Monthly Car Sales Prediction using Internet Word-of-Mouth (eWOM)

Chaochang Chiu
Information Management Department
Yuan Ze University
Chungli Dist., Taoyuan City, Taiwan
imchiu@saturn.yzu.edu.tw

Chia-Houng Shu
Information Management Department
Yuan Ze University
Chungli Dist., Taoyuan City, Taiwan
nickefarm@yahoo.com.tw

***Abstract*—Internet's mature has changed the behavior of consumers. Most of consumers before purchase will queries opinion on the Internet. Vehicle is high priced and durable merchandise, so consumer would be more prudent to view Internet opinion before they buy. Past research has pointed out eWOM (electronic Word-of-Mouth) and customer satisfaction will influence purchasing decision, 70% of consumers believe eWOM. In this study, we utilizes economic indexes, Internet "word of mouth", and Google Trends variables created by key word searches to forecast the monthly sales volumes of a given model of car. The experiment results demonstrate that the proposed GA/KNN model has the highest predictive power in terms of Mean Absolute Percentage Error (MAPE).**

Keywords—Electronic Word-of-Mouth; Genetic Algorithm; Car Sales Prediction

I. INTRODUCTION

In the midst of globalizing competition, the automotive industry is becoming more and more customer care oriented in its business model. In order to strengthen its own competitiveness and improve the quality of the service it offers, the automotive industry cannot but seek to understand consumer needs and recognize the importance of those needs in maintaining their own enterprises.

The present study investigates expansion trends, combines Internet resources with forecast related research results, and attempts to research how the automotive industry may build the best sales volume forecast model. Since the rise of the Internet, people no longer rely solely on traditional media to transmit and share information. Every kind of information is now transmitted to consumers through the medium of the Internet. Through the internet consumers also now have a forums, discussion threads, blogs, and self-organized social media promotional groups, wherein people can express their opinions, feelings, experiences, and carry out so-called "word of mouth" advertising.

Kanamori & Kimura (2003) point out that 15% of people worldwide perform searches in social media for discussion of information regarding products and services. Based on Google's own research, 97% of Google users get on the internet every day to search for information, and up to 76% of those who do searches for product related information actually end up buying through the internet. Additionally, InsightXplorer (2010) reported that 24% of Taiwanese people searched the Internet for automobile related information when considering to buy a car.

Car sales prediction is a crucial issue for automotive industry. Huang et al. (2015) has provided a novel trigger model for car sales prediction. Zhang et al. (2017) adopted text mining approach to forecast yearly national car sales. In the past few years some researchers have begun to notice the forecasting capabilities of online automotive sales data, but have rarely seen Internet "word of mouth" utilized in analysis and forecasting (Hulsmann et al., 2012).

The present study utilizes economic indexes, text mining technology, Internet "word of mouth", and Google Trends variables created by key word searches in order to forecast the monthly sales volumes of various makes and models of cars. As such, it is distinct from the rest of the body of research on automotive market sales volume.

The data were taken from the Mobil01 discussion forum, Google Trends, and overall statistics database of the DGBAS (Directorate General of Budget, Accounting and Statistics). The present study utilized: Genetic Algorithm/K-Nearest Neighbor (GA/KNN), the well-known Support Vector Regression (Vapnik, 1995); Classification and Regression Trees (Breiman et al., 1984); Neural Network (Rumelhart & McClelland, 1986); as well as four different types of estimation methods to do comparisons. It also used well-known machine learning to create the best automotive sales volume-forecasting model. The results of the experiment demonstrate that among the four methods mentioned above, the GA/KNN model has the lowest predictive power in terms of Mean Absolute Percentage Error (MAPE).

II. LITERATURE REVIEW

In the modern world, people publish their personal buying experiences on the Internet. The literature has shown that such online feedback is very useful. It can be used to predict the

978-1-5090-5795-5/17/$31.00 ©2017 IEEE

sales volumes of various products or social trends. In recent years many have begun using data analysis technology to aid their research, bringing about successful data analysis that can be used in many different applications.

Further, web opinions on social media and blogs are propagated throughout the web by the online social network, giving rise to the so-called Internet word of mouth phenomenon. 70% of online consumers trust such word of mouth opinion (e-WOM) (Nielsen, 2012), and 76% of Taiwan searches the Internet for information on products and goods based on consumer experience (MIC, 2012).

Internet word of mouth is now the main source for consumers to get product related information, and so numerous companies use it as a marketing tool (Zhu & Zhang, 2010). Many researchers have discussed the relationship between sales volume and Internet word of mouth, Gu et al. (2012). According to research done by Amazon, the amount of good or bad word of mouth directly affects book sales volumes.

