Preliminary study of tourist experience on theme parks using big data mining techniques

Jing Han* and Kaige Zhang**

*Department of Journalism, Jinzhong University, Jinzhong, China[1]

**Department of Computer Science, University of Minnesota - Twin Cities, MN, USA[2]

Abstract

Theme park industry has created a market yielding tens of billions of dollars each year in the United States and the market size is expanding at a compound annual growth rate (CAGR) of 5.8%. The top 10 theme park corporations have seen an annual attendance of over five hundred million visitors which can generate a huge amount of costumer experience data that could be used to advance the upgrading of park infrastructures and the related management systems, and direct the future development plans. In this paper, we present a preliminary study of collecting and exploring tourists experience feedback of theme parks using data mining technology. Taking one of the most popular social media apps, Twitter, as the example, we show how to collect the relevant user data from social media technically in a computer science manner. In the aspect of acknowledging user personalities, we present the data distribution according to specific user attributes of interest, such as age, gender, dates, etc. To aggregate the semantic structure of the user data, we introduced machine learning technique and the knowledge in nature language processing to study the visitor sentiment based on the user comments data on social media. The experimental results and methodological discussions demonstrate the significance of using data mining technology to guide the developing of theme park industry.

Keywords: *Theme park, Amusement park, Data mining, Social media, Sentiment analysis.*

1. Introduction

Amusement parks such as theme parks has been a big, profitable industry all over the world, making hundreds of billions of dollars each year,

1 Jing Han is an Assistant Professor at Jinzhong University. Email: hanjing9c@163.com, corresponding author

2 Kaige Zhang is a Postdoctoral Research Associate at the University of Minnesota Twin Cities. Email: zhan7704@umn.edu

servicing hundreds of millions of visitors and generating a huge amount of user data every year. The ability to make use of such data can improve the park management efficiency, advance the infrastructure upgrade, and develop new ideas for novel entertainment projects with promising profit feedbacks. With the recent development of BIG DATA technologies, data mining and data analysis have been used for amusement park data exploration to advance the development of the industry. The former works mainly focused on using the data collected inside the theme park to explore visitors' dynamics. Wood et al. (2011) utilized data mining method to study bike rental data near the theme parks to gain insight of the utilization of rental infrastructure which direct the budget plan.

Ferreira et al. (2013) and Chu et al. (2014) made use of the taxi trajectory data to get acknowledge about the information of the people's mobility. Beardsley et al. (2016) developed a system to track visitors' trajectories in Disney; there, the computer vision technique is utilized to take images of visitors' shoes and built an image matching system for identity localization. Under such scenario, IEEE hold a competition called The Visual Analytics Science and Technology (VAST) Challenge in 2015 where the related data were collected near DinoFun World, a typical modest-sized amusement park which sited on about 215 hectares and hosted thousands of visitors each day. Based on the VAST-2015 data, Ye et al. (2015) performed visitors' behavior study via tracking the localization records and rebuilt the trajectory traces; the authors also implemented a collaborative visualization system for their work. Similarly, a visualization system named ParkVis was developed by Puri et al. (2015) to detect unusual events in the park which could be used as an alarm to provide instant information to the park patrons. Steptoe et al. (2017) designed a system to characterize and visualize behavior of crowds and infrastructure usage in DinoFun World. Other works also tried to make use of the data analytic technology to benefit the theme park industry (Andrienko et al. 2013; Chen et al. 2016; Silva et al. 2016; Ma et al. 2019).

