Analyze the Trend of Product Characteristics by the Extraction of Review Comments to Provide Concise Information to Enterprises

Teng-Hsien HUANG, Yu-Jie LAN, Yuan FENG, Yu-Chieh WU and Ming-Chuan CHIU[1]
National Tsing Hua University, Taiwan

Abstract. With the improvement of technology and internet, innovation of the smartphone design, a majority of customers will make their comments through the internet after using the product. The enterprises could make their products better and develop the new product according to these comments. However, complex info might exist in the comments and thus companies have to spend much time in analyzing comments characteristics. In order to realize the product characteristics which customers focus on, our research is to collect the customer reviews. By the capture of customer reviews and the time series analysis, we could understand the characteristics of the product over time, and thus provide enterprises with reference to improve the product. To figure out product characteristics, we develop the system which is used to analyze the comments developed in this study can quickly filter out the noise of comments. Furthermore, it can not only save the time to read comments and reduce the labor costs but also provides the new product with right information to meet customer needs.

Keywords. Text Mining, Product Characteristics, Product Comments, Time Series Analysis

Introduction

With the population of community networks, people often post their feelings of life and comments on community network platforms, including interesting life, comments of restaurants, and product using experience. Besides express the opinion of products, customers also search the related product information through the community network platform. These comments usually show the potential needs of customers and companies can use the comments written by customers to grasp the current market trends to improve their products to correspond with customer needs. However, the comments made by customers often entrain too much noise and useless information of the product, companies have to spend too much time reading the comments and cannot quickly understand what customers want.

Therefore, the aim of this study is to develop a system to grasp the product characteristics through these product comments. With the time series analysis, we want to find the trend of product characteristics. In order to develop a grasp system, this study collects related product comments on the network, and then observe the

[1] Department of Industrial Engineering and Engineering Management
National Tsing Hua University, Hsinchu, Taiwan, 30013, R.O.C
e-mail: mcchiu@ie.nthu.edu.tw

comments to analyze some product characteristics which customers usually focused on. After analyzing product characteristics, we use SQL to build the relevant vocabulary of the product characteristics, and then establish a matching system of the product characteristics by JSP to analyze the comments. The filtered results would be analyzed by using time series analysis to understand the characteristics of each product with the time trend of the growth and decline.

2.Literature Review

2.1 Text mining

Web content mining is a mining method based on the Web content, it is a process of finding message and useful knowledge through the huge Web data. These data include text file, in-text file, also include media data like picture, photography, sounds [1]. The data can be structured data for the database, or HTML tagged semi-structured data, or Unstructured text. Through text classification, the text message processing, including the English document Stemming processing and Chinese document word processing. Because the Chinese document sentences in the separation of the word without a fixed blessing, the Chinese word frequency document statistics before the need for Chinese word processing word segmentation, that is, to join the separation between the word blessing, the beginning of knowledge into a dispersed word form. Patent documents are an ample source of technical and commercial and patent analysis has long been considered a useful vehicle for R&D management and techno economic analysis. [2] They note that citation analysis is subject to some crucial drawbacks and propose a network-based analysis, an alternative method for citation analysis. The analysis is departed for two parts, includes Network analysis and Text mining, Text mining is a rather new technique that has been proposed to perform knowledge discovery from collection operations are performed on the labels. The document in text format can be features by keywords that are extracted through text mining algorithm. In recent years, psychologists have tried to use the systematic method and psychological characteristics of the Big Five personality model, which are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, to describe personality traits. [3] proposed a system to measure the five personality traits of community users to help informers quickly and accurately measure the top five personality traits of social network users. The study also uses the Linguistic Inquiry and Word Count, MRC Psycholinguistic Database and General Inquirer personality tools to analyze user's posts on these tools that contain defined features. Finally, this study adopts Pearson Correlation Analysis to examine the significance of the result that obtained from user's five major personality traits questionnaire and post contents. Furthermore, two Regression Analysis-Gaussian Process and ZeroR are used to predict inference result of user's Big Five personality traits. In the Internet and information Age, online data is growing explosively. Most of the web data is unstructured and difficult for people to decipher. Hence, [4] use sentiment analysis and text mining approaches to detect and predict the hotspot online forums. This study established an approach to analyze text sentiment and digits. Then, this approach integrated with K-means clustering and support vector machine (SVM) to develop the unsupervised text mining approach. According to the result, SVM forecasting achieves highly consistent results with K-means clustering. There have been many studies on text extraction