Qi et al. (2012) did research on the Chinese tourism industry. They investigated the relationship between Internet word of mouth and hotel booking rates using regression analysis. The results of this research demonstrated hotel booking rates and word of mouth are apparently closely related. Therefore, the tourism industry should place more importance on Internet word of mouth.

Park & Chung (2012) on the other hand discuss how the number of times a comment is shared on Twitter is related to sales volume. They discovered that the number of Twitter shares does indeed affect sales volume. When Cai et al. (2014) researched the effects of internet word of mouth on vendor reputation within the market, they likewise found that the internet played a very important role in shaping perceptions of vendors in terms of customer care, market holdings, business domain, and professionality. Asur & Huberman (2010) analyzed the critical reception of Twitter users toward movies and discovered that the amount of criticism could be used to successfully predict box office sales. Bollen et al. (2011) discuss the apparent relationship between the mood on Twitter and the stock market. They were able to use Twitter tweets to create a mood index, and using the time series method and Self-organizing Fuzzy Neural Network (SOFNN) they constructed a forecasting model for the Dow Jones index.

In recent years an overwhelming amount of research has demonstrated that Internet word of mouth may influences the decision making process of consumers. Therefore, the present study will use text mining to extract consumer sentiment from word of mouth and turn that into a mood index that can be used as an automotive sales forecasting model.

Due to the increasing popularity of the internet, many consumers get product related information through the internet before making purchases in order to better understand the ins and outs of the goods they seek to buy and to be able to comparative shopping. Since some consumers only search for comments but never make any of their own, key words searches are more relevant to individual consumer spending than comments (Vosen & Schmidt, 2011). Google Trends is very popular search tool that suggests related/trending key words. It allows uses to enter key words and then makes real time suggestions for related key word search data.

III. THE RESEARCH METHOD

Apart from taking important economic indices into account, the present study also added Internet automotive discussion boards and customer sentiment indices. Additionally Google Trends website information was included and taken as a variable to do automotive sales forecasting.

First, as shown in Figure 1, the system used Crawler to collect and extract Mobile01 discussion forum comments, Google Trends, and the PC-AXIS overall statistics database of the DGBAS (Directorate General of Budget, Accounting and Statistics).

Second, in doing data pre-processing, the present study calculated the sentiment of commenters on various car makes and models and used relevant key word searches in Google Trends. Specific content and search data was obtained, the data was organized, and the quarterly index was changed into a quarterly average in order to sales data for variable input.

In this research framework, Shift1 is to move all the input variables of the original month one month backward. Shift2 is to move all the input variables of the original month two month backward. Shift3 is to move all the input variables of the original month three month backward. Shift1~3 then combines the shift1, shift2, and shift3.

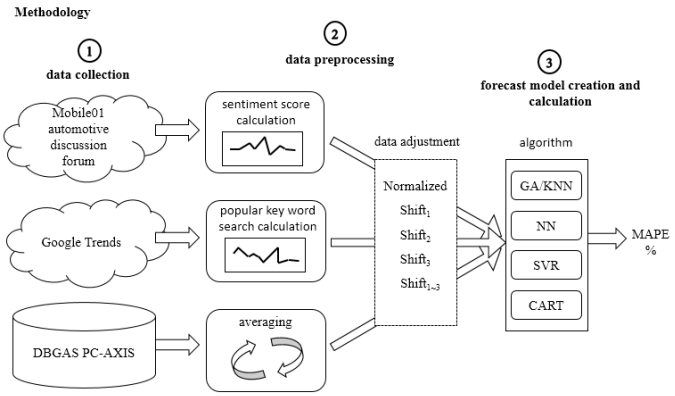

Fig. 1. The research flowchart

A. *The Experimental Data*

Since the online community has begun to flourish, many automotive discussion forums have appeared on Taiwan's internet like the Mobile 01 discussion forum, the Yahoo automotive discussion forum, the Ellie automotive discussion forum, U-Car and the like. Based on the reported findings of the Institute for Information Industry, Taiwanese Internet users often visit the Mobile 01 discussion forum. The content of these discussions also emphasizes automobiles at the consumer level. Other, by gathering Google Trends data one can understand the key word searches relating to Mazda3, Mazda 3, 馬 3, 馬自達 3 Taiwan and internet users around the globe.

B. *Creating and Assessing the Forecasting Model*

Yang & Honavar (1998) suggested the advantage of using GA in identifying a suitable minimum set of variables to improve the classification results requires the use of variable screening. Other, Rawat &Upadhyay (2012) also used the GA/KNN algorithm to analyze customer care related management data bases, and as a result were able to suggest various ways of improving customer retention. The present study proposed the Genetic Algorithm/K-Nearest Neighbor (GA/KNN) to create an automotive sales forecasting model. Furthermore, it also used GA/KNN to do comparative analysis of classification and the following well-known and effective data mining approaches: Neural Network, Support Vector Regression, Classification and Regression Trees.

