinion information on the Internet, so as to understand the condition of the people, correctly guide public opinion.

With the development of the Internet, users can easily participate in the information publishing and sharing. According to the 31-th China Internet development situation statistic report [1], China micro-blog user continues to grow up to 3.9 billion, releases 1 billion per day by the end of 2012 in December. The emerging social media information wide coverage, in many people, strong real-time, has become an important channel for people to obtain information. But the cyber word updating quickly, non-standard and fewer words bring difficulties of word polarity mining. According to the situation, we propose a word polarity analysis method based on the topic model [2]. Our method uses the a few given polarity dictionary to infer the polarity of other words by utilizing the sentiment analysis technique.

Word Polarity Analysis Based on Topic Model

Because the network language updates faster, the traditional library cannot be marked polarity for it without delay that the supervised algorithm based on unable to get the ideal corpus, resulting in positive and negative inference of word accuracy drops. With the prevalence of Twitter and micro-blog, informal text, and the influence of the network vocabulary is bigger, so the tendency of non-standard words analysis becomes particularly important.

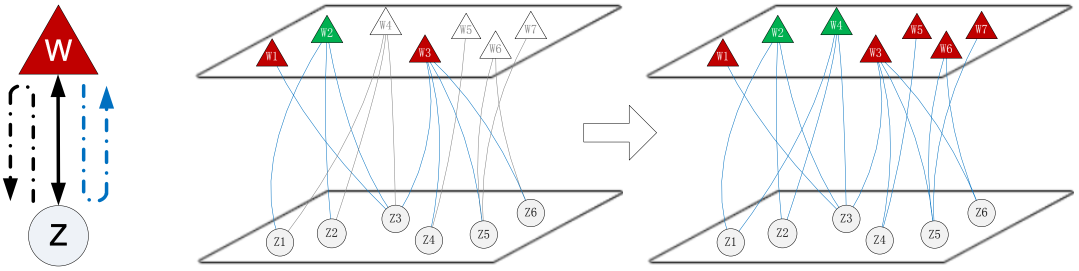
Since word tendency and context have the close relationship, direct idea is to obtain the relation between document and tendency of that words. In order to obtain that relationship, the method is based on lexical syntactic analysis and connected, this method requires a large amount of prior information, and the meanings of vocabulary and word order has a strong relation. To avoid the complex rules and a priori information, a topic model meet the requirements based on probability. The Latent dirichlet allocation [3], one of probabilistic topic model, without extra prior information, can discompose the text matrix to the probability distribution of the corresponding word in different topics.

The topic is based on the word co-occurrence characteristic. If we utilize the topic word co-occurrence features to the analysis of word polarity, we can get a similar conclusion. If an article is negative tendency, which may contain some negative polarity words and these words are often found in common conversation topics, which can be identified as negative. A need to deal with the problem is how to infer the semantic meaning of the ambiguous words in a text context, such as one person's name in a negative topic, so this name with negative tendencies in the current context. It fitted ordinary commonsense observation. To the same effect, if this name occurs in a positive many times, it should be regarded as a positive case to refer. In short, we can conclude that the words within the same topic have more probability to get the same polarity. The word co-currency can prove this phenomenon directly.

Polarity Score by Random Walk

By utilizing the potential topic tendency known strong vocabulary as a priori vocabulary and the topic variable decomposition by LDA infer the polarity of informal vocabulary. Specifically, the words and topics are abstracted into node in the graph, because the type of words and topic nodes are different, the graph is composed of two kinds of nodes of the graph, which is called bipartite graph. We can get the matrix ϕ from LDA to generate our word - topics transition probability matrix and perform a random walk process in the graph. The random walker starts from the known words through the topic nodes get that word nodes without the priori polarity in the vocabulary. Note that, there are several different comparing with traditional random walk method. The first one, traditional random walk method such as PageRank [4] runs in directed connected graph, our method is based on undirected graph. The second, traditional random walk algorithm employ in homogeneous network, out random walker runs on bipartite network [5] which belongs to heterogeneous information network [6]. The third one, traditional random walk method starts from one given node and then do the random walk process, but our graph belongs to bipartite graph, we should give the direction of random walker in advance, that is to say, random walker starts from one type nodes(such as word) to another type nodes(such as topic) or the opposite. From another view, the topic generated by LDA is mutual independence, meanwhile, based on a bag of word rule, the word nodes is also mutual independence. So, we can assume the transmission probability among word nodes is zero, transmission probability among topic nodes as well.

