Location Estimation in Location Based Social Networks

Huseyn Valiyev, Elvin Mammadli
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, 34220 Istanbul, Türkiye
{l1115907, l1116904}@yildiz.edu.tr

***Özetçe* —İnsanların internete erişebildiği zamandan günümüze kadar terabaytlarca kullanıcı verisi oluştu.Bu verilerin içerisinde kullanıcıların yer bildirimleri çok büyük bir önem arz etmektedir. Bu bilgileri reklam platformaları ve alışveriş siteleri kullanıcılarına daha iyi hizmet verebilmek için aktif olarak kullanıyorlar. Böylelikle kullanıcıların aktif lokasyonlarını tahmin etmek bu sistemler için daha değerli hale geliyor. Bu geniş kapsamlı kullanımından dolayı tahmin sistemleri, hali hazırda milyarlarca dolarlık bir endüstrinin kalbinde yer almaktadır.**

Bu çalışmada, tahmin sistemlerinden biri olan lokasyon bazlı sosyal ağlarda yer tahmini sistemi üzerinde çalışılmıştır. Çalışmanın amacı, kullanıcıların eski kategorisel bildirimlerini inceleyerek geleceğe yönelik tahmin yapmaktır. Projenin yürütülmesi için açık olarak sunulan Foursquare veri seti kullanılmıştır. Çalışmada Item-Based ve User-Based olmak üzere 2 farklı İşbirlikçi filtreleme yöntemleri kullanılmıştır. Ayrıca sistemin gerçek kullanıcılarla etkileşime girebileceği bir arayüz tasarlanmıştır.

Anahtar Kelimeler—Sosyal Medya, Lokasyon tahmini, İşbirlikçi filtreleme, Lokasyon Bazlı Sosyal Ağlar, Yer Bildirimleri.

***Abstract*—Terabytes of user data have been created from the time when people can access the internet. The user's location check-ins are very important in this data. Advertising platforms and shopping sites actively use this information to better serve their users.Thus, estimating the active locations of users becomes more valuable for these systems.Due to this wide use, estimation systems are already at the heart of an industry of billions of dollars.**

In this study, the location estimation system in location-based social networks, which is one of the prediction systems has been studied. The study aims to make predictions by examining the old categorical check-in of the users. To carry out the study, the open Foursquare data set was used. Two different Collaborative Filtering methods, Item-Based and User-Based, were used in the study.Also, an interface is designed in which the system can interact with real users.When the literature is analyzed, we can observe many methods are used for position prediction in LBSA.

Keywords—Social Media, Location estimation, Collaborative filtering, Location Based Social Networks, Location check-in.

I. Introduction

The explosive growth in the amount of available digital information and the number of visitors to the Internet has created a potential challenge of information overload which hinders timely access to items of interest on the Internet. This has increased the demand for estimation systems more than ever before[1]. Estimation systems are information filtering systems that deal with the problem of information overload by filtering vital information fragment out of large amounts of dynamically generated information according to the user's preferences, interest, or observed behavior about items. The estimation system has the ability to predict whether a particular user would prefer an item or not based on the user's profile.

The location estimation system in location-based social networks is an important part of estimation systems. Location Based Social Networks (LBSA) is used by millions humans which can connect to the internet network.LBSA allows to users that they can make notifications ("check-in") in their locality and billions of information which accumulated in LBSA give the opportunity to learn the user's spatial social behavior.

II. Related Works

Estimation system is defined as a decision making strategy for users under complex information environments.This systems handle the problem of information overload that users normally encounter by providing them with personalized, exclusive content and service recommendations. When the literature is analyzed, we can observe many methods are used for position prediction in LBSA.Before following the user location, to measure the relationship between Backstrom and Sun geography and friendship humans can estimate user location by used location information provided by users and the friendship network on Facebook .

