World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:11, No:1, 2017

Application of Data Mining Techniques for Tourism Knowledge Discovery

Teklu Urgessa, Wookjae Maeng, Joong Seek Lee

***Abstract*—**Application of five implementations of three data mining classification techniques was experimented for extracting important insights from tourism data. The aim was to find out the best performing algorithm among the compared ones for tourism knowledge discovery. Knowledge discovery process from data was used as a process model. 10-fold cross validation method is used for testing purpose. Various data preprocessing activities were performed to get the final dataset for model building. Classification models of the selected algorithms were built with different scenarios on the preprocessed dataset. The outperformed algorithm tourism dataset was Random Forest (76%) before applying information gain based attribute selection and J48 (C4.5) (75%) after selection of top relevant attributes to the class (target) attribute. In terms of time for model building, attribute selection improves the efficiency of all algorithms. Artificial Neural Network (multilayer perceptron) showed the highest improvement (90%). The rules extracted from the decision tree model are presented, which showed intricate, non-trivial knowledge/insight that would otherwise not be discovered by simple statistical analysis with mediocre accuracy of the machine using classification algorithms.

***Keywords*—**Classification algorithms; data mining; tourism; knowledge discovery.

I. Introduction

DATA mining can be defined as the process of accessing, selecting, exploring, and modeling large amount of data to uncover previously unknown patterns that are potentially useful [1]. Classification functionality is a data mining functionality that assigns dataset instances in a collection to target categories or classes or target attribute values [2]. The goal of classification technique is to accurately predict the target class for each case in the dataset. There are different classification algorithms that would be appropriate for particular domain with varying degree of accuracy of their performance on particular domain [3], and hence, rigorous research is required for the applicability in each domain. At the same time, not all attributes (features) in the dataset are equally relevant for the classification purpose. Some of attributes just consume the computational resources for some of the algorithms. These attributes might be considered as noises or overfitting for machine to learn from training dataset [4]. There are some algorithms that tolerate noise and these ones are greedier than the others [5]. Thus, it is important to check different scenarios and algorithms for a particular domain. One of the scenarios used in this paper is to apply the information gain-based attribute selection method and to build model for the compared algorithms before and after selection of the attributes. This helps to select or to pick top relevant attributes to target attribute for classification. It is also important to find out which algorithm is more noise tolerant than the others to show better performance for a given noisy dataset. Therefore, this research is framed in such a way that classification models are built using the selected algorithms before and after attribute selection based on information gain to compare the performance of each scenario. The techniques selected for comparison are Decision Tree and Support Vector Machine (SVM). From decision tree category, C4.5 (J48 in Weka), Random Forest, and Projective Adaptive Resonance Theory (PART) were experimented. From Artificial Neural Network, Multilayer Perceptron (MLP) model is experimented. From SVM, Sequential Minimal Optimization (SMO in Weka) implementation is built. All implementations were experimented before and after attribute selection based information gain. The algorithm best performance is reported before and after attribute selection.

The models were built on the tourism data with the aim of finding out the noise tolerant classification algorithm to be applied for knowledge discovery in the domain. Several emerging applications in information-providing services call for various data mining techniques to better understand user behavior, to improve the service provided, and to increase business opportunities [6]. Tourism which is one of these domains is becoming important sector in every nation's national Growth Domestic Product (GDP) [7]. Tourist arrivals in South Korea averaged 582848 from 1993 until 2016 [8]. Tourism service is data intensive with complex relationships of the attributes paralleled with the increasing volume of tourism service [9]. Better decision making and service improvement needs to be supported by the insights and knowledge discovered from the service data or obtained from the visitors through survey mechanisms. On the other hand, only statistical analysis cannot reveal non-trivial insight as a result of intricate relationship among the features/attributes in tourism data. Therefore, this complex relation and voluminous nature of tourism data calls for necessitates use of state of the art data analytics methods, tools, and techniques beyond simple statistics analysis [10]. This research aimed at applying data mining classification algorithms to explore the potential benefits in discovering hidden knowledge from tourism data. It used various algorithms and scenarios to come up with the findings as an insight for decision making purpose in the

Teklu Urgessa is with the Seoul National University, Graduate School of Convergence Science and Technology, Suwon, South Korea, Phone: +8226886659; fax: 82-31-888-9148; e-mail: tek2013@ snu.ac.kr).

