Preprint accepted at IEEE GCIoT 2019. You can find the published version at: https://ieeexplore.ieee.org/document/9058397

Detecting Tourist's Preferences by Sentiment Analysis in Smart Cities

Zahra Abbasi-Moud
Faculty of Electrical and Computer Engineering
University of Birjand
Birjand, Iran
zahra.abbasi@birjand.ac.ir

Hamed Vahdat-Nejad
Faculty of Electrical and Computer Engineering
University of Birjand
Birjand, Iran
vahdatnejad@birjand.ac.ir

Wathiq Mansoor
Department of Electrical Engineering
University of Dubai
Dubai, United Arab Emirates
wmansoor@ud.ac.ae

***Abstract*—Smart tourism is one of the significant areas of smart cities and IoT. One of the important tasks of smart tourism systems is to intelligently and automatically extract the users' preferences. This paper investigates the automatic extraction of users' preferences in smart tourism. To this end, users' comments published in social networks are used. The proposed method extracts users' preferences through semantic clustering the comments and sentiment analysis. Evaluation results indicate high values of precision, recall, and f-measure parameters in tourist preference extraction.**

Keywords—***Smart tourism, Tourist preferences, Sentiment analysis, Semantic similarity.***

I. Introduction

Tourism is currently considered as one of the prominent industries of the world, which increases economic prosperity and social welfare. The key factor to be successful in tourism is to gain satisfaction of tourists; Therefore, extracting implicit tourists' preferences is important. Tourist preference extraction not only flourishes the tourism market, but also is regarded as the main part of tourism recommender systems. One solution to increase tourists' satisfaction and extract their preferences is to use the Internet of Things (IoT), particularly in tourism recommendation systems [1]. Smart city and IoT bring intelligence and offer personalized services to people [2].

Previously, various methods have been proposed to extract tourists' preferences. For example, the ratings given to tourist attractions visited by users have been exploited to extract their preferences [3]. In some methods, individual preferences are extracted based on tourists' movement pattern, which has been extracted using their published geo-tagged photos [4]. As social networks contain the users' reviews, the accuracy of preference extraction can be improved by investigating those reviews [5]. Accordingly, some other methods have extracted the keywords of the texts published by tourists as their preferences [6]. The method proposed in this paper extracts tourists' preferences based on their reviews in the cyberspace and performing sentiment analysis. To this end, nouns existing in tourists' reviews are firstly extracted and then clustered according to the maximum semantic similarity. Afterwards, a score is dedicated to each cluster based on its word frequency as well as sentiment analysis. The cluster with the highest score indicates the tourists' preferences. The proposed method has been evaluated based on a dataset of user's comments from the TripAdvisor[1] social network. The results show a high value of precision, recall and f-measure parameters.

Rest of the paper is organized as follows: Section II reviews the literature. Section III introduces the proposed preference extraction method. Section IV gives the evaluation details, and Section V concludes the paper.

II. Related work

In this section, tourist preference extraction methods used in the literature are investigated. In most of the studies, users with similar behavior are clustered and their preferences are assumed to be the same. For example, according to the geo-tagged photos published in Flicker, users that have traveled similar paths are clustered [7]. The user's visit records are actually considered as his/her preferences. PSiS is also a recommendation system that uses tourists' visit records as their preferences [8].

Nowadays, social networks are considered as a rich source of user reviews. Therefore, recent research has used users' reviews to extract their preferences. For example, in order to extract tourists' preferences, a private chat page is created and several questions are asked [9]. Special words within the intended ontology are extracted from the responses, and tourists' preferences are predicted accordingly. In another research, a knowledge graph is created based on available words in tourist' comments, and is used to infer their preferences [10].

A major deficiency of the mentioned methods is the use of tourists' movement patterns and/or words available in tourists' comments regardless of their sentiments. Therefore, the present proposed method extracts tourists' preferences using semantic methods as well as sentiment analysis.

