2013 IEEE 10th International Conference on e-Business Engineering

Knowledge Discovery of Service Satisfaction Based on Text Analysis of Critical Incident Dialogues and Clustering Methods

Charles Trappey
Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan
trappey@faculty.nctu.edu.tw

Hsin-Ying Wu*
Department of Technology Management, Open University of Kaohsiung, Taiwan
cindywu@ouk.edu.tw
*Corresponding author

Kuan-Liang Liu
Institute of Information Management
National Cheng Kung University, Tainan, Taiwan
kliangliu@gmail.com

Feng-Teng Lin
Department of Finance
Shu-Te University, Kaohsiung, Taiwan
ftlin@stu.edu.tw

***Abstract*—Text mining of consumer's dialogues regarding their service experiences provides a direct and unbiased feedback to service providers. This research proposes an analysis process to analyze unstructured input from consumer dialogues. The goal is to apply the critical incident and text mining methods to discover factors that contribute to customer satisfaction and dissatisfaction. The critical incident method is used to construct an open-ended questionnaire to collect customer's positive and negative opinions toward the service provided. Valid and reliable text mining techniques are used to cluster significant text to help analyze incidents that customers care about. A case study of consumers riding the Kaohsiung Mass Rapid Transit System (KMRT) was cased to evaluate the proposed analysis process. Based on dialogues collected from the open-ended questionnaires, the analysis process extracts key phrases related to consumer's best and worst service experiences, creates significant dialogue clusters, and derives meaningful trends, baselines, and interpretations of consumer satisfaction and dissatisfaction. The results of this case study can be used as a basis for building more complete analytical methods to understand consumer experiences and provide strategic feedback for service providers.**

Keywords - customer satisfaction, critical incident techniques, text mining, cluster analysis, CKIP, KMRT

I. INTRODUCTION

An e-business company daily faces the questions, "How can I better satisfy my customers? What do my customers need from me to better improve on products and services?" When dealing with consumer satisfaction, the most popular solution that researchers adopt may be to use a scaled questionnaire. However, a re-occurring problem for analysts of customer satisfaction is to devise ways to collect data from consumers that are not culturally bound and do not require preconceptions of what is or is not important to consumers [1]. When evaluating consumer experience using structured questionnaires, researchers may have a general idea of what causes satisfaction or dissatisfaction. However, their preconceptions may be very limited, biased, and completely different from the factors and variables that truly underlie the consumer's experience. In addition, the causes of satisfaction and dissatisfaction continuously change with the operations environment. Thus, the critical incident technique, a well documented and validated data collection technique [2] [3] is often used under dynamic service conditions to collect a detailed account of a customer's experience using their own words and explanations. There is no a priori determination of what will be important to consumers regarding their experience since they are simply asked to relate critical satisfying and dissatisfying experiences in the form of a short written dialogue.

However, the interpretation of these dialogues into meaningful statistics, especially when attempting to model, cluster, and analyze the critical elements of satisfaction and dissatisfaction over time, requires new techniques and data analysis methodologies. Previous research [4] [5] has demonstrated the feasibility of text mining the content of patent documents to cluster patents into homogenous groups and then use technology forecasting to evaluate these clusters for market opportunities (Figure 1).