Wookjae Maeng and Jooong Seek Lee are with the Seoul National University, Graduate School of Convergence Science and Technology, Suwon, South Korea (e-mail: ideaplayer@snu.c.kr, joonlee8@snu. c.kr).

domain sector. It is also a typical interdisciplinary research that can be a corner stone pertaining to information technology research in the domain of tourism. The data mining research itself involves machine learning, statistics, computer science, information theory, and domain implication working knowledge.

II. Related Works

Bach et al. [11] indicated that forecasting, personalization's tourism management, tourism systems (such as recommendation systems), and machine learning techniques such as support vector, regression, multi agent systems, particle swarm optimization are the common research areas in tourism data mining research. Their finding was based on keyword analysis of existing researches. Our research is a continuation based on recommendation in the sense that we focused on empirical knowledge discovery process from tourism data using classification algorithms. Olmeda and Sheldon [12] have published a paper on "Data Mining Techniques and Applications for Tourism Internet Marketing". In their analysis, the potential uses of data mining techniques and technologies in tourism internet marketing and electronic customer relationship management were discussed. According to Olmeda and Sheldon, tourism industry provides the consumer with experiences, and those experiences need increasingly to be customized. In their research, they concluded that data mining techniques can provide tools to discover insight for customization of user experiences based on literature survey on data mining research papers on travel industry data. The paper provided background evidence for business understanding in our work. However, our work is an empirical research, not only literature review and to fill the research gap that they pointed out. Bose [13] has published a paper entitled "Data Mining in Tourism" with the objective of researching and discussing the data mining techniques applicable for tourism using literature review from the perspective of data mining application in the other domains. The techniques discussed in this paper is classification learning, which are the most commonly used machine learning in most data mining researches in another domain. From Bose's paper, we observed the supportive evidence that classification data mining problem should be researched in tourism domain as a recommendation of their research based on the experience of other applications like health care, and banking.

Aghdam et al. [14] have conducted a research on tourism entitled "Finding Interesting Places at Malaysia". The objective of the study was applying data mining techniques in tourism data. Data for their research was obtained or crawled from tripadvisor.com. They made quantitative analysis using Weka and qualitative using Nvivo. They used Cross Industry Standard Process-Data Mining (CRISP-DM). Their paper is similar in using data mining on actual tourism data and in using Weka as a machine learning tool, but they used only dataset related to places and association rule mining alone. We believe that including data of visitors, that include all tour related issues like shopping, travel expenses, service satisfaction and users' own opinions with the visited places, might reveal important insight and hence aimed at including the whole tourist experience which has something to do with knowing the whole tastes and preferences of tourists. The dataset in our case is more comprehensive, and the experiments are rigorous.

We tried to fill the research gap identified through review of the related works as acknowledged by other researchers to be explored using data mining. We figured out from the review of the related works that there is a research gap where applying data mining classification to discover important knowledge from tourism data is important. Our research is at least new in one aspect as discussed in this section either in data mining problem selection, or dataset considered or data source or algorithms employed or tools used for data mining researches of the previous works in tourism. Hence, we hope that we would contribute in finding out the best classification algorithm and process model for knowledge discovery and shade the light on the potential applicability of data mining classification technique in tourism domain.

