Clustering of Text Documents using Features from Latent Semantic Analysis

SHU-CIH TSENG[†1] YU-CHING LU[†2]
GOUTAM CHAKRABORTY[†3] LONG-SHENG CHEN[†4]

Abstract: Text documents could be classified using words as features. As the number of words in the vocabulary is large, the dimension of the document space will be very high. In that case, the feature vector for a document is too long, and clustering and classification algorithms fail. There are various ways to reduce this dimension by topic modeling. In this work, we used Latent Semantic Analysis (LSA), which is actuated by Singular Value Decomposition (SVD). After SVD, we have a compact representation of the documents, which are clustered. The ground truth is verified manually.
In this work, we used tourists' comments as documents. Tourists visit to a place is influenced by comments from previous visitors. In this work, we first cluster the comments into two, and investigate the factors behind these two classes. It is verified, that the documents are automatically separated into groups of positive comments and negative comments.
Our final goal is to extract factors that lead to positive comments and those leading to negative comments. That would help promoting tourist business by focusing on the factors that really matters for the customers.

Keywords: Text mining, Semantic orientation, Tourism, SVD, NNMF,

1. Introduction

In recent years, tourisms have big growth every year [1]. Due to rapid development of social media, many websites about tourism mushroomed in short period of time. A few examples are TripAdvisor, Hotels.com, Trivago or booking.com etc. Those sites, their comment section, greatly influence tourism [2] [3]. Customers usually review other customers' comments, and choose the hotel while making a reservation [4]. After their trip, some customers will write a comment for the hotel they lived [5], especially if the stay were enjoyable or if they hate the service. Those comments are important to influencing hotel's business, because those are reviewed all the time by other customers, and influence customers' selection of hotel, or in general visiting a particular destination.

In order to find the important factors which are influencing customer selection, many study survey customers' revisit intention, behavior, satisfaction or word-of-mouth by getting questionnaire filled out by customers [6] [7] [8] [9] [10]. However, it is both time consuming and costly. In addition, one is restricted to express oneself within the boundary of the questions, and can not express freely their feelings. The questionnaire already decides the items of importance [11]. On the other hand, text comments have successful survey in other studies, and at a reduced cost than questionnaire [12] [13] [14]. For customer reviews on the web-site website, text comments are better compared to other reviews, such as star ranking [15].

In this study, we will use text comments to investigate comment classes and finally influencing factors. We will first verify the ground truth of clustering, whether the clustering cold successfully divide the comments into two classes or positive and negative comments. Finally, we will extract the key factors, from individual comment for a specific hotel, the help promoting the business.

†1 Faculty of Software and Information Science, Iwate Prefectural University, Iwate, Japan
†2 Faculty of Software and Information Science, Iwate Prefectural University, Iwate, Japan
†3 Faculty of Software and Information Science, Iwate Prefectural University, Iwate, Japan
†4 Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan

2. Motivation of the work

2.1 Semantic orientation by text mining

Text mining is editing, organizing, and analyzing large number of documents. Its another purpose is to identify potentially useful and key information from a document [16]. When customer makes a negative comment, it contains the potential action of positive discussion, expectations, suggestions, content improvement and warnings. It may even prevent other potential customers to visit the destination [17]. However, how to find a negative comment is not easy. As the number of comments grow, it is also difficult to manually go through all comments to find out important factors.

For the semantic orientation, it is a technique of text mining. It focuses on determining the polarity of a text, sentence, or feature (positive or negative). The main purpose is to compile or customize a single word into a positive or negative category [18]. This study will use text mining with semantic orientation to classify an article, here comment text. based on the amount of positive or negative words, try to verity the intention. We want to do it automatically, without looking into words, and labelling words into positive or negative category.

2.2 Dimension reduction

The original comment corpus consists of many words. When arranged as a matrix, with rows as documents and columns as terms, the number of columns, i.e., the dimension of a document vector will be too large. Clustering this high dimension vectors will not be successful. Therefore, we need use some method to do dimension reduction, and to find potential semantic information, rather than individual word. In the study, we adopt two methods, Singular Value Decomposition(SVD) and

Non-Negative Matrix Factorization(NNMF).

2.2.1 Singular Value Decomposition

The purpose of SVD is to factorize a matrix into 3 matrices, the first one U is column orthonormal, the last one V^T is row orthonormal, and the middle one is diagonal matrix. It frequently deals with image processing, to reduce unnecessary information or eliminate noise to clean data [19].

