A Fuzzy-Based Personalized Recommender System for Local Businesses

Chun-Hua Tsai
School of Information Sciences
University of Pittsburgh
135 N Bellefield Ave, Pittsburgh, PA 15213
cht77@pitt.edu

ABSTRACT

On-line reviewing systems have become prevalent in our society. User-provided reviews of local businesses have provided rich information in terms of users' preferences regarding businesses and their interactions in reviewing systems; however, little is known about how the reviewing behaviors of users can benefit businesses in terms of suggesting potential collaboration opportunities. In the current study, we aim to build a recommendation system for businesses to provide suggestions for business collaboration. Based on historical data from Yelp that shows two businesses being reviewed by the same users within a same season, we were able to identify businesses that might attract the same customers in the future, and hence provide them with a collaboration suggestion. Our results suggest that the evidence—two businesses sharing reviews from same users—can provide recommendations for businesses to pursue future collaborative marketing opportunities.

Categories and Subject Descriptors

I.5.2 [**Computing Methodologies**]: Design Methodology—*classification, feature selection, recommender system*

Keywords

Fuzzy Logic, Reviewing Network, Recommender System

1. INTRODUCTION

The effect of agglomeration economies is always an interesting research question for urban economic studies. Agglomeration economics is defined as businesses that benefit from their location, shared histories, or workers found near each other [3]. As a result, businesses can share commercial interests, due to economic scale and network effects. Hence, complementary and substitute businesses tend to cluster near each other. For example, the gas stations may locate themselves close to supermarket to attract drivers to refill their cars and perform grocery shopping in the same area. Alternately, similar restaurants may open in the same areas to attract customers with the same tastes. Through city urbanization, these clusters usually gather naturally over a long term, but local businesses can benefit more if we can distinguish potential collaboration opportunities in advance. This turns the task of suggesting collaboration among local businesses into an interesting research problem.

Online reviewing systems like Yelp have become popular destinations in our information society, where people can search for and review businesses, as well as interact with friends. The rich data from these systems provides plenty of opportunities to provide meaningful personalized recommendation results, based on users' historical data. Different approaches have arisen to study personalized recommender systems like these, such as those that are based on user interests [11], social factors [5] and heterogeneous networks [16] to help users browse, search, and explore the system, given their own preference of information. However, little attention has been paid to the recommendations from the business side. As a crowd-sourced local business review and social networking site [9], Yelp provides users with options to share reviews on businesses. These shared reviews allow us understand the latent relationships among different businesses. If two businesses receive overlapping reviews from the same group of users, we can infer that they share the same targeted consumer groups. This provides a different angle to observe and analyze the economic behavior that occurs between businesses.

In this study, we build a review network among businesses, based on user reviews on Yelp. The goal of this study is to predict the likelihood that two businesses will attract the same user's review in the future. The key to fulfilling such a prediction task is the predictor selection [14]. This study further considers various prediction features, including geographical distance, reviewing network, the fuzzy-businesses vector, and content similarity. The experimental result supports the effectiveness of the proposed features, as the prediction model is improved by 31% in AUC value. Based on the result, we aim to build a recommendation system for businesses on Yelp that is dedicated to making suggestions for businesses that share the same target consumers. The system will allow local businesses to easily reach out for potential collaboration opportunities.

The remainder of this paper is divided into six sections. We firstly review the related work of recommender system in section 2. In section 3, we describe the dataset and its preprocessing procedures. In section 4, we discuss the approach of feature selection and experiment setting. The experimen-

HT '16, July 10–13, 2016, Halifax, Nova Scotia, Canada.
© 2016 ACM. ISBN 978-1-4503-4247-6/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2914586.2914641

tal result and prototype system will be shown in section 5. Finally, we summarize our findings and future research directions.