For location suggestions, Leung and others advise Collaborative Location Recommendation(CLR). This method considers users' classes and activities to create more sensitive and refined recommendations. Authors also used Community-based Agglomerative-Divisive Clustering (CADC) for group users by similar locations and similar events. Also, because the large dimensionality(that is, the user, location, activity, etc.) like the LBS has tensor-based approaches for estimation. For example, Biancalana and others can identify user preferences and information needs and potentially interesting points, that give social advice based on a tensor suggesting personalized recommendations, and they become implemented in the system

III. Collaborative Filtering

Recently, various approaches for building estimation systems have been developed, which can utilize either collaborative filtering, content-based filtering, or hybrid filtering [2], [3], [4]. The collaborative filtering technique is the most mature and most commonly implemented...The fundamental assumption of CF is that if users X and Y rate n items similarly, or have similar behaviors (e.g., buying, watching, listening), and hence will rate or act on other items similarly[5].CF techniques use a database of preferences for items by users to predict additional topics or products a new user might like. However Collaborative Filtering algorithms are commonly used, there are several limitations for the memory-based CF techniques, s as the fact that the similarity values are based on common items and the are unreliable when data are sparse and the common items are therefore few[6].Since we have sufficient user data for the location based estimation system to be implemented, thanks to Foresquare and collaborative filtering method has been used in the project.

In this study, similarity determination was made in order to understand how much the user's tastes overlap with the other users.Then, the most similar users were selected and the weighted average of their rates for all locations was obtained.

A. Problems in Collaborative Filtering

There are many challenges for collaborative filtering tasks . CF algorithms are required to have the ability to deal with highly sparse data, to scale with the increasing numbers of users and items, to make satisfactory recommendations in a short time period, and to deal with other problems like synonymy (the tendency of the same or similar items to have different names), shilling attacks, data noise, and privacy protection problems.

1) Cold Start: Today's estimation systems, especially the ones that use collaborative filtering method, require users to interact with content to produce recommendation. The situation in which this cannot be accomplished, when the content or users that have just entered the system, is defined as cold start.

2) Lack of data: The nature of recommendation systems requires that the success of the system depends on the amount of user interaction, as the system produces its recommend.

B. Foresquare Data Set

Applications of collaborative filtering typically involve very large data sets.In this project, we use Foursquare dataset which shared by Dingqi Yang.This data set includes check-ins in NYC and Tokyo, collected for about 10 months(From April 12, 2012 to February 16, 2013).Includes 573,703 check-ins. in Tokyo and 227,428 check-ins in NewYork.Every check-in associated with a timestamp, GPS coordinates, location and categories.In Tokyo dataset we have 342 users and 342 different categories, In Newyork dataset, we have 235 users and 342 different category.This dataset is used to examine the spatial of user activity regularity.

C. Similarity functions

1) Jaccard Similarity: Although Jaccard similarity is a similarity function that produces fast results, it ignores the rating values and just evaluates whether or not users vote.

$$J(u,v) = |u \cap v| / |u \cup v|$$

where A, set of rate which rated by a , B set of rate which rated by b.

2) Cosine Similarity: The cosine similarity is an approach that the cosine between the vectors will give the similarity between the two users if we consider the users' votes as vectors. In this approach, missing ratings are threated as negative.

In the calculation of cosine similarity, it is essential that each line is taken as a vector and the cosine distance between these multidimensional vectors is taken as a similarity criteria.

$$sim = (u,v)\frac{u \cdot v}{|u||v|} = \frac{\sum_{i=1}^{N} u_i v_i}{\sqrt{\left(\sum_{i=1}^{N} u_i^2\right)\left(\sum_{i=1}^{N} v_i^2\right)}}$$

where A vector of rates which rated by a, B vector of rates which rated by b.

3) Pearson Correlation Similarity: In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and -1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

$$r = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2}\sqrt{\sum_{i=1}^{n}(v_i - \bar{v})^2}}$$

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

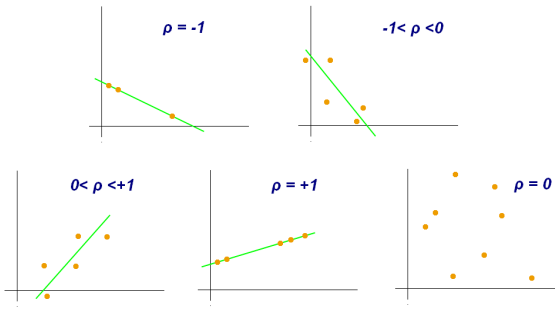

Figure 1 Pearson Correlation

D. Filtering

1) User Based Filtering: User-based collaborative filtering predicts a test user's interest in a test item based on rating information from similar user profiles. As illustrated in Fig. 3, each user profile (row vector) is sorted by its dis-similarity towards the test user's profile.Ratings by more similar users contribute more to predicting the test item rating.