III. Proposed method

The proposed method consists of four stages, and the input is the set of a tourist's reviews. At first, sentences are pre-processed. A semantic graph is then created based on the nouns extracted from the first stage. Then, nouns are semantically clustered, and finally, tourists' preferences are extracted. Fig. 1 depicts the general procedure. These stages are explained in the following:

[1] www.TripAdvisor.com

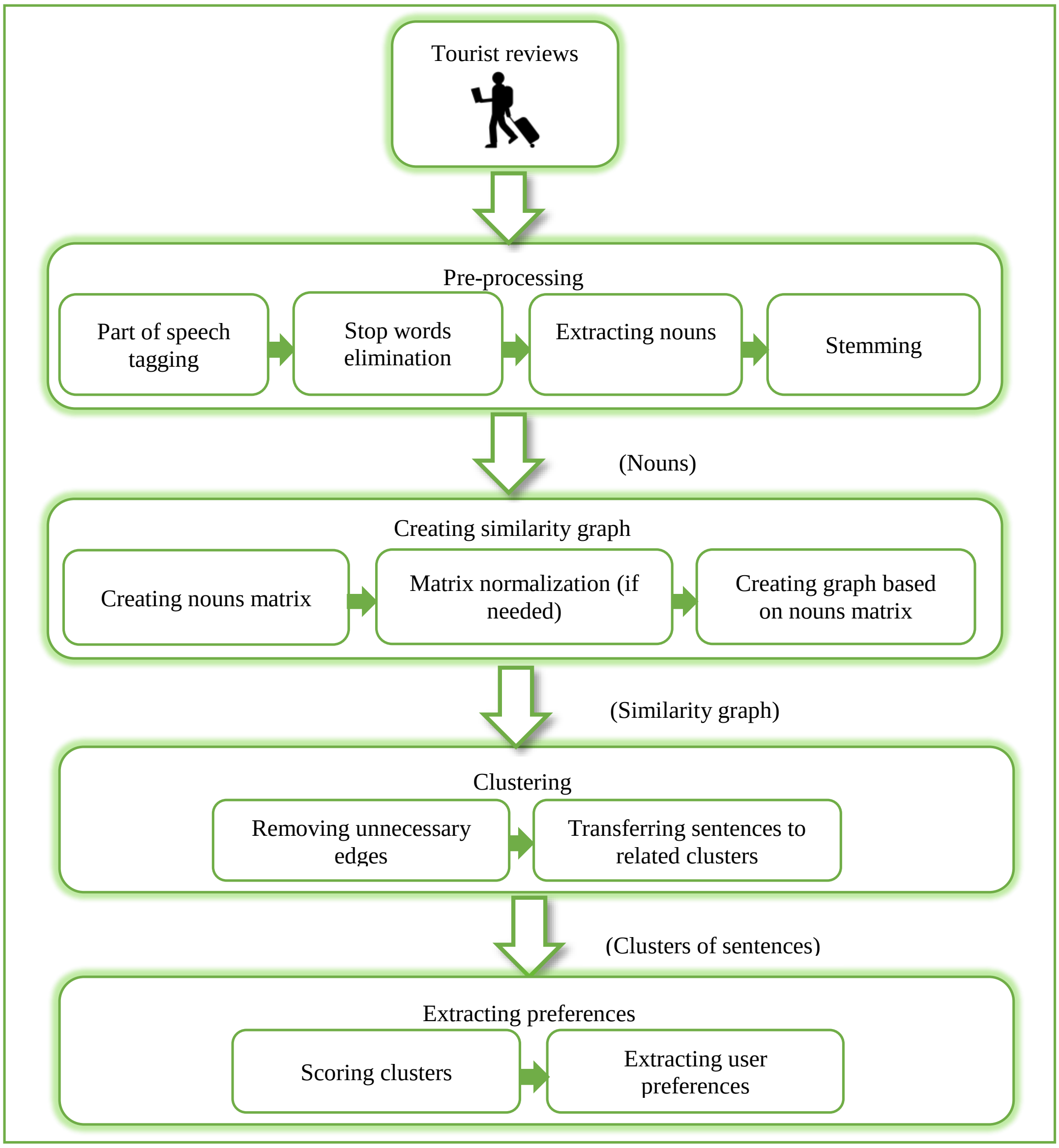

Fig. 1. The proposed method of extracting preferenes

Stage 1: Pre-processing

In this stage, the role of parts of speech is firstly determined by Part of Speech Tagging, and frequent but useless words (Stop Words) are omitted. Then the words that are tagged as noun are extracted. The reason to do so is to reduce the number of words and omit the insignificant ones [11]. Finally, the rest of the words are stemmed, and only one word is kept rather than several cognates.