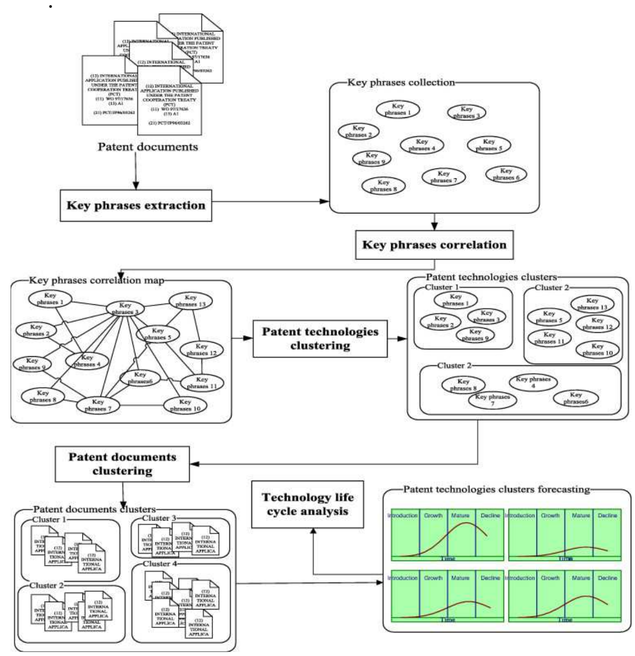

Figure 1. Patent content analysis process [4]

978-0-7695-5111-1/13 $26.00 © 2013 IEEE
DOI 10.1109/ICEBE.2013.40

Knowledge discovery can be classified as data mining and text mining. [6-8] Data mining is used to extract useful information from databases, while text mining is to discover the underlying opinions or comments in documents. Text mining is a recent discipline of information retrieval and computational linguistics studies the underlying meanings and relations of text in context. The ideal text mining tool process a set of search results for a given item, generates a list of product attributes (quality, features, etc.) and aggregates opinions about each of them (poor, mixed, good)."[9].

This research applies the critical incident and text mining techniques to discover factors underlying customer satisfaction and dissatisfaction. The critical incident dialogues collected from customers require key phrase extraction and the preprocessing of the dialogues into a standard format. Thus, the Chinese word segmentation system CKIP (Chinese Knowledge and Information Processing) is used. The CKIP group is a research team formed by the Institute of Information Sciences and the Institute of Linguistics at Academia Sinica in 1986. The resources include Chinese electronic dictionaries, Mandarin Chinese corpora, and processing technologies for Chinese texts. Utilizing and incorporating with these technologies, we extend our previous research in the area of patent content analysis and apply critical incident dialogues to Chinese language opinion mining.

The Chinese word segmentation system developed by CKIP is an online tool resource used to process text content through word detection and extraction. The returned results include a word segmentation list, an unknown word list, and details of the text processing, classification, and possible alternative meanings for unknown characters. CKIP is a client end service script that allows users to connect to the remote segmentation system server and request segmentation services (http://ckipsvr.iis.sinica.edu.tw/resources/CKIPClient.rar). A key phrase counting script, CKIP helps users count and summarize key phrases from the segmentation results. (http://ckipsvr.iis.sinica.edu.tw/resources/CountWordFreq.rar). The full Chinese is available online at http://parser.iis.sinicaedu.tw. The critical incident dialogues are processed in batch mode by CKIP prior to cluster analysis. The resulting frequency of keywords and their correlations are used as input for the clustering software.

Clustering is widely used for text mining, pattern recognition, webpage analysis, and marketing analysis [10, 11]. Clustering is used to separate a heterogeneous population into a number of homogeneous subgroups without predefined classes [12]. The purpose of clustering is to select elements that are as similar as possible within groups but as different as possible between groups. Groups are clustered based on entities' similarity according to specified variables and the meanings of clusters depend on the context of the analysis.

Clustering algorithms are usually used for exploring data structures based on defined data format. Whenever a clustering algorithm is used, it is necessary to evaluate clustering result and decide upon a finalized clustering structure. That is, how many clusters should be defined? For unsupervised learning processes, cohesion and separation are used to analyze variation within and between clusters [13]. Cohesion is defined as a sum of weighted links within each cluster and shows variations between items separated in clusters. Separation is defined as the summary links between every item from each cluster pairs. In this study, we will apply these indexes as the clustering criteria to evaluate cluster structure and size.

In order to evaluate service satisfaction for the Kaohsiung Mass Rapid Transit System (KMRT) case, a facility field study was conducted. The KMRT is a rapid transit system covering metropolitan Kaohsiung, Taiwan, which is made up of orange and red lines with 36 stations covering a distance of 42.7 km (26.5 mi). The current operating lines were opened for service on September 14, 2008. Since Kaohsiung city is expanding, the KMRT system network in the near future, feedback including both positive and negative aspects from riders is very important to future strategic planning. Data was collecting using hand distributed questionnaires to KMRT riders at stations. The Kaohsiung MRT route map and transit system used as the sample target can be found at http://www.krtco.com.tw/en/stationGuide_map.aspx.

