International Journal of Computer Applications Technology and Research
Volume 8–Issue 07, 291-295, 2019, ISSN:-2319–8656

An Enhanced Association Rule Mining Method for Processing Network Comments

Yang Di
Chengdu University of Information Technology
Institute of Communication Engineering
Chengdu, China

Wen Chengyu
Chengdu University of Information Technology
Institute of Communication Engineering
Chengdu, China,

Abstract: In order to facilitate the processing and understanding of text by computer, this paper uses a triple tuple containing entity and entity relationship to represent the text's fact in a more formal and concise way. The knowledge base (KBs), which contains a large number of facts, is used in various fields related to natural language processing. KBs usually integrates information from different places, such as manually edited encyclopedia, news articles, and social networks. In this paper, a natural language enhanced association rule mining method (NEARM) combined with KBs is used to deal with network comments. The fragments of fact are found from the pure text, and then the emotion is classified according to the fragments of fact found by the classifiers. Firstly, NEARM clusters the original data containing the pairs of related entities into clusters with different granularity from the data in KBs, and then excavates the rules in each cluster. These rules contain a large number of relational facts, which can reflect the relationship between plain text data, and can be effectively used in emotional classification of text. The experimental results show that the method is feasible. NEARM can deduce the relational triple facts and improve the accuracy of emotional classification.

Keywords: mining association rules; triple relational facts; natural language processing; knowledge base

1. INTRODUCTION

In the Internet age, people like to make all kinds of comments on the Internet, and there is a lot of valuable information hidden in these online comments. Through emotional analysis, we can grasp the trend of information, which can be used as information prediction, decision making, opinion mining, public opinion monitoring, product improvement, commodity recommendation and so on. However, the number of network comments is large and the content is complex, it is unrealistic to rely on manual monitoring and processing of text, and the poor computing power of the computer and the small storage capacity will also affect the data processing. Therefore, it is urgent to carry out in-depth study of text emotion analysis. Emotional analysis mainly distinguishes the emotional tendency of the text, and in the early days, Riloff and Shepherd conducted related research on the construction of semantic dictionary on the basis of text data [1]. McKeown finds the restrictive effect of conjunctions on the semantic expression of adjectives in large-scale text data sets, and then studies the emotional tendency of adjectives and conjunctions in the text [2]. Since then, more and more studies have begun to consider the relationship between characteristic words and emotional words. Narayanan et al. proposed a classification scheme based on clause, result sentence and whole sentence, which combines various features and related information, and has achieved good results [3]. Therefore, association rules can be used to mine the correlation between comment data, and some association features of comment objects can be represented in the form of rules, and then more relational facts can be excavated from the text.

2. RELATED WORK

KBs contains a large number of relational facts. Combining the knowledge base with plain text to mine association rules can not only obtain new relational facts, but also enrich the knowledge base. Similar text fragments are represented in a model way, that is to put similar pieces of text together. The traditional word bag (BoW) model ignores the grammar and word order of the text, and regards the text fragment as a collection of several words. The appearance of each word in the document is independent, and the position information of the word can not be saved. RLSW [7] proposes a distributed word bag (BoD) model for text, which uses Beta distribution to fit the position of each word in different text segments, and then converts the text set into a set of Beta distribution. However, BoD only models the words between the subject and the object, ignoring the words before and after the subject, which may lead to the loss of some important information. In this paper, the BoD* model is used to transform the natural language text fragments into probabilistic models, which contain more word information than the BoD model.

Amie [4] is a rule mining system in KBs by using some operators to extend the rules, and a new confidence measure is proposed. However, Amie's search strategy is valid only on small KBs. Therefore, scholars launched Amie+ [5], which is more efficient in mining rules. RDF2Rules [6] mining rules from the RDF KBs by mining frequent predicate loops(FPCs). These methods are based on KBs, and supplement the relational facts from the KBs. In fact, pure text contains a lot of information. It is helpful to mine more valuable information from pure text by complementing relational facts from knowledge base and mining rules. In the first step, the sentences are divided into different granularity clusters, and then they are integrated with the facts in KBs to mine the rules in each cluster. The second step is to capture the relationship between text and facts. An enhanced association rule mining method(NEARM) can generate rules that contain relational triple facts. This paper mainly includes the following contents:

- Find the fragment of information from the plain text.

