A Comparison Study of Clustering Models for Online Review Sentiment Analysis

Baojun Ma[1,3], Hua Yuan[2,⋆], and Qiang Wei[1]

[1] School of Economics and Management, Tsinghua University, 100084 Beijing, China
{mabj.03,weiq}@sem.tsinghua.edu.cn
[2] School of Management and Economics, University of Electronic Science and Technology of China, 610054 Chengdu, China
yuanhua@uestc.edu.cn
[3] School of Economics and Management, Beijing University of Posts and Telecommunications, 100876 Beijing, China

Abstract. In this work, we conduct a comparison study of the online review sentiment clustering problem from a combined perspective of data preprocessing, VSM modeling and clustering algorithm. To that end, we first introduce some methods for data preprocessing. Then, we explore the impacts of the term weighting models for review representation. Finally, we present detailed experiment results of some review clustering techniques. The conclusions would be valuable for both the study and usage of clustering methods in online review sentiment analysis.

Keywords: Online review, sentiment analysis, term weighting model, clustering algorithm.

1 Introduction

In recent year, online reviews have become an important resource for people to provide product information and recommendations from the customer perspective [1]. The customer reviews are so useful that almost all the e-commerce related organizations, such as Amazon and Google, have accumulated a huge amount of reviews data, and the analysis of these data to extract latent public opinion and sentiment is a challenging task. To automate the sentiment analysis, different approaches [2–4] in literature have been applied to predict the sentiments of words, expressions or documents.

Clustering, which tries to find the natural clusters in the data by calculating the distance from the centers of the clusters, is especially useful for organizing documents to improve information retrieval [5]. However, the analysis results generated by clustering method would be affect heavily by some intermediate steps, such as preprocessing strategy, term weighting model and clustering algorithm. It is valuable for online review sentiment analysis to make clear that: which kind of clustering algorithm is more effective for sentiment analysis? and

⋆ Corresponding author.

J. Wang et al. (Eds.): WAIM 2013, LNCS 7923, pp. 332–337, 2013.

which types of term weighting models are more suitable for review representation? In this work, we conduct comparison study of the online review sentiment clustering problem from a combined perspective of data preprocessing, VSM (Vector Space Model) modeling [6] and clustering algorithm.

2 Related Work

The task of sentiment analysis is to judge whether a review expresses a positive, neutral or negative opinion and a lot of efforts have been devoted into this area in literature [7]. The typical work is method that presented by Pang and Lee of sentiment classification on the document level [3].

Also, few efforts have been devoted to the study of sentiment analysis with clustering. Agarwal et al. Li & Liu [8] proposed a method to choose solid polarity reviews to generate a positive seed set and a negative seed set to solve this problem. Zhai et al. [9] studied the problem of product feature clustering for opinion mining applications, in which they casted the problem as a semi-supervised learning task. An approach of semi-automatic public sentiment analysis for opinion and district is proposed in [10], which includes automatic data acquiring, sentiment modeling, opinion clustering, and district clustering, and manual threshold setting and result analysis. [11] investigated the effect of feature weighting on document clustering, including a novel investigation of Okapi BM25 feature weighting. Especially, [12] presented the results of some common document clustering techniques. However, there are not impressive research on comparing the different clustering performance under various environments.

3 The Methodology

The research framework consists of four parts: data preprocessing, VSM modeling, clustering and results evaluation.

3.1 Data Preprocessing

First, a part-of-speech tagger developed by Stanford University is used to tag the reviews [13]. Under the impact of the work of the first step, the words which are not tagged as being either an adjective or adverb would be eliminated. Then, the words stemming is done by applying Porter's algorithm [14]. Finally, we utilize the stop-word list to remove stop words which were built by Gerard Salton and Chris Buckley for the experimental SMART information retrieval system.

3.2 Term Weighting Models

Let $D = \{d_1, d_2, ..., d_N\}$ be a set of documents/reviews and $T = \{t_1, t_2, ..., t_M\}$ be the complete term set of D, in which, $d_i = [w_{i1}, w_{i2}, ..., w_{im}]$ where w_{ij} is the weight of term t_j to document d_i. Table 1 illustrates six term weighting models selected to be utilized in this study.

