Cultural Distance awared Collaborative Filtering Algorithm in Recommendation System

Zhipeng Gao, Yuan Liu, Kaile Xiao and Yang Yang

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China.
Email: lylyly@bupt.edu.cn

Abstract. Recently most of the existing studies on recommendation system are based on historical rating matrix and specific characteristics, such as location, season and weather. Different from these studies, we introduce an abstract feature about cultural backgrounds and values, called cultural distance, into the recommendation system to understand user intent better and improve the precision of recommendation results. We design a novel similarity representation which combine the item-based collaborative filtering and cultural distance to recommend items for users. We also propose a collaborative filtering-based missing cultural distance prediction algorithm to improve the precision of recommendation further. To evaluate the performance of our proposed algorithm, we execute experiments based on a large-scale real-world dataset, the results show that our algorithm can improve the precision by 10% accurate compared to existing recommendation approaches.

1. Introduction

In the past ten years, the speed of data generating is becoming faster and faster, and we have already entered the big data era. The trend of big data has enabled us to use data mining to analyze and find patterns or information from existing data piles. The recommendation system is one of the solutions to alleviate the information overload problem. The recommendation system is a process of extracting information and filtering useful information for users by giving a "rating" to each item [1]. The recommendation system utilizes information to provide users with recommendations for a given number of choices. In recent years, the recommendation system has become more important. With the advent of new recommendation methods, the quality of recommendations has been improved gradually. The recommendation system has been used in numerous of applications, such as Google, YouTube, MovieLens, Netflix and Facebook [2].

The traditional technologies in the recommendation system include collaborative filtering, content-based filtering, and knowledge filtering. Collaborative filtering is the most popular recommendation technology due to its accuracy and stability [3]. However, the traditional collaborative filtering technology does not consider the attributes of users and items. Therefore, context-aware collaborative filtering technology has recently been developed to address recommended system issues and improve recommendation quality. There are many studies on the context-aware collaborative filtering recommendation technology, combining user preferences, evaluation projects and situational awareness (e.g., location, time of day, season) [4][5][6][7]. Although the complex context-aware collaborative filtering technology improve the accuracy of recommendation, the accuracy of recommendation system is still a key issue.

In this paper, we proposed a collaborative filtering recommendation based on cultural distance. As well known, the decisions of users are closely related to his cultural background and values, especially in many cultural related fields, such as movies, music, travel recommendations etc. At present, context-aware collaborative filtering algorithms are mostly based on surface data, and do not consider the essential reasons for users to select certain items. Based on this discovery, we introduce Hofstede's cultural distance model in collaborative filtering recommendation technology, which is a concept to measure cultural differences and value orientation in different countries.

The contributions of our cultural distance based collaborative filtering recommendation are four-fold:

1) To express the user preference and improve the accuracy of recommendation, we introduce cultural distance model into the traditional item-based collaborative filtering algorithm.
2) To predict the missing cultural distance of the items, we propose a missing cultural distance prediction algorithm based on the similarity matrix between items and collaborative filtering. The cultural distance prediction improves the accuracy of recommendation further.
3) We design a novel similarity representation which combine the item-based collaborative filtering and cultural distance, predict the rating of a user to the unrated items using the cultural distance-based similarity matrix, and then recommend the items with high predicted rating to the users.
4) We provide experimental evidence that our algorithm is feasible on a real dataset (MovieLens 1M) and provides better accuracy than traditional methods.

The rest of this paper is organized as follows: Section 2 introduces the background including cultural distance and related work. Section 3 presents the cultural distance-based collaborative filtering algorithm. Section 4 shows the experiments and Section 5 concludes the paper.

2. Background

This section aims to discuss the background of the cultural distance-based collaborative filtering algorithm, including the concept of cultural distance and related work with recommendation system.

2.1. Cultural Distance

Cultural distance is an important conception for studying cultural differences and consultants in many fields relating to international business and communication. The theory has been widely used in several fields as a paradigm for research, particularly in cross-cultural psychology, international management, and cross-cultural communication [8]. Culture and values influence people's daily lives and decisions in a subtle way, although many people accustom to it and do not realize what prompted him to make the decision.

