A hybrid approach of feature extraction for content-based recommender system

Andreea Salinca
Faculty of Mathematics and Computer Science
University of Bucharest
Bucharest, Romania
andreea.salinca@fmi.unibuc.ro

***Abstract*—Recommendation systems have gained popularity in the last decades, having a big impact on business models and consumers. This article describes a hybrid approach on feature extraction on a recommender system for business recommendations using a large scale dataset provided by Yelp, one of the most popular review websites. We extract features from the provided dataset using a hybrid technique and we use content-based models in order to identify future users' preferences. We evaluate and compare the systems' performances using Root metrics Mean Squared Error.**

Keywords-recommendation systems; personalization; content based; feature extraction;

I. INTRODUCTION

Recommendation systems have gained popularity in the last years, and are becoming increasingly used in various domains, having a social, economic and demographic impact. Due to the rapid growth of Internet, people have great difficulties in finding the desired information and recommendation systems aim to overcome it.

The development and improvement of recommendation systems has become a demand in the last years, the most popular of them being used in various domains such as e-commerce, businesses, financial services or social networking.

Building efficient recommender systems is a priority in helping people finding local business. Yelp has a very large database of general information provided by thousands of users about any business, reviews and ratings of them. It is a difficult task for users to make a good choice based only on raw data about businesses. The amount of information and Yelp data format is not efficient for users in finding the desired information.

In this paper, we present an approach of a content-based recommendation system of Yelp businesses for predicting users' ratings. We propose an approach to a feature extraction method based on clustering and two predictive models for the recommendation task.

In the first section we present an introduction of recommendation systems and prior work, next we describe the dataset provided by Yelp and the predictive models we built for recommending businesses. Further, we present the obtained results, conclusions and future work.

II. BUSINESS RECOMMENDATION SYSTEM

A. Prior work

When designing recommender systems, two main approaches have emerged: content-based recommender systems, which try to recommend items to a user similar to other items the user has liked in the past, based on a profile of the user's preferences, and collaborative recommender systems which try to find users whose preferences are similar to the given user and recommend the items they liked.

Hybrid recommender systems combing the two filtering approaches have been proposed, Netflix movie recommender system approach being the most popular.

The advantages of content-based recommender systems are: user independency - collaborative approaches need ratings from other users in order to find the nearest neighbors, transparency due to the feature characteristics which are indicators used in recommendation and new item recommendation is possible - this type of systems doesn't suffer from "first-rater" problem as collaborative approaches [1].

However, content-based recommender systems have some drawbacks: the content analysis limitation and over-specialization. The analyzed content has to contain sufficient information in order to be able to distinct items and to make suitable recommendations [1].

The content-based recommendation approach has roots in information retrieval and information filtering domains. Beside the traditional approach, which is based on information retrieval methods, other learning techniques such as Bayesian Classifier, decision trees [2], cluster analysis [3] and artificial neural networks have been used in designing content-based recommendation systems.

These model-based techniques differ from information retrieval-based traditional approaches in that they do not make recommendations based on a heuristic formula, e.g. cosine similarity measure or TF-IDF (Term Frequency-Inverse Document Frequency), but rather are based on a model learned from the underlying data using machine learning techniques [2].

Several techniques have been proposed in building hybrid recommender systems such as collaborative, content-based, demographic techniques, utility-based and knowledge-based techniques. One of the hybridizations techniques can combine features from different data sources into a single recommendation algorithm [4].

In this paper we propose a content-based recommendation system and a hybrid method of feature extraction. Feature extraction plays an important role in building accurate recommendation systems. It is also important for the reduction of computational cost and dimensionality reduction.

B. Dataset

The dataset used for creating the business rating predictions was provided by Yelp in a recommendation system competition organized during ACM RecSys 2013: RecSys Challenge on Kaggle which is a specialized platform for competitions in predictive modeling and analytics [10]. The theme of the competition was to obtain personalized business recommendation for each Yelp user. The task was to predict future ratings on a business-user pair having a rate scale from 1 to 5, where 5 stars represent the highest rating.

The dataset including training dataset and test dataset contains over 10,000 businesses, 8,000 check-in sets, 40,000 users and 200,000 reviews from the Phoenix, AZ metropolitan area (Table I).

Each business set contains detailed information: business ID, business name, neighborhood, full address, city, state, latitude, longitude, star rating (rounded to half-stars), review count, categories names and open status.

Each review set contains detailed information: business ID, user ID, star rating, review text, date and votes ('funny', 'cool' and 'useful' counts). For users there is also the following information: user ID, name, review count, average stars (floating point average) and votes ('funny', 'cool' and 'useful' counts).