III. Methodology

In this section, data source, data description and selection, data mining process model used for the research, classification framework and data mining tool used for the implementation of classification algorithms are briefly explained.

q1	q2b	wq2b	q2c	q2c1	q3a2	mq3a2
3	2	2	2	10	3	3
1	1	1			3	3
1	1	1			3	3
2	1	1			3	3
1	1	1			2	2
1	3	3	2	10	2	2
1	1	1			3	3
1	4	6	12	12	3	3
1	3	3	9	12	3	3
2	1	1			3	3
3	1	1			3	3
3	1	1			3	3
3	2	2	3	10	3	3
1	4	8	11	12	4	4
1	1	1			3	3
1	1	1			3	3
1	2	2	3	10	3	3
1	4	6	12	12	3	3
1	4	10	13	13	3	3
2	1	1			3	3
2	3	3	9	12	3	3
3	4	6	12	12	3	3
1	2	2	9	12	3	3
3	4	4	8	11	4	4
3	4	10	10	12	3	3
2	1	1			3	3

Fig. 1 Sample screenshot original coded data in CSV format

A. Data Source, Description, and Preprocessing

Tourism survey data were accessed and downloaded in CSV file format from Korean Tourism and Culture website [8]. These are open data for research purpose. The dataset contained 12030 instances and 134 attributes before selection and preprocessing. After selection and preprocessing, the attributes were reduced to 56 attributes. Data features categories can be summarized as: type of visiting conditions,

socio-demographic attributes, visiting areas, shopping conditions, expenses related to the visit, satisfaction levels of tourists with different services while visiting, their personal opinion on their likelihood to recommend Korea to others as a tourist destination, the opinion of their likelihood to consider visiting Korea again within three years, the opinion of tourists on the level of change in impression during their current visit.

Different data description, selection and preprocessing activities including statistical summary measures, visualization, detecting outliers and fixing missing values, discretization, and conceptual hierarchy feature generation were performed before modeling the classification models. Finally, 56 attributes and 12030 records were used for model building. The original CSV EXCEL date file is obtained from the source mentioned above as in Fig. 1.

All the original data are coded. Without defining it, no one can say something about it. Therefore, getting the definitions for the codes was necessary. The data definition document and the questionnaire, with which the tourist's survey was conducted, were available by the owner organization (the Korean Tourism and Culture Institute). Based on the description document and the questionnaire, the meaning of each data item is presented as the preliminary business understanding to the research framework. The description of the whole attributes and their values are annexed at the end.

B. Data Mining Process Model, Tools, and Algorithm Selection

The data mining process model selected for this research is Knowledge Discovery in Database (KDD) Process model [15]. As it is indicated in Fig. 2, it has five steps with each step having its own deliverable. The knowledge discovery process used in this research has been modified a little bit by introducing one new step: access to the data source at the very beginning, and merging data processing and data transformation into one step-data processing. The modification is made for fitting to the practical procedures followed for tourism data mining application in this research.

TABLE I
COMPARISON BEFORE SELECTION OF ATTRIBUTES

Implementations	Accuracy (%)	Time (in Minutes)	Rank by Performance
Random Forest	76	7.36	1
SMO	75	105.38	2
J48(C4.5)	72	2.04	3
MLP	71	181.62	4
PART	69	83.55	5

The data source is determined first, then means of data acquisition was sought, and we found that it was free and open for research by the institute. There was no challenge regarding data access. However, the entire data were messy, so careful data selection was needed to come up with target data set. Judgmentally those attributes or features with 10% and more of its values were missing were excluded, and then, we left with 56 attributes out of 134 attributes. So, our target dataset selected for mining is 56. There was further preprocessing on the target data too. The data preprocessing activities include making statistical summary measures for each attribute to see the distribution of their values, fixing missing values, discretization of large distinct values, e.g. in age attribute. Transforming the data was performed using conceptual hierarchy generations and discretization through binning methods for continuous attributes like average expenses of visitors, number of day. Changing the data type into the one which can be handled by the mining algorithm and tools like J48 need nominal class labels. The code values 1 to 5 Likert scale was considered as the numeric values by Weka, so it needed to change the data type from numeric to nominal using the NumericNominal feature of Weka on the fly.