technology, and the method of research can make the system learn to extract the rules by increasing the training data to build the model. However, this method faces incorrectness of text extraction, long time of text extraction and lack of professional knowledge. [5] proposed an extracting keywords method of multi-steps and extending TF weights. This method doesn't require a lot of training data, and based on the keyword frequency to evaluate the importance of a text which is contained in the document. The sentimental analysis model Construct is an important topic in text mining. [6]They first collect a lot of product reviews and divided these reviews in different clusters by their release time. The research use Alchemy API method to capture the key words, then match each comments to its corresponding sentiment category. Using support vector machine method(SVM) train each comment's by their sentiment category and its key words, and construct a sentiment classifier. Then use Feature selection to improve the classifier. Each product reviews can use this classifier to judge its sentiment, and through the results to figure out the cause and effect about the comment. Finally use the Tweet comment to test the model's performance. The result shows that the sentiment models can inference the comment's sentiment efficiently, and let the company know the market trend. To capture the key word of the post in the community platform, the researchers often capture the Structured Data's important information trough Ontology – Based Clustering, rather than Unstructured Data is capture the key word by Natural Language Processing (NLP). [7] proposed a NLP and Ontology Based Clustering TVC Algorithm to anylize the community platform's key words. The research first collects data from the community plat form and break the sentence to get the vocabulary. Then, do the Morphological Analysis, Syntactic Analysis, Semantic Analysis and Pragmatic Analysis to understand the importance of these vocabularies. The result shows that the proposed method can capture the post's key words to help the Information requester understand the post's importance. The problem of text mining, i.e. discovering useful knowledge from unstructured or semi-structured text, is attracting increasing attention. [8] suggests a new framework for text mining based on the integration of Information Extraction (IE) and Knowledge Discovery from Databases (KDD), a.k.a. data mining. KDD and IE are both topics of significant recent interest. KDD considers the application of statistical and machine-learning methods to discover novel relationships in large relational databases. IE concerns locating specific pieces of data in natural-language documents, thereby extracting structured information from free text. However, there has been little if any research exploring the interaction between these two important areas. In this paper, we explore the mutual benefit that the integration of IE and KDD for text mining can provide.These researches only focus on the text mining. They didn't consider about the application in real world and the revenue about the product. It is important to combine these two concepts.

2.2 Product analyse

Clustering is a powerful technique for large-scale topic discovery from text. [9]. It involves two phases: first, feature extraction maps each document or record to a point in high-dimensional space, then clustering algorithms automatically group the points into a hierarchy of clusters. Document clustering help tackle the information overload problem in several ways. They develop a fast, scalable document clustering system. This tool is designed to discover topic hierarchies in gigabytes of documents per day, and they use the algorithms for feature extraction and clustering of extracted feature.

Then, evaluating the impact of different algorithms on the quality of the generated cluster hierarchies and on the processing time required to build those hierarchies. With the community platform to flourish, Enterprise marketers often analyze the content of the users of the community, in order to understand the different preferences of customers that focus on the product or service-oriented.A research used the Decision Tree Algorithm to attribute the corresponding category by the user's vocabulary of the post in the community platform to let the marketers understand their customer. [10] The research fisrt collect the posts from the Reality Mining, Gowalla,Brightkite and Twitter during the President election. Capture those post's characteristics and its frequency. Through the default Thresholding the research filter the important characteristics in the posts. Use these characteristics construct the Social Tree to attribute the post. Finally, the research calculates the number of the posts to attribute the author. The results show that the decision Tree Algorithm can help the enterprise marketer recognize the author's category and provide some important information. These researches only focus on the grouping. They didn't consider about the application in real world and the revenue about the product.