Prior polarity vocabulary is easy to obtain, and only fraction of the positive and negative vocabulary be covered can get the polarity of other words through random walk. Certainly, the more priori vocabulary perfect is the higher infer accuracy gets. Random Walk score idea is that negative vocabulary by a random walk factor spread to some connected topics, so as to the topics into the negative tendency, and then starting from these negative topics will also spread negative tendency to other words in these topics. So it is also for positive evaluation. Finally, the neutral topic will be more close to the median of overall score, which is mixed score from the positive topics and negative topics.



(a) random-walk process (b) algorithm scoring process

Fig.1 Random walk among words and topics

For the sake of convenience,the algorithm formulation as following:

1) Define the word set in prior vocabulary $W^P = \{w_1^p, w_1^p, \dots w_P^p\} \in PriorSet$, where P is the size of prior vocabulary. Because prior vocabulary is consist of positive and negative parts, we can further note $w_i^p \in PosSet$ or $w_i^p \in NegSet$, and $PosSet \cup NegSet = PriorSet$.

2) The probability transmission matrix for random walk is noted as $M^{(Z,W)}$, where Z is the topic variable, W is word variable. $|Z|$ and $|W|$ is represented as the number of topic and the number of word. Because our method utilizes the LDA generating topic distribution, we note that $M^{(Z,W)}$ as the ϕ matrix from LDA method.

3) $Ra(W) \in [-1,1]^{1\times|W|}$ is the vector of word score, $Ra(Z) \in [-1,1]^{1\times|Z|}$ represents the vector of topic score.

The Algorithm Framework

The algorithm has 2 random walk process and the first as following:

Using prior vocabulary $W^P \in PriorSet$, construct the score vector of priori word $Ra(W)$. For the positive word, we have $w_\mathrm{i}^p \in PosSet, Ra(w_\mathrm{i}^p) = 1$, and negative word is $w_\mathrm{i}^p \in NegSet, Ra(w_\mathrm{i}^p) = -1$.

1) $Ra(\mathrm{Z}) = 0$. Initialize all topic with zero score.

2) Utilize matrix ϕ generated from LDA to build $M^{(Z,W^P)}$ matrix for random walk process. Due to $\phi_{w_i}^{(z)}$ represents the probability word $\mathrm{w_i}$ beloing topic z, we can seen this as the trasnsition probability starting from word node $\mathrm{w_i}$ to topic node z in bipartite graph .Further, removing the isolation node from the graph, we can get a matrix M only obtaining the word in vocabulary W^P.

3) Random walker starts from the word nodes W^P transmit to topic node Z, and then travel back to the starting nodes, until the score $Ra(Z)$ get convergence, the formula shown as following:

$$Ra(Z) = M^{(Z,W^P)} \times Ra(W)^T \tag{1}$$

$$Ra(W^P) = Ra(Z) \times M^{(Z,W^P)} \tag{2}$$

We run (1) (2) process iteratively, until $Ra(Z)$ and $Ra(W^P)$ get convergence. The $Ra(Z)$ results in topic tendency score.

The second random walk process is similar to the first, but the second cover all of words, $W \in Corpus$,and the first random walk result $Ra(Z)$ is needed to join the further calculation.

1) $Ra(W) = 0$, the initial value is assigned to zero, $Ra(Z)$ is inherited from previous random walk result.

2) Utilize matrix ϕ generated from LDA to build matrix$M^{(Z,W)}$. Note that W∈Corpus, is to say, W contain all words in corpus, not only containing the prior vocabulary word.

3) Random walker starts from the topic nodes Z to word nodes W, and then back to the starting nodes, until the score $Ra(W)$ get convergence, the formula shown as following:

$$Ra(W) = Ra(Z) \times M^{(Z,W)} \tag{3}$$

$$Ra(Z) = M^{(Z,W)} \times Ra(W)^T \tag{4}$$

When the $a(W)$ reveice convergency, we can get the word tendency score, $Ra(W) \in [-1,1]^{1\times|W|}$. The score close to -1 represent negative trend is significant. As the contrary, closing to 1 represents the word with obviously positive polarity.