While the aforementioned works have paid efforts mainly on the data collected inside the amusement parks, the data were unable to reflect peoples' tourist experience and cannot provide objective feedbacks about the infrastructure from the customers' point of views, which we think are very important information to guide the development plan for the theme park industry. Based on such consideration, we propose to make use of the BIG DATA technology to mine the social media data that are related to amusement/theme parks to abstain direct experiential feedback from the visitors' opinion. Data mining from social media has been a mature technique and has been used widely in the data science community to explore user personality. Amid the works in such area, sentiment analysis

has been one of the most important and effective method to provide feedback of user experience for products and services. Liu et al. (2010) performed a review on sentiment analysis which discussed the strong and the weak points of the traditional methods on sentiment analysis. Zhao et al. (2018) performed sentiment analysis by using a word embedding method, where unsupervised machine learning method was introduced based on the platform from large Twitter corpora. As a popular research direction, there has also been open source software packages which provide well-trained model by utilizing large scale data from the social media (Joshi et al. 2018; Zhang et al. 2019). We will utilize the off-the-shelf model to obtain the sentimental information for the analysis.

In terms of the task related to theme park tourists experience analysis in this paper, we perform the sentiment analysis by using the publically available software package which provide the pre-trained model that can be used to sort the text message into happy, neutral, or sad of which we will further focus on digging deeper into those "sad" information to find out the factors that resulted in the customers' bad experience and try to find solution to address the potential problems and provide a better user friendly environment. We also present the data distribution according to specific user attributes of interest, such as age, gender, dates, etc. to give out an overview of the user's personality which could be used to guide visitor management facilities.

2. Method

In this section, we first present how to collect data of specific interest according to setting up the favorite key words from the social media app Twitter. Then we discuss sentiment analysis using the collected user comments data based on a machine learning method. Figure 1 is a flowchart for the method.

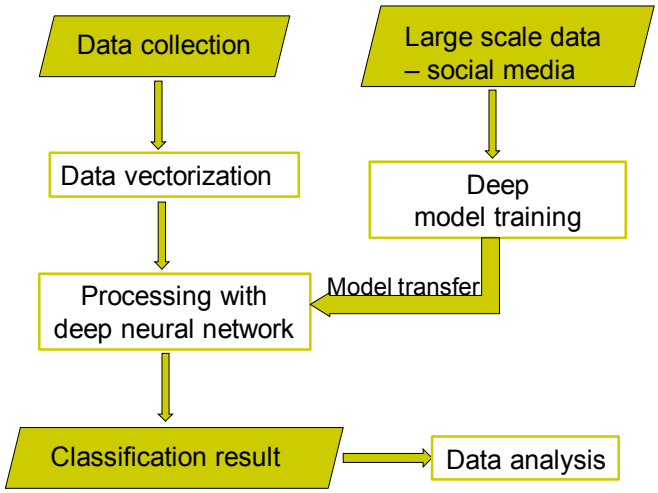

Figure 1. Flow chart of the proposed method

2.1 Data Collection from Social Media

2.1.1 Data collection from Twitter

Twitter provide the interface for accessing user data for qualified developers of interest. To that end, open the Twitter Apps website, sign in (or create an account if you need to) and then click on the 'Create new app' button to create an API key. Then click the 'Keys and Access Tokens' tab to get: consumer key, consumer secret, access key and access secret. Once you got the access key and the passcode, you can make use of those public code to download the data at your preference by setting proper key words to ask the computer to search accordingly. Here, the key words are "Theme, Parks, Amusement." More guidelines can be found on the official website: https://developer.twitter.com/en.

2.1.2 Data collection from Facebook

Facebook is another popular social media app where people could share their opinions by making comments and posting status of their lives. Performing data mining from Facebook has been another popular activity in the data science community, and Facebook provides the relevant tool for access the data freely. The Facebook API is also known as Graph API. The Graph API is an interface with REST (Representational State Transfer). It implies that Facebook calls functions by using remote methods, like HTTP, GET, POST to send messages and echo back REST service. In order to use this service, we need to register the app from the Facebook development website. Then go to the "Myapp" menu and click to the relevant app you

are going to develop, which will require you to provide your ID and the passcode to move forward. Then supposed you are using Python for the program development, you will need to first download the software package "facebook-sdk", and then set up your token accordingly. Once the environment is set up, you can communicate with the Facebook graph API and set up the key words for the data download, similar as the way discussed at last subsection. We use the same key words to collect the comments related to theme parks.