Product cycle life has been widely discussed. The research is aimed at clarifying the concept of maintainability, distinguishing the confusion between repair and maintenance, defining the object of study, and studying the maintainability of products. [11] Meanwhile, based on the product life cycle theory, the influence of each stage of the whole life cycle on the maintainability of products is discussed in depth so as to change the evaluation mode of a single indicator such as economic benefits and environmental protection. Evaluation system in this paper which is put forward to measure product maintainability can be more accurate on the evaluation result and helpful to specify evaluation process of product maintainability, contributing to optimization of product design and improvement of product quality. Official price indexes are usually constructed using samples of products rather than a census. In particular, the standard approach is to select a sample in some base period and follow the items through time until they disappear from the market, in which case they are replaced. What this means is that the sample will tend to have a different – and older – age composition than the set of items available in the market. As has been noted, such as [12], this leaves these indexes open to bias if there are systematic changes in prices related to the age of products. We have shown that life cycle price trends are important for a range of products and importantly, that the prices do not change at the same rate across the life cycle. This means that the age of the sample is important for measured price change. The research emphasizes that attention needs to be paid to the construction of samples for elementary price indexes and ideally they should be updated as frequently as is practicable.These researches only focus on the product life itself. They didn't consider about the comment feature. It is important to combine the concept about the product comment and product's revenue.

2.3 Summary

This research is aim to capture the characteristics about the comment, and find the trend of these characteristics. Also, we will compare the characteristics' trend to the sales of the product, to see whether there exits the correlation between them.

After the review of some literature, we found that in the past the research usually construct the characteristics capture only or analysis the product life only, they didn't combine the concept of characteristics capture and apply the in the real application.

This research apply the cellphone comment's characteristics capture and compare these characteristics with the real sales of the cellphone.

3.Methodology

As introduction mentioned, nowadays customer or enterprise can know the comments about some products. These comments usually show the potential needs by customers and companies can use the comments written by customers to grasp the current market. This research applies the method which is proposed by [13] in the cellphone area and compare the characteristics that mostly been mentioned in the iphone6 and Samsung note4 with their sales through a year. Finally, we can see whether there is a correlation between the sales and the comment's characteristics.

✓ Notation

DE_i ： the *i*th content of comment.

$\mathrm{CS}(\mathrm{DE}, j)$ ：the *j*th clause of the content of comment DE.

CT_k ：the *k*th topic type of the lexicon of domain.

$\mathrm{El}_t(\mathrm{S}, j)$ ：the *j*th element of congregation S.

$\mathrm{KP}(\mathrm{CS})$ ：the congregation of key points of content CS.

$\mathrm{STV}(\mathrm{CT})$ ：the corresponding lexicon of a subject CT.

Step (A1): Comments on the sentences.

In order to capture the characteristics of the comments posted on the Internet for specific products and services, this step is to punctuate the content of the comment, and then break the sentences after the sentence are punctuated. In general, the comment before and after the space, comma, Chinese back-sloping comma, period, exclamation mark and question mark are two different sentences, so that we use space or punctuation marks to separate the comments in this step.

Specifically, for a content of comment DE_i. In this step, it will be punctuated according to punctuation marks in order to form a number of clauses. After that, mark each of the clauses in the sequence according to the content of comment. Finally, we can get the set of multiple marked clauses. The set of marked clauses is represented by the following equation (1):

DE_i={CS(DE_i, 1), CS(DE_i, 2), ..., CS(DE_i, *j*), ...} (1)

After that, we can obtain the table

Table 1. comment's subset

The content of comment	the set of multiple marked clauses
DE_1	{CS(DE_1, 1), CS(DE_1, 2), ..., CS(DE_1, j), ...}
DE_2	{CS(DE_2, 1), CS(DE_2, 2), ..., CS(DE_2, j), ...}
DE_3	{CS(DE_3, 1), CS(DE_3, 2), ..., CS(DE_3, j), ...}
…	…
DE_i	{CS(DE_i, 1), CS(DE_i, 2), ..., CS(DE_i, j), ...}

…	…

Step (A2): Judge the key points of the comments.