TABLE I. YELP DATASET

Name	Training set	Test set
Businesses	11,537	2,797
check-in sets	8,282	1,796
Users	43,873	9,522
Reviews	229,907	36,404

III. PREDICTING FUTURE RATING OF BUSINESSES

In creating the model of the content-based recommender system that would predict future business ratings, we focus on feature extraction on Yelp dataset in order to find the user's relation to a business, using clustering techniques. We customize traditional techniques to comprise distinct features unique to the dataset, such as location, number of check-ins, user gender, state, and review counts.

A. Feature extraction

Several hybridization techniques can be combined to build a recommendation system. A common technique in building recommender systems is to extract and to combine features from different knowledge sources. Knowledge-based recommender systems can contain: user knowledge, such as demographical information about the user or his preferences, or knowledge on the features of the objects and that are being recommended [4].

We use the demographic information about the users and their ratings of items in our approach of building a content-based recommender system. Demographic recommender systems categorize users based on their attributes and make recommendations based on the associated classes. However, we use the demographic information for feature extraction.

We use a density-based clustering algorithm OPTCIS (Ordering points to identify the clustering structure) [5] to cluster locations of businesses using latitude and longitude coordinates. OPTICS is an algorithm similar with DBSCAN, which uses the definition of density as a number of points within a specified radius in order to create the cluster, but it overcomes the shortcoming of DBSCAN in detecting clusters of various density data.

The parameters used for clustering business location are: 5 MinPts cluster size, 200 meter radius for neighborhood consideration (ε-maximum radius parameter) and 100 meter threshold (ε –Neighborhood parameter). After clustering the business training data, Fig. 1, we obtain 106 clusters.

Features derived from multiple knowledge sources are combined together to be used in a single recommendation model. Further, we analyze business categories data and observe that the data is very sparse: from over 300 categories, no more than 10 are used to describe any business.

Some of the most frequent names of categories include: restaurants, food, hotels, apartments, shopping, beauty, health. We factorize all category business names and choose 5 of the most frequent of them as features to express businesses. Due to high dataset sparsity, a clustering technique applied on business categories would prove ineffective in this case.

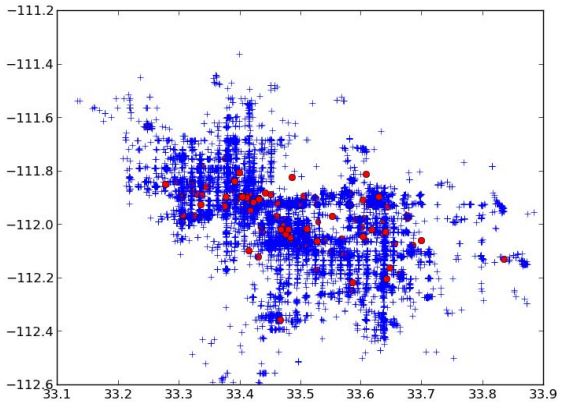

Figure 1. Clusters of business training data.

We also choose, as features, the number of reviews given to a business by users and the open state of a business.

Considering the idea that people with similar genders will give similar ratings on a business, we compute a gender list for each business based on the user name provided in the dataset. We use a list with most popular given names to match the gender of users in Yelp dataset.

Next, we analyze the number of check-ins sets (containing hours and number of counts) and express a business in terms of it. For each business in the dataset we extract 24 different features corresponding to the time of check-in, computing the average number of check-ins, Fig. 2. We combine these features to describe each business for predicting user's ratings.

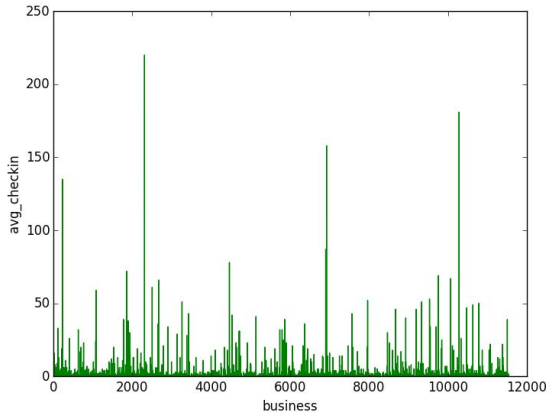

Figure 2. Business check-in features.

B. *Predictive model approaches*

In designing the proposed recommendation system we use two model approaches for predicting future business ratings.

The recommendation system uses *Decision Regression Trees* to create a model that predicts the value of the ratings (1) by learning decision rules inferred from data features characteristics.

$$\hat{R}(user_{id}, business_{id}) \quad (1)$$

We use CART (Classification and Regression Trees) algorithm for regression which constructs binary trees using the feature and threshold that contains the largest information acquired at each node instead of computing rule sets [5]. The decision trees will recursively partition the training space in order to group similar items until a maximum depth is reached and the selected parameter minimizes the equation (2).

$$\theta^* = argmin_\theta G(Q, \theta) \quad (2)$$

Where the data at a node j is Q.