II. Research Methods and Materials

This section introduces the methodology and approach for the proposed research. Fig. 2 depicts the proposed data analysis process. The first step is to distribute questionnaires at the Kaohsiung MRT stations. Besides recording general population features, e.g., age, gender, we also ask respondents to provide feedback concerning their positive and negative MRT experiences. The respondents were asked to write several sentences to describe the experiences in Chinese. In order to analyze these questionnaires and the related positive and negative comments, user comments were extracted for key phrases, and then transformed into a key phrase matrix. Based on the key phrase matrix, the clustering algorithm was utilized to group similar key phrases together for further analysis. In order to decide the number of clusters within each set, we implemented an iteration process to calculate a clustering quality index composed of cohesion and separation statistics. The detailed method is described in the following sections.

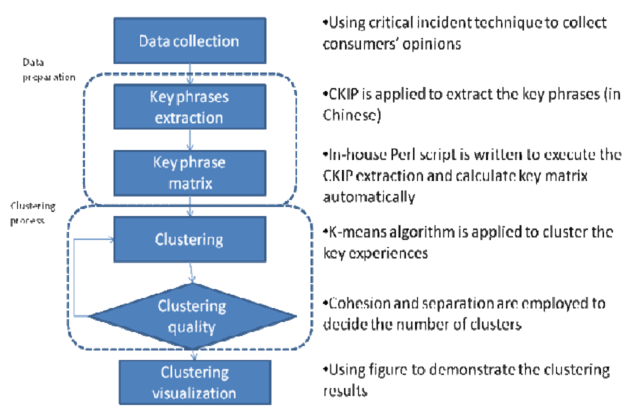

Fig 2. Proposed data analysis process

A. Data Collection: Questionnaire formation and distribution

The critical incident (CI) methodology is used to analyze the KMRT customer satisfaction data and to create a formal representation schema. Figure 3 shows the form of the CI survey. The positive incident open-ended query reads:

Please describe your best experience riding the Mass Rapid Transit System (MRT) in Kaohsiung. Describe the ride in detail and explain exactly what you liked about the ride (please include any details about the station, the passengers, bathrooms, or atmosphere that made the experience enjoyable). Describe the experience of riding the MRT so that I will know exactly how you felt, what you enjoyed most, and say why the MRT ride made you more satisfied than other forms of public transportation (e.g., boat ferries, trains, high speed rail, taxis, and buses). Tell me where you took the ride from and to, tell me about your experience paying the fare if it made you happy (the people or machines providing tickets or directions, and tell me about any special skills or knowledge that was given to you that made the trip easy for you). When you describe your best MRT ride experience, make me feel like I was there riding along with you.

The negative incident repeats the same open-ended question except with a reversal of terms. For example, "please describe your worst experience riding" and so forth.

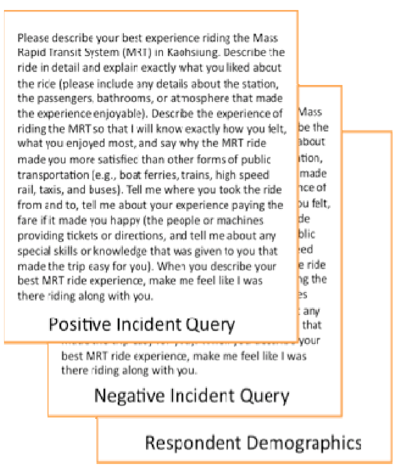

Figure 3. The Kaohsiung MRT critical incident three page survey form.

Fifty questionnaires were distributed at convenient and well-spaced Kaohsiung MRT stations to passengers who were asked to write down their worst and best MRT experiences. The surveys were collected and analyzed to further understand the text content and clusters of the consumer dialogues.