- Choose the word sequence to represent the relationship.

- Obtain the relational facts related to the sequence of relational words through the KBs.

- Implement hierarchical clustering of relational word sequences.

- Text modeling and rule mining for different granularity clusters.

Finally, the relational facts in the rules are used in the emotional tendency analysis, and the effectiveness of the rules excavated by NEARM is proved by several experiments of different classification methods.

3. FRAMEWORK

According to the ratiocination that similar texts may contain the same relational facts, similar sentences can be grouped together to mine relational facts, and the unified form can be used to represent such clusters. The specific algorithm framework is shown in figure 1. The relational word sequences obtained from the KBs and data set are first clustered, then several clusters are modeled separately, and finally the mining rules of each cluster are excavated separately.

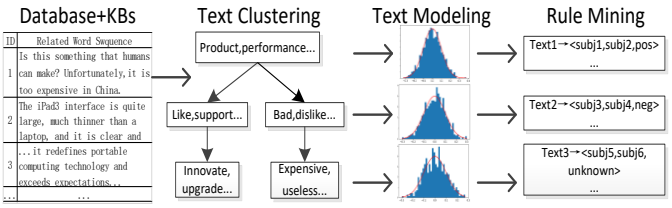

Figure. 1 The framework diagram of NEARM

Input a subset of sentences S in the frame, it contains a large number of related entity objects $e \in E$, output the facts of related entities e_1 and e_2, or the entity pair (e_1, e_2), which also called relational fact. As defined in formula (1):

$$f(e_1, e_2) = < e_1, e_2, rel > \in F \qquad (1)$$

Then, the excavated rule definition is shown in formula (2):

$$(ptn, e_1, e_2) \rightarrow f(e_1, e_2) \qquad (2)$$

ptn is a pattern of matching text in S.

For example, the text segment of comment, "XX phone, the hardware is really bad, the battery is bad". And match the pattern ptn, you can deduce what commentators think is the fact: " XX phone, use experience, bad ". If KBs contains some subjective facts about xx phones, it can also be used to excavate new facts. For example, there are "XX phone, main point, take a photo " in KBs that may lead to new facts " XX phone, game performance, bad >. Then, the definition of the rule is extended to formula (3):

$$(ptn, e_1, e_2) \wedge f(e_1, e_2) \rightarrow f'(e_1, e_2) \qquad (3)$$

The facts in the text matching mode and the facts in the knowledge base jointly derive new rules.

3.1 TEXT FRAGMENT CLUSTERING

In order to quickly find valuable fragments of information from plain text, similar text fragments in sentences need to be gathered together. It mainly includes the words around the entity object, and the word sequence composed of this word is also called relational word sequence. NEARM can aggregate all the relational word sequences in the cluster, and a new relational word sequence can convert the word sequence into vector representation through the fitness of the associated words. This idea was first used in RLSW [7].

3.1.1 Relational word sequence collection

Relational word sequence is a combination of words around an entity, which can be defined as three words before the first entity and three words after the second entity. For a sentence $S = \omega_1, \omega_2, \ldots, \omega_n$, the sequence of relational words is defined as formula (4):

$$m = \{\omega_i \mid i > idx(e_1) - 3 \,\&\, i <= idx(e_2) + 3\} \qquad (4)$$

$idx(e_1), idx(e_2)$ are the position of e_1 and e_2 in S.

In addition, predefined entity pairs can be collected from encyclopedia, social networking sites, and sentences containing such entity pairs can be collected from relevant articles.

3.1.2 Text similarity calculation

The similarity of the text is determined according to the value of the position sensitive word roaming distance, and the distance between the words is re-defined by combining the semantic distance and the position distance. The calculation method is as shown in the formula (5) and (6):

$$d(\omega_1, \omega_2) = \alpha * ed(\omega_1, \omega_2) + (1-\alpha) * |loc(\omega_1) - loc(\omega_2)| \qquad (5)$$

$$loc(\omega_i) = \frac{1}{n} * (i - 0.5) \qquad (6)$$

$loc(\omega_i)$ is the position value of the i[th] word, $ed(\omega_1, \omega_2)$ is the Euclidean distance between the vector representations of ω_1 and ω_2. But this only focuses on the words between the subject and the object, in order to make the word sequence contain more related words, redefine such as formula (7):

$$loc(\omega_i) = 2 \times \frac{i - idx(e_1)}{idx(e_2) - idx(e_1)} - 1 \qquad (7)$$

$idx(x)$ represents the position of x in the word sequence, the position values of e_1 and e_2 set to -1 and 1.

