An Expert System for Summer Tourism in Turkey by Using Text Mining and K-Means++ Clustering

Yunus Doğan and Alp Kut

Dokuz Eylul University, Department of Computer Engineering, Tinaztepe Campus, Buca, 35100 Izmir, Turkey
{yunus,alp}@cs.deu.edu.tr

Abstract. This study has an aim to support tourism sector in Turkey by using an expert system; thus, tourists will be able to select the most suitable holiday places for themselves. Before the tourists go to a holiday place which they have not visited before, they make a research about this place. Also, some surprises in this place are learnt before the tourists go and many tourists do not like this situation. Therefore, an operation of text mining is preferred in this study. Thus, tourists do not need a research about the holiday places. All expert system will be returned a decision according to users' preferences. Our expert system has an aim to return more decisions than one. When a tourist uses the system; only one place is not returned, sorted places from the most suitable place to the least suitable one are given. Therefore, a clustering structure is needed. After the system decides the most suitable place for the tourist; the cluster where this suitable place locates finds and the all holiday places in this cluster are recommended in order from the most suitable to the least suitable.

There will be lots of features as attributes from text collection although there will be a low number of holiday places; thus, a large dataset is obtained. Therefore, K-Means clustering algorithm as both simple and fast clustering algorithm is preferred. However, K-Means has problem about deciding the space of clusters, because K-Means can give a different space of clusters with same dataset at each working. The cause of this situation is that K-Means starts clustering with random initial center points. Therefore, K-Means++ clustering is used as a new approach to K-Means without random initial center points and with consistent result spaces.

This study has four steps briefly. Firstly, the most preferable places for summer holiday in Turkey are decided. According to a research on web pages of Cultural and Tourism Ministry of Turkey about tourism, the most important places are Alanya, Ayvalık, Bodrum, Çeşme, Datça, Didim, Dikili, Fethiye, Kaş, Kuşdası, Marmaris, Side and Yalova. These places are preferred by both foreign and regional tourists a lot because of both common and unique features of these places. Therefore, secondly, the features must be determined. For this step, a research with rich documents about these places is done on web and these documents are collected in a text file for each place. These text files will be used for text mining operations in the next steps.

Thirdly, a dictionary is created for each place from the collection of text files. These dictionaries are too large to process, because these dictionaries content stop-words and unnecessary words for tourism. Therefore, some words are

determined to delete from the dictionaries and they are deleted; thus, the satisfactory dictionaries are obtained for each holiday places. A data warehouse must be created from these dictionaries for mining operations. Therefore, preprocess with vector space model is needed; thus, a dataset is obtained with tuples and their attributes. In last step, this dataset is used by K-Means++. It gives a space of clusters where there are the places. Finally, an expert system is ready to use and holiday places are recommended according to these clusters and the expectations of tourists.

Keywords: Expert Systems, Data Mining, Text Mining, Vector Space Model, K-Means++