Stage 2: Creating similarity graph

In this stage a symmetric matrix is created as is shown in Fig. 2 The nouns extracted from the first stage are used as the rows and columns of the matrix. Semantic similarity measurements between each pairs of nouns are considered as the matrix entries.

In this matrix, given the type of the semantic similarity measurement, values with different ranges are possible. Therefore, matrix normalization is performed, if required, in order to allow comparison of data with different measurements.

Then, similarity graph is drawn based on this matrix. The graph vertices consist of nouns, and its edges indicate the degree of semantic similarity of incident vertices (nouns). No edge is drawn between nouns with no semantic similarity.

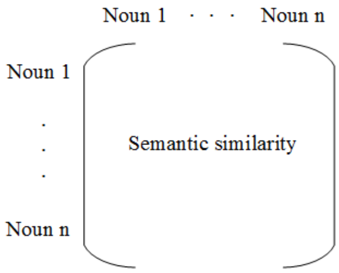

Fig. 2. Similarity matrix of nouns

Stage 3: Clustering

In this stage, edges with weights less than a specified threshold are omitted. Then, several clusters (connected sub-graphs) are yielded, each of which consists of a set of nouns with high semantic similarity. Afterwards, sentences are assigned to clusters in an overlapping manner in which each sentence is assigned to the clusters that contain each of the nouns of it. For example, if Cluster 1 includes "*a, b, c*" nouns and Cluster 2 includes "*d, f*" nouns, then the sentence "*abd*" is assigned to both clusters and the sentence "*fgh*" to Cluster 2.

Stage 4: Extracting preferences

The first step in this stage is to specify the score of each cluster according to its word frequency and sentiment analysis score. Thus, (1) is used to calculate the score of each cluster.

$$Score_{cluster_i} = TF_{cluster_i} \times Score_{Sentiment\ Analysis}(cluster_i) \tag{1}$$

Due to the importance of repetitive words of clusters, $TF_{cluster_i}$, in the above equation, denotes the sum of word frequencies in the ith cluster. For example, if Cluster *i* includes the words *a* and *b*, and their frequencies are 3 and 5, respectively, then $TF_{cluster_i}$ is equal to 8. Moreover, $Score_{Sentiment\ Analysis}(cluster_i)$ is equal to the sentiment analysis score of the ith cluster. Sentiment analysis refers to the process of extracting the attitudes and sentiments of people toward other people, events, topics and their characteristics [7]. Given the significance of this analysis in extracting individuals' preferences, tourists' sentiment analysis is done semantically using SentiWordNet.

Since words in different situations bear various meanings and emotional loads, the average of the positive and negative loads of the synset of a word is considered as its positive and negative load, respectively. For example, there are three synonyms for the word "*poor*". The positive score of all them is 0. The negative score of the synonyms are 0, 0.5, and 0.125. Therefore, the average of positive and negative loads of these three synonyms are used to represent the emotional load of the word "*poor*". The score of each sentence is computed by deducting the negative scores from the positive ones. The emotional load of emoticons is also taken into account, and positive emoticons of each sentence are scored +1; while negative emoticons are scored -1 [12].

Equation (2) calculates the sentiment analysis score of each cluster.

$$Score_{Sentiment\ Analysis}(cluster_i) = \frac{\sum score\ of\ each\ sentence\ of\ (cluster_i)}{Total\ number\ of\ sentences\ in\ (cluster_i)} \tag{2}$$

Finally, the cluster with the highest score denotes users' preferences. In other words, the set of words of that cluster indicates the user's preferences.

IV. EVALUATION

The proposed method has been implemented using Python3 programming language. Besides, the dataset has been extracted from tourists' comments in the TripAdvisor social network. The training data includes the opinions of 100 tourists with different age and nationality about various attractions within 6 months (from January to June, 2018). The testing data includes those tourist's comments in the first trip after June, 2018. Sentiment analysis has been conducted using SentiWordNet 3.0 [13]. The threshold used in the clustering stage (the third stage) has been specified through trial and error as 80%.