As shown in Fig. 2, GA/KNN is based on the KNN calculation method. It uses genetic algorithms to calculate each variable with the most accurate weight and to improve the certainty of estimates. The weighted value represents the importance of each variable. The greater the value is the greater the weight. When adjusting the weight, the target GA value is entered to reach the mean absolute percentage error. The fitness function value was calculated as show in the following formula.

Fitness Function

$$\textit{Minimize} \qquad MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \times 100\ \%$$

where A_t is the actual value and F_t is the forecast value of the tth month.

In the creation of initial weight, the initial weight value was set within a range of 0 ~ 1,000. Then, the data matrix multiplied the initial weight, and then data forecasting was done through the estimate algorithm in order to estimate the MAPE the smallest possible percentage of absolute minimal differential average.

To assess the effectiveness of the forecasting models the Mean Absolute Percentage Error (MAPE) was used for evaluation.

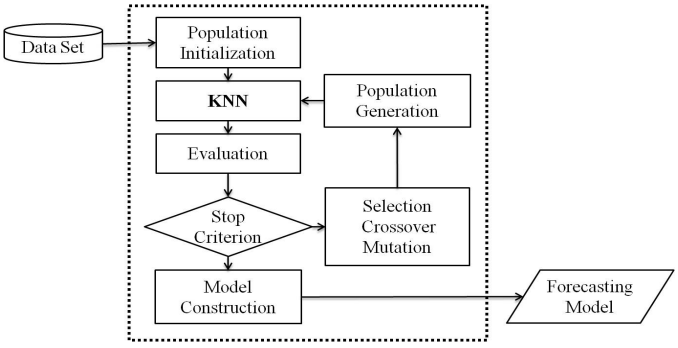

Fig. 2. GA/KNN computation process

IV. The Experiment Results

The present study tried various algorithms. Apart from GA/KNN, it also tried SVR, CART, and NN. Taken together with 49 months of data and 35 variables this made for a six-fold cross-validation. GA's control parameters were set with chromosome rate of 0.5 and mutation rate of 0.1, with chromosome size of 100. The stopping criteria of evolution computation were set when the fitness value improvement is less than 0.001% after 10 generations. When only 1 month of mean shift data was used, the best results of MAPE were 23.86% of GA/KNN (K=3); when only 2 months of mean shift data were used, the best results of MAPE were 26.18% of GA/KNN (K=3); when only 3 months of mean shift data were used, the best results of MAPE were 28.83% of GA/KNN (K=4); and when 1 ~3 months of mean shift data were combined, the best results of MAPE were 25.62% of GA/KNN (K=4). Based on these results, the SVR, CART and NN experimental data was not ideal. During the experiment it was discovered that various indices created interference, and the resulting effect on the sales volume was various. Therefore, this study suggests using the GA/KNN algorithm, the forecasting results of which are shown in Table 1. It is clearly evident that the results produced by the GA/KNN algorithm surpass other methods of calculation. Through GA each input variable is given a different weighted value. Overall, the results indicated for shift1 are the best.

In regards to the mood indices, economic indices, and popular keyword search indices, when doing forecasting it was using the negative and positive Internet automotive discussion forum comments as well as the automotive related keyword search data that had the greatest impact on the sales of the following month.

TABLE I. Experimental Test Results for 6-fold Cross Validation Average Error

Calculation Method	Test Results of MAPE based on 6-fold Average			
	Shift1	Shift2	Shift3	Shift1~3
GA/KNN (K=1)	29.85	39.44	36.65	35.98
GA/KNN (K=2)	26.25	35.35	32.89	31.69
GA/KNN (K=3)	23.86	26.18	31.91	27.11
GA/KNN (K=4)	27.04	30.60	28.83	25.62
GA/KNN (K=5)	28.66	33.10	32.50	28.91
SVR	34.96	34.87	35.41	34.14
CART	33.53	30.59	32.20	29.65
NN	25.27	51.87	40.63	105.77

V. Conclusion and Future Development

Based on the experimental results, we see that the combination of web search data and economic indices can be used to improve the degree of accuracy of forecasting.

Therefore, this chapter addresses the use of indices, the small amount of GA/KNN prediction data, real time forecasting data for decision makers, and the fact that some of the predictions are less accurate for certain months.

The results of the six-fold experiments were that the GA/KNN (K=3) attained the best MAPE, the 15 highest average weighted variables of which are shown in Table 1.

It is apparent that he higher the weight, the greater the influence of the indicated variables on automotive sales forecasting. Furthermore, based on the findings reported in Table 4 the following three sentiment related indices affect automotive sales: sentiment score, the given number of negative comments per instance vs. the total number of negative comments over the course of the study.

