Creation of Semantic Location Profiles using Bayes, Rule-Based, Trees and Meta Classification Approaches

S.C. Ganiachchi
Faculty of Information Technology,
University of Moratuwa,
Sri Lanka.

S.S. Wijenayake
Faculty of Information Technology,
University of Moratuwa,
Sri Lanka.

J.C. Wijekoon
Faculty of Information Technology,
University of Moratuwa,
Sri Lanka.

S. Ahangama
Faculty of Information Technology,
University of Moratuwa,
Sri Lanka.

***Abstract*— Travel recommender systems are gaining a higher popularity in the society due to their capability of planning trips in a short time period. The challenge of providing the most accurate recommendations has become a complex and difficult task due to the numerous variations in user preferences. In order to provide the most accurate recommendations, it is necessary to consider additional parameters like the prevailing weather condition, which has a direct influence on the user preference to a particular location. Therefore, when providing travel recommendations considering the weather context, it was identified that the ability to correctly identify a location as an indoor or outdoor attraction plays a vital role in improving the accuracy of the recommendations. Considering the millions of locations available in Google Places, it is difficult to manually tag the location status as indoor or outdoor. This paper provides a novel approach to determine the status of a particular location based on the identified attributes of the location. Moreover, the experimental results and the accuracy of the predicted outputs have been discussed by using different classifiers under the Bayes, Rule-Based, Trees and Meta classification approaches.**

Keywords- Travel recommendation, weather context, location status, decision tree, J48, decision stump, zeroR, oneR, Regression, Naïve-bayes, indoor/outdoor classification

I. Introduction

Travel recommender systems are one of the major research topics at the present where researchers aim to increase the accuracy of recommendations by using different parameters. When providing accurate recommendations, traveler's preference to different locations is considered as one of the most important parameters [1]. A traveler's personal preference could depend on many different aspects and determine the most significant aspect would not be straightforward. Especially, prevailing weather condition could impact on the traveler's mood and satisfaction. According to Ettema et al., changes in the weather condition have a direct impact on the traveler preferences related to the transportation mode and satisfaction [2]. Petrovic et al. have revealed that the transportation demand depends on different weather conditions [3].

Therefore, it is clear that the weather has a direct effect on changing people's everyday activity patterns which are inferred such as visited places and the duration it has taken for the visit [4]. It is defined as the prevailing condition observed and the time duration a person stayed at a particular location has a direct association with the weather condition of that location [5].

Depending on the weather condition (rainy or sunny), travelers could visit indoor or outdoor locations. For an example, when it is warm and sunny, people tend to spend the leisure time outdoors [1]. Thus, to provide accurate recommendations based on the prevailing weather conditions, it is important to determine whether a specific location is an indoor or outdoor location.

The rising use of smartphones and low cost access to the internet has allowed finding the current weather condition of any location in any country. This could be included as a parameter to increase the accuracy of travel recommender systems. A location will be recommended as indoor or outdoor locations based on the current weather conditions. Therefore, in this study, a mechanism is developed to determine a particular location's status, (that is, whether it is indoor or outdoor).

One of the best and accurate methods for identifying a location's status is labeling the locations manually as indoor or outdoor. Even though this approach is highly accurate, labeling a large number of tourist locations will be time-consuming and tedious.

As a solution, automation methods are used in location labeling as indoor or outdoor. To apply automation techniques, there should be a mechanism to identify some attributes belonging to the target location and based on those attributes, could predict the status of the location as indoor or outdoor. The ideal solution should be able to handle even new locations. Moreover, the location status prediction should be in higher accuracy with minimum human intervention. Location status prediction (as a binary output) could be performed using different classification techniques.

978-1-5386-2165-3/17/$31.00 ©2017 IEEE

This paper discusses using different classifier models such as Naïve-Bayes classification, Decision Tree classification (J48, Decision Stump), OneR classification, ZeroR classification and Regression classification and the accuracies when predicting the status of a location, indoor or outdoor.