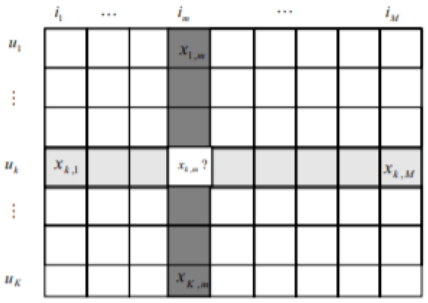

Figure 2 The user-item matrix

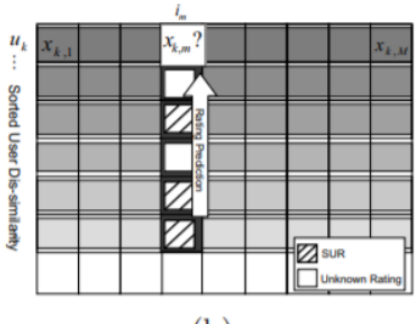

Figure 3 Rating prediction based on user similarity

2) Item Based Filtering: Item-based approaches such as apply the same idea, but use similarity between items instead of users. As illustrated in Fig. 4, the unknown rating of a test item by a test user can be predicted by averaging the ratings of other similar items rated by this test user . Again, each item (column vector) is sorted and re-indexed according to its dis-similarity towards the test item in the user-item matrix, and, ratings from more similar items are weighted stronger.[7]

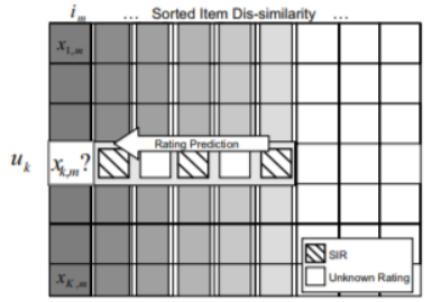

Figure 4 Rating prediction based on item similarity

E. Error Calculation

1) Root Mean Square Error: The Root Mean Square Error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed.RMSE is always non-negative, and a value of 0 (almost never achieved in practice) would indicate a perfect fit to the data. In general, a lower RMSE is better than a higher one.

$$\sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}$$

2) Mean Absolute Error: Some researchers have recommended the use of the Mean Absolute Error (MAE) instead of the Root Mean Square Deviation. MAE possesses advantages in interpretability over RMSE. MAE is the average of the absolute values of the errors. MAE is fundamentally easier to understand than the square root of the average of squared errors. Furthermore, each error influences MAE in direct proportion to the absolute value of the error, which is not the case for RMSE.

IV. Conclusion

In this study, the location-based estimation system, as an example of estimation systems was implemented with different collaborative algorithms and the results were compared with different statistical methods and a user the interface was developed in which the system could interact with the user.

As a result of the comparisons made with different collaborative algorithms, the best results were obtained by user-based collaborative filtering method.After we apply RMSE correctness of item-based and user-based filtering methods are 0.98 and 0.99 respectively for Tokyo.This values for Newyork is 0.78 and 0.98 respectively.The correctness of Tokyo values is more precise because our Tokyo dataset included more user and more category than Newyork dataset.

Also, datasets that we used aren't comprehensive and for more straight values these datasets should contain more user and category information. In our condition, it is so difficult to work with enormous information and handle it.

References

[1] J. R. J.A. Konstan, "Recommender systems: from algorithms to user experience," pp. 101–123, 2012.

[2] A. A. A.M. Acilar, "A collaborative filtering method based on artificial immune network," in *Exp Syst Appl, 36 (4)*, 2009, pp. 8324–8332.

[3] M. C. Y. H. L.S. Chen, F.H. Hsu, "Developing recommender systems with the consideration of product profitability for sellers," pp. 1032–1048, 2008.

[4] M. S. A. M. M. Jalali, N. Mustapha, "An efficiency comparison of document preparation systems used in academic research and development," *Exp Syst Applicat*, vol. 9, no. 37, pp. 6201–6212, 2010.

[5] J. Wang, A. P. de Vries, and M. J. T. Reinders, "Unifying user-based and item-based collaborative filtering approaches by similarity fusion," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 501–508. [Online]. Available: https://doi.org/10.1145/1148170.1148257

[6] P. Resnick and H. R. Varian, “Recommender systems,” *Commun. ACM*, vol. 40, no. 3, p. 56–58, Mar. 1997. [Online]. Available: https://doi.org/10.1145/245108.245121