2. RELATED WORKS

A recommender system is used to process, digest, and provide users with meaningful personalized suggestions. One line of research focuses on developing personalized recommendation systems. For instance, [5, 8] have tried to build recommender systems based on the social network of Yelp and user profile information. [5] proposes a unified and personalized recommendation model that is based on probabilistic matrix factorization to explore three social factors in the Yelp network: personal interest, interpersonal interest similarities, and interpersonal interests as a whole. [8] uses the side information of users, based on a set of sparse linear methods, to improve the performance of conventional recommender systems. In the meantime, [13] has adopted a global search approach to collect external data to obtain a people-collaboration recommender system in conference.

The variant information can be combined into heterogeneous networks for recommendation tasks. [16] builds an attribute-rich heterogeneous information network of Yelp reviews and combines various related information from the network with user feedback to provide high-quality recommendations for users. While much prior research has focused on providing recommendations for users based on their similar interests in businesses, social structure, or participation in reviewing systems, little attention has been paid to the relationships between businesses. However, understanding relationships between businesses and further developing a recommendation system for businesses could be beneficial in many different domains, especially in collaborative marketing and location-based recommendations.

Precise user modeling is the key to providing meaningful recommendation results. However, in a real-world recommender system, user preferences are usually vague or may be incomplete. For example, the preference of a movie might be represented as a rating (1-5 stars) and a genre (scary, action, and so forth). As a result, it is challenging to combine the two features into user modeling. The study of [17] proposed the fuzzy-set theoretic method (FTM), which offered a series of methods to manage non-stochastic uncertainty. This approach defined the representational method, aggregation method, and similarity measures for a content-based recommender system. Fuzzy logic is widely used in [10, 18, 15] for recommender systems.

3. DATA

We adopted the Yelp Dataset Challenge dataset[1]. This dataset contains 61,184 businesses and 12 million reviews from 10 cities across 4 countries. The type of businesses included is diverse, and includes restaurants, bars, clubs, and many others. The dataset also includes business metadata, such as geo-location, categories, open hours, and other attributions. We filtered out the businesses who shared fewer than 2 reviews with any other businesses. Based on this criteria, we obtained a shared review network with 24,593 nodes and 2.17 million edges. Each edge is included with comment text from users.

We conducted a preliminary exploration of shared review networks within the dataset. Figure 1(a) shows the scatter plots of geographic distance and user comment overlap. In this plot, we can observe the relationship of comment overlap over geographic distances. Unsurprisingly, businesses that are close together seem to share more reviews. The distribution follows the power law, which means the businesses share reviews in a near distance. However, note the second spike, around 190 to 300 miles. It is interesting that a group of users tend to comment on the same stores over a long distance (exceeding normal urban territory). When further exploring these data points, we find that most of the reviews land in Las Vegas and Phoenix area (See Figure 1(b)). This pattern indicates review behavior across two popular tourist cities. This pattern also implies that people travel between these cities and write a review on Yelp, even from a long distance. These results support the cycle of agglomeration economies over either a short or long distance.