In this work, we collected 1,200 related comments. The distributions are listed in Table 1 according to user gender, ages.

2.2 Sentiment Analysis with Machine Learning

2.2.1 Data vectorization

The raw text data cannot be processed directly by computer. The goal of this data vectorization is to find the high-quality word vectors to represent the comments obtained from the social media apps by using the data loading method discussed in last section. To achieve this, we adopt the publically available model directly based on the transfer learning theory (Pan et al. 2010). Transfer learning has been widely used for training deep neural networks which intends to transfer knowledge learned in previous tasks to alleviate the training.

Depending on situations, there are different transfer learning strategies according to "what knowledge to transfer" and "how to transfer the knowledge". Yosinski et al. (2014) discussed the knowledge transferability of different layers in deep neural networks. Oquab et al. (2014) transferred the mid-level knowledge for nature image processing. Zhang et al. (2018) transferred the generic knowledge learned from ImageNet (Deng et al. 2009) to ease the training of a crack detection network.

Since we are focusing on a specific topic on theme park comments analysis, we are not interested in building a state-of-the-art model for words representation learning, but we do need a model to translate the raw text words into the most representative format for machine learning. Firstly, a well-built model need to be trained from huge data sets with billions of words, and with millions of words in the vocabulary. In this work, the model is obtained based on transfer learning that makes use of the well-trained model knowledge for sentiment estimation/classification. We first use Word2Vec (Mikolov et al. 2013) to translate the text data into a vector where a shallow neural network was used to produce the word embedding represented by a one-dimensional vector. Then the method measures the quality of the translated representations from the raw comments words, and

tuning the network so that similar comments will tend to have a close similarity difference with the other; however, the words can have a similarity degree with difference scales. A deep neural network was trained to realize this function with the help of the publically available resources. Afterwards, the vectors representing original comments point-by-point are put into the well-trained neural network for sentiment classification as described in next subsection. The Word2Vec algorithm provides different models with the transformation input of 25, 50, 100, 200 and 300 based on 2, 6, 42, 840 billion tokens. Here, we use the 100 vectorization output to translate the raw data into feature representation.

Table 1 Settings of the neural network for sentiment classification

Property / Ser. #	Layer type	# of input	# of output
1	Fully connected	100	1024
2	Relu	1024	1024
3	Fully connected	1024	2048
4	Relu	2048	2048
5	Fully connected	2048	4096
6	Relu	4096	4096
7	Fully connected	4096	2048
8	Relu	2048	2048
9	Fully connected	2048	2048
10	SoftMax	2048	3

2.2.2 Classification with pre-trained model

For the sentiment analysis, we use Cloud Natural Language API which provides a powerful artificial neural network that is well trained on a huge amount of data collected from internet world widely. The data were downloaded by using the open software package which could search the internet through pre-defined settings, such as key words. It can filter the unnecessary stuff and reserved the needed the information accordingly. Google web browser was used to collected the large scale data directly from the website with the help of the powerful google search engine and data base.

After the big data was collected, the deep learning model from the google cloud was used to train the model. Before that, the data was aligned to the same format and transferred mathematically same in the dimension so that they can be used as training samples directly by the neural network. By using the big data technique, a model can be trained to classify the

semantic meaning of the samples according to the pre-defined labels. In this work, we are not supposed to complete this task since it is not our target. We borrow the well-trained model from google cloud and used the model directly to classify the vectored sample comments. The pre-trained model can be used directly to classify a sentence (in the form of a vector by the Word2Vect algorithm) into positive, negative, and neutral. The results are present in Figure 2 in terms of the distributions on different ages.

As shown in Table 1, we adapt the 100 vectorization model and the input dimension from the first layer is 100. Since the data is sequence data, the neural network is composed mainly with fully connected layers, and the rectified linear unit is as the activation layer to introduce the non-linearity to model the complex mapping relationship.