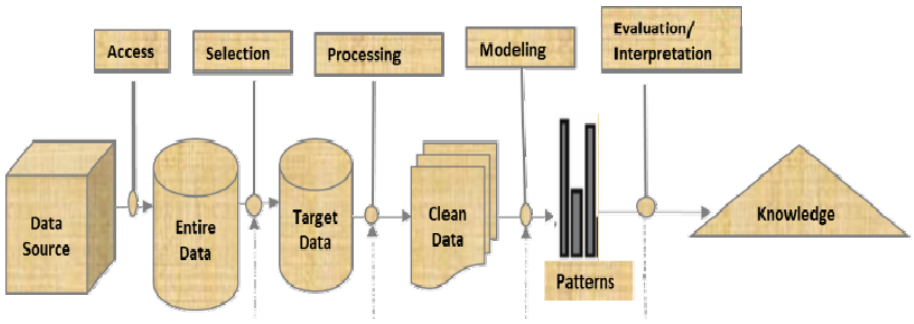

Fig. 2 The process of data mining applications in tourism

The target attribute for classification is opinion of visitors of their likelihood to recommend Korea to others as a tourist destination. This target attribute has values (1. very unlikely, 2. unlikely, 3. Neutral, 4. Likely, and 5. Very likely). The rest of the 55 attributes are independent variables (predictors). The specific attributes or also called features in the dataset include nationality, sex, age, education, touring condition, number of visits in Korea, number of companion, expenses, places visited, shopping places, shopped items, and so on. The high level categories of independent attributes or predictors are presented in Table I.

The machine learning software selected for this research is Weka. Weka is a well-tested and most commonly used open source machine learning tool for general purpose data mining researches [16]. It is a collection of machine learning algorithms for solving real world data mining problems. It is written in Java and runs on almost any platform [17].

Classification algorithms selected for comparison in this

research are C4.5 (J48 in Weka), Random Forest, SVM (SMO in Weka), PART, and MLP as mentioned in the introduction part. These algorithms are described in section III in detail with examples and visual illustrations.

After final dataset was ready for model building, five algorithms of classification mining were built both before and after information gain attribute selection and after selection and performance of each algorithm in each scenario is compared. 10-fold cross validation was used for testing. Accuracy measures based on confusion matrix were reported in terms of correctly classified instances, and other detailed accuracy measures like Recall, Precision, F-measure, ROC area, PRO area, etc.

Finally, the sample tree structure is illustrated and same sample rules from decision tree rules were extracted as a showcase of knowledge discovered, which otherwise would not have been possible without data mining.

IV. Experiments and Results

As stated in the introduction and methodology sections, the aim of this research is to compare the performance of the selected classification algorithms implementations in Weka on both the entire data dataset and the selected top 10 relevant attributes based on information gain. Therefore, in this section, summary of the classification algorithms models and their analysis results are presented in Tables I and II.

TABLE II
Top 10 Relevant Attributes Based on Information Gain

S No	Attribute code original data	Meaning of the code	Rank
1	q19	the interest of the visitor to visit Korea again within the coming three years	1
2	q16b	the overall satisfaction level of a visitor	2
3	q21	change of impression during the current visit	3
4	q1606	satisfaction level by appeal of tourist spot	4
5	q1607	satisfaction level of a tourist with tourist information services	5
6	q1604	satisfaction level of a tourist with food	6
7	q1610	satisfaction level of a tourist with a security services	7
8	q1605	satisfaction level of a tourist with shopping service	8
9	q1602	satisfaction level of a tourist with public transportation service	9
10	q1609	satisfaction level of a tourist with travel expense	10
11	q20	tendency to recommend Korea to others as a tourist destination	

As it can be seen from Table I, Random Forest outperforms (76%) the rest on the entire attributes. SMO is the second best in scenario with 75%. In terms of training time, C4.5 (J48) is the fastest algorithm, and ANN (Multilayer perceptron) is the slowest to build the model.