4. APPROACH

4.1 Problem

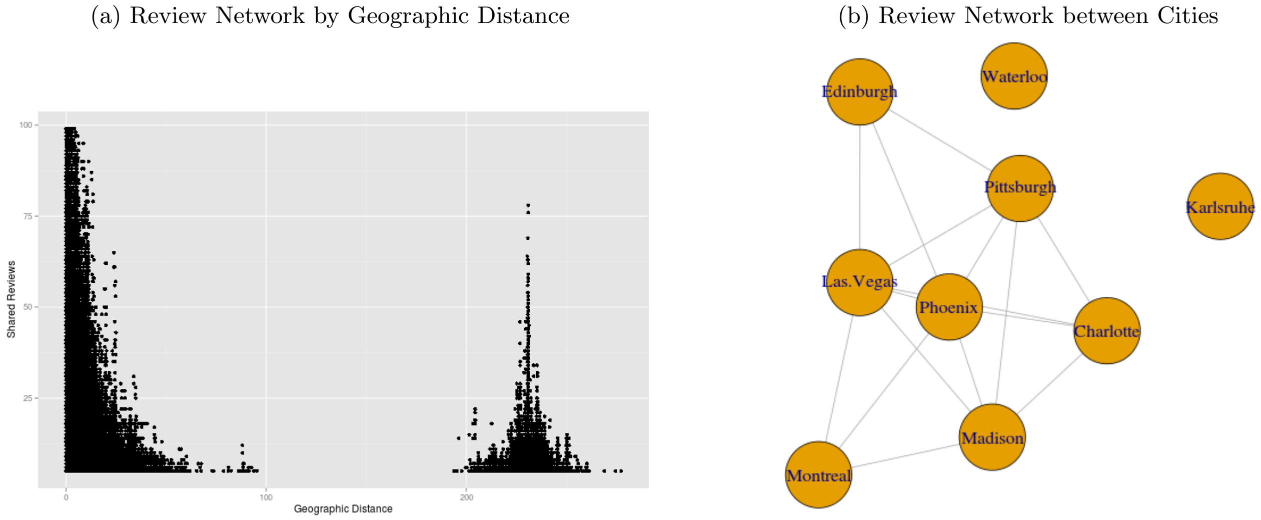
We focus on the link prediction problem for a shared review network among businesses with Yelp reviews. We define the review network as G = (V, E), in which each edge $e = (u, v) \in E$ is an interaction between stores u and v at a particular time, t. Here, the interaction is defined as a connection in terms of sequential user reviews. We can construct a shared review network, based on user review information. Our goal is to predict future reviews at $t' > t$. In other word, the goal is to find a review link that will be formed at a future time t', based on data observed at time t. We can treat this as a binary classification problem to distinguish the positive/negative links at the future time t'.

4.2 Baseline

The content-based approach has been widely used in many different recommender systems [13, 11, 15, 18, 17]. In this study, we adopt the content similarity of Yelp reviews as a baseline. For each business, we aggregate all its reviews into a long text, called comment text. The similarity between two businesses will be represented by the degree of their comment text similarity, which can be calculated with a vector space model (VSM) [12].

The VSM is an algebraic model for representing text documents as vectors. Each document is presented by a high-dimensional vector in the space of words, where each entry corresponds to a different word, and entry value is the number where that word appears in the document. Using VSM, we could obtain a term-document matrix. However, not all terms (or words) are equally important; for example, 'as' is less important than 'aspect' in context. We want to put less discriminatory power on the term in the collection that occurs more frequently. Therefore, we have adopted a classic VSM, which is known as a term frequency-inverse document frequency (TF-IDF) model. The weighting for document d is d is $V_d = [w_{1,d}, w_{2,d}, ..., w_{N,d}]^T$, where $w_{t,d} = tf_{t,d} \cdot \log \frac{|D|}{|d' \in D | t \in d'|}$ and $tf_{t,d}$ is the term frequency of term t in document d. $\log \frac{|D|}{|d' \in D | t \in d'|}$ is the inverse document frequency. $|D|$ is the total number of the document set; $|d' \in D | t \in d'|$ is the number of documents containing the term t.

[1] http://www.yelp.com/dataset_challenge


(a) Review Network by Geographic Distance

(b) Review Network between Cities

Figure 1: (a) Scatter plots of geographic distance and user comment overlap of all Yelp datasets. This distribution shows the two spikes through geographic distances between businesses. The distance within 100 miles follows the power law distribution; namely, that the closer businesses share more customer reviews. The distance of between 190 to 260 miles show a normal distribution that two or more cities shared a review over a long distance; (b) shared review network between 10 cities within the Yelp dataset. The cities that shared more reviews will generally be close to each other. For instance, Las Vegas and Phoenix shared more reviews between each other, while Karlsruhe and Waterloo have limited shared reviews with other cities farther away.

The specific steps that we used to adpot VSM on businesses' comment text are shown as below: 1) Stopwords (such as a, about, an, are, etc.) were removed, because they help construct sentences but do not represent any specific context of reviews. We also did a word-stemming to remove the suffix of each word. For example, 'eat' and 'eating' could be substituted by the single term 'eat'. The remaining words were used to create a term frequency matrix. Second, the comment text of each business could be presented by a high-dimensional vector in our word space. Each entry corresponds to a different word, and entry value is the frequency of that word repeating in this comment text. Third, we used TF-IDF to calculate the re-weighted term frequency vector for each business. Finally, the similarity of comment text between the two businesses is calculated by the Euclidean distance of their vectors.