3.2 Text fragment modeling

An improved model, BoD *, which is used to represent a cluster, the words in the cluster and their position values are extracted together, and then fit the position value of each word by Gaussian distribution. The definition is the formula (8):

$$BoD^*(c) = \{(\mu_i, \sigma_i, p_i) \mid \omega_i \in W_c\} \qquad (8)$$

c is a cluster that contains a sequence of relational words, W_c represent the all words in c, $p_i = \frac{count(\omega_i)}{\sum_i count(\omega_i)}$ is the probability of ω_i, μ_i and σ_i which are mean and variance of sets $loc(\omega_i)$ for Gaussian Distribution modeling.

Then the association rule $(ptn, e_1, e_2) \rightarrow f(e_1, e_2)$ can be instantiated as $BoD^*(c) \rightarrow f(e_1, e_2)$. For any plain text fragment S, if it contains entity pair (e_1, e_2) and match the model $BoD^*(c)$, the rule means the existence of a fact $f(e_1, e_2)$. This fact may be a relationship between e_1 and e_2, or a relationship property value about them.

3.3 Rule mining

Since the relational word sequences of the same cluster are similar, the entities involved may share similar attribute values. To find the relationship between the relational word sequence and the attribute value, it must first build a transaction that contains the relational facts in each relational word sequence, as well as the corresponding entity pair. For example, the relational

word order is "xx mobile phone is the most popular game phone". In the corresponding KBs, all attribute values of "xx mobile phone" and "game" should be included in the transaction of this word sequence.

3.3.1 Rule mining process

Apriori algorithm is used in rule mining, which only uses frequent 1-length itemsets and frequent 2-length itemsets, corresponding to first-order rules and second-order rules respectively. For each fact in the set of frequent 1-length itemsets of clusters , the rule $BoD^*(c) \to f$ will be added to the first order rule set; For each fact in the set of frequent 2-length itemsets of clusters , two rules $BoD^*(c) \wedge f \to f'$ and $BoD(c) \wedge f' \to f$ will be added to the second order rule set.

Finally, calculate the value of support(sup) and confidence(conf) for each rule, and keep the rule above the threshold. The threshold is set to 0.8 times the maximum. The calculation formula are (9) and (10):

$$sup(r) = \frac{|\{t \mid f, f' \in t \,\&\, t \in T\}|}{|\{t \mid f \in t \,\&\, t \in T\}|} \qquad (9)$$

$$conf(r) = \frac{|\{t \mid f, f' \in t \,\&\, t \in T\}|}{|\{t \mid f' \otimes t \neq \varnothing \,\&\, t \in T\}|} \qquad (10)$$

r represents rule $(BoD^*(c) \wedge f \to f')$, t is a collection of facts. $f' \otimes t \neq \varnothing$ indicates that regardless of whether f' exists in t, its attribute value exists in t.

4. EXPERIMENT

If the rules excavated by NEARM can deduce the facts of ternary relationship, the effectiveness of NEARM can be proved. If the derived relational triple facts can be used to improve the accuracy of emotional analysis of the text, it can reflect the applicability of association rules in the emotional classification of the text.

4.1 Data preparation and setup

The KBs used in the lab consists of data that is manually collected on the network, plain text data is obtained by crawling the network interface of the relevant entity. There are 8121 plain text data, which are divided into training set and test set, and the ratio is 7:3. In NEARM, hierarchical clustering is used to mine multi-granularity rules. Because of the use of plane clustering, one cluster always accounts for a large part of the data, while other clusters usually contain only a few samples. Therefore, the experiment selects three levels,and in a deeper level, there are often not enough data items in the cluster.