Hofstede's cultural dimension model is one of the most famous culture studies. It first realized the quantitative calculation of abstract cultural concepts [9]. This quantitative calculation of cultural dimensions allows researchers to visually calculate cultural values as data to compare differences between cultures and behaviors. Hofstede's cultural dimension theory proposed six dimensions along which cultural values could be analyzed: individualism-collectivism; uncertainty avoidance; power distance; masculinity-femininity; long-term orientation and indulgence versus self-restraint. Based on the Hofstede's cultural dimension theory, [10] uses a simple mathematical formula to define the cultural distance as follows:

$$cd_j = \sum_{i=1}^{m} \left\{ \left(I_{ij} - I_{i0} \right)^2 / V_i \right\} / m , \tag{1}$$

where cd_j represents the cultural distance between country j and the host country, I_{ij} represents Hofstede's i-th cultural dimension score of the j-th country, the I_{i0} represents the ith cultural dimension score of the host country, V_i represents the variance of the i-th cultural dimension scores and m represents the number of cultural dimension.

2.2. Related Work

A large number of recommendation system approaches have been proposed in the literature, we only review some notable here.

At present, uniform standard for the classification of recommendation systems doesn't exist, and many researchers have carried out different divisions of the recommendation methods in different ways [11-13]. The mainstream recommendation methods include the following: collaborative filtering recommendation, content-based recommendation and hybrid recommendation.

Collaborative filtering is the process of filtering or evaluating items using the opinions of others [14]. It predicts ratings of new users on other topics or products based on training data [15]. The main idea is that similar users may like similar items. There are two types of methods, including memory-based method and model-based method [16]. The memory-based method attempts to establish correlations based on the user's user space or users in the item space [17]. This method is easily to modify when new data is added. Model-based method uses user databases to estimate or learn models from neighbor preferences [14]. This method does not focus on the similarities between users, but rather helps people make choices based on the opinions of others with similar interests.

The context-aware recommendation system integrates context information as additional valid data or information into the recommendation process to solve modeling and predict user preferences [17]. In context-aware recommendation system, user preferences and interests are not only modeled as the function of the project and the user, but also as a function of context-aware information.

A method was proposed by Xu et al. for recommending tourist locations based on the distribution of tourism history S (season and weather) [18]. They use the user travel history to build a user-user similarity model and use post-filtering to filter out recommendations that do not conform to context constraints. Savage et al. [6] developed a context-aware location recommendation system called "I feel LoCo", which aims to design a location recommendation algorithm by inferring user preferences, automatically considering time, geography and similarity measures.

However, using only seasonal, weather conditions and locations does not represent the optimal feature set for the recommendation process. These context-aware factors only consider the specific characteristics, and do not grasp the underlying reasons why people make a decision or what does he thought. We believe that people are more willing to accept things that have similar cultural backgrounds and values, especially in certain areas (such as movies, music, books, etc.). It is easier to accept such recommendations due to the similar culture. Therefore, we propose a collaborative filtering model based on cultural characteristics, paying more attention to the underlying reasons for users to make decisions.

3. Cultural Distance Based Recommendation Algorithm

In our cultural distance based recommendation algorithm, we use item-based collaborative filtering as the primary recommendation technique, due to its popularity and widespread adoption in commercial systems (e.g., Amazon). As shown in Figure 1, the cultural distance based recommendation algorithm is divided into four processes. The first is the basic similarity computation based on the Pearson correlation coefficient. We compute the basic similarity matrix containing the similarity score of each item pairs, which is the same with the similarity matrix in item-based collaborative filtering. The second is the missing cultural distance prediction for the items from the countries which do not have a cultural distance value. The prediction algorithm is based on collaborative filtering. The third process is the cultural distance-based similarity computation. We design a novel similarity which combine the item-based collaborative filtering and cultural distance. The last recommendation process is to predict the user's possible ratings for items that have not been rated, and recommend one or several items most likely to be enjoyed by the user based on the results of the prediction ratings.