A hybrid semantic similarity measure is used to form the semantic similarity matrix [14]. In order to increase the accuracy of similarity calculation in this measure, all direct and indirect relationships among concepts are taken into account in WordNet. The measure eliminates the deficiencies of Wu-Palmer semantic similarity [15] (which does not yield accurate results due to missing all direct relationships among concepts), and the gloss-based similarity method for concepts in WordNet [16] (due to the failure of this method in detecting similarity among concepts that have no similarity in their definition, but have a direct relationship in WordNet structure).

Five repetitive words are extracted from all tourists' comments about each attraction visited after June, 2018. A trip to a place where its repetitive words have more than 80% similarity with the tourist's preferences is regarded successful. Precision and recall values are calculated as follows:

Precision = What percentage of attractions that are visited are similar to preferences?
Recall = What percentage of attractions that are similar to preferences are visited?

Fig. 3 shows the results of precision measure evaluation in case of using different semantic similarity measures in clustering. The results are indicative of high precision of the proposed method (93.05%). The highest precision value is observed in hybrid semantic similarity measure compared with the other two measures. Fig. 4 shows the comparison of the recall value. It expresses a high recall value for the proposed method that yields the optimal value for the hybrid semantic similarity measurement.

Finally, Fig. 5 demonstrates f-measure value for the proposed method in different states. It indicates similar results to the previous figures, and exhibits the superiority of the hybrid measure. Given all the above-mentioned results, the proposed preference extraction method indicates high values of evaluation parameters.

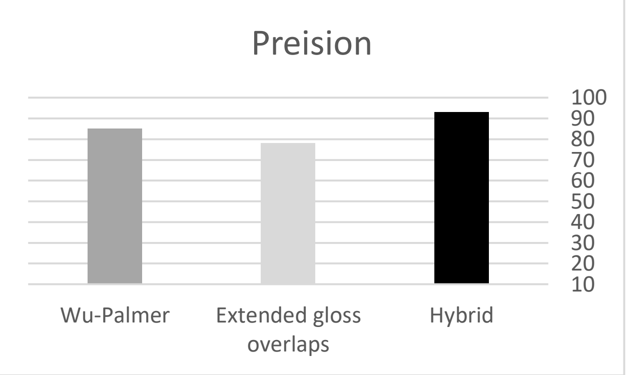

Fig. 3. The precision of the proposed method using different semantic similarity measures (in percentage).

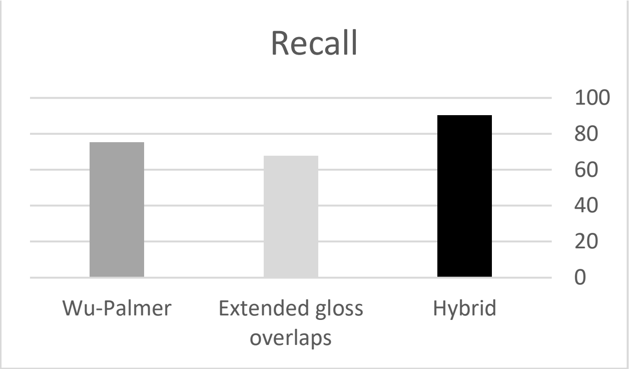

Fig. 4. The recall of the proposed method using different semantic similarity measures (in percentage).

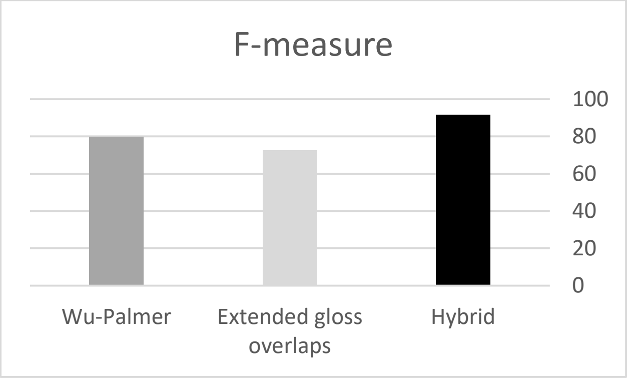

Fig. 5. The f-measure of the proposed method using different semantic similarity measures (in percentage).