B. Key phrases extraction and summarization

Language processing applications, such as data mining, machine learning and information retrieval have to organize the words in the text so that the text can be processed. In order to process these questionnaires automatically, the phrases, terms and words used by each questionnaire respondent were extracted and separated into good and bad experience sets. A Chinese sentence, however, contains no delimiters, such as a space or separate words. Therefore, preprocessing of these questionnaires is important to determine alternative word combinations of a sentence by comparing it with a lexicon to eliminate ambiguities. The CKIP (Chinese Knowledge and Information Processing) system is a Chinese word segmentation method with unknown word identification and part-of-speech (pos) tagging (http://godel.iis.sinica.edu.tw/CKIP/engversion/index.htm). The system contains a 100,000-entry lexicon with pos tags, word frequencies, and pos tag bigram information, etc. The word segmentation process is based on the lexicons, morphological rules for quantifier words and reduplicated words. Pos tagging is for both known and unknown words.

We integrated the CKIP client end script provided by CKIP with our in-house PERL script to batch process questionnaires. After phrase extraction and ordering, each questionnaire was transformed into a phrase list and each phrase contains its related phrase category identified by CKIP. Table 1 shows a snapshot of a transformed pretest questionnaire.

TABLE 1. TRANSFORMED QUESTIONNAIRE

要(D) 去(D) 搭(VC) 左營(Nc) 高鐵(Na) (PERIODCATEGORY) 從(P) 高雄(Nc) 火車站(Nc) 直接(VH) 可以(D) 通往(VCL) 高雄(Nc) 捷運站(Nc) (PERIODCATEGORY) (PERIODCATEGORY) 不用(D) 再(D) 轉車(Na) (PERIODCATEGORY) (PERIODCATEGORY) 且(Cbb) 省時(VH) 間(Ng) (PERIODCATEGORY) (PERIODCATEGORY) 班次(Na) 也(D) 很多(VH) (PERIODCATEGORY) (PERIODCATEGORY) 對於(P) 外縣市(Nc) 的(DE) 人(Na) (PERIODCATEGORY) (PERIODCATEGORY) 省(VH) 了(T) 等 (Cab) 車 (Na) 及 (Caa) 轉車 (Na) 的 (DE) 麻煩(Na) (PERIODCATEGORY) (PERIODCATEGORY)

As shown in table 1, extracted phrases include two parts, phrase and phrase category tag. Since Chinese sentences and words usually do not have clear timing status, a single word might have different meaning or even different phrase categories within different context. Tag information is necessary for solving these ambiguities. Table 2 shows the simplified tag used in CKIP and the related phrase category.

TABLE 2. PHRASE CATEGORIES.

Simplified phrase category	**Corresponding CKIP phrase tag**	
A	A	/*adjective*/
C	Caa	/*conjunctions, e.g. and, or*/
ADV	Daa	/*quantity adverb*/
ADV	Dfa	/*adverb before verb*/
ADV	Dfb	/*adverb after verb*/
Vi	VA	/*intransitive verb*/
Vt	VC	/*transitive verb*/
N	Na	/*general noun*/
N	Nb	/*specific noun*/
N	Nc	/*location noun*/
N	Ncd	/*position noun*/

An in house PERL script is used for extracting phrases from questionnaire text file and the descriptive statistics are then summarized based on phrases' frequencies. After phrase extraction and ordering, each questionnaire is transformed into a phrase matrix as shown in Table 3.

TABLE 3. PHRASE MATRIX

Questionnaire	Key phrase			
Questionnaire 1	f11	f12	…	f1m1
Questionnaire 2	f21	f22	…	f2m2
…	…	…	…	…
Questionnaire p	fp1	fp2	…	fpmp

C. Clustering algorithm: K-means

Based on the phrases frequency matrix and their similarity, we use K-means clustering analysis with the key phrase dimensions. The software Waikato Environmental for Knowledge Analysis (WEKA) was used for text mining and clustering analysis. WEKA is well known computer software that is often used to implementing data mining algorithms [14]. The k-means algorithm is one of clustering algorithm implemented within WEKA, which is a method of cluster analysis aims to partition n samples into k clusters in which each sample belongs to the cluster with the nearest mean. Although clustering analysis is most often used for sample partition, it can also be used for key phrase features partition if we transpose the data matrix.