4.3 Fuzzy Business Vector

The similarity of two businesses can be obtained from shared characteristics [4]. There are various attributes to describe a business, with which we can construct a high-dimensional space to hold all businesses, and each business can be represented as a vector. Given the hypothesis that similar businesses share more similar descriptions than different businesses, similar attributes can be extracted from similar businesses, and will give a short Euclidean distance for these two business vectors. Based on the Yelp data set, selected attributes are: business category, review count, business star, open hours, and an attribute combo, which includes WIFI, parking, credit card acceptance, and other attributes of the business. The same weights are given to all dimensions.

For dimensions with numerical values (like review count), we can simply use their numerical values to calculate the Euclidean distance. On the other hand, to calculate two businesses' Euclidean distance in a categorical dimension, like business categories, we adopt an FLM (fuzzy set theoretic method) [17] as a way to represent the business vector similarity between two businesses. FLM is characterized by its membership function, which is defined as $u_a(x) : x \in X \to [0, 1]$, where X is a domain space. We use two businesses' attribution intersection divided by attribution union, which can be defined as $Sim_{BusinessVector} = \frac{A\cup B - A\cap B}{A\cup B}$, where A and B are the business attribution value sets.

4.4 Review Network

Finding the right features is the key for the prediction model. For link prediction, it is necessary to extract the features that represent some properties between two paired nodes in a network. In Yelp, users can review their comments of the visited businesses. In previous studies [13, 11, 15, 18, 17], the network is defined as a user's social network or coauthorship network. In this paper, the goal is to recommend potential collaboration opportunities for local businesses. Hence, the review network is built by shared reviews. For example, if 2 users have reviewed businesses A and B at the same time, then we define a link between A and B with weight equal to 2. Based on this network, we consider 4 classic network proximity features as predictors [6].

1) Common Neighbors (CN): The CN [7] of two stores x and y is computed as $Sim_{CN}(x, y) = \Gamma(x) \cap \Gamma(y)$, where $\Gamma(\dot{})$ indicates the set of neighbors of the given stores x and y. Here we define the set of neighbors as all stores with shared comments observed at t.

2) Jaccard Coefficient (JC): The JC [2] measures similarity between finite neighbor sets. Here we defined neigh-

bor sets as a shared review at time t. For any two given stores, it is the intersection of their shared review divided by the union of their shared review. It is computed as $Sim_{JC} = \|\Gamma(x) \cap \Gamma(y)\| / \|\Gamma(x) \cup \Gamma(y)\|$, where x or y is the given business and $\Gamma(\cdot)$ represents the shared reviews they have.

3) Adamic/Adar (AA): The AA [1] is a typical local network similarity measurement that considers a weighted parameter between network nodes. The weighted parameter w is defined as $w_{x,y} = \frac{1}{log(z)}$, where z is the shared neighbors between businesses x and y. We then extend the function to $Sim_{AA}(x,y) = \sum \gamma(x) \cap \gamma(y) \cdot w_{x,y}$, where we sum the shared reviews between businesses x, y and multiply by the weight w.

4) Geographic Distance (GD) [14]: The GD is used to measure the actual geographic distance between two businesses. We used the Haversine formula to compute the geographic distance between two points on earth, based on longitude and latitude data. Then the formula can be defined as:

$$Sim_{GD}(x,y) = 2r\ arcsin*$$

$$(\sqrt{haversin(\phi_y - \phi_x) + \frac{cos(\phi_x)cos(\phi_y)}{haversin(\lambda_y \lambda_x)}})$$

where r is the radius of the sphere, ϕ_x and ϕ_y are the latitude of businesses x and y, and λ_x and λ_y are the longitude of businesses x and y.