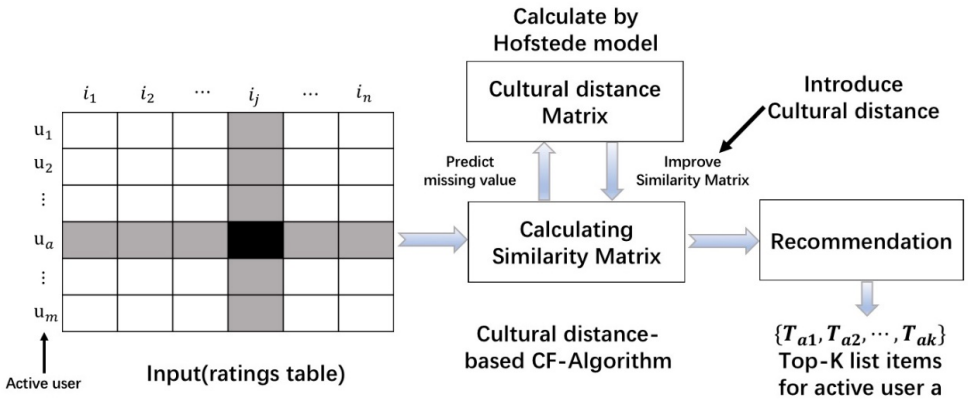


Figure 1. Our cultural distance based recommendation algorithm

3.1. Similarity Computation

Assume that there are a set of N users $U = \{u_1, u_2, u_3, \cdots, u_N\}$ and a set of M items $I = \{i_1, i_2, i_3, \cdots, i_N\}$. Each user expresses opinions about a set of items, and the opinions can be a numeric rating (e.g., the MovieLens scale of one to five stars). In this subsection, we compute the similarity matrix containing the similarity score of each item pairs based on the user-item rating matrix.

There are many methods to calculate the similarity, such as Cosine, Pearson correlation coefficient and Adjusted Cosine. We use the Pearson correlation coefficient similarity in our algorithm due to its precision and popularity. The formulation of the basic similarity score sim_{mn} for each pair of i_m and i_n that have at least one common rating by the same user (i.e., co-rated dimensions) is as follows

$$sim_{mn} = \frac{\sum_{u \in U_{mn}} (r_{um} - \overline{r_m})(r_{un} - \overline{r_n})}{(\sum_{u \in U_{mn}} (r_{um} - \overline{r_m})^2)^{1/2} (\sum_{u \in U_{mn}} (r_{un} - \overline{r_n})^2)^{1/2}}, \quad (2)$$

where U_{mn} represents the subset of ratings where i_m and i_n are both rated by user u, $\overline{r_m}$ represents the average of ratings of i_m obtained by the users in U_{mn}, $\overline{r_n}$ represents the average of ratings of i_n obtained by the users in U_{mn}. The value of sim_{mn} is between -1 and 1, where -1 represents completely negative linear correlation and 1 represents completely positive linear correlation. The larger value of sim_{mn} indicates the higher similarity between i_m and i_n. The similarity computation is an $O(R^2 / U)$ process, where R and U are the number of ratings and users.

3.2. Missing Cultural Distance Prediction

In this subsection, we introduce an abstract feature of items, called cultural distance, to improve the precision of collaborative filtering recommendation algorithm. And we proposed a cultural distance prediction algorithm to predict the missing cultural distance values.

In the process of introducing cultural distance, there will be a small number of countries whose cultural distance cannot be calculated due to the data missing. If the cultural distances of these countries are set to constant or the mean of the other country's cultural distances, the items of these countries not only would have accuracy error while calculating the similarity, but also affect the recommendation of other items. Therefore, we proposed a collaborative filtering based approach to predict the missing cultural distance values.

In missing cultural distance prediction approach, we first set a loose threshold th. to filter items with low similarity, and then select the remaining items which are the top K most similar items as neighbors. The selected neighbors of i_m can be described as follows:

$$S_m = \{n \mid sim_{mn} < th, r_{mn} < K\}, \tag{3}$$

where S_m represents the neighbors of i_m, r_{mn} represents the ranking of sim_{mn} among similarities between i_m and other items. The formulation means that i_n is a neighbor of i_m if i_n belongs to the top K similar items of i_m, and sim_{mn} is less than th.

Then, we predict the missing cultural distance of i_m by the similarity neighbors as follows:

$$cd_m = \overline{cd_{S_m}} + \frac{\sum_{n \in S_m} sim_{mn} * (cd_n - \overline{cd_{S_m}})}{\sum_{n \in S_m} sim_{mn}}, \tag{4}$$

where cd_n is the cultural distance between i_n and the host country computed by equation (1), and $\overline{cd_{S_m}}$ is the mean cultural distance between the items in S_m and the host country. The formulation means that the predicted culture distance value of i_m is $\overline{cd_{S_m}}$ plus the sum of $(cd_n - \overline{cd_{S_m}})$ weighted by sim_{mn} and normalized by the sum of sim_{mn}.

