Acta Technica *62* (2017), No. 5B, 1033–1038

Feature fusion analysis of online user reviews

Heyong Wang[2,4], Rong Cui[2], Wei Lv[3]

Abstract. Word of Mouth (WOM) attracts lots of attention in recent years, and it's necessary for us to discover valuable information from it, which is important for business to make product strategies and operation managements under the background of e-business and big data. In traditional analysis, only one type of features (features from structured data or unstructured data such as text data) is considered. Firstly, this paper analyzes the influences of fusion features (a combination of features from structured data and unstructured data). Features from structured data are used to build classifiers of Support Vector Machine (SVM). Secondly, fusion features are used to build new SVM classifiers. Finally, experimental results indicate that SVM classifiers using fusion features achieve better accuracies than those using features only from structured data, which indicates that fusion features have positive effects on online products classification.

Key words. online user reviews, topic discovery, feature fusion, support vector machine.

1. Introduction

In recent years, WOM becomes increasingly popular and it is an important step to discover useful information in user reviews to help users make product and management strategies under the background of e-business and big data. WOM has great influences on consumer behaviours [1]. With the rapidly development of electronic technology, WOM was taken into account by consumers gradually [2].

Most studies of WOM only focused on structured data [3]-[4] (e.g. ranking and location of users, scores of items given by users) or unstructured data [5]-[7] (e.g. user reviews). This paper tries to analyse the influences of feature fusion on online products classification.

[1]Acknowledgment - This research was supported by the National Sciences Foundation, Grant No. 71731006; the Fundamental Research Funds for Guangdong Province Software Sciences, Grant No. 2016A070705009.

[2]Workshop 1 - Department of E-Business, South China University of Technology, Guangzhou, 510006, China

[3]Workshop 2 - School of Information Technology, Beijing Normal University, Zhuhai Campus, 519085, China

[4]Corresponding author: Heyong Wang

Text mining is used for analysing unstructured text data. Feldman proposed knowledge discovery in textual databases to discover the potential information in text which made the foundation for future studies [8]. Ghanem focused on the model of text mining and put forward a distributed model for mixture data and text mining [9]. Based on semantic relations, Dekang suggested an unsupervised text finding reference rule which extended the algorithm of text mining [10]. Nowadays, researchers pay more attention to the application of text mining [11]- [12]. Text mining extends the way of knowledge discovery.

The rest of this paper is organized as follows. Section 2 introduces sample data used in experiments including data source and data processing. Experimental results are covered in Section 3. Finally, Section 4gives the conclusions.

2. Data Source and Processing

2.1. Data Source

The data used in this paper is from DataTang (http://www.datatang.com/data/15516). 1,000 samples are chosen from data randomly. The attribute, meaning and type of sample are shown as table 1.The attribute 'summary' is unstructured text data of user reviews. Short reviews as well as reviews containing only repeated words have been removed before experiment. The rest of the attributes such as 'score', 'usrtype', 'usrloc' and 'protype' are structured features.

Table 1. The attribute, meaning and type of sample

Attribute	**Meaning**	**Type**
prdid	Product ID	numeric
score	Score marked by users	numeric
summary	Given by users	text
usrtype	Users' type	text, measurable
usrloc	Users' location	text, measurable
protype	Products' type	text, measurable

2.2. Data Processing

2.2.1. Quantification of Structured Data Attributes 'usrtype', 'usrloc' and 'protype' must be quantified, and the results are shown as table 2.

Table 2. The results of measure

usrtype		**usrloc**		**protype**	
Register	0	Blank	0	Computers	0
Level 1	1	Westland	1	Household appliances	1
Level 2	2	Midland	2	Daily use articles	2
Level 3	3	Eastland	3	Electronics	3
Level 4	4				
Level 5	5				
Level 6	6				

2.2.2. Word Segmentation and Dimension Reducing R packages 'Rwordseg' and 'tm' are used for segmentation. After text pre-processing, a matrix with 691 topic words (nouns or adjectives) was built. Words with frequencies no larger than 2 are removed to avoid high dimensional sparsity. Part of DTM$[X_1, X_2, ..., X_{208}]_{1000\times208}$ is shown in table 3.208 topic words are selected by IG. Part of the topic words are shown in table 4.

Table 3Document Term Matrix

keyboard	**white**	**office**	**packaging**	**shelf life**
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	1	0

Fig. 1

Table4 Scores by using IG

Key words	**IG score**
good	0.2664
not bad	0.2634
price	0.1609
quality	0.1490
cheap	0.1125
thing	0.1018

Fig. 2

3. Feature Fusion

3.1. Classification on Structured Data

Clementine 12.0 is used in this section. The inputs are 'usrtype', 'usrloc' and 'protype'. The output is 'score'.