Table 1. Term weighting models used in our experiments

Weighting	Equation	Reference
Binary	$w_{ij} = \begin{cases} 1, \text{ if } tf_{ij} > 0; \\ 0, \text{ otherwise.} \end{cases}$	B. Ricardo et. al,(1999)
TF	$w_{ij} = tf_{ij}$	Salton et al.(1981; 1983)
TF_IDF	$w_{ij} = tf_{ij} \times \log \frac{N}{df_j}$	Jones(1972)
BM25	$w_{ij} = \frac{tf_{ij}(k_1+1)}{tf_{ij}+k_1((1-b)+b\cdot\frac{dl_i}{avgdl})} \times \log \frac{N}{df_j}$	S.E. Robertson et. al,(1994)
DPH_DFR	$w_{ij} = \begin{cases} 0, \quad \text{if } tf_{ij} = 0; \\ \frac{(1-\frac{tf_{ij}}{dl_i})^2}{tf_{ij}+1} \cdot \{tf_{ij} \cdot \log(\frac{tf_{ij} \cdot avgdl}{dl_i} \cdot \frac{N}{tf_j}) \\ +0.5\log[2\pi \cdot tf_{ij} \cdot (1-\frac{tf_{ij}}{dl_i})]\}, \quad \text{otherwise.} \end{cases}$	G. Amati, et. al,(2007)
H_LM	$w_{ij} = \log(1 + \frac{\lambda \cdot tf_{ij} \cdot sigmatf}{(1-\lambda) \cdot df_j \cdot dl_i})$	Hiemstra (2001)

3.3 Clustering Algorithms and Sentiment Recognition

We have selected a collection of 18 clustering algorithms to conduct a relative complete comparison. Table 2 lists all these algorithms in detail.

Table 2. The clustering algorithms used in our experiments

Algorithm	Short	Reference
K-means	Kmeans (Direct-I2)	Lloyd (1982)
Repeated bisecting k-means	RB-Kmeans (RBR-I2)	Steinbach, et al. (2000)
Partition around medoids	PAM	Kaufman & Rousseeuw (1990)
Clustering Large Applications based upon RANdomized Search	CLARANS	R. T. Ng & Han (2002)
Unnormalized spectral	Spect-Un	Luxburg (2007)
Random walk spectral	Spect-RW	Shi & Malik (2000)
Symmetric spectral	Spect-Sy	A. Y. Ng et al. (2001)
Principle component analysis + K-means	PCA-Kmeans	Pearson (1901)
Non-negative matrix factorization	NC-NMF	Xu et al. (2003)
Unweighted pair group method	UPGMA (Agglo-upgma)	Kaufman & Rousseeuw (1990)
Single linkage	Slink (Agglo-Slink)	Jain et al. (1999)
Complete linkage	Clink (Agglo-Clink)	Jain et al. (1999)
Repeated bisecting H1 with global optimization	RBR-H1	Zhao and Karypis (2004)
Direct H1	Direct-H1	Zhao and Karypis (2004)
Agglomerative I2	Agglo-I2	Zhao et al. (2005)
Agglomerative H1	Agglo-H1	Zhao et al. (2005)
Agglomerative cluster-weighted single-link	Agglo-WSlink	Zhao et al. (2005)
Agglomerative cluster-weighted complete-link	Agglo-WClink	Zhao et al. (2005)

3.4 Evaluation

To evaluate the clustering effectiveness, a confusion matrix could be constructed as shown in Table 3. Cluster 1 is the positive cluster if $(a + d) \geq (b + c)$. Otherwise, Cluster 2 is the positive cluster. Consequently, $Accuracy = \frac{a+d}{a+b+c+d}$, if $(a + d) \geq (b + c)$; else $Accuracy = \frac{b+c}{a+b+c+d}$.

Table 3. The confusion matrix

	Cluster 1	Cluster 2
Actual # of positive reviews	a	b
Actual # of negative reviews	c	d

4 Experimental Results

In this work, we introduce eight datasets in the following experiments (Table 4).