Given a set of key phrases (k1, k2, … , kn), where each key phrase is a d-dimensional real vector, k-means clustering aims to partition the n key phrases into k sets ($k \le n$) S = {S1, S2, … , Sk} so as to minimize the within-cluster sum of squares (WCSS):

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2 \tag{1}$$

where μi is the mean of points in Si.

D. Clustering quality evaluation: cohesion and separation

In order to decide the number of clusters within an automatic process flow, we implemented a clustering quality criteria composed of cohesion and separation. These two values would help us review the clustering result and decide the final structure. Cohesion and separation calculations help us understand the variations between clusters within cluster structure whereas cohesion evaluates variations within clusters and separation evaluates variations between clusters. The cohesion equation for each cluster is as the following:

$$Cohesion(C_i) = \frac{\sum_{x \in C_i, y \in C_i} similarity(x, y)}{\frac{|C_i|(|C_i| - 1)}{2}} \tag{2}$$

where C_i denotes each cluster within clustering result and similarity was defined by Euclidean distance. x and y represent key phrase pairs found within the cluster. The algorithm reviews each cluster and aggregates the cluster cohesions values as a synthetic clustering quality criterion. The larger cohesion value denotes a better clustering result.

If there is only one key phrase within a cluster, the cohesion of the cluster will be 1. Finally the average cohesion of clustering result is calculated as follows:

$$Cohesion = \sum_{i=1}^{K} Cohesion(C_i) / K \tag{3}$$

Separation reviews how close each cluster is within the resulting structure. A better clustering result will have a larger separation value. The equation follows:

$$Separation(C_{i,j}) = \sum_{\substack{j \ne i \\ i=1}}^{K} \frac{\sum_{x \in C_j, y \in C_i} similarity(x, y)}{|C_j| \times |C_i|} \tag{4}$$

Each pair of clusters within the result will have a separation value. After summation of every cluster pair separation and dividing to the number of cluster pairs, the average separation of the clustering result is computed:

$$Separation = \sum_{i=1}^{K} Separation(C_i) / K \tag{5}$$

For comparing cluster results, we use Cartesian coordinates to display both values.

To combine cohesion and separation as an integrate index, we combine them with the following formula:

$$Cohesion - Separation = \frac{Cohesion}{\sqrt{Separation}} \tag{6}$$

III. CASE STUDY: KAOHSIUNG MASS RAPID TRANSPORT CUSTOMER DIALOGUES

For the case study, open-ended questionnaires were distributed at the Kaohsiung MRT Station to passengers that were asked to write down their worst and best MRT experiences. Sixty valid surveys were retrieved and analyzed to further understand the text content and clusters of the consumer dialogues.

The collection from the questionnaire is a statement regarding passenger's best or worst experiences. These statements might consist of several sentences or even just one word, which means the data matrix parsed from these questionnaires was very sparse. However, we believe that even single word descriptions contain very important information. Therefore, no questionnaires were excluded from our analysis since all relayed relevant data. A single word in Chinese can have very significant meaning.

According to key phrases extracted from satisfactory and dissatisfactory questionnaires, we implemented the k-means clustering algorithm to explore relationships between these key phrases. Figure 4 shows the result of different clustering number parameter settings (2~10) from satisfactory key phrases. The x-axis denotes the cluster number and the y-axis denotes the value of separation and cohesion. The smaller separation value denotes a better clustering result, i.e., the smaller the separation value, the more dissimilar each cluster pair. Figure 4 shows that the lowest separation value 0.014 is reached with 9 clusters set. The pattern also reveals that when clustering size increases to 5, the separation value becomes more stable.