In order to clarify the key points included in comments, this step compares the commentary content of the punctuated sentence with the vocabulary items of the sub-lexicon of each subject type in the subject vocabulary of the domain subject. Therefore, to determine the key points included in comments.

According to the above concept, to determine the key point of the clause for the *j*th clause CS (DE_i, *j*) of a commentary content DE_i. In this step, the *k*th topic type CT_k sub-lexicon store STV (CT_k) in the domain subject vocabulary is compared with the clause CS (DE_i, *j*). If a term El_t (STV (CT_k), m) in the sub-lexicon store exists in STV (CT_k) which appears in the clause CS (DE_i, *j*), the topic type CT_k is the key point of the commentary content DE_i, and the subject type CT_k can be aggregated into the set of key point KP (CS (DE_i, *j*)) of the clause CS (DE_i, *j*). As shown in the mathematical formula (2):

$$\textbf{IF } \exists \mathrm{Elt}(\mathrm{STV}(\mathrm{CT}_k), j) \in \mathrm{CS}(\mathrm{DE}_i, j) \textbf{ THEN } \mathrm{CT}_k \in \mathrm{KP}(\mathrm{STV}(\mathrm{CT}_k), j) \quad (2)$$

Through the above method, the congregation of key points KP(CS(DE_i , *j*)) corresponding to the *j*th clause of the commentary content DE_i (DE_i, *j*) can be obtained in this step. Furthermore, the *j*th clause CS (DE_i, *j*) of the commentary content DE_i is compared with the vocabulary items of the sub-lexicon of each subject type CT_1, CT_2, ..., CT_k, ... in the domain subject vocabulary. That is, the respective topic types CT_1, CT_2, ..., CT_k, ... can be aggregated into the congregation of key points KP(CS(DE_i, *j*)) of the clause CS (DE_i, *j*). As shown in the mathematical formula (3):

$$\textbf{IF } \exists \mathrm{Elt}(\mathrm{STV}(\mathrm{CT}_k), j) \in \mathrm{CS}(\mathrm{DE}_i, j) \textbf{ THEN } \mathrm{CT}_k \in \mathrm{KP}(\mathrm{STV}(\mathrm{CT}_k), j), k = 1,2, \ldots \quad (3)$$

This step is to further compare the clauses CS(DE_i, 1), CS(DE_i, 2) , …, CS(DE_i, j), … of the comment DE_i with the lexical items of the sub-lexicon of each topic type in the CT_1, CT_2, …, CT_k, … which can be the practice of the various topic types CT_1, CT_2, …, CT_k, … will be aggregated into the respective clause CS(DE_i, 1), CS(DE_i, 2), …, CS(DE_i , *j*), … corresponding to the congregation of key points KP(CS(DE_i , 1)), KP(CS(DE_i, 2)), …, KP(CS(DE_i, j)), ….

Through the above method, in this step, we can obtained each clause CS(DE_i, 1), CS(DE_i, 2), …, CS(DE_i, *j*), … of a comment DE_i corresponding to a congregation of key points KP(CS(DE_i, 1)), KP(CSDE_i, 2)), …, KP(CS(DE_i, *j*)), …, and this step is also to further compare the clauses CS(DE_i, 1), CS(DE_i, 2) , …, CS(DE_i, *j*), … of the comment DE_i with the lexical items of the sub-lexicon of each topic type in the CT_1, CT_2, …, CT_k, … which can be the practice of the various topic types CT_1, CT_2, …, CT_k, … will be aggregated into the respective clause CS(DE_i, 1), CS(DE_i, 2), …, CS(DE_i, *j*), … corresponding to the key point set KP(CS(DE_i, 1)), KP(CS(DE_i, 2)), …, KP(CS(DE_i, *j*)), ….each clause of each comment corresponding to the congregation of key points as shown in the table 2.