Much previous research has used economic indices for forecasting variables for automotive sales. Moreover, the addition of Mobile01 discussion forum comments and Google Trends popular keyword searches will most certainly add to the effectiveness of automotive forecasting data.

The results of the study show that using only one month of data, where k=3, the MAPE of GA/KNN provides the most reliable results at 23.86% as compared to three other calculations methods. It also shows that GA/KNN can be used to create a successful forecasting model. Furthermore, when GA/KNN is applied to sales data predictions, the results are more precise. However, when it is applied to certain months the results are rather dividend from real sales volume. It could be that there were insufficient variables incorporated, or it could be that there were unpredictable factors such as storms, earthquakes, and etc, and that these made for poor prediction results.

The present study on Mobile01 only collected data from automotive discussion forum comments to use as sentiment index values. In the future, it might also be wise to use Facebook group comments as one way to accomplish the goal of this study of creating input variables for forecasting model.

ACKNOWLEDGMENT

This study was supported by the Ministry of Science and Technology, Taiwan, R.O.C., under contract no. This research is supported by MOST 104-2410-H-155-029-MY2

REFERENCES

[1] Bollen, J., Mao, H. and Zeng, X., "Twitter mood predicts the stock market," Journal of Computational Science, Vol.2, Issue 1, pp. 1-8, 2011.

[2] Breiman, L., Freidman,J.H., Olshen, R.A., Stone, C.J., Classification and Regression Trees, Chapman & Hall, New York, 1984.

[3] Cai, Hongbin, Jin, Ginger Zhe, Liu, Chong, Zhou, Li-an., "Seller Reputation: From Word-of-mouth to Centralized Feedback," International Journal of Industrial Organization (2014).

[4] Holland, J.H., Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, 1975.

[5] Hulsmann, M., Borscheid, D., Christoph, M., Friedrichl., and Reith,D., " General Sales Forecast Models for Automobile Markets and their Analysis," Transactions on Machine Learning and Data Mining, Vol. 5, No. 2, pp. 65-86, 2012.

[6] Huang, W. , Zhang, Q., Xu, W., Fu, H., Wang, M., Liang, X., "A Novel Trigger Model for Sales Prediction with Data Mining Techniques," Data Science Journal, Art. 2015, pp. 1-8.

[7] InsightXplorer Report, URL: www.ixresearch.com/reports/, Data Accessed: 2011/10/2.

[8] Kanamori, T., Kimura, A., "Net Communities in Brand Marketing," NRI Paper, No. 63, pp. 1-10, 2003.

[9] Liu, Y., "Word-of-Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," Journal of Marketing, Vol. 70, No. 3, pp. 74-89, 2006.

[10] Liu, Y., Huang, X., An, A., Yu, X., "ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs," The 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 607-614, 2007.

[11] Liu, Y., Chen, Y., Lusch, R.F., Chen, H., Zimbra, D., Zeng, S., "User-Generated content on social media: Predicting Market Success with Online Word-of-Mouth," IEEE Intelligent Systems, Vol. 25, No. 1, pp. 75-78, 2010.

[12] Rawat, M.K., Upadhyay, D.C., " Cluster Detection Using GA-KNN Conjunction Approach," *Journal of Global Research in Computer Science*, Vol. 3, No. 5, pp. 7-10, 2012.

[13] Rumelhart, F., McClelland, J.L., Parallel Distributed Processing : Explorations in the Microstructure of Cognition, MIT Press Cambridge, USA, 1986.

[14] Vosen, S., Schmidt, T., " Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends," Journal of Forecasting, Vol.30, Issue 6, pp. 565-578, 2011.

[15] Zheng, Xin, et al. "Research on Forecasting the Car Sales in Mainland of China Based on Improved Particle Swarm Optimization Algorithm and Gray Model." Information-An International Interdisciplinary Journal 15.4 (2012): 1461-1476.

[16] U-Car, 「 Taiwan Car Sales Report 」, http://www.u-car.com.tw/search.asp?keywords=%E8%87%BA%E7%81%A3%E6%B1%BD%E8%BB%8A%E5%B8%82%E5%A0%B4%E9%8A%B7%E5%94%AE%E5%A0%B1%E5%91%8A, Data accessed 2012/8/1

[17] Yang, J.H., Honavar, V., " Feature Subset Selection Using a Genetic Algorithm," *IEEE Intelligent Systems*, Vol. 13, No. 2, pp. 44-49, 1998.

[18] Zhang, Q., Zhan, H., Yua, J., "Car Sales Analysis Based On the Application of Big Data," Procedia Computer Science, Vol. 107, 2017, pp. 436 – 441.