2.2.2 Non-Negative Matrix Factorization

NNMF is a very popular tool in fields, such as document clustering, data mining, machine learning and image analysis [20]. In NNMF, all elements of the original matrix are required to be non-negative, then the matrix can be decomposed into the product of two smaller non-negative matrices. As our document matrix elements are all positive, we will also use NNMF to decompose the data, and compare the performance with SVD.

2.3 K-means

K-means is a popular clustering algorithm widely used [21] [22] [23]. First, decided to divide the k-group and randomly select k-points to do the cluster center. Next, classify each point to the nearest cluster center. Recalculate the cluster centers for each group by average value [24]. Finally, according to the clustering results, we can further identify to any group potential relationship. This will continue until convergence. We will adopt K-means for clustering the data. We hypothesize that the documents will automatically be partitioned into two groups, positive and negative, without any manual intervention. After clustering, we will check the result, to verify whether our hypothesis was correct or not.

3. Experiment Procedure

3.1 Data collection

We collect the data from tourism website, and all the comments are texts in English. Such as Fig. 1 for an example comment.

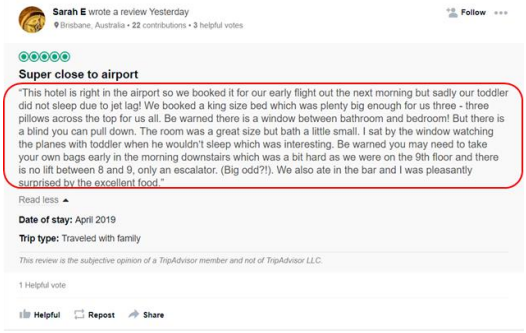

Fig. 1 Sample of data collected

3.2 Cleaning of the Data

In this step, we have 4 sub-steps. First, we do the uni-gram that let all comments remove any symbols and words become terms.

Then remove the stop words, any word without specific meaning, like articles, such as “a”, “an”, “the”, “and”, ”that” etc., are removed.

Next, we do stemming and lemmatization. All of words will be changed to their formal dictionary form. For example, “goes”, “gone” and “went” become “go”. “studies” and “studied” become “study”.

Finally, we extract nouns and adjectives. Nouns are the factors, and adjective qualifying the nouns will give it positive or negative orientation, like “clean dishes” and “dirty dishes”.

3.3 Normalization

In this step, we build term-document matrix (TDM). As shown in Fig. 2, column means document, row means terms. m is the number of documents, and n is the number of terms. After TDM is formed, we do normalization. Sum of elements for each row of document is normalized to 1, so that there is no effect from the length of the document.

$$TDM = \begin{array}{c} \\ D_1 \\ D_2 \\ D_3 \\ \vdots \\ \\ D_m \end{array} \begin{array}{c} \begin{array}{cccccccc} T_1 & T_2 & T_3 & \cdots & \cdots & \cdots & \cdots & T_n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & \cdots & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & \cdots & \cdots & \cdots & \vdots \\ a_{31} & a_{32} & a_{33} & \cdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & & & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & & & \vdots \\ \vdots & \vdots & \vdots & & & \ddots & & \vdots \\ a_{m1} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & a_{mn} \end{bmatrix}_{m \times n} \end{array}$$

Fig. 2 TDM established

3.4 Dimension Reduction

We tried both SVD and NNMF for matrix factorization and evaluate them on the basis of computational complexity and quality of result by checking square error after recomposition of the matrix from its factors. First, we use SVD to decompose the TDM. When decomposed, it will get 3 matrices (Fig. 3). The number of columns of U, shown as k, is the number of latent semantic formed by linear composition of different words with close meanings.

$$\left[\ A\ \right]_{m \times n} = \left[\ U\ \right]_{m \times k} [\Sigma]_{k \times k} \left[\ V^T\ \right]_{k \times n}$$

Fig. 3 SVD decomposition

V^T matrix shows the relation between k semantic compositions and n terms. Σ matrix is the strength of relations, the diagonal elements are actually eigenvalues of TDM. Rows of U\Sigma are documents vectors, expressed in terms of the semantic composites. Thus, we have document vectors whose dimension is now reduced from n to k. We can further reduce k as follows. According to the Eq. (1), we get the k-value, which is much less than original, but retaining most of the information (90%) in TDM. As the diagonal elements of \Sigma decreases, from top left to bottom right element, \Sigma will be reduced to a much smaller matrix.

$$\frac{\alpha_{11} + \alpha_{22} + \alpha_{33} + \cdots + \alpha_{ii}}{\sum_{i=1}^{N} \alpha_{ii}} \cong 0.9 \qquad (1)$$