Experiment

Accuracy Study

The data set is from Lillian Lee and Bo Pang. We select 60 articles randomly, 49870 words, and the prior vocabulary is labeled with word polarity. If the word in the vocabulary gets the wrong polarity through our algorithm, this word should be seen as a bad case. The accuracy measure is adopted as the traditional precision measure in data mining field.

$$Accuracy = \frac{(TP+TN)}{TP+FN+FP+TN} \tag{5}$$

TP, TN, FP, FN is the same as confusion matrix [7].

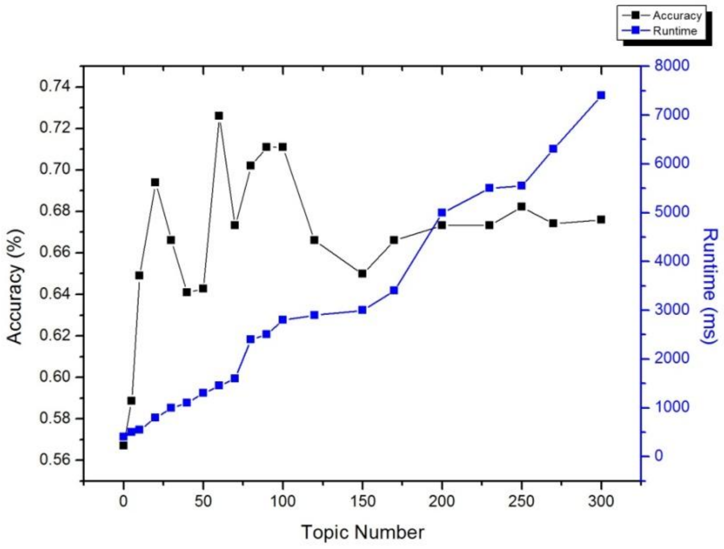


Fig. 2 the accuracy with different number of topics

From the figure 2, when topic number is 60, the algorithm gets the best accuracy 72%. When we have few topics, the word with opposite polarity has more probability to be assign in the same topic. This makes topic polarity confusion so that mislead our algorithm. When we have too much topics, many prior word with polarity will be isolation, it is to say, words hard to be assigned in the similar topic and result in a bad accuracy.

Now we do case study for our algorithm.

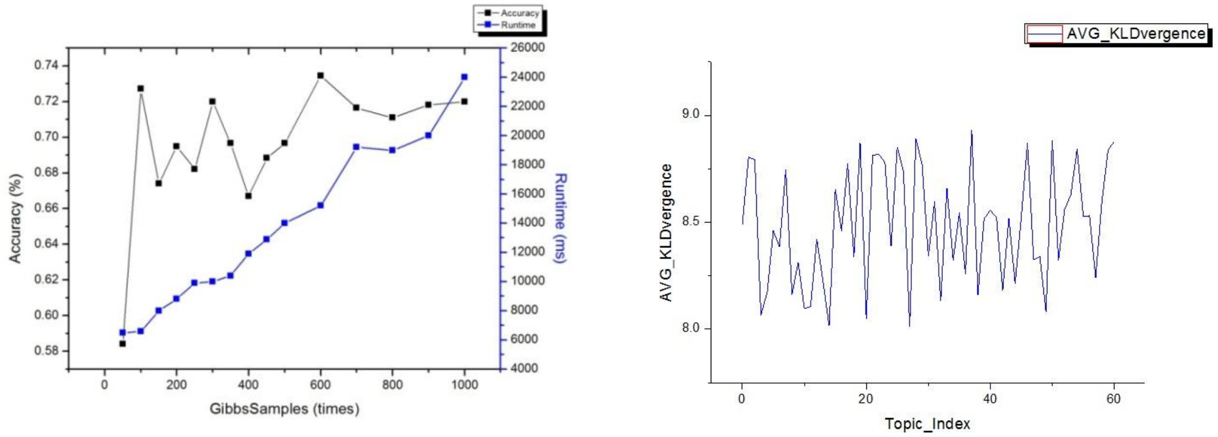
Table 1 Word_Score of Non prior vocabulary