The detailed accuracy measures for these models before attribute selection is presented in Fig. 3 Even though the difference in other measures is a little bit blared, the difference in the ROC area and PRC area is more visible. That Random Forest is at the top and followed by J48 in terms of ROC and PRC areas. In terms of the weighted average recall, MLP and J48 outperformed others with (75.3% and 75.2%, respectively).

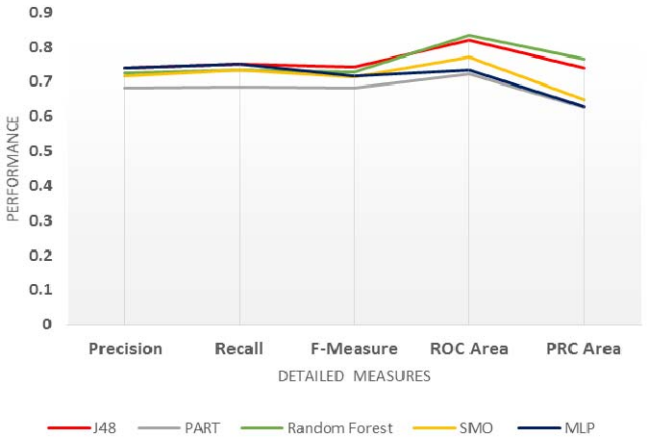

Fig. 3 Comparison of algorithms by detailed accuracy measures before attribute selection

It is visible from Fig. 4 that Random Forest is higher in ROC curve than others in the models built on the entire final data sets. It is of course in every aspect that Random Forest outperformed the other models.

To see whether selecting attributes with more information gains affect the performance algorithms or not, top ten attributes were selected. Those models were rebuilt with similar properties with the selected attributes. The top ten selected attributes are as presented in Table II. The top attribute is the opinion of the visitors on the likelihood to visit Korea again. The second ranked attribute is the overall satisfaction level of the visitors.

TABLE III
Comparison Algorithms After Attribute Selection

Algorithms	Performance (%)	Time (in minutes)	Rank by Performance
J48(C4.5)	75	1	1
Random Forest	74	5	2
SMO	74	3.68	2
PART	73	6	4
MLP	73	17.71	4

The performance result of the algorithms after attribute selection is presented in Table III. From the table, it is clear that attribute selection based on information gain improved the performance of C4.5 (J48) from 72% to 75%, PART from 69% to 73% and MLP from 71% to 73% but degrade the performance of Random forest from 76% to 74% and SMO from 75% to 74%. This indicates that entropy based information gain attribute ranking does not help to improve the performance of Random Forest and SVM (SMO). In terms of training time, all of the implementations showed decrease but for MLP, it is big (90%). This shows that Random Forest and SMO perform better respectively in rank on noisy data than the others in the comparison.

The detailed accuracy of the selected implementations after

attribute selection is presented in Fig. 4 In terms of average weighted precision and recall, random forest and SMO have higher scores respectively. In terms of ROC and PRC area values, random forest and MLP outperformed others, even though random forest was much higher, as first and second, respectively, with values of 0.83 and 0.82 for ROC area and 0.765 and 0.74 for PRC area. The least in these measures is PART. Knowledge/insight hidden within the dataset, which cannot be discovered using simple statistical analysis, was revealed.

The limitation of this study is that the rules were not evaluated by experts in the domain. In practical sense, rules should be evaluated by the experts in the domain for decision making purposes.