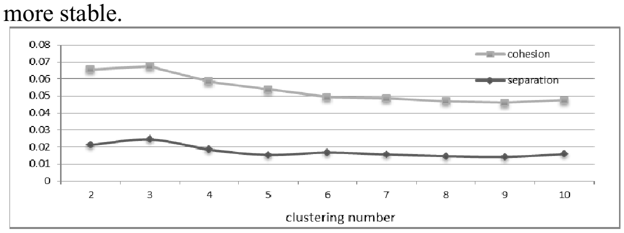

Figure 4. Cohesion and separation index from satisfactory key phrases clustering result

Figure 5 shows the cohesion-separation value of different clustering number settings with satisfactory key phrases. The x-axis denotes the separation value and the y-axis denotes the cohesion value. The number of clusters with cohesion-separation value within parentheses was labeled beside each point. From figure 5, the point of cluster number 5 has the highest cohesion-separation value. Therefore we choose 5 as the final clustering number for exploring satisfactory questionnaire data as discussed in the next section.

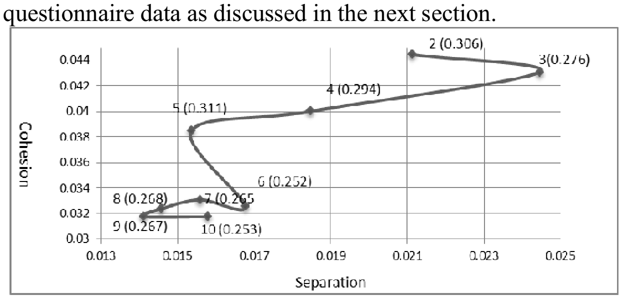

Figure 5. Cohesion and separation index from satisfactory key phrases clustering result. Cluster number (cohesion-separation)

Figure 6 shows the result of different clustering number parameter settings from 2 to 10 utilized for clustering dissatisfactory key phrases. The separation value was around 0.02 for clustering number 2 to 8. But when the clustering number goes to 9 and 10, the separation drops dramatically to 0.01 and 0.009 respectively. The other value, cohesion, denotes similarity of items within each cluster. Therefore a larger cohesion denotes a better clustering result. Fig. 6 shows that the result with 2 clusters has the best cohesion value 0.086. The cohesion also drops dramatically to around 0.023 when the clustering number goes to 9 and 10.

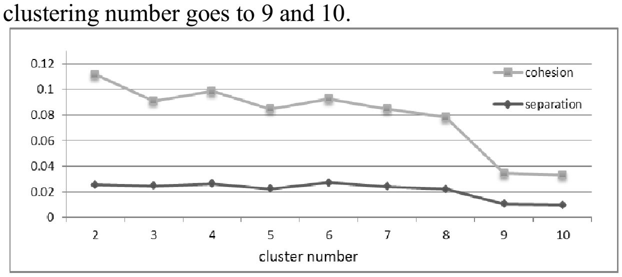

Figure 6. Cohesion and separation index from dissatisfactory key phrases clustering result

Figure 7 shows cohesion-separation values of different clustering number settings with dissatisfactory key phrases. The clustering result has the highest (0.535) and second highest (0.442) cohesion-separation value when the clustering number was set to 2 and 4 respectively. Here we regard the results for 2 clusters as an outlier. Therefore the result of 4 clusters becomes our final clustering result for exploring dissatisfactory questionnaire data

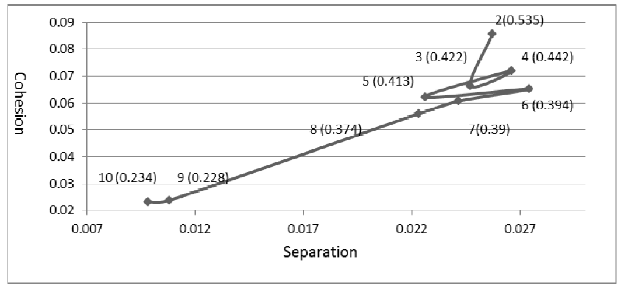

Figure 7. Cohesion and separation index from dissatisfactory key phrases clustering result. Cluster number (cohesion-separation)