II. Background

Using image processing methods is one of the main approaches that can be used to detect the status of a location. By analyzing an image, it is possible to identify the status of the image location as indoor or outdoor.

According to Yiu [6], indoor and outdoor scenes can be classified using nearest neighbor and support vector machine approaches based on the color information and dominant orientation of an image. However, finding images of each and every location would be difficult. Moreover, a large database is required to store multiple images for representing each location which could be difficult to maintain as it grows in size.

According to the studies of Lipson, the general scene query approach in image processing can be used to detect the status of a location. Here, the scenes are described by the graphs, which represent relationships between the image regions. The considered relationships are consistent with the relative color, spatial location, and high-pass frequency content [7]. However, manual construction of templates for each scene layout is a drawback of the Lipson's approach [9].

Yu has implemented a mechanism by computing vector quantized color histograms for each sub blocks of the image and trained one dimensional Hidden-Markov model along the vertical or horizontal segments of specific scene layouts like mountains, skies and river scenes [8]. However it is identified that one dimensional model is unable to handle the spatial relationships properly [9].

III. Implementation Details

Several implementation steps were followed in this research study to predict the location status.

Firstly, the location data (including location attributes) was extracted from the Google Places API. Secondly, the training data set, test data set and the evaluation data set were generated based on the extracted location data. Thirdly, the classification model was implemented to predict the location's status based on the location attributes. Finally, the test accuracy and evaluation accuracy of Naïve-based classification, Decision Tree classification (J48 and Decision Stump), OneR classification, ZeroR classification and Regression classification were calculated using Weka.

IV. Data Corpus

A. Data Collection

The Google Places API has been used to extract location details. This API gets data from the same database used by the Google Maps and Google+. For each location it provides an array of location types which could be a combination of aquarium, airport, point_of_interest, establishment, amusement_park, aquarium, art_gallery, lodging, bar, food, restaurant, cafe, shopping_mall, place_of_worship, church, health, spa, bowling_alley, casino, park, hindu_temple, mosque, library, liquor_store, movie_theater, museum, night_club, stadium, zoo, etc. Apart from the location types, Google Places API provides other location details including place_id, place name, location rating (if any), location's opening hours (if any), location address etc.

After extracting the location data, the dataset was prepared such that each location was represented using its location types. Subsequently, the location types were used as the attributes to predict the location status.

B. Data Preparation

After the data extraction process, the dataset was properly prepared to train, test and evaluate the classification model. When preparing the dataset, the status of each and every location is manually checked with the location images and it can be considered as one of the unique tasks that have been performed in this paper. At the end of the data preparation, a balanced dataset was prepared for the classification where the dataset was composed of 59% of indoor locations and 41% of outdoor locations.

Furthermore, for each location in the extracted data set, there can be several location types and this number is different from one location to another location. It is important to note that, while some locations have several location types, some locations only have one location type. There can be locations without having a single location type and in the preprocessing step, it is needed to remove that type of data, since the locations' types are been considered as the attributes of the classifier to predict the output as indoor or outdoor.

After the data has been prepared according to the requirement, then the data set was divided into three main sets as the training set, testing set, and the validation set. At the end of the training process, the final model should be able to predict the output i.e. the location status based on the location attributes. In order to have accurate predictions, it is important to generalize the data well. Otherwise, it may lead to overtraining, which could make the classifier unable to predict the outputs correctly for the patterns that may not exist in the training data set. This is known as the bias and variance dilemma and it is important to balance the minimal bias and the minimal variance of a model by splitting the data appropriately. In order to overcome this problem, a common technique that can be used is the cross-validation [10].

In cross-validation, it divides the main data set into two subsets and taking one subset to train the model while leaving the other subset to be used in evaluating the performance of the model later. The main purpose of the cross-validation is to achieve a stable and confident estimate of the model performance [11]. In this paper, it uses the cross validation method to divide the data set into two main subsets randomly where 60% of the data is used to train the model and 40% of the data is used to test and evaluate the model. Once the 40% of the data is separated, it is divided into another two subsets equally using the cross-validation method as the test data set and the validation dataset.