4.2 Method comparison

4.2.1 Deduction of relational triple facts

For an unknown sequence $S = \omega_1, \omega_2, ..., \omega_n$, there are three ways to deduce the relational triple facts.

1) $BoD^*(c) \to f(e_1, e_2)$ (NEARM), according to formula (11):

$$\hat{c} = \underset{c}{argmax} \sum_{i=1}^{n} p_i * \int_b^e \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} dx \qquad (11)$$

Assign S to $BOD*(c)$, and p_i, μ_i, σ_i are probabilities which indicate the standard deviation of ω_i in $BOD*(c)$. The formulas for calculating parameters b and e are shown in formula (12), (13):

$$b = 0.5 * (loc(\omega_{i-1}) + loc(\omega_i)) \qquad (12)$$

$$e = 0.5 * (loc(\omega_i) + loc(\omega_{i+1})) \qquad (13)$$

If ω_i not in $BOD*(c)$, $p_i = 0$.

2) $BoD(c) \to f(e_1, e_2)$, the formulas for calculating $\hat{c}$ is shown in formula (14):

$$\hat{c} = \underset{c}{argmax} \sum_{i=1}^{n} p_i * \int_b^e Beta(\alpha_i, \beta_i) \qquad (14)$$

α_i and β_i are the parameters of the $Beta$ distribution model.

3) $BoW(c) \to f(e_1, e_2)$, the traditional word bag model is used to model the unknown relational word sequence. And the nearest cluster is directly selected as $\hat{c}$ for the unknown relational word sequence.

Finally, all the rules of the cluster $\hat{c}$ are applied to the word sequence S. For the first order rule, the rule is directly used to deduce the triple facts. For second-order rules, first look at the KBs and confirm whether the entity in S conforms to the corresponding facts, and then determines whether the derived triple facts are added to the result.

4.2.2 Emotional classification

The general mood classification model only inputs the pure text data to achieve the classification. However, the relational facts in the rules can be used to improve the classification effect by adding association rules to the model input. The experimental results show that the relational triple facts include two word entities and an emotional tag (pos, neg, unknown), pos indicates that the fact matches the positive text, neg indicates matching negative text, and unknown represents the emotion of uncertain relational phrases. According to the relational facts in each text, there are three main classification methods:

1) SVM, support vector machine is a two-classification model, and its basic model is the linear classifier with the largest interval defined in the feature space.

2) TextCNN, text convolutional neural network, the text is transformed into a word vector by using the Vocabulary Processor that from the Tensorflow, then transformed into a matrix graph, and the convolution neural network is used to convert text classification into image classification.

3) TextRNN, text recurrent neural network, it's commonly used in natural language processing, which can better express context information.

4.3 Performance comparison

4.3.1 Performance in the number of rules

From the experimental data in Table 1, it can be seen that with the deepening of clustering, more rules will appear in the detailed clusters. This indicates that the depth of the cluster has a certain effect on the number of rules, and with the deepening of clustering depth, the total number of rules will show a trend of growth, especially the proportion of the rules of the "unknown" label is greatly reduced.

Table 1. the number of rules in different clustering depths

Emotion	Rules		
	Level 1	Level 2	Level 3
pos	1127	3653	4106
neg	739	2790	3489
unknown	3340	2179	953

4.3.2 Performance of relational fact derivation

The number of relational facts derived from the experiment is shown in Table 2. Because there are many effective rules excavated by the second and third layer clusters, only the rules in the two layers are used to deduce the relational facts in the experiment.

Table 2. the number of relational facts derived from three methods at different levels

Emotion	BoD		BoD*		BoW	
	L 2	L3	L2	L3	L2	L3
pos	2557	2874	2739	3079	1826	2053
neg	1953	2442	2092	2616	895	1744
unknown	1525	667	1634	714	1089	476

4.3.3 Performance of emotional classification

In the experiment, the third layer cluster with the most effective rules is selected, and the relational facts derived by the BoD* method are used for the final emotional classification, and the experiments are compared with the classifiers without adding relational facts. The experimental results are shown in Table 3, and the accuracy, recall rate, F1 value (Pre,Rec,F1) are the performance indicators.