A total of 5 and 4 clusters were derived from positive and negative experiences respectively. Referring to Table 6, the positive experiences of the consumers are grouped into 5 different clusters. The first cluster relates to people riding on the MRT and do not drive cars. The second cluster is labeled as MRT station environment where people are satisfied with the bright bathrooms and platforms. The third cluster features the services and amenities. Passengers are satisfied with the service provided for people with special needs or for people on holiday traveling with bicycles. The third cluster also features the station environment including the artistic interior and exterior designs and clear station layout and signage. The forth relates to the convenience and cleanliness of the station. The fifth positive consumer dialogue cluster is labeled as sitting space within carriage and relates to the comfort of the MRT train where consumers feel satisfied with the comfortable seats and pricing.

TABLE 4. CLUSTER FEATURE FOR POSITIVE EXPERIENCE

Cluster number	Cluster features	Representative phrases
1	General terms	Ride(VC), car(Na),
2	MRT station facilities	time(Na), Taipei(Nc), toilet(Nc), birght(VH), satisfy(VK), bus(Na), station affairs(Na), convenient(VH), platform(Nc), family(Na), connect(VC), abroad(VA), parking(VA), safe(VH),
3	MRT services and amenities	people(Na), experience(Na), space(Na), personnel(Na), Formosa boulevard(Nb), comfortable(VH), speed(VH), feeling(Na), Taiwan high speed rail(Na), MRT station(Nc), serve(VC), get in(VA), crowd(VH), happiness(VH), Zuoying(Nc), usage(VC), station(Nc), elevator(Na), first(Neu), like(VK), ride(VC), friend(Na), scooter(Na), clear(VH), beauty(VH), rail(Na), international(Nc), to Airport(Nc), shopping area(Nc), chewing gum(Na), enthusastic(VI), design(Na), eat(VC), stable(VH), Sanduo(Nb), feeling(VK), warming(VHC), get off(VA), frequency(Na), crowd(Na), air condition(Na), starting point(Na), attitude(Na), arriving station(VA), view(Na)
4	Convenient and clean	Convenient(VH), Clean(VH), environment(Na),
5	Sitting space within carriage	carriage(Na), passenger(Na), sit(VA), fare(Na), airport(Nc), seat(Na)

Table 7 lists the negative experiences of the consumer complaints with the cluster features and corresponding descriptions. The first cluster is labeled as peak hours, which is describes the station rush hours. The second cluster refers to the personnel services when passengers complain about the services they received. The third cluster relates to unpleasant transit experiences and station environment whereby customers complain about the bad transit system to bus or other transportation authorities. The fourth cluster describes complaint concerning riding etiquette. Some users are not satisfied about the riding culture in Taiwan since some people rush into the trains without passengers forming a waiting queue.

TABLE 5. CLUSTER FEATURE FOR NEGATIVE EXPERIENCE

Cluster number	Cluster features	Representative phrases
1	Peak hours	stand(VA), terrible(VH), station(Nc)
2	MRT personnel service	ride(VC), Kaohsiung(Nc), personnel(Na), public(Nh), traffic(Na), MRT station(Nc), Service(VC), tool(Na), eat(VC), convenient(VH), comfortable(VH),
3	Transit system and station environment	carriage(Na), population(Na), time(Na), transport(VC), system(Na), Taipei(Nc), crowd(Na), bike(Na), inconvenient(VH), fare(Na), platform(Nc), loud(VH), connect(VC), international(Nc), plan(VC), scooter(Na), clear(VH), Xiao gang(Nc), Experience(Na), bus(Na), happiness(VH), unable(D), like(VK), process(VA), expensive(VH), speed(Na)
4	riding etiquette	Passenger(Na), get off(VA), get on (VA), advocacy(VC)

IV. DISCUSSION AND CONCLUSION

When evaluating customer satisfaction, market researchers often apply quantitative questionnaires using scales rather than open-ended questionnaires that provide more information directly from respondents. The reason is that there are few analysis techniques that rapidly and automatically process open-ended and unstructured questionnaires. Therefore, this research proposes an analysis process using the critical incident technique, text mining, and clustering methods to analyze open-ended questionnaires. The method is a valid and reliable approach to discover the underlying causes of satisfaction and dissatisfaction. The critical incident technique discovers the experiences that influence customer's satisfaction and dissatisfaction. Text mining and clustering is utilized to generate the key factors of the experiences.