V. Handling the Problem of Detecting the Status of Locations

It is important to handle the problem of detecting the status of a location when the travel recommender systems consider the user preferences based on the weather conditions. In this paper, a fresh and unique approach is introduced to determine the status of a location, whether it is indoor or outdoor, based on the location types that are extracted from the Google Places API. All the different location types that came under the process of the data extraction are considered as the attributes of the locations to predict the final output, which is the status of the location.

Each location extracted from Google Places is tagged with a set of "types" which can explain certain characteristics of that location. These tags are the major features that are assumed to have a direct influence in determining the location's status in this research. Therefore in order to train the classifier, each location is represented as a vector of its location tags along with the manually tagged location status (as indoor or outdoor). The classifier has to be trained well before using it to predict the outputs of testing or validating data.

The approach which is used to determine the status of the locations is directly based on the attributes (location types) which are extracted from the Google Places API. Therefore when a new location is provided to the classifier as a vector of its location "types" the classifier needs to predict it as indoor or outdoor.

The Figure1 illustrates the process model used to predict the status of the locations.

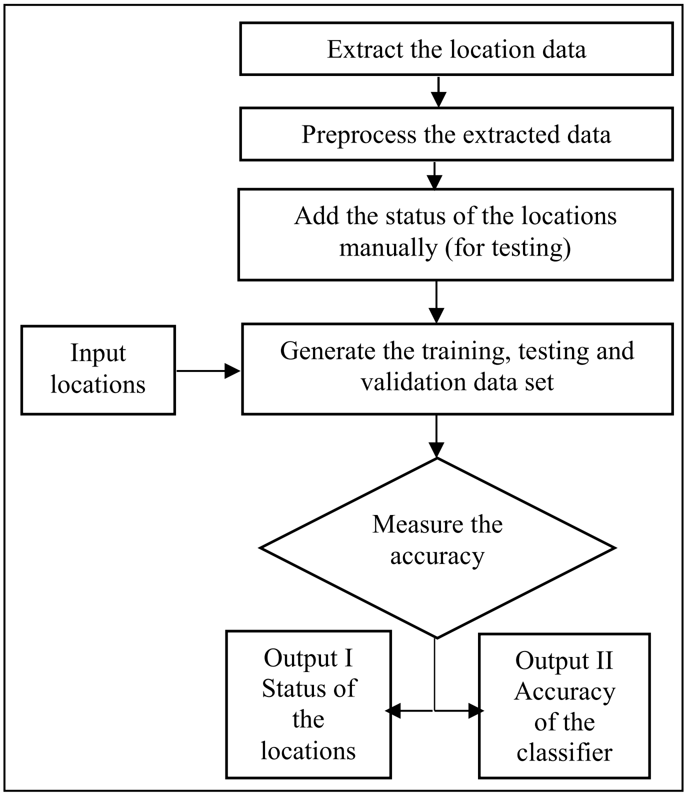
Extract the location data
Preprocess the extracted data
Add the status of the locations manually (for testing)
Input locations
Generate the training, testing and validation data set
Measure the accuracy
Output I Status of the locations
Output II Accuracy of the classifier

Figure 1. The Process model to detect the status of locations

VI. Experimental Results AND ANALYSIS

This paper discusses the accuracy of predicting the status of locations using different classifiers based on Bayes, Rule-Based, Trees and Meta classification approaches. Naïve Bayes classifier is one of the simplest and, most commonly used methods, which is capable of giving effective results in classification. When using the Bayesian approach, it is required to have independent attributes in the dataset [12]. Since, the attributes (location types) of a particular location, extracted from the Google Places API are independent attributes, the Naïve Bayes classification could be used to predict the status of a particular location as the benchmark approach.

As another classification approach, the decision tree approach has been used to predict the location status. The decision tree method is a powerful statistical methods that can be used for classification and prediction, as it is capable of simplifying the complex relationships between the input variables and the target variables by dividing the original input variables into significant subgroups [13].