Table 2. Each clause of each comment corresponding to the congregation of key points

Comment	Key point	Clause	Key point of clause
DE_1	$\mathrm{KP}(\mathrm{DE}_1)$	$\mathrm{CS}(\mathrm{DE}_1, 1)$	$\mathrm{KP}\big(\mathrm{CS}(\mathrm{DE}_1, 1)\big)$
		$\mathrm{CS}(\mathrm{DE}_1, 2)$	$\mathrm{KP}\big(\mathrm{CS}(\mathrm{DE}_1, 2)\big)$
		…	…
		$\mathrm{CS}(\mathrm{DE}_1, j)$	$\mathrm{KP}\big(\mathrm{CS}(\mathrm{DE}_1, j)\big)$
		…	…
DE_2	$\mathrm{KP}(\mathrm{DE}_2)$	$\mathrm{CS}(\mathrm{DE}_2, 1)$	$\mathrm{KP}\big(\mathrm{CS}(\mathrm{DE}_2, 1)\big)$
		$\mathrm{CS}(\mathrm{DE}_2, 2)$	$\mathrm{KP}\big(\mathrm{CS}(\mathrm{DE}_2, 2)\big)$
		…	…
		$\mathrm{CS}(\mathrm{DE}_2, j)$	$\mathrm{KP}\big(\mathrm{CS}(\mathrm{DE}_2, j)\big)$
		…	…
…	…	…	…
DE_i	$\mathrm{KP}(\mathrm{DE}_i)$	$\mathrm{CS}(\mathrm{DE}_i, 1)$	$\mathrm{KP}\big(\mathrm{CS}(\mathrm{DE}_i, 1)\big)$
		$\mathrm{CS}(\mathrm{DE}_i, 2)$	$\mathrm{KP}\big(\mathrm{CS}(\mathrm{DE}_i, 2)\big)$
		…	…
		$\mathrm{CS}(\mathrm{DE}_i, j)$	$\mathrm{KP}\big(\mathrm{CS}(\mathrm{DE}_i, j)\big)$
		…	…
…	…	…	…

4.Results

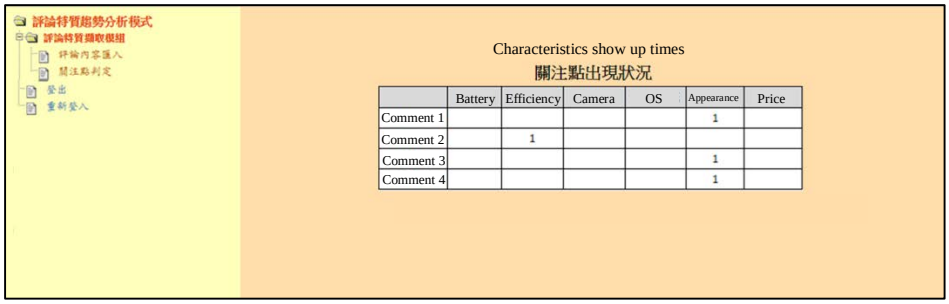
4.1 The comments analysis

We search 120 comments from the websites includes ‘Mobile01’ and ‘ePrice’. According to the methodology, we design the grasping system that its main page would show ‘Welcome to use the comments capturing system’, and the left-frame shows the functions the system can do, which are ‘upload comments’ and ‘capture the characteristics’. Then we can choose the comment we want to analyze to upload to our system. When we upload the comments success, the system will show the comments which we uploaded in a table. After uploaded the comments, we can choose the function ‘capture the characteristics’ on the left-frame, then the system can automatically help us to capture the characteristics. The result of ‘capture the characteristics’ is shown as **Figure.1**. We search the comment through the internet from third quarter, 2015 to second quarter, 2016. When we analyze all the comments we searched from the internet through our system, we can use statistical methods to calculate the amounts of each characteristic, then we can draw the result charts. We Also search the revenue information of Samsung, and the Samsung note4 revenue in

2014/2015. As the analysis of Samsung note4, we use the same way to analysis iPhone6 and also search the revenue information from the website.