Then, we use NNMF to decompose the raw data (Fig. 4). W matrix means number of k-topic for document. H matrix means number of k-topic for terms. However, it is difficult to find the best k-value from the NNMF method, because it is decomposed randomly every time we run the algorithm. We use the k-value, which is found by SVD, to decompose TDM by NNMF.

$$\left[\ A\ \right]_{m \times n} = \left[\ W\ \right]_{m \times k} \left[\ H\ \right]_{k \times n}$$

Fig. 4 NNMF decomposed

3.5 Computation cost and Performance of RMSE by SVD and NNMF

SVD is faster compared to NNMF. We use root-mean-square error(RMSE) to compare the performance of SVD and NNMF results. After dimension reduction, we recompose the matrix from factors created by SVD and NNMF. We get the new matrices as shown in Fig. 5 and Fig. 6.

$$\begin{bmatrix} U' \end{bmatrix} * [\Sigma'] * \begin{bmatrix} V'^{T} \end{bmatrix} = \begin{bmatrix} A' \end{bmatrix}$$

Fig. 5 Return matrix via use best k-value from SVD

$$\begin{bmatrix} W' \end{bmatrix} * \begin{bmatrix} H' \end{bmatrix} = \begin{bmatrix} A' \end{bmatrix}$$

Fig. 6 Return matrix via use best k-value from NNMF

We use Eq (2) to evaluate the accuracy.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=0}^{n} \sum_{j=1}^{m} (a_{ij} - a'_{ij})^2}{n * m}} \quad (2)$$

3.6 K-means

We will cluster document vectors expressed in terms of latent semantic features. We get matrix A* as shown in Fig. 7. Rows of A* are document vectors. We cluster them (unsupervised clustering) into two classes using k-means, with k=2.

$$\begin{bmatrix} U' \end{bmatrix} * [\Sigma'] = \begin{bmatrix} A^* \end{bmatrix}$$

Fig. 7 Documents expressed as vector of semantic compositions

3.7 Evaluation of the Result

After clustering, we verify two groups of documents, whether they are truly separated into positive and negative comments, by manually reading and finding the ground truth.

4. Experiment Results

4.1 Experiment Results

In this study, in order to verity different customers' purchase habits, we choose two types of hotels, luxury hotel and economical hotel. We collect the data from top 5 hotels ranked by TripAdvisor in Hong Kong. All comments are in English. The date of collected comments is from Jan. 2015 to Dec. 2017. Number of collected comments for luxury hotel is 6939, and for economical hotel is 2069, as shown in table 1.

After cleaning data, we get the terms for each hotel's data. For luxury hotel, number of original terms is 9610, the number of nouns and adjective terms is 3666, and the k corresponding to 90% value, calculated from elements of matrix \Sigma is 281. For economical hotel, number of original terms is 5343, the number of nouns and adjective terms is 2323, and k for 90% of elements of \Sigma is 234. The results are shown in table 2.

Table 1 Data distribution

Type	Source of the data	Language	Date of collected comments	Number of collected comments
Luxury hotel	Top 5 hotels ranked by TripAdvisor in Hong Kong	English	2015/01 ~ 2017/12	6939
Economical hotel				2069

Table 2 Exacted terms and dimension reduction by SVD

Type	Number of original terms	Number of Nouns and Adjective terms	k-value of after reducing dimensions by SVD
Luxury hotel	9610	3666	281
Economical hotel	5343	2323	234

4.2 Performance of RMSE by SVD and NNMF

Although we already found the best k-value from SVD, but in order to compare SVD and NNMF performance, we try to use different k-value and compare the performance (this study use k-value from 1 to 700). As shown in Fig. 8 and Fig. 9, for each hotel's data, when k is more than 20, it is apparent that SVD performs better than NNMF. And for same k-values to training the model, NNMF spent a lot more time than SVD.