The next classification approach that has been used in this paper is the Rule-Based classification. A Rule-Based classifier uses a set of IF-THEN rules for classification. Rule: IF (Condition) – THEN conclusion, where Condition is a conjunction of attributes and conclusion contains the class prediction [14]. The Rule-Based classification approach is used as it is highly expressive as the decision trees and has the capability of classifying the new instances quickly [14]. In this paper, under the Rule-Based classification approach, mainly OneR and ZeroR classification methods have been used to derive the predictions and analyze the results.

The Meta classification analysis offers a mechanism to estimate the magnitude of effects in terms of a statistically significant effect size or pooled odds ratio. As a result of combining data from several studies, this approach increases the statistical power. Thus, it allows to completely assess the impact of a procedure or variable [15]. This paper uses the regression classification as a Meta classifier.

Weka which is a well-known open source machine learning tool has been used to apply different classification approaches mentioned above [16].

A. *Data Preprocessing*

The extracted location details from the Google Places API, include the location types, ratings, place_id, opening hours, location name and location address. Under the data preprocessing stage, all the locations that did not have at least one tag in its location "types" array was removed, since the predictions are based on the "types" tags of the given location.

B. *Analysis of Results*

The test results and validation results obtained by applying Naïve-Bayes classification, Decision Tree classification (J48, Decision Stump), OneR classification, ZeroR classification and Regression classification are discussed below.

1) Results of Naïve-Bayes Classification:When evaluating the accuracy of an algorithm, Naïve-Bayes is considered as *the baseline. This is a Bayes classification approach. The following TABLE I shows the testing and the validating results that were obtained by using the Naïve-Bayes classification in Weka.*

TABLE I. Experimental Testing and validation Results Using the Naïve-Bayes Classification

Scenario	**Test Results**	**Validation Results**
Correctly Classified Instances	92.9 %	88.4 %
Incorrectly Classified Instances	7.1 %	11.6 %
Mean absolute error	0.0861	0.1057
Root mean squared error	0.1956	0.2332

2) Results of Decision Tree Classification (J48): This is a Trees classification approach. The following TABLE II shows the testing and the validating results that were obtained by using the Decision Tree (J48) classification in Weka.

TABLE II. Experimental Testing and validation Results Using the Decision Tree (J48) Classification

Scenario	**Test Results**	**Validation Results**
Correctly Classified Instances	99.6 %	98.8 %
Incorrectly Classified Instances	0.4%	1.2 %
Mean absolute error	0.0075	0.0154
Root mean squared error	0.061	0.1081

3) Results of Decision Tree Classification (Decision Stump): This is a Tree classification approach. The following TABLE III shows the testing and the validating results that were obtained by using the Decision Tree (Decision Stump) classification in Weka.

TABLE III. Experimental Testing and validation Results Using the DECISION TREE (DECISION STUMP) CLASSIFICATION

Scenario	**Test Results**	**Validation Results**
Correctly Classified Instances	89.5 %	84.5 %
Incorrectly Classified Instances	10.5 %	15.5 %
Mean absolute error	0.1777	0.2164
Root mean squared error	0.3021	0.3599

4) Results of OneR Classification: This is a Rule-Based classification approach. The following TABLE IV shows the testing and the validating results that were obtained by using the OneR classification in Weka.

TABLE IV. Experimental Testing and Validation Results Using the OneR Classification

Scenario	**Test Results**	**Validation Results**
Correctly Classified Instances	89.5 %	84.5 %
Incorrectly Classified Instances	10.5 %	15.5 %
Mean absolute error	0.1047	0.155
Root mean squared error	0.3235	0.3937

5) Results of ZeroR Classification: This is a Rule-Based classification approach. The following TABLE V shows the testing and the validating results that were obtained by using the ZeroR classification in Weka.

TABLE V. Experimental Testing and validation Results Using the ZeroR Classification

Scenario	**Test Results**	**Validation Results**
Correctly Classified Instances	70.9 %	66.3 %
Incorrectly Classified Instances	29.1 %	33.7 %
Mean absolute error	0.4063	0.4271
Root mean squared error	0.4543	0.4767

6) Results of Regression classification: This is a Meta classification approach. The following TABLE VI shows the testing and the validating results that were obtained by using the Regression classification in Weka.