After obtaining the missing culture distance values of items, we can calculate the cultural distance value of the corresponding country by the following formulation:

$$cd_c = \frac{1}{|C|} \sum_{i=1}^{|C|} cd_m, \tag{5}$$

where cd_c is the cultural distance value of the country c corresponding to i_m, C is the set of items whose country is c. The meaning of the above formulation is that the predicted cultural distance value of a country is the average value of all the items belonging to the country.

3.3. Cultural Distance-based Similarity Computation

In this subsection, we calculate the cultural distance-based similarity matrix of i_n on i_m. First, we calculate the cultural distance factor c_{mn} between the i_m and the i_n based on the predicted cultural distance values as follows:

$$c_{mn} = 1 - \frac{|cd_m - cd_n|}{\max_{cdm} - \min_{cdm}}, \tag{6}$$

where c_{mn} represents the cultural distance factor of i_m to the i_n, $\max_{cdm}$ represents the maximum value among the culture distance values of neighbors of i_m, $\min_{cdm}$ represents the minimum value among the culture distance values of neighbors of i_n.

In our basic similarity computation, we get the basic similarity between i_m and i_n as sim_{mn}, and in our previous steps, we get the cultural distance factor c_{mn} represents the cultural distance of i_m and i_m. So we design a novel similarity representation which can consider not only the basic similarity of neighbors but also the culture distance factor. The formula is as follows:

$$Simc_{mn} = \alpha * sim_{mn} + (1 \quad \alpha) c_{mn}, \tag{7}$$

where α represents the weight between the basic factor sim_{mn} and the cultural distance factor c_{mn}.

3.4. Recommendation Generation

In this subsection, we predict the rating of a user to the unrated items using the cultural distance-based similarity matrix, and then recommend the items with high predicted rating to the users.

The predicted rating of user u to an unrated i_n can be calculated by the following formula:

$$P_{un} = \frac{\sum_{m \in L_u \cap S_n} Simc_{mn} * r_{um}}{\sum_{m \in L_u \cap S_n} |Simc_{mn}|}, \tag{8}$$

where L_u represents the set of items rated by the user u. The predicted rating is the sum of r_{um} weighted by the cultural distance-based similarity $Simc_{mn}$ and normalized by the sum of similarity scores. The more similar an item is with an interest item in the user's history, the higher ranking it will have in the user's recommendation list.

4. Experiments

This section evaluates the collaborative filtering algorithm based on cultural distance, and the dataset of the evaluation is MovieLens 1M. This dataset describes 5-star rating and free-text tagging activity from MovieLens, a movie recommendation service. It contains 100004 ratings and 1296 tag applications across 9125 movies.

Based on the dataset, we compare the cultural distance-based collaborative filtering algorithm (CDCF) with two other recommendation approaches.

- **Standard item-based collaborative filtering algorithm (Item-based CF)**. This standard algorithm computes the similarity matrix using the Pearson correlation coefficient directly, and then predict the user's possible ratings based on the similarity matrix.
- **CDCF***. This algorithm is a variation of CDCF. It only introduces the cultural distance to compute the similarity, but not consider the missing cultural distance values.

Unless mentioned otherwise, the default value of α is 0.08, the number of item neighbors K in user interest calculation and missing cultural distance feature prediction is 10. The rest of this section evaluates recommendation quality (Section 4.1) and parameter analysis (Section 4.2).

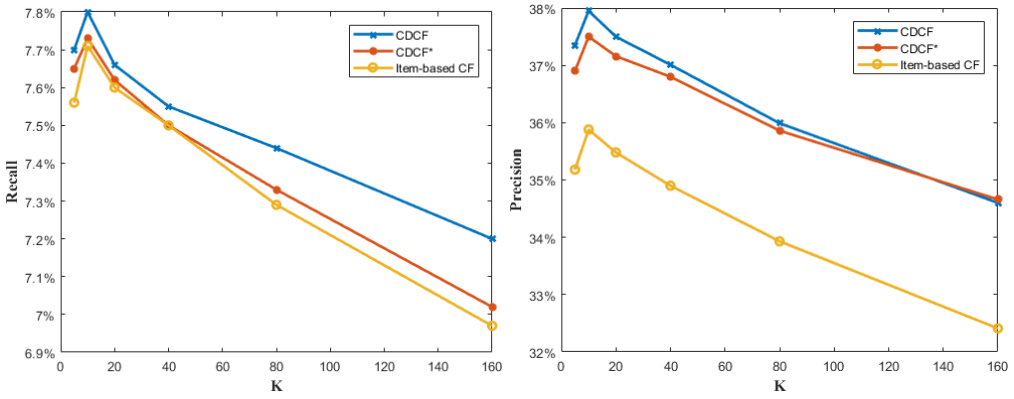
4.1. Recommendation Quality for Recall and Precision