V. CONCLUSION

The present paper proposes a method to specify the personal tourism preferences, in which tourists' comments published in cyberspace as well as their sentiment analysis are exploited. Evaluation results show the success of this method in obtaining high values of evaluation parameters including precision, recall, and f-measure. The proposed method can significantly contribute to the tourism systems of the Internet of Things and smart cities paradigm. In this regard, the development of a tourism recommender system in the context of a smart city is considered as the main future work of this research.

REFERENCES

[1] K. Kaur, and R. Kaur, "Internet of things to promote tourism: An insight into smart tourism" International Journal of Recent Trends in Engineering & Research, vol.2, no. 4, pp. 357-361, 2016.

[2] S. Cha, M. P. Ruiz, M. Wachowicz, L. H. Tran, H. Cao, and I. Maduako, "The role of an IoT platform in the design of real-time recommender systems" In 3rd world forum on internet of things, IEEE, Reston, pp. 448-453, 2016.

[3] R. P. Biuk-Aghai, S. Fong, and Y. W. Si, "Design of a recommender system for mobile tourism multimedia selection" In 2nd International Conference on Internet Multimedia Services Architecture and Applications, IEEE, Bangalore, pp. 1-6, 2008.

[4] I. Memon, L. Chen, A. Majid, M. Lv, I. Hussain, and G. Chen, "Travel recommendation using geo-tagged photos in social media for tourist." Wireless Personal Communications, vol. 80, no. 4, pp. 1347-1362, 2015.

[5] X. Zheng, Y. Luo, L. Sun, J. Zhang, and F. Chen, "A tourism destination recommender system using users' sentiment and temporal dynamics," Journal of Intelligent Information Systems, vol. 51, no. 3, pp. 557-578, 2018.

[6] E. Marrese-Taylor, J. D. Velásquez, F, Bravo-Marquez, and Y. Matsuo, "Identifying customer preferences about tourism products using an aspect-based opinion mining approach" Procedia Computer Science, vol. 22, no. 1, pp, 182-191, 2013.

[7] A. Majid, L. Chen, G. Chen, H. T. Mirza, I. Hussain, and J. Woodward, "A context-aware personalized travel recommendation system based on geotagged social media data mining," International Journal of Geographical Information Science., vol. 27, no. 4, pp. 662–684, 2013.

[8] R. Anacleto, L. Figueiredo, A. Almeida, and P. Novais, "Mobile application to provide personalized sightseeing tours," Journal of Network and Computer Application, vol. 41, no. 1, pp. 56–64, 2014.

[9] S. Loh, F. Lorenzi, R. Saldaña, and D. Licthnow, "a Tourism Recommender System Based on Collaboration and Text Analysis," Information Technology & Tourism, vol. 6, no. 3, pp. 157–165, 2004.

[10] P. Yochum, L. Chang, T. Gu, and M. Zhu, "Intelligent Information Processing IX," In International Conference on Intelligent Information Processing, Springer, Cham, pp. 80–85, 2018.

[11] S. Fodeh, B. Punch, and P. N. Tan, "On ontology-driven document clustering using core semantic features," Knowledge and information systems, vol. 28, no. 2, pp. 395–421, 2011.

[12] S. Fodeh, B. Punch, and P. N. Tan, "On ontology-driven document clustering using core semantic features," Knowledge and information systems, vol. 28, no. 2, pp. 395–421, 2011.

[13] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.," Lrec, vol. 10, no. 2, pp. 2200–2204, 2010.

[14] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains," Expert Systems with Applications, vol. 42, no. 4, pp. 2264–2275, 2015.

[15] Z. Wu and M. Palmer, "Verb Semantics and Lexical Selection," In 32nd annual meeting of the association for computational linguistics, New Mexico, 1994.

[16] S. Banerjee and T. Pedersen, "Extended gloss overlaps as a measure of semantic relatedness," International Joint Conference on Artificial Intelligence, Acapulco, pp. 805–810, 2003.