TABLE VI. Experimental Testing and validation Results Using the Regression Classification

Scenario	**Test Results**	**Validation Results**
Correctly Classified Instances	95.3 %	94.6 %
Incorrectly Classified Instances	4.6%	5.4 %
Mean absolute error	0.0741	0.0746
Root mean squared error	0.2062	0.2141

.

Based on the results that were obtained from the Naïve Bayes, Decision Tree (J48), Decision Tree (Decision Stump), OneR, ZeroR and Regression classification methods, it can be confirmed that the implemented approach of detecting the location status as indoor or outdoor based on the location's "types "tag is accurate.

The accuracy of the results obtained from the above mentioned classifiers are evaluated using the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). By analyzing the results from TABLE I – TABLE VI, it is clear that the most accurate results can be obtained by using the Decision Tree (J48), Naïve Bayes and Regression classifiers.

Furthermore, using Weka the accuracy results for different cross-validations can be obtained. This paper provides the accuracy results for 5-fold cross-validation and 10-fold cross-validation using Naïve Bayes, Decision Tree (J48) and Regression classifiers.

7) Results of Naïve-Bayes classification for 5-fold cross-validation and 10-fold cross-validation: The following TABLE VII shows the accuracy results that were obtained by using the Naïve-Bayes classification for 5-fold cross-validation and 10-fold cross-validation in Weka.

TABLE VII. 5-FOLD AND 10-FOLD CROSS-VALIDATION RESULTS USING THE NAÏVE-BAYES CLASSIFICATION

Scenario	**5-Fold Cross-Validation**	**10-Fold Cross-Validation**
Correctly Classified Instances	92.5%	92.6 %
Incorrectly Classified Instances	7.5%	7.4 %
Mean absolute error	0.0927	0.0903
Root mean squared error	0.2024	0.2006

8) Results of Decision Tree (J48) classification for 5-fold cross-validation and 10-fold cross-validation: The following TABLE VIII shows the accuracy results that were obtained by using the Decision Tree (J48) classification for 5-fold cross-validation and 10-fold cross-validation in Weka.

TABLE VIII. 5-FOLD AND 10-FOLD CROSS-VALIDATION RESULTS USING THE DECISION TRESS (J48) CLASSIFICATION

Scenario	**5-Fold Cross-Validation**	**10-Fold Cross-Validation**
Correctly Classified Instances	99.6%	99.6 %
Incorrectly Classified Instances	0.4%	0.4 %
Mean absolute error	0.0073	0.0075
Root mean squared error	0.0632	0.0634

9) Results of Regression classification for 5-fold cross-validation and 10-fold cross-validation: The following TABLE IX shows the accuracy results that were obtained by using the Regression classification for 5-fold cross-validation and 10-fold cross-validation in Weka.

TABLE IX. 5-FOLD AND 10-FOLD CROSS-VALIDATION RESULTS USING THE REGRESSION CLASSIFICATION

Scenario	**5-Fold Cross-Validation**	**10-Fold Cross-Validation**
Correctly Classified Instances	96.8%	96.8 %
Incorrectly Classified Instances	3.2%	3.2 %
Mean absolute error	0.0592	0.0594
Root mean squared error	0.1704	0.1708

The summarized accuracy results of testing and validation along with the cross-validation results indicates that it is possible to determine the status of a location as indoor or outdoor using classification techniques with a significant accuracy.

VII. CONCLUSION

This paper has introduced a novel approach which can be used to determine the status of a location as indoor or outdoor. The main purpose of developing a mechanism to detect the location's status is, increasing the accuracy of the travel recommender systems which consider the user preferences based on the current weather context of the region in consideration. It is important to extract the location data, including its location types, where the location types are used as the classification attributes to predict the final location status as indoor or outdoor.