5. EXPERIMENT

5.1 Design

Let B_t be the set of businesses who shared reviews in year t, and $B_{t'}$ be the set of businesses who shared at least two reviews in year $t' = [t+1]$. To predict a shared review link, we divide B_t into two non-overlapping partitions. The first partition is selected as the training data set and the later partition as the testing dataset. We use $B_{t'}$ as a ground truth to generate the positive and negative link. The positive link is established among the businesses that have at least two shared reviews in B_t' and the negative link is established among those who have no shared reviews record in B_t'.

We divide the whole dataset into 4 seasonal partitions from 2013 to 2014. We will predict the shared review link in t' years, based on the feature information in time t. We also limited the ego network to 4 hops, because this covers the most possible review links. However, the positive and negative links are unbalanced; the negative links are much greater than the positive links. Hence, we randomly choose 1:1 positive and negative links to represent the performance of our proposed model. All the performance measures will be reported by the averaged values of 10-fold cross-validation.

We adopt logistic regression as a classifier in this experiment. The classifier performance is measured by F-Score and AUC as merits to evaluate the model performance. First, the F-Score is calculated by the harmonic mean of precision and recall, for which the formula is $F_1 = 2 * \frac{precision*recall}{precision+recall}$. Second, the area under the curve (AUC) that is the merit of the classifier will rank a randomly chosen positive instance than a negative instance. A higher AUC value means a higher accuracy rate of the classification model. We divide the proposed features into three classification models: 1) content-based baseline; 2) network-based baseline; and 3) a hybrid model. The content-based model is considered to have two features: a fuzzy business vector(BV) and content similarity(CS). The network-based model includes common neighbors (CN), the Jaccard coefficient (JC), Adamic/Adar (AA) and geographic distance (GD). The hybrid model is combined with the two models above.

5.2 Results

In Figure 2, we present the seasonal experiment results with three classification models. Similar performance patterns are shown across four different quarters. There is a major performance improvement between the content-based model and the hybrid model. In detail, the hybrid model is better than the content-based model by 16-17% and the network-based models by 11-15% in its F-score. Moreover, the hybrid model is better than the content-based model by 31-33% and the network-based models by 4-5% in AUC. The hybrid model outperforms the other two models and supports the effectiveness of this prediction model.

Hence, this result helps to predict potential collaborative opportunities between businesses, based on empirical data. Our experiment proves that the business can predict future review links, based on the hybrid model. The best case is shown on quarter 2 with 86% AUC value. This evidence indicates that the classifier is both effective and useful. This finding suggests a further analysis of the economic behavior from a different angle and to further design a local business recommender system. In other words, the hybrid features are effective to fulfill a recommender system for the local businesses that shared the same group of user reviews. This finding may help local businesses to discover potential collaborative opportunities and benefit from common commercial interests.

5.3 Prototype of Recommendation System

Figure 3 shows the demo recommendation system based on the proposed prediction model. We ranked the recommendation result list, based on predicting their likelihood, and plotted them in a map layout. In our demo system, we bundled a Google Map service to provide an AJAX-style user interface. The recommendation result can be used to discover potential commercial opportunities between businesses. For example, a local restaurant might find a link with a nearby nightclub, supermarket, or dessert shop. The businesses may initiate collaboration through the system suggestions. Moreover, in a map view, the business can explore collaborative opportunities by scaling the geographic distance, to find collaborative opportunities across cities. For ecample, note the potential for business collaboration between Las Vegas and Phoenix. Some of the businesses may actually link through the same group of customers. Based on this recommender system, businesses can discover a hidden value, even over a long distance.

6. DISCUSSION

In the current study, we propose an effective methodology to predict the likelihood that any two businesses attract the same group of target reviews by using historical Yelp data. We build a review network that connects two businesses if they received reviews from the same users, and further present a prediction model that integrates networks,