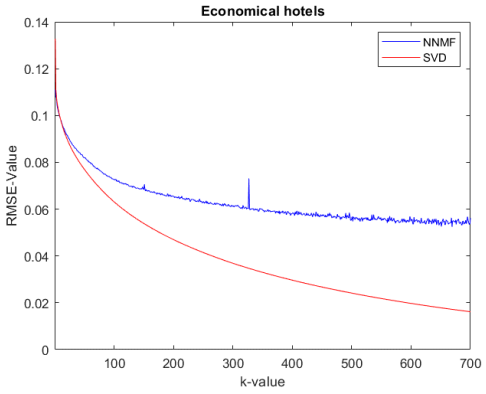

Fig. 8 RMSE performance for economical hotel

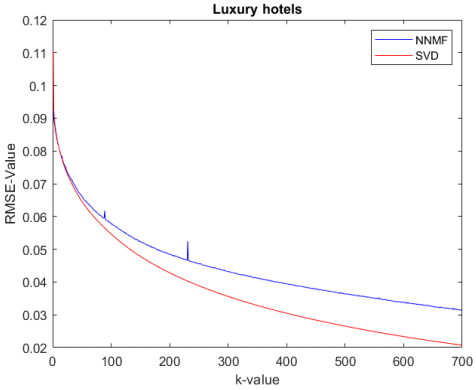

Fig. 9 RMSE performance for luxury hotel

4.3 Clustering of Comments Data

We use K-means to cluster the data from SVD. We use 10% of comments from each data to verity the members of two groups, which are positive or negative. We check the text manually to verify whether the comments are really grouped into positive and negative comments. In Fig. 10 and Fig. 11, the red group means positive comments, the blue group means negative comments.

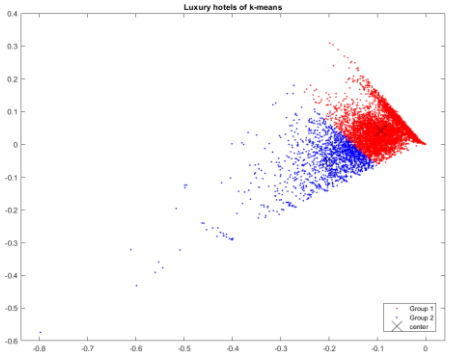

Fig. 10 2-means result for economical hotel

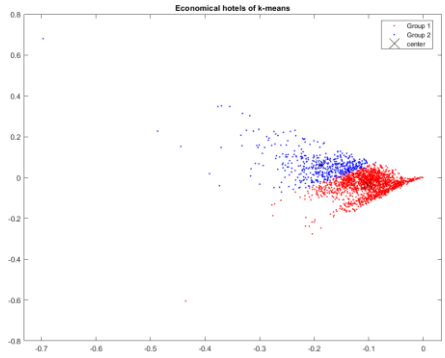

Fig. 11 2-means result for luxury hotel

4.4 Sample of Clustering Results

In table 3, we show some classified samples which we actually read. For the positive comments, if customers review those comments, it will be positive influencing customers' visit intention. For the negative comments, then negative influencing.

5. Conclusion

The purpose of this study is unsupervised classification of text comments into two groups, hypothesizing that they will be grouped into positive and negative comments. Considering words as features, the document (here comments) vector will have too many elements. Clustering at that high dimension failed. To reduce dimension, we used latent semantic analysis, which use Singular Value Decomposition. Clustering is done using k-means. By manual inspection, we verified that the comments are, in fact, grouped into positive and negative.

For the future work, we will add feature selection methods to exact the important factors, which are important to influence customers' review about the service. We also plan to collect data from different countries, and compare visitors' priorities.

Table 3 Sample of clustering results by ground truth

Positive comments	Negative comments
Excellent service, extremely comfortable and spacious room. Good food included in my package as well. Most importantly, the staff is very friendly, kind and helpful. In fact, it has been the 10th year that I book this hotel. Its services have always been excellent throughout the years. I will definite book it again and I really recommend this hotel	I started staying here when it was Langham Place Hotel Mong Kok about 8 years ago. 2 years ago next month management decided, in its wisdom, to rebrand as Cordis That involved pitching to a different market. This time I stayed here as my preferred hotel was fully booked for the time I wanted. Club staff remain faithful to trying to serve customer needs. In my perception, it is indeed a pity the management decisions have significantly dumbed down the hotel What a pity
I had an issue and it was quickly resolved by the on duty managers. I really appreciate their efforts!!! The Hotel is in a great location and the concierge is very helpful daily in helping me arrange my side trips. In summary, a great hotel.	never stay again, staff was not ability to be helpfull. especially receptionist lady whom name was not memorable by me was enough knowledge. when we joined breakfast, the staff also not hospitable enough and they were approach on wellcome as dirctive and like commander. tea and coffee service were very bad

References

[1] Statista 2019: Direct and total contribution of travel and tourism to the global economy from 2006 to 2017, Statista (online), available from <https://www.statista.com/statistics/233223/travel-and-tourism--total-economic-contribution-worldwide/>, (accessed 2019-05-20).