Table 4. The benchmark datasets

ID	Data set	URL
D1	Polarity Dataset V2.0	www.cs.cornell.edu/People/pabo/movie-review-data
D2	Sentence Polarity Dataset v1.0	www.cs.cornell.edu/People/pabo/movie-review-data
D3	Amazon reviews (Books)	
D4	Amazon reviews (DVDs)	www.cs.jhu.edu/~mdredze/datasets/sentiment/
D5	Amazon reviews (Electronics)	
D6	Amazon reviews (Kitchen)	
D7	TripAdvisor-15763	http://patty.isti.cnr.it/~baccianella/reviewdata/corpus/
D8	Amazon-83713	http://patty.isti.cnr.it/~baccianella/reviewdata/corpus/

4.1 Results on Term Weighting Models

We compared the clustering results on six weighting models (the results are shown in Table 5). We see that, on average, BM25, DPH_DFR and H_LM weighting models are somewhat better than TF_IDF, and notably better than Binary as well as TF. However, TF_IDF is actually sometimes better than H_LM and DPH_DFR (D7), and performs similar results to BM25 on D2 and D7.

Table 5. The percentage difference in average accuracy

Dataset	Binary	TF	TF_IDF	BM25	DPH_DFR	H_LM	Max
D1	-3.12%	-5.16%	-8.17%	-0.95%	0	-5.91%	0.585
D2	-1.65%	-1.39%	-0.63%	-0.45%	0	-0.60%	0.521
D3	-0.42%	-2.09%	-0.25%	0	-0.19%	-0.79%	0.529
D4	-1.69%	-1.48%	-0.82%	0	-0.62%	-0.98%	0.528
D5	-1.83%	-5.15%	-3.01%	-2.92%	0	-1.80%	0.552
D6	-3.69%	-4.57%	-2.15%	-0.51%	-0.18%	0.00%	0.548
D7	-5.10%	-6.05%	-0.05%	0	-1.96%	-1.71%	0.649
D8	-2.33%	-4.08%	-2.72%	0	-0.79%	-1.12%	0.576
Overall	-2.07%	-3.36%	-1.76%	-0.10%	0.00%	-1.16%	0.558

4.2 Results on Clustering Algorithms

In this section, we try to find out which kinds of clustering algorithms are more effective for clustering-based sentiment analysis. The results are illustrated in Fig. 1, in which the lines of Kmeans and RB-Kmeans are almost coinciding with each other, so as are those of RBR-H1 and Direct-H1.

(a) The average performance. (b) The highest performance.

Fig. 1. The average (a) and highest (b) performance of each clustering algorithm

To see more clearly the differences between clustering performances, we make a comparison of the 18 clustering algorithms by dataset and average over all clustering algorithms for that dataset. Fig. 1 and Table 6 indicate that, on average, four algorithms of Kmeans, RB-Kmeans, RBR-H1 and Direct-H1 show clear advantage over the other 14 methods on clustering accuracy. However, CLARANS algorithm is actually somewhat better than the above four best methods and substantially better than other 13 algorithms for D2 and D3.