Negative	Word_Score	Positive	Word_Score
foreign-guy-who-mispronounces-english	-0.0430	superman	0.0260
ditched	-0.0619	mature	0.0247
ditched	-0.0619	mature	0.0247
little-known	-0.0540	Superior	0.0298
booby-trapped	-0.0430	rushmore	0.0060
Blindly	-0.0283	hollywood	0.0032
simple	-0.0619	compelling	0.0260
seemingly	-0.0633	fun-filled	0.0247
squirm-inducing	-0.0619	actor-wise	0.0247
mushy	-0.0602	fully-animated	0.0261
poorly-integrate	-0.0619	star	0.0046
without	-0.0602	down-to-earth1	0.0258
copycat	-0.0602	spielberg	0.0012

Because the cyberword is informal, prior vocabulary only can cover limited words. For example "foreign-guy-who-mispronounces-english" have clear polarity but this word cannot be collected in prior vocabulary because of its informal usage. The meaning word "star" need to be decided according the context. Table 1 is shown the word score without prior labeled and the rank score is highly related the word real semantic meaning such as the famous director "spielberg" or "superman" with the positive score by our algorithm. Besides, the irony or sarcasm will induce a certain degree of error.

Parameter Study

LDA hyper parameter is according to Griffiths, Thomas's experiment α =50/K; β =0.01,topic number 50.



(a) the accuracy with different Gibbs sampling frequency (b) the average value of KL-Divergence

Fig.3 Parameter study

Now we discuss the impact of times of Gibbs sampling on our algorithm. From figure 3(a) shown, when Gibbs sampling time is low, topic distribution is very close to random allocation distribution and has a bad accuracy. With the sampling times growing, the accuracy will promote to 0.78, because the topic distribution close to convergence.

KL—Divergence is used to measure similarity among topic distributions.

$$KL(k_1,k_2)=\frac{1}{2}\sum_{w_t=1}^{W}\phi_{w_t}^{(k_1)}\,log_2\frac{\phi_{w_t}^{(k_1)}}{\phi_{w_t}^{(k_2)}}+\frac{1}{2}\sum_{w_t=1}^{W}\phi_{w_t}^{(k_2)}\,log_2\frac{\phi_{w_t}^{(k_2)}}{\phi_{w_t}^{(k_1)}} \qquad (6)$$

From figure 3(b), we can see the average KL-Divergence when topic is 60. We can conclude that bigger KL Divergence represents the low similarity among topics and topic is more dispersed. So the intersection between positive topic and negative is less and induce the better result.

Summary

This paper presents words polarity analysis method based on the topic model, which contribute to sentiment information filter from mass information on the internet, and decision making effectively.

We can employ LDA method to generate θ and ϕ matrix, and the words and topics are abstracted into nodes in bipartite graph. The ϕ matrix is used to generate our word - topics transition probability matrix. For running random walk process on the bipartite graph, we propose the polarity score by random walk method. The basic idea is that negative vocabulary by a random walk factor spread to some connected topics, so as to the topics into the negative tendency, and then starting from these negative topics will also spread negative tendency to other words in these topics. So it is also for positive evaluation. Finally, the neutral topic will be more close to the median of overall score, which is mixed score from the positive topics and negative topics. Experiments validate the practicability of our algorithm.

References

[1] The 31-th China Internet development situation statistic report. Beijing: China Internet Network Information Center. 2013.1.15

[2] Steyvers M, Griffiths T. Probabilistic topic models [J]. Handbook of latent semantic analysis, 2007, 427(7): 424-440.

[3] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. the Journal of machine Learning research, 2003, 3: 993-1022.

[4] S. Brin, L. Page, The anatomy of a large-scale hyper textual web search engine, Comput.

Netw. ISDN Syst, 30(1-7):1757 – 1771(1998).

[5] I. Dhillon, Co-clustering documents and words using bipartite spectral graph partitioning,

In KDD, 269-274(2001).

[6] J. Han, Mining heterogeneous information networks: the next frontier, Keynote speech.

In KDD, (2012).

[7] Confusion matrix on http://en.wikipedia.org/wiki/Confusion_matrix[EB/OL]