[2] Valdivia, A., Hrabova, E., Chaturvedi, I., Luzón, M. V., Troiano, L., Cambria, E., and Herrera, F.: Inconsistencies on TripAdvisor Reviews: a Unified Index between Users and Sentiment Analysis Methods, *Neurocomputing* (2019).

[3] Vyas, C.: Evaluating state tourism websites using Search Engine Optimization tools, *Tourism Management*, Vol.73, pp.64-70. (2019).

[4] Li, J., Xu, L., Tang, L., Wang, S., and Li, L.: Big data in tourism research: A literature review, *Tourism Management*, Vol.68, pp.301-323 (2018).

[5] Daniszewski, H.: Smartphones changing tourism business–smartphones the latest travel tool (2012).

[6] Stylos, N., Bellou, V., Andronikidis, A., and Vassiliadis, C. A.: Linking the dots among destination images, place attachment, and revisit intentions: A study among British and Russian tourists, *Tourism Management*, Vol.60, pp.15-29 (2017).

[7] Taher, S. H. M., Jamal, S. A., Sumarjan, N., and Aminudin, N.: Examining the structural relations among hikers' assessment of pull-factors, satisfaction and revisit intentions: The case of mountain tourism in Malaysia, *Journal of outdoor recreation and tourism*, Vol.12, pp.82-88 (2015).

[8] Zhang, H., Wu, Y., and Buhalis, D.: A model of perceived image, memorable tourism experiences and revisit intention, *Journal of Destination Marketing & Management*, Vol.8, pp.326-336 (2018).
[9] Han, H., and Hyun, S. S.: Impact of hotel-restaurant image and quality of physical-environment, service, and food on satisfaction and intention, I*nternational Journal of Hospitality Management*, Vol.63, pp.82-92 (2017).
[10] Cantallops, A. S., and Salvi, F.: New consumer behavior: A review of research on eWOM and hotels, *International Journal of Hospitality Management*, Vol.36, pp.41-51(2014).
[11] Wright, K. B.: Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services, *Journal of computer-mediated communication*, Vol.10, No.3, JCMC1034 (2005)
[12] Sohail, S. S., Siddiqui, J., and Ali, R.: Feature extraction and analysis of online reviews for the recommendation of books using opinion mining technique, *Perspectives in Science*, Vol.8, pp.754-756 (2016).
[13] Chen, L., Jiang, T., Li, W., Geng, S., and Hussain, S.: Who should pay for online reviews? Design of an online user feedback mechanism, *Electronic Commerce Research and Applications*, Vol.23, pp.38-44 (2017).
[14] Amaro, S., Duarte, P., and Henriques, C.: Travelers' use of social media: A clustering approach, *Annals of Tourism Research*, Vol.59, pp.1-15 (2016).
[15] Bhole, B., and Hanna, B.: The effectiveness of online reviews in the presence of self-selection bias, *Simulation Modelling Practice and Theory*, Vol.77, pp.108-123 (2017).
[16] Sullivan, D.: Document warehousing and text mining: techniques for improving business operations, marketing, and sales, John Wiley & Sons, Inc. (2001).
[17] Vásquez, C.: Complaints online: The case of TripAdvisor. Journal of Pragmatics, Vol.43, No.6, pp.1707-1717 (2011).
[18] Liu, B.: Sentiment Analysis and Subjectivity, *Handbook of natural language processing*, Vol.2, No.2010, pp.627-666 (2010).
[19] Golub, G. H., and Reinsch, C.: Singular value decomposition and least squares solutions, In Linear Algebra, pp. 134-151, Springer, Berlin, Heidelberg (1971).
[20] Lee, D. D., and Seung, H. S.: Learning the parts of objects by non-negative matrix factorization, *Nature*, Vol.401, No.6755, pp.788 (1999).
[21] Liu, X. et al.: Multiple kernel k-means with incomplete kernels. *IEEE transactions on pattern analysis and machine intelligence* (2019).
[22] Schmidt, M., Schwiegelshohn, C., and Sohler, C.: Fair coresets and streaming algorithms for fair k-means clustering, arXiv preprint arXiv:1812.10854 (2018).
[23] Jeong, Y., Lee, J., Moon, J., Shin, J. H., and Lu, W. D.: K-means data clustering with memristor networks, *Nano letters*, Vol.18, No.7, pp.4447-4453 (2018).
[24] Zhang, G., Zhang, C., and Zhang, H.: Improved K-means algorithm based on density Canopy, *Knowledge-Based Systems*, Vol.145, pp.289-297 (2018).