3. Results

In this section, we provide the related results for data analysis. First, the we present the user distribution of the collected data according to the age and gender. For the gender information, the male users take 47.08% of the total users, and the female users are with 52.92%. The female users are 5 percentages higher than the males, probably because that the female are more likely to take care of their children which are the main population visiting the theme parks. For the age distribution, those under 16 have 96 which is a relative small number; it is probably because the population of the users of social media are adult people. The population between 26-43 occupies the main part of the users, and this is understandable because they could be the parents of most children at the ages to visit theme parks.

For the sentiment classification result, most people are with positive feedback which shows that the people enjoy the visiting of the theme park. Neutral also occupied a large population that shows us that many people may just talk about things related to theme park. And there are comments with negative sentimental results and those user comments are the ones we should take seriously. Figure 3 present the distribution of the comments from January 2019 to August 2020. We could see that the theme park visiting peak appears around August 2019 which is a summer time; however, the comments are few after March 2020 which probably because of the COVID-19.

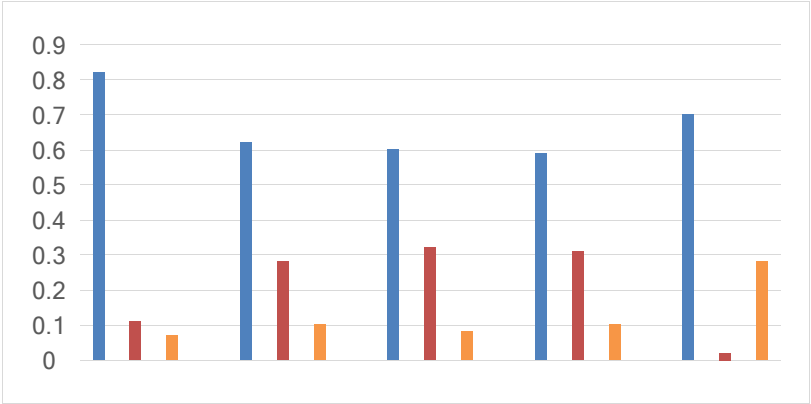

Figure 2. Sentiment proportional distribution of different ages: blue is positive/happy, red is neutral, and orange is negative/sad.

Table 2 Data distribution of age and gender

Category	Number	%
Gender		
Male	565	47.08
Female	635	52.92
Age		
<16	96	8.00
16-25	102	8.50
26-35	560	46.67
36-43	412	34.33
>43	30	2.50