Another approach in designing the predictive model was to use a generalized linear model: *Ridge Regression,* which adds a penalty on the size of coefficients (3) in the problem of ordinary least squares [6].

$$min\|\mathrm{Xw} - \mathrm{y}\|_2^2 + \alpha\|\mathrm{w}\|_2^2 \quad (3)$$

Where $w = (w_1, \ldots, w_p)$ and y is the predicted value.

In the next section we present performance results of the business recommendation system using the two modeling approaches with the extracted features.

IV. RESULTS

As evaluation metric we use a popular [8] metrics used to measure the performance of recommendation systems: Root Mean Squared Error (RMSE) (4).

$$\sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \quad (4)$$

where

- n is the total number of review ratings to predict.
- p_i is the predicted rating for review i.
- a_i is the actual rating for review i.

First we evaluate the RMSE of the decision tree model approach for the business recommendation system.

Using a part of the characteristics described in the previous section, excluding features such as clustering of locations and gender, we obtain a RMSE of 1.32302 when choosing a maximum depth of the decision tree of 2. We obtain an improvement on the RMSE score when using clustering technique on feature characteristics: with maximum depth 10, we obtain 1.31367 (Table II).

TABLE II. TABEL II. RMSE SCORE DECISION TREE

RMSE	Max tree depth
1.31367	10
1.31467	8
1.31820	5

After using the gender feature in the model, we obtain a better RMSE score: with maximum depth 10 of the decision tree we obtain a decrease of 0.07.

We also compute the feature importance in the decision tree, known as Gini importance [9], as the normalized total reduction of the criterion brought by each feature.

The highest importance features in the proposed model are business categories, location and gender.

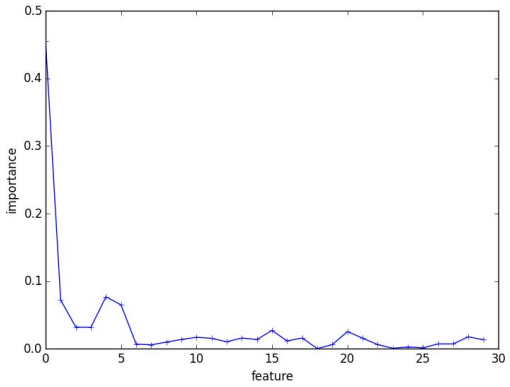

Figure 3. Feature importance in the model.

When using Ridge Regression in the design of the predictive model we obtain a RMSE score of 1.31604. The result is close to the RMSE score obtained using Decision Regression Tree approach. Therefore, Decision Tree seems to perform better in predicting the business ratings.

V. CONCLUSIONS AND FUTURE WORK

The content-based business recommender system using two predictive model approaches has proven the best RMSE score of 1.24 when using Decision Regression Trees. We show that the proposed feature extraction method on YELP dataset has proven efficient in predicting future business ratings.

In the ACM Yelp RecSys challenge, [10] the best RMS score obtained was 1.21251 using a hybrid recommender system combining multiple filtering approaches as collaborative filtering and content-based filtering. In [11] the author uses as predictor GBM and random Forests for the business recommendation system and obtains a 1.21552 RMSE score placing on the second place in the leader board. Other approaches such as [12] use nearest neighbor, matrix factorization and clustering, obtaining a recommender system with accuracy of 1.24039 RMSE.

However, the feature extraction method is very important in building accurate recommender systems. In future research we could extract more features and process comments on businesses using information retrieval techniques in order to improve RMSE score. Also, we could apply dimensionality reduction on the extracted features.

REFERENCES

[1] P. Lops, M. de Gemmis and G. Semeraro. "Content-based Recommender Systems: State of the Art and Trends", in Recommender Systems Handbook: A Complete Guide for Research Scientists & Practitioners, P. Kantor, F. Ricci, L. Rokach and B. Shapira, Berlin: Springer, 2010, pp. 73-105.

[2] Cho, Yoon Ho, Jae Kyeong Kim, and Soung Hie Kim. "A personalized recommender system based on web usage mining and decision tree induction." Expert Systems with Applications 23.3 , 2002, pp. 329-342.

[3] Adomavicius, Gediminas, and Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." Knowledge and Data Engineering, IEEE Transactions on 17.6, 2005, pp. 734-749.

[4] R. Burke: "Hybrid recommender systems: Survey and experiments." User modeling and user-adapted interaction, vol. 