In this subsection, we compare the algorithms in terms of recall rate and precision rate. The definitions of recall rate and precision rate are as follows:

$$Recall = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |T(u)|}, \tag{9}$$

$$Precision = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|}. \tag{10}$$

As shown in Figure 2a) and Figure 2b), our algorithm is better than item-based CF and CDCF* in terms of recall and precision. In terms of recall rate, our algorithm is flat with CDCF*, which are more precise than item-based CF. However, the calculation of CDCF* is based on the multi-layer/level pyramid model, and the time-cost is much larger than our algorithm. Therefore, our method performs better in time cost and also performs better in scalability.



a)Comparison results in terms of recall b) Comparison results in terms of precision

Figure 2. Comparison results with respect to the number of item neighbors K

4.2. Parameter Analysis

Figure 3 describes the relationship between α and precision, where α is the coefficient of calculating the cultural distance interest matrix of user u for j. As shown in Figure 3, when α is closer to 0, our algorithm approximates the item-based collaborative filtering algorithm. When α is closer to 1, we consider the similarity between items less and less, and the effect is also fall down, and α=0.08 has the best performance in precision.

Figure 4 depicts the relationship between K and precision. As shown in Figure 4, precision increases with the increase of K.When K is greater than 10, the increase of precision is obviously slowed down, we suspect that this trend is related to the size of the rating data and the size of the cultural distance data. The amount of calculation also grows fast with the increase of K, so we set the default value of K is 10.

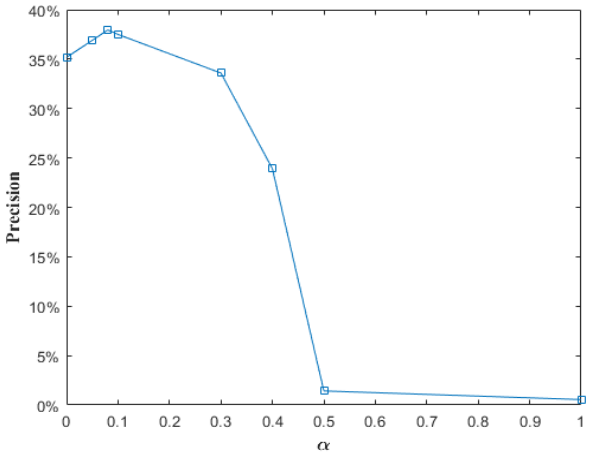


Figure 3. The relationship between α and precision

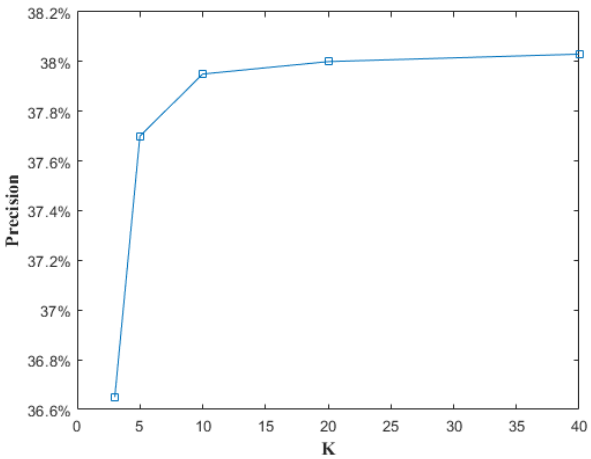


Figure 4. The relationship between K and precision

5. Conclusion

The recommender systems are very popular in current web environment, the precision of the recommendation system and whether the recommendation results meet the user's preferences is very important. In this paper, we introduce the cultural distance feature into the recommendation system, in order to understand user intent better and produce more precision recommendation results. At first, we compute the basic similarity matrix containing the similarity score of each item pairs. And then we

predict the missing cultural distance for the items from the countries which do not have a cultural distance value. And then we compute the cultural distance-based similarity which combine the item-based collaborative filtering and cultural distance. Finally, we predict the user's possible ratings for items that have not been rated, and recommend one or several items most likely to be enjoyed by the user based on the results of the prediction ratings. Experiments using large-scale real-world data reveals that our algorithm can improve the precision by 10% accurate compared to existing recommendation approaches.