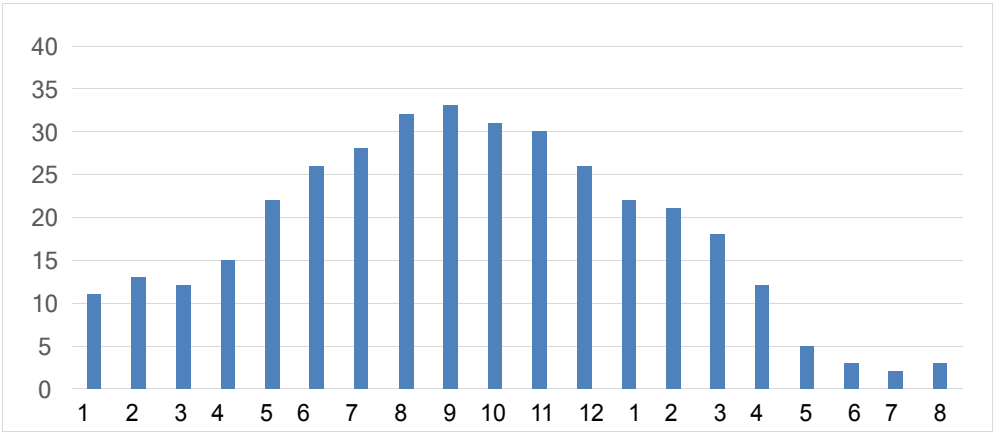

Figure 3 Time of posted comments by month

4. Conclusion

In this paper, we conducted a preliminary study for tourist experience analysis on theme parks from a novel data mining perspective based on social media comments. It provides a new idea for collecting theme park-relevant information that can be used to advance the theme park industry development with data mining. The method introduced machine learning technology to perform sentimental analysis to show how satisfactory people are with the current infrastructures, and what the people's concerns are. The experimental results indicate that people from different ages has different sentimental opinions on their tourist experience which is worth to be further explored to figure out the reasons behind it, and that could be used to improve the infrastructure or the management systems at the users' favours. The time of the posts provides the information that could be used to explore how to improve tourist experience at different seasons. More detailed data analysis based on the results of this paper will be further performed to provide useful information to guide and improve the development quality of theme park industry.

References

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. In Proceedings of ACL 2019.

Joshi, P. (2019). Comprehensive hands on guide to Twitter sentiment analysis with dataset and code. Retrieved Oct. 26, 2020, from https://www.analyticsvidhya.com/blog/2018/07

Zhao et al. (2018). Deep convolution neural networks for twitter sentiment analysis. IEEE Access.

Liu, B. (2010) Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing.

Wood, J., Slingsby, A., and Dykes, J. (2011) Visualizing the dynamics of London's bicycle-hire scheme. Cartographica 46, 4 (2011), 239–251.

Ferreira, N., Poco, J., Freire, J., and Silva, C.T. (2013). Visual exploration of big spatio-temporal urban data: A study of New York city taxi trips. IEEE Transactions on Visualization and Computer Graphics, 19 (12), 2149–2158.

Chu, D., Sheets, D. A., Zhao, Y., Wu, Y., Yang, J., Zheng, M., and Chen, G. (2014). Visualizing hidden themes of taxi movement with semantic transformation. In IEEE Pacific Visualization
Symposium, 137–144.

Beardsley, P., and Taneja, A. (2016). System and method using foot recognition to create a customized guest experience, 19 (07).

Ye, T., Hao, Y., Wang, Z., Lai, C., Chen, S., and Yuan, X. (2015). Behavior analysis through collaborative visual exploration on trajectory data. In Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on, 131–132.

Puri, A., Liu, D., Chen, S., Fu, S., Wang, T., Chan, Y., and Qu, H. (2015). ParkVis: A visual analytic system for anomaly detection in DinoFun World. In 2015 IEEE Conference on Visual Analytics Science and Technology (VAST), 123–124.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In Proc. NIPS, Montreal, Canada.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations. In Proc. IEEE CVPR, New York.

Steptoe, M., Kruger, R., Garcia, R., Liang, X., and Maciejewski, R. (2017). A visual analytics framework for exploring theme park dynamics. In 2015 IEEE Conference on Visual Analytics Science and Technology (VAST).

Ma, X., Zhang, K., Yang, X. (2019). Pose-invariant facial expression recognition based on 3D face morphable model and domain adversarial learning. In proceedings of ICIG 2019, LNCS.

Andrienko, G., Andrienko, N., and Wrobel, S. (2013). Visual analytics of movement. Springer.

Mikolov et al. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781

Chen, S., Yuan, X., Wang, Z., and Zhang, J. (2016). Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. IEEE Transactions on Visualization and Computer Graphics, 1 (2016), 270–279.

Silva et al. (2016). Online clustering of trajectory data stream. In IEEE International Conference on Mobile Data Management (MDM), Vol. 1. 112–121.

Zhang, K., Cheng, H. D., and Zhang, B. (2018). Unified approach to pavement crack and sealed crack detection using pre-classification based on transfer learning. J. Comput. Civil Eng., vol. 32, no. 2 (04018001).

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. IEEE Trans. on Knowl. Data Eng., vol. 22, no. 10, pp. 1345-1359.

Deng, J., Dong, W., Socher, R., and Li, F. F. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In Proc. IEEE CVPR.

Google Cloud. (2012). Natural language API basics. Retrieved Oct. 2020 https://cloud.google.com