A KMRT case study tests the proposed analysis process. The results of the case study can be used as a basis for managing the factors underlying customer satisfaction and dissatisfaction, especially for a mass rapid transit system. Therefore, as a customer-centric company, KMRT can use the results to improve their personnel services and derive solutions to reduce crowds during peak hours. The KMRT can also construct a more convenient transit and connection system and educate passengers about riding etiquette to create a more comfortable riding experience for all passengers. As for the positive experiences, the light and clean station environment and the artistic interior and exterior designs create pleasant memories. KMRT should maintain the comfortable environment and public art atmosphere because these are factors that passengers care about. KMRT can also advertise customer's good experiences to improve the frequency of use.

Therefore, the proposed analysis process can be utilized to discover the factors behind consumer's comments. The long-term research goal is to apply the proposed process to analyze comments or opinions on social networks. Social networks are popular information sharing platforms where people discuss experiences providing information that can be analyzed as unstructured text dialogues. The proposed process can be used to generate important strategic information based on the voice of the consumer.

REFERENCES

[1] J.C. Flanagan, "The Critical Incident Technique," Psychological Bulletin, vol. 51, pp. 327-358, 1954.

[2] R.A. Feinberg, K. DeRuyter, C.V. Trappey, and T.Z. Lee, "Consumer-Defined Service Quality in International Retailing," Total Quality Management, vol. 6, no. 1, pp. 61-67, 1995.

[3] C.V. Trappey, T.T. Hsio, T.C. Chang, M.H. Che, W.J. Chiu, "Consumer Driven Game Design," Proceedings of DiGRA 2005 Conference: Changing Views – Worlds in Play, 2005.

[4] C.V. Trappey, H.-Y. Wu, T. Taghaboni-Dutta, A.J.C. Trappey, "Using Patent Data for Technology Forecasting: China RFID Patent Analysis," Advanced Engineering Informatics, vol. 25, pp. 53-64, 2011.

[5] F. Taghaboni-Dutta, A.J.C. Trappey, C.V. Trappey, and H.-Y. Wu, "An Exploratory RFID Patent Analysis," Management Research News, vol. 32, no.12, pp. 1163-1176, 2009.

[6] Feldman, R, and Dagan, I., "Knowledge Discovery in Textual Database (KDT) ," Proceedings of the first ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.112-117, 1995.

[7] Tsai, Ho-Cho, "Text Mining System on Customer Complaint Document," Master Thesis, Institute of Information Management, National Cheng Kung University, Tainan, Taiwan, 2003.

[8] Chen, Liang-Chu and Ku, Yi-Ping, "The Study on Exploring Electronic Commerce Topics Network Based on the Term Analysis," International Journal of Commerce and Strategy, vol. 3, No. 3, pp. 137-154, 2011.

[9] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," Proceedings of WWW, pp. 519–528, 2003.

[10] Berkhin, P. (2002), "Survey of clustering and data mining techniques," Technical Report, Accrue Software, Inc.

[11] Chen, B., Tai, P.C., Harrison, R., and Yi, P. (2005), "Novel hybrid hierarchical-K-means clustering method (H-K-means) for microarray analysis," Computational Systems Bioinformatics Conference, Aug. 8-11, Standford CA, USA.

[12] Berry, M. J. A. and Linoff, G.. (1997), Data Mining Techniques: For Marketing, Sale, and Customer Support, John Wiley & Sons Inc, New York, NY, pp. 55.

[13] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). Introduction to Data Mining. Addison Wesley, 1 edition.

[14] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H., "The WEKA Data Mining Software: An Update," ACM SIGKDD Explorations Newsletter, vol. 11, No. 1, pp. 10-18, 2009.