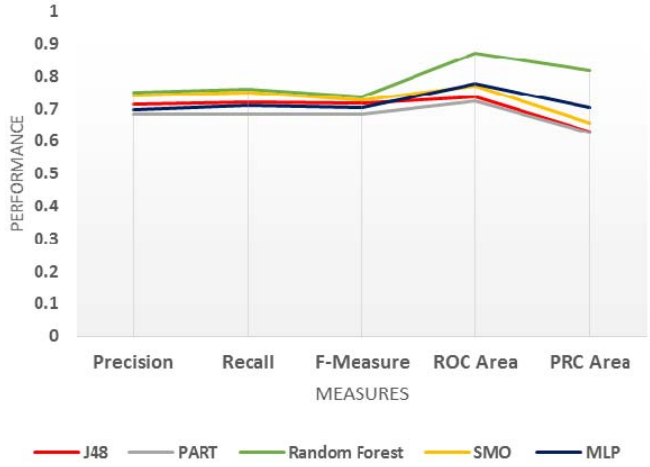

Fig. 4 Detailed accuracy of the algorithms after attribute selection

Let us see some of the rule generated from this decision tree based on the principle of decision tree interpretation using *IF condition THEN outcome* starting from the root node to the decision leaf.

- Rule#1: IF interest of a visitor to visit Korea within the coming three years (q19) is "Highly likely" and IF impression of the visitor during current visit(q21) is "feeling much better" THEN it is "highly likely" that a particular visitor recommends Korea as a tourist destination
- Rule#2: IF interest of a visitor to visit Korea within the coming three years (q19) is "likely" THEN it is "likely" that a particular visitor recommends Korea as a tourist destination.
- Rule#3: IF interest of a visitor to visit Korea within the coming three years (q19) is "Very likely" and IF impression of the visitor during current visit (q21) is "Neutral" and IF the over satisfaction with services received is "very satisfactory" THEN it is "highly likely" that a particular visitor recommends Korea as a tourist destination.
- Rule#4: IF interest of a visitor to visit Korea within the coming three years (q19) is "Highly likely" and IF impression of the visitor during current visit (q21) is "feeling better" THEN it is "likely" that a particular visitor recommends Korea as a tourist destination.
- Rule#5: IF interest of a visitor to visit Korea within the coming three years (q19) is "Very likely" and IF impression of the visitor during current visit (q21) is "unlikely" THEN it is "unlikely" that a particular visitor would recommend to others.

To see more specific rules, only at satisfaction level, visited area and demographic information is used to construct the decision tree using J48 algorithm. The following rules were extracted as a sample:

- Rule #6: IF a visitor is "very satisfied" with security service and IF he/she is "very satisfied with travel expenses THEN he/she would 'very likely' to recommend Korea as a tourist destination.
- Rule #7: IF a visitor is "very satisfied" with security service and IF he/she is "satisfied" with travel expenses THEN he/she would be 'Neutral' to recommend Korea as a tourist destination.
- Rule #8: IF a visitor is "very satisfied" with security service and IF he/she is "neutral" with travel expenses and he/she is "satisfied" with the public transport THEN it is "likely" that he/she recommends Korea as a tourist destination to others.
- Rule #9: IF a visitor is "very satisfied" with security service and IF he/she is "neutral" with travel expenses and IF he/she is "Neutral" with the public transport THEN he/she would be "neutral" to recommend Korea as a tourist destination to others.
- Rule #10: IF a visitor is "satisfied" with security service and IF he/she is "Japanese" and IF he/she is "satisfied' with 'Shopping" THEN he/she would be "neutral" to recommend Korea as a tourist destination to others.
- Rule#11: IF a visitor is "satisfied" with Security and IF he/she is "Japanese" and IF he/she is "unsatisfied" with shopping THEN it is "unlikely" that he/she would recommend Korea as a tourist destination to others.
- Rule#12: IF a visitor is "satisfied" with Security and IF he/she is from "Taiwan" and IF he/she is "satisfied" with communication THEN it is "likely" that she/he would recommend Korea as a tourist destination to others.