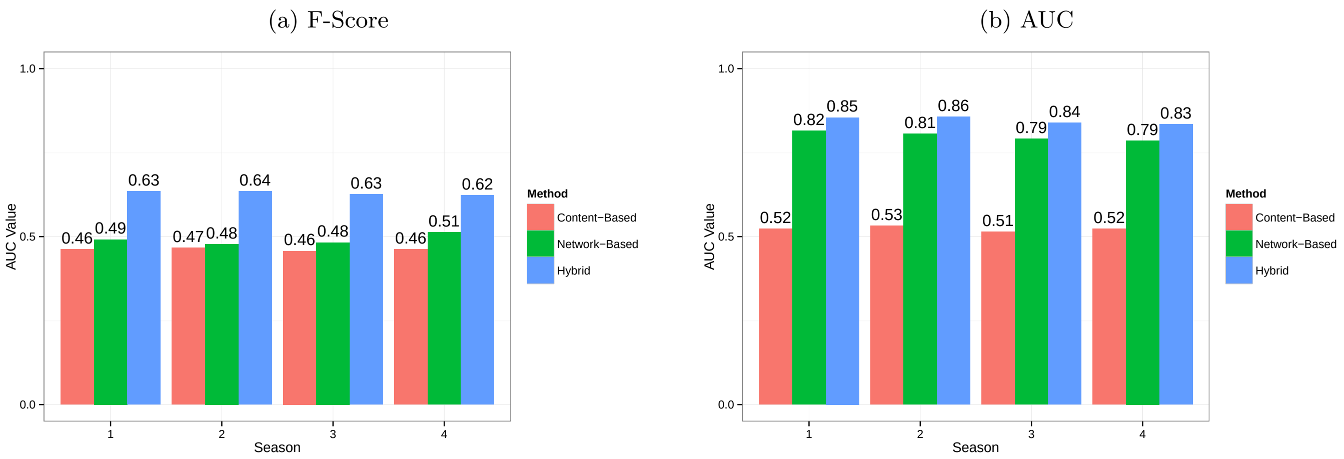


Figure 2: (a) The experiment result of F-Score: in all seasons, the performance of Hybrid model is better than content-based and network-based models by 11%; (b) The experiment result of AUC: in all seasons, the performance of Hybrid model is better than content-based model by 31% and network-based model by 4%.

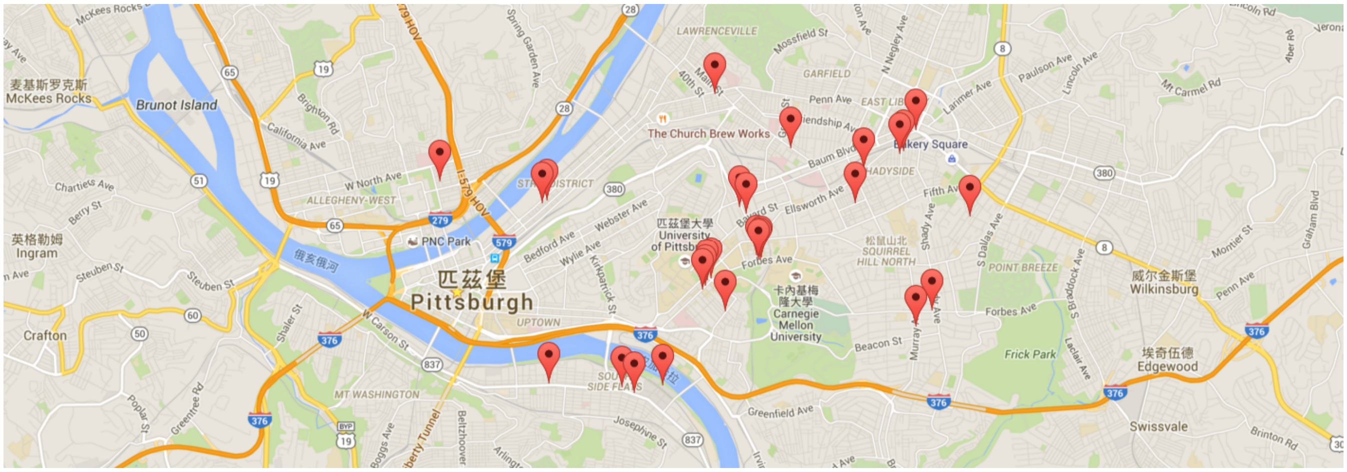


Figure 3: A location-based view of the recommendation result for a local business in Pittsburgh. This result provides the information of potential collaborative opportunities to the local business.

fuzzy business vectors, and content similarity features. Our experimental results suggest that the model performs relatively well in predicting future links between businesses. The best case is outperforming the baseline model by 33%. Furthermore, our prediction model can be used as a core model in a business recommendation system that can generate a list of businesses that will potentially attract the same target customers to each business.