As shown in table 5, there are four kernel functions for SVM.RBF kernel shows high accuracies on structured data. RBF kernel has an adjustable parameter$y$, which value is generally between $3/k$ and $6/k$. Where k is the number of input attributes. In experiment, $k$is 3 because attributes 'subtype', 'usrloc' and 'protype' are used. Generally, the classification accuracies improve as the value of $y$increases but high values of y may lead to over fitting of SVM classifiers. Experimental results demonstrate that the values of $y$have little influences on classification accuracy, so 1.0 is chosen for y. The classification accuracies chart is shown in figure 3. (The horizontal axis shows proportions of training set)

Table 5 Kernel functions and parameters

Kernel function	Formula	Parameters
Liner	——	——
RBF	$\exp(-\gamma\times\lvert\mu-v\rvert^2)$	γ
Sigmoid	$\tanh(\gamma\times\mu\times v+coef0)$	$\gamma, coef0$
Polynomial	$(\gamma\times\mu\times v+coef0)^D$	$\gamma, coef0, D$

Fig. 3

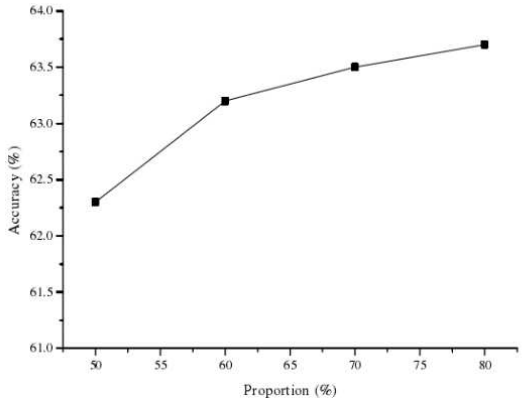

Fig. 4. Classification accuracies of structured data based on RBF kernel

Figure1shows that the classification accuracies based on RBF have improved as proportions of training sets increase and the accuracies are higher than 62%, which indicates that the structured data have influence on 'score' and supervised learning is effective.

3.2. Classification on Fusion Data

The classification accuracies are shown in figure 2(The horizontal axis shows the number of topic words, and different lines represents classifiers using different kernel functions).

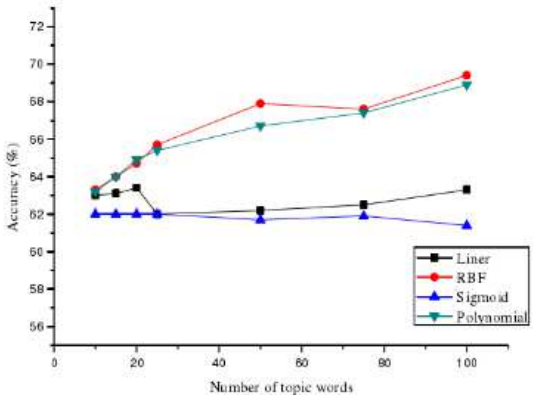

Fig. 5. Classification accuracies with different proportions using the same kernel function

Experimental results indicate that classifiers using RBF and polynomial kernel achieved better classification accuracies. The number of topic words have significant positive effects on classifiers using RBF and polynomial kernel but little or no positive effects on classifiers using linear and sigmoid kernel.

Above all, for classifiers using RBF or polynomial kernel, the classification accuracies on fusion data are higher than those on structured data. The results show that unstructured topic words selected by IG can be considered during classification and fusion data can improve the classification accuracy. Topic words such as 'price' and 'quality' selected by IG are able to improve the classification accuracy.

4. Conclusion

It is one of the most important steps to discover valuable information and knowledge for product strategies and operation managements under the background of e-business and big data. This paper focused on online products classification and analysed the influences of fusion data on classification accuracies. Experimental results indicate that topic words selected by IG are able to improve classification accuracy. In summary, feature fusion can improve the classification accuracies and can be used to discover the knowledge about online user reviews.

But there are still some limits in this paper. For example, emotion is not separately considered in this paper when doing text mining, which is an important feature of reviews, and how to use the results for products' recommendation is not discussed. For future studies, emotion scores and products' recommendation will be

considered on the study of feature fusion analysis.

References

[1] J. Goldenberg, B. Libai, E. Muller: *Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth*. Marketing Letters *12* (2001), No. 3, 211–223.

[2] C. M. Cheung, D. R. Thadani: *The impact of electronic word-of-mouth communication: A literature analysis and integrative model*. Decision Support Systems *54* (2012), 461–470.

[3] H. Y. Wang, X. H. Wu: *Research on correlation analysis of customer online reviews on association rules*. Logistics Engineering and Management *2* (2014), 81–84.

[4] H. Y. Wang, X. H. Wu: *The application of customer online reviews using on clustering analysis*. Logistics Engineering and Management *4* (2014), 86–89.

[5] M. David, Y. N. Andrew: *Michael I. Jordan. Latent Dirichlet allocation*. Journal of Machine Learning Research *3* (2003), 993–1022.

[6] A. Pons-Porrata, R. Berlanga: *Topic discovery based on text mining techniques*. Information Processing& Management *43* (2007) 752–768.

[7] X. T. Han, W. Li, Q. W. Shen: *Extracting and clustering product features from user reviews*. Computer system and applications *5* (2013), 188–192.

[8] R. Feldman, I. Dagan: *Knowledge discovery in textual data-bases (KDT). Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD- 95)*. Montreal *20* (1995), No. 21, 112–210.

[9] M. Ghanem, A. Chortaras, Y. Guo, A. Rowe: *A grid infrastructure for mixed bioinformatics data and text mining*. Computer Systems and Applications *34* (2005), No. 1, 116–130.

[10] L. Dekang, P. Pantel: *DIRT- Discovery of Inference Rules from Text*. Journal of Natural Language Engineering *12*, (2001), 22–31.

Received November 16, 2017