V. Conclusion

This research showed a clear research method for applying classification algorithm for tourism knowledge discovery on comprehensive survey data. Related works were discussed and research gap is pointed out procedurally. Data mining process model for tourism is proposed. The best performing algorithm is identified from the experiment. Top important factors that determine the opinion of visitors on their likelihood of recommending Korea as a tourist destination were identified as presented in Table II. Five models were built using the selected classification algorithms; namely, J48, Random Forest, PART, SMO, and MLP. The experimental result showed that Random Forest and SVM (SMO) respectively are more noise tolerant than the other algorithms as it showed better performance on entire attributes respectively. After, the

entropy based attribute selection performance of J48 is improved, while that of random forest is degraded. The research clearly showed that it possible to extract useful insights from tourism data with fair level of performance. Decision tree interpretation is presented as indicator of the hidden knowledge that can be discovered using data mining techniques. Those rules can be used as a base for decision making to target the service demands of tourists. The research findings can be taken as important base for the further research in the area to apply more techniques like association rule mining on tourism data. In terms of training time, MLP is the slowest in both cases before and after attribute selection though the improvement in training time for MLP is above 90% after attribute selection. So, MLP needs careful attribute selection for better time efficiency. However, selection of top relevant attributes reduced computational time for all algorithms overall focusing on high level insights. The future work would be applying association rule mining to see any linkage that exists within the dataset, which could be used as a base for building recommendation systems and possibility of service improvement to visitors' expectation. It is believed that the discovered insights will enable the service provider to give priority to what is most important to the visitors based on their demographics and tourism preferences.

REFERENCES

[1] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
[2] Glusman, G., Bahar, A., Sharon, D., Pilpel, Y., White, J., & Lancet, D. (2000). The olfactory receptor gene superfamily: data mining, classification, and nomenclature. Mammalian genome, 11(11), 1016-1023.
[3] Wu, Xindong, et al. (2008). "Top 10 algorithms in data mining." Knowledge and information systems 14.1: 1-37.
[4] Hong, T. P., Kuo, C. S., & Chi, S. C. (2001). Trade-off between computation time and number of rules for fuzzy mining from quantitative data. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 9(05), 587-604.
[5] Caruana, R., & Freitag, D. (1994, July). Greedy Attribute Selection. In ICML (pp. 28-36).
[6] Gupta, A. K., & Gupta, C. (2012). Analyzing Customer Behavior using Data Mining Techniques: Optimizing Relationships with Customer. Management Insight, 6(1).
[7] Rodríguez, I., Williams, A. M., & Hall, C. M. (2014). Tourism innovation policy: Implementation and outcomes. Annals of Tourism Research, 49, 76-93.
[8] South Korea Tourist Arrivals 1993-2016 available at http://www.tradingeconomics.com/south-korea/tourist-arrivals accessed on 2016.08.09
[9] OECD Tourism Trends and Policies 2014 accessed http://www.keepeek.com/Digital-Asset-Management/oecd/industry-and-services/oecd-tourism-trends-and-policies-2014_tour-2014-en#page1 on 2016.08.10
[10] Sabou, M., Onder, I., Brasoveanu, A. M., & Scharl, A. (2016). Towards cross-domain data analytics in tourism: a linked data based approach. Information Technology & Tourism, 16(1), 71-101.
[11] Bach, M. P. (2003, June). Data mining applications in public organizations. In Proceedings of the 25th international conference on information technology interfaces (pp. 211-216).
[12] Olmeda, I., & Sheldon, P. J. (2002). Data mining techniques and applications for tourism Internet marketing. Journal of Travel & Tourism Marketing, 11(2-3), 1-20
[13] Bose, I. (2009). Data Mining in Tourism. Encyclopedia of Information Science And Technology.
[14] Aghdam, A. R., Kamalpour, M., Chen, D., Sim, A. T. H., & Hee, J. M. (2014, August). Identifying places of interest for tourists using knowledge discovery techniques. In Industrial Automation, Information and Communications Technology (IAICT), 2014 International Conference on (pp. 130-134). IEEE.
[15] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.
[16] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016
[17] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.