The major contributions of this study are: 1) to the best of our knowledge, this is the first work to build a review network for local businesses to enable collaborative prediction and recommendation. We use user-centric community data to analyze the economic and business value. Second, our proposed prediction features largely outperform the baseline model by 33%. This finding indicates the need to further build up a recommender system for local businesses. Finally, we provide a beta version of a location-based recommender system that local businesses can use to explore the distance-free potential collaborative opportunities, and that may realize a benefit through shared commercial interests.

We acknowledge the current work also has some limitations. Although we present the prototype of business recommendation system using the prediction model, the performance of the recommendation system needs to be further evaluated. In this case, we do not have a baseline to compare our system with, so we plan to conduct a user study to evaluate the overall usefulness and satisfaction of the recommendation results. Also, in the current study, we built the review network based only on the binary variable that a same user comments on both businesses, but that does not consider the variance of shared reviews. It is possible that the positive or negative relationship in the shared-comments network can also contribute to the prediction task. Hence, further investigation into the different ways of building review networks is needed to improve the prediction model. Beyond that, understanding the geographic difference in consuming patterns and mining the costumer shopping sequence are also some future directions for our research.

7. REFERENCES

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[2] G. Chowdhury. *Introduction to modern information retrieval*. Facet publishing, 2010.

[3] P.-P. Combes, G. Duranton, L. Gobillon, and S. Roux. Estimating agglomeration economies with history, geology, and worker effects. In *Agglomeration Economics*, pages 15–66. University of Chicago Press, 2010.

[4] M. Ehrig, A. Koschmider, and A. Oberweis. Measuring similarity between semantic business process models. In *Proceedings of the fourth Asia-Pacific conference on Comceptual modelling-Volume 67*, pages 71–80. Australian Computer Society, Inc., 2007.

[5] H. Feng and X. Qian. Recommendation via user's personality and social contextual. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1521–1524. ACM, 2013.

[6] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559. ACM, 2003.

[7] M. E. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001.

[8] X. Ning and G. Karypis. Sparse linear methods with side information for top-n recommendations. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 155–162. ACM, 2012.

[9] J. M. O'Brien. Yelp's ambitious plan to take over the local ad market. *CNNMoney.com*, 2, 2007.

[10] B. Ojokoh, M. Omisore, O. Samuel, and T. Ogunniyi. A fuzzy logic based personalized recommender system. *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, 2:1008–1015, 2012.

[11] X. Qian, H. Feng, G. Zhao, and T. Mei. Personalized recommendation combining user interest and social circle. *Knowledge and Data Engineering, IEEE Transactions on*, 26(7):1763–1777, 2014.

[12] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[13] C.-H. Tsai and P. Brusilovsky. A personalized people recommender system using global search approach. *iConference 2016 Proceedings*, 2016.

[14] C.-H. Tsai and Y.-R. Lin. Tracing and predicting collaboration for junior scholars. In *Proceedings of the 25th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2016.

[15] S. K. Verma, N. Mittal, and B. Agarwal. Hybrid recommender system based on fuzzy clustering and collaborative filtering. In *2013 4th International Conference on Computer and Communication Technology (ICCCT)*, 2013.

[16] X. Yu, X. Ren, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, and J. Han. Recommendation in heterogeneous information networks with implicit user feedback. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 347–350. ACM, 2013.

[17] A. Zenebe and A. F. Norcio. Representation, similarity measures and aggregation methods using fuzzy sets for content-based recommender systems. *Fuzzy Sets and Systems*, 160(1):76–94, 2009.

[18] Z. Zhang, H. Lin, K. Liu, D. Wu, G. Zhang, and J. Lu. A hybrid fuzzy-based personalized recommender system for telecom products/services. *Information Sciences*, 235:117–129, 2013.