Table 3. Comparison of Classification Results

Method	Pre	Rec	F1
SVM	0.787	0.785	0.779
SVM+BoD*+L3	0.801	0.792	0.789
TextCNN	0.805	0.780	0.792
TextCNN+BoD*+L3	0.824	0.806	0.811
TextRNN	0.850	0.836	0.823
TextRNN+BoD*+L3	0.874	0.856	0.848

From the experimental results, it can be seen that the emotional classification effect of the three classifiers using the improved BoD* method is obviously improved in the third layer cluster. Among them, the CNN and RNN methods for text classification are the best.

5. CONCLUSION

In the study of emotional analysis, text is the basis for realizing emotional classification, and this paper combines the pure text and the data in the related knowledge base, and uses the NEARM framework based on association rules to mine the data. NEARM can well capture the relationship between word and emotion in the text and output the relational triple fact with emotional label. Then the emotional classification is carried out according to the relational facts and their labels existing in the text. The experimental results show that compared with input pure text to emotional classification, the accuracy of this method is obviously improved. At the same time, it also shows that the NEARM framework excavates valuable association rules. However, the data set used in the experiment is limited, only three levels of clustering are carried out, and the change of clustering level will also affect the performance of classification. It is hoped that mining the association rules in a larger database in the future.

6. ACKNOWLEDGMENTS

Thanks to the guidance of my research supervisor, and he points out the problems existing in the paper for me. There is also the concern of my classmates, so that I insist on not giving up. I would like to thank them for their help in helping me to complete this paper successfully.

7. REFERENCES

[1] Riloff E, Lehnert W. Automated dictionary construction for information extraction from text[P]. Artificial Intelligence for Applications. Proceedings Ninth Conference on 1993.

[2] Hatzivassiloglou V, McKeown KR. Predicting the semantic orientation of adjectives. In Proceedings of EACL-1997. Morristown: ACL, 1997: 174-181.

[3] Qingliang Miao, Qiudan Li, Ruwei Dai. AMAZING: A sentiment mining and retrieval system. Expert Systems with Applications: An International Journal, 2009, 36(3): 7192-7198.

[4] Galrraga L A, Teflioudi C, Hose K, et al. AMIE: association rule mining under incomplete evidence in ontological knowledge bases[C] //Proceedings of the 22nd international conference on World Wide Web. ACM, 2013: 413-422.

[5] Galrraga L, Teflioudi C, Hose K, et al. Fast rule mining in ontological knowledge bases with AMI[J]. The VLDB Journal, 2015, 24(6): 707-730.

[6] Wang Z, Li J.RDF2Rules: Learning Rules from RDF Knowledge Bases by Mining Frequent Predicate Cycles[DB/OL]. (2015-12-24) [2018-08-20].

[7] Yang X, Ren S, Li Y, et al. Relation Linking for Wikidata Using Bag of Distribution Representation[C] //National CCF Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2017: 652-661.

[8] Lipika Dey, Sk Mirajul Haque. Opinion mining from noisy text data[J]. International Journal on Document Analysis and Recognition (IJDAR), 2009, 12(3): 197-201.

[9] Li Q, Ji H. Incremental joint extraction of entity mentions and relations[C]//Proceedings of the 52ndAnnual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014, 1: 402-412.

[10] Ramanathan Narayanan, Bing Liu, Alok Choudhary. Sentiment Analysis of Conditional Sentences. In:

Proceedings of the 2009 Conference on EMNLP. Morristown, USA: ACL, 2009: 180-189.

[11] Ren X, Wu Z, He W, et al. CoType: Joint extraction of typed entities and relations with knowledge bases[C] //Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017: 1015-1024.

[12] Hang C, Mittal V, Datar M. Comparative experiments on sentiment classification for online product reviews [C] //Proceedings of the 21 National Conference on Artificial Intelligence. New York: Mountation View, 2006: 1265-1270.

[13] John D. Holt, Soon M Chung. Multipass Algorithms for Mining Association Rules in Text Databases[J]. Knowledge and Information Systems, 2001, 3(2): 433-437.

[14] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[J]. 2014: 253-258.

[15] Huosong Xia, Min Tao, Yi Wang. Sentiment text classification of customers reviews on the Web based on SVM[P]. Natural Computation (ICNC), 2010 Sixth International Conference on, 2010: 3636-3633.