4.2 The Comparison between Samsung and Iphone6

By Observing the analyzing result of comments of Samsung note4 and iPhone6, the characteristics of Samsung note4 would be the first to discuss are ‘Appearance’ and ‘Battery’. With the time go through, the ‘Appearance’ discuss would decrease, and the ‘Efficacy’ would increase. ’Price’ is discussed often at first, but decreased in the year 2015. The characteristics of Iphone6 would be first to discuss are ‘Appearance’ and ‘Efficacy’. With the time go through, the ‘Appearance’ and the ‘Efficacy’ discuss would increase. On the other hand, we found that if characteristics ‘Appearance’, ‘Efficacy’ and ‘Price’ discussed by customers increased, the Samsung note4’s revenue would increase. Also, if characteristics ‘Appearance’, ‘Battery’ and ‘Price’ discussed by customers increased, the Iphone6’s revenue would increase. According to these results, we could conclude that the same characteristics to influent revenues are ‘Appearance’ and ‘Price’. The results are shown as **Figure.2** and **Figure.3**.



	Battery	Efficiency	Camera	OS	Appearance	Price
Comment 1					1	
Comment 2		1				
Comment 3					1	
Comment 4					1	

Figure 1. capture the characteristics

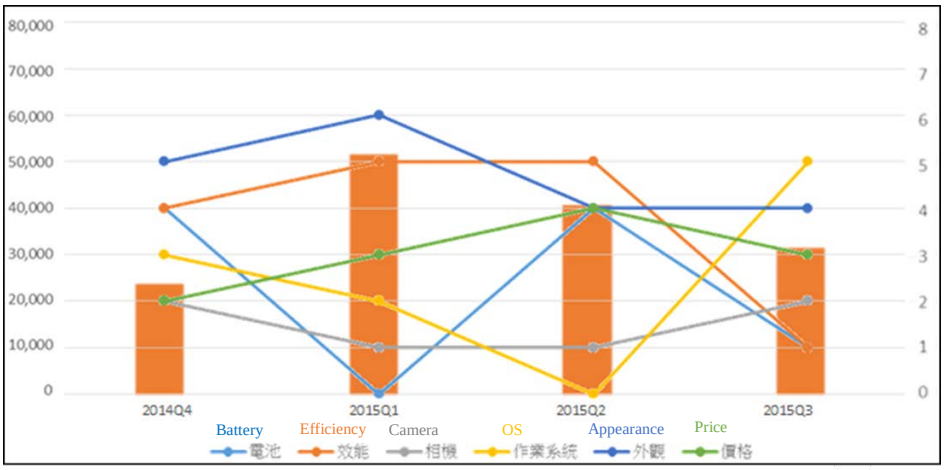


Figure 2. comment’s characteristics trend (iphone6)

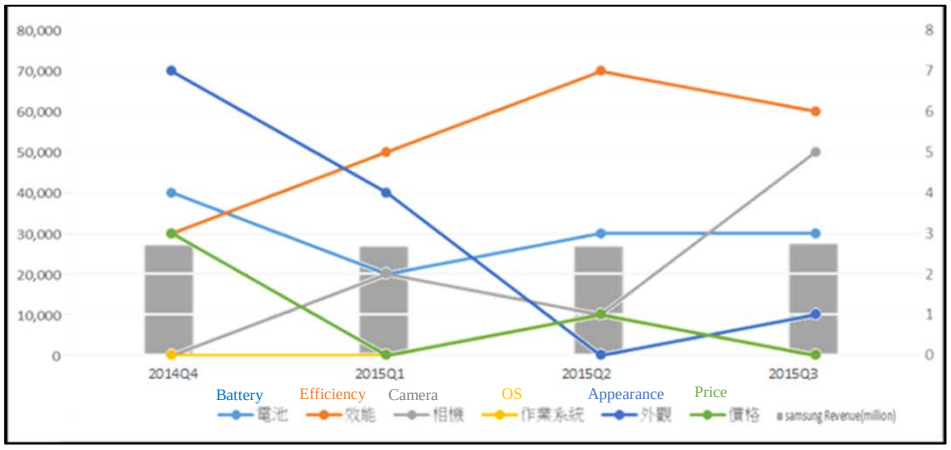


Figure 3. comment's characteristics trend (Samsung note4)