Table 6. The percentage difference in average accuracy

Algorithm	D1	D2	D3	D4	D5	D6	D7	D8	Overall
Kmeans	0	-3.9%	0	0	0	0	0	0	0.0%
RB-Kmeans	0	-3.7%	-0.2%	0	-0.2%	0	0	0	0.0%
PAM	-22.4%	-8.5%	-7.7%	-7.0%	-14.7%	-9.6%	-25.1%	-17.1%	-14.5%
CLARANS	-17.0%	0	0	-2.5%	-10.0%	-7.2%	-32.9%	-13.7%	-11.5%
Spect-Un	-22.1%	-10.8%	-12.9%	-12.0%	-20.1%	-18.1%	-41.9%	-26.3%	-21.5%
Spect-RW	-16.7%	-11.0%	-12.9%	-12.0%	-20.1%	-18.1%	-41.9%	-26.3%	-20.9%
Spect-Sy	-15.9%	-11.0%	-12.9%	-12.0%	-20.1%	-18.1%	-41.6%	-26.3%	-20.7%
PCA-Kmeans	-20.6%	-11.0%	-12.9%	-11.8%	-19.3%	-18.1%	-41.4%	-25.6%	-21.1%
NC-NMF	-15.3%	-7.5%	-7.8%	-4.7%	-9.6%	-10.8%	-12.3%	-10.4%	-9.7%
UPGMA	-25.8%	-11.0%	-12.9%	-12.0%	-19.9%	-18.1%	-42.0%	-26.3%	-22.0%
Slink	-25.8%	-10.8%	-12.9%	-12.0%	-20.1%	-18.1%	-42.0%	-26.3%	-22.0%
Clink	-23.7%	-9.9%	-11.8%	-10.0%	-18.5%	-16.7%	-39.4%	-24.7%	-20.3%
RBR-H1	-5.2%	-4.6%	-2.1%	-3.9%	-3.8%	-1.6%	-0.8%	0	-2.2%
Direct-H1	-5.3%	-4.6%	-2.1%	-3.9%	-4.1%	-1.8%	-0.8%	0	-2.3%
Agglo-I2	-17.3%	-10.8%	-11.3%	-10.2%	-17.7%	-15.2%	-18.2%	-15.7%	-14.5%
Agglo-H1	-18.1%	-8.2%	-8.2%	-8.6%	-16.7%	-12.6%	-16.5%	-13.1%	-12.7%
Agglo-WSlink	-23.7%	-9.9%	-11.8%	-10.0%	-18.5%	-16.7%	-39.4%	-24.7%	-20.3%
Agglo-WClink	-25.8%	-10.8%	-12.9%	-12.0%	-20.1%	-18.1%	-42.0%	-26.3%	-22.0%
Max	0.675	0.563	0.575	0.569	0.627	0.612	0.862	0.68	0.643

5 Conclusion

In this work, we find averagely the following experimental conclusions for online review sentiment clustering:

- BM25, DPH_DFR and H_LM weighting models are somewhat better than TF_IDF, and notably better than Binary as well as TF.
- Kmeans, RB-Kmeans, RBR-H1 and Direct-H1 show clear advantage over the other 14 methods on clustering accuracy. However, CLARANS algorithm is actually somewhat better than the above four best methods and substantially better than other 13 algorithms for D2 and D3.

The experiment methods and conclusions would be valuable for both the study and usage of clustering methods in online review sentiment analysis.

Acknowledgments. The work was partly supported by the National Natural Science Foundation of China (71271044/U1233118/71072015/71110107027), and the Tsinghua University Initiative Scientific Research Program (20101081741).

References

1. Zhu, F., (Michael) Zhang, X.: Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. Journal of Marketing 74(2), 133–148 (2010)
2. Yi, J., Nasukawa, T., Niblack, W., Bunescu, R.: Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: Proceedings of the ICDM 2003, Florida, USA, pp. 427–434 (2003)
3. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
4. Prabowo, R., Thelwall, M.: Sentiment analysis: A combined approach. Journal of Informetrics 3, 143–157 (2009)
5. Anick, P., Vaithyanathan, S.: Exploiting Clustering and Phrases for Context-Based Information Retrieval. In: Proceedings of the 20th ACM SIGIR, pp. 314–323 (1997)
6. Salton, G., Wong, A., et al.: A vector space model for automatic indexing. Communication of the ACM 18(11), 613–620 (1975)
7. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan and Claypool Publishers (May 2012)
8. Li, G., Liu, F.: Application of a clustering method on sentiment analysis. Journal of Information Science 38(2), 127–139 (2012)
9. Zhai, Z., Liu, B., Xu, H., Jia, P.: Clustering product features for opinion mining. In: Proceedings of the WSDM 2011, New York, USA, pp. 347–354 (2011)
10. Wang, D., Feng, S., Yan, C., Yu, G.: An approach of semi-automatic public sentiment analysis for opinion and district. In: Wang, L., Jiang, J., Lu, J., Hong, L., Liu, B. (eds.) WAIM 2011. LNCS, vol. 7142, pp. 210–222. Springer, Heidelberg (2012)
11. Whissell, J.S., Clarke, C.L.: Improving document clustering using Okapi BM25 feature weighting. Information Retrieval 14(5), 466–487 (2011)
12. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining (2000)
13. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the EMNLP/VLC 2000, Hong Kong, pp. 63–70 (2000)
14. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)