Through this paper, it is proved that the location types, which are added by people when introducing a location to Google Places can be used as classification attributes to predict whether a location is indoor or outdoor. The approach which has been discussed in this paper can be used to make accurate travel recommendations incorporating the current weather context to accurately identify user preferences.

REFERENCES

[1] Bazarova, N. N., Choi, Y. H., Sosik, V. S., Cosley, D., & Whitlock, J., 2015. Social Sharing of Emotions on Facebook: Channel Differences, Satisfaction, and Replies. CSCW 2015, March 14-18, Vancouver, BC, Canada, pp. 154 – 162.

[2] Ettema, D., Friman, M., Olsson, L. E., & Gärling, T., 2017. Season and Weather Effects on Travel-Related Mood and Travel Satisfaction. Frontiers in Psychology, 8, 140. http://doi.org/10.3389/fpsyg.2017.00140.

[3] Petrovic, D., Ivanovic, I., and Vladimir, D., 2015 "Does Weather Impact on Commuters' Travel Demand - Empirical Case Study of Belgrade." AET PAPERS - European Transport Conference.

[4] Horanont, T., Phithakkitnukoon, S., Leong, T. W., Sekimoto, Y., & Shibasaki, R., 2013. Weather Effects on the Patterns of People's Everyday Activities: A Study Using GPS Traces of Mobile Phone Users. PLoS ONE, 8(12), e81153. http://doi.org/10.1371/journal.pone.0081153.

[5] Becken, S., 2010. Home | Lincoln University | Christchurch, New Zealand. THE IMPORTANCE OF CLIMATE AND WEATHER FOR TOURISM. PP 146-183.

[6] Yiu, E. C., 1996. Image classifcation using color cues and texture orientation. Master's Thesis, MIT, dept EECS.

[7] Lipson, P. R., 1996. Context and Configuration Based Scene Classification, PhD thesis, MIT. EECS dept.

[8] Yu, H. H., Wolf, W., 1995. Scenic classification methods for image and video databases. In Proc. SPIE, Digital Image Storage and Archiving systems, PP 363–371. http://www.ee.princeton.edu/heathery/

[9] Szummer, M., Picard, R. W., 1998. Indoor-Outdoor Image Classification. MIT Media Laboratory Perceptual Computing Section Technical Report No.445, Appeared: IEEE Intl Workshop on Content-based Access of Image and Video Databases. http://www-white.media.mit.edu/szummer/

[10] Reitermanov, Z. 2010. Data Splitting, WDS'10 Proceedings of Contributed Papers, Part I. pp. 31–36.

[11] Reed, R. D. and Marks, R. J., 1998. Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, MIT Press, Cambridge, MA, USA.

[12] Vijayarani, S., & Deepa, S. (2014). Naïve Bayes Classification for Predicting Diseases in Haemoglobin Protein Sequences. International Journal of Computational Intelligence and Informatics, ISSN: 2349 – 6363 Vol. 3(No. 4), January - March , PP 278-283.

[13] SONG, Y., & LU, Y. (2015). Decision tree methods: applications for classification and prediction. Shanghai Archives of Psychiatry, 27(2), PP 130–135. http://doi.org/10.11919/j.issn.1002-0829.215044.

[14] Thangaraj, M., & Vijayalakshmi, C. R. (2013). Performance Study on Rule-based Classification Techniques across Multiple Database Relations. International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868, Volume 5(No.4), march, 1-13. Retrieved from www.ijais.org.

[15] THACKER, S. B., Modern Methods of Clinical Investigation: Medical Innovation at the Crossroads: Volume I. Washington (DC): National Academies Press (US); 1990. 8, Meta-Analysis: A Quantitative Approach to Research Integration. Available from: https://www.ncbi.nlm.nih.gov/books/NBK235484/

[16] Al-Khoder, A., Harmouch, H., 2013. Evaluating four of the most popular Open Source and Free Data Mining Tools, IJASR International Journal of Academic Scientific Research, ISSN: 2272-6446 Volume 3, Issue 1 (February - March), PP 13-23.