4. Discussion

The focus of our reach was that we developed a system to catch these key characteristics of comments from the website. Under the assumption that the authenticity of these comments, the result showed us that there are some characteristics would affect the revenue. Sometimes, when the customers talked about one of the characteristics, the situation will lead a trend and reflection on the company's sale, and also inflected the revenue. Furthermore, it can not only save the time to read comments and reduce the labor costs but also provides the new product with right information to meet customer needs. However, through the literature we found that there are few papers relatively trends and the company's revenue to do with the combination, so we can provide the company some value information. But our system cannot tell which is the real message, which is a malicious message. So this need to add the authenticity of the reviewers and other considerations.

5.Conclusion and future research

With the capture of characteristics, we can see that some characteristics really affect the company's sales. Therefore, the company can use the system to capture the keywords they want to analyze, and control the company's marketing strategy for a real benefit. Furthermore, the company can capture the comments leave under the advertisements, to analyze if the advertisement is benefit to company to do that sales or strategy.

In the future, we can develop a system that can draw the trend picture automatically, as to be more efficiently. Adding positive and negative judgments, you can let the company refer to those marketing strategies to bring positive returns, those who may bring negative evaluation.

Reference

[1] C.X. Tu, M.Y. Lu and Y.C. Lu, Research on Web Content Mining. *Application Research Of Computers*, vol. 20, 2003, pp. 5-9.

[2] B. Yoon and Y. Park, A Text-Mining-Based Patent Network: Analytical Tool for High-Technology Trend. *The Journal of High Technology Management Research*, vol. 15, 2004, pp. 37–50.

[3] J. Golbeck et al., Predicting Personality from Twitter, In: *IEEE International Conference on Social Computing*, Boston, 2011, pp.149-156.

[4] N. Li and D. D. Wu, Using Text mining and Sentiment Analysis for Online Forums Hotspot Detection and Forecast, *Decision Support Systems*, vol. 48, 2009, pp. 354-368.

[5] B. Hong and D. Zhen, An Extended Keyword Extraction Method, *Physics Procedia*, vol. 24, 2012, pp. 1120-1127.

[6] S. Desai et al., Efficient Regression Algorithms for Classification of Social Media Data, In: *International Conference on Pervasive Computing*, Vadgaon, 2015, pp. 1-5.

[7] D. Shabina et al., NLP and ontology based clustering – An Integrated Approach for Optimal Information Extraction from Social Web, *International Conference on Computing for Sustainable Global Development*, 2016, pp. 1765-1770.

[8] X. Jia., S. Cai., Q. Chen, A Study on The Evaluation of Product Maintainability Based on the Life Cycle Theory, *Journal of Cleaner Production*, 2016, pp. 481-491.

[9] B. Larsen et al., Fast and effective text mining using linear-time document clustering, In: *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, 1999, pp. 16-22.

[10] P.G. Preethia et al., Temporal Sentiment Analysis and Causal Rules Extraction from Tweets for Event Prediction, *International Conference on Intelligent Computing, Communication & Convergence*, 2015, pp. 84-89.

[11] D. Melser and I. A. Syed, The Product Life Cycle and Sample Representatively Bias in Price Indexes, *Applied Economics*, vol. 49, 2016, pp. 573-58.

[12] R. -J. Mooney and U. -Y. Nahm, Text Mining with Information Extraction, In W. Daelemans, et al (eds.): *Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium*, pp.141-160, 2005.

[13] H. C. Yang and C. L. Hou, The model for Sentiment Analysis of Comment. National Tsing Hua University Industrial Engineering and Management Master thesis. 2016.