Future work can be carried out in the following two aspects: 1) Try to introduce cultural distance in the user-based collaborative filtering algorithm, and then we can design a hybrid recommendation algorithm. 2) We can split the culture from the original six dimensions and add new dimensions (e.g., economic distance, social distance) combined according to a certain weight, experiment and compare with the results of this paper.

6. Acknowledgment

This work is supported by National Science & TechnologyPillar Program (2015BAH03F02), and National Key Researchand Development Program of China (2016YFE0204500).

7. References

[1] Ricci F, Rokach L and Shapira B, “Recommender systems: introduction and challenges,” *Recommender systems handbook*, pp. 1-34, 2015.
[2] Linyuan L, Medo M, Chi-ho Y, Yi-Cheng Z and Zi-Ke Z, "Recommender Systems," *Physics Reports*, vol. 519, no. 1, pp. 1-49, 2012.
[3] Aggarwal and Charu C, "An introduction to recommender systems," *Recommender Systems*, pp. 1-28, 2016.
[4] Bagci H and Karagoz P, "Context-aware location recommendation by using a random walk-based approach," *Knowledge and Information Systems*, vol. 47, no. 2, pp. 241-260, 2016.
[5] Fung W, Shijun L and Quanrui W, "Points of Interest Recommendation Based on Context-aware," *Int. Journal of Hybrid Information Technology*, vol. 8, no. 3, pp. 55-62, 2015.
[6] Norma S, Maciej B and Norma Elva C, “I'm feeling LoCo: A Location Based Context Aware Recommendation System,” *Advances in Location-Based Services*, pp. 37-54, 2012.
[7] Haosheng H, "Context-Aware Location Recommendation Using Geotagged Photos in Social Media," *ISPRS Int. Journal of Geo-Information*, vol. 5, no. 11, pp. 195, 2016.
[8] Laszlo T, David A G and Craig J R, "The effect of cultural distance on entry mode choice, international diversification, and MNE performance: a meta-analysis," *Journal of Int. Business Studies*, vol. 36, pp. 270-283, 2005.
[9] Reformat M, Dengming L and Cuong L, "Approximate reasoning and Semantic Web Services," in *Proc. of. IEEE Annual Meeting of the Fuzzy Information*, pp. 413-418, 2004.
[10] Kogut B and Singh H, "The Effect of National Culture on the Choice of Entry Mode," *Journal of Int. Business Studies*, vol. 19, pp. 411-432, 1988.
[11] Balabanovic M and Shoham Y, “Fab: Content-based collaborative recommendation,” *Communications of the ACM*, vol. 40, no. 3, pp. 66-72,1997.
[12] Yingyuan X, Pengqiang A, Ching-Hsien H and Jiao X, “Time-Ordered Collaborative Filtering for News Recommendation,” *China Communications*, vol. 12, pp. 53–62, 2015.
[13] Terveen L and Hill W, “Beyond recommendation systems: Helping people help each other,” in *Proc. of HCI in The New Millennium*, pp. 487-509, 2001.
[14] J. Ben S, Dan F, Jon H and Shilad S, “Collaborative filtering recommender systems,” *The adaptive web*, pp. 291-324, 2007.
[15] John S B, David H and Carl K, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proc. of the Fourteenth conference on Uncertainty in artificial intelligence*, pp. 43-52, 1998.

[16] Hsin-Hui L and Yi-Shun W, "An examination of the determinants of customer loyalty in mobile commerce contexts," *Information & management*, vol. 43, no. 3, pp. 271-282, 2006.
[17] Gediminas A and Alexander T, "Context-aware recommendation systems," *Recommendation Systems Handbook*, pp. 217–253, 2011.
[18] Zhenxing X, Ling C and Gencai C, "Topic based context-aware travel recommendation method exploiting geotagged photos," *Neurocomputing*, vol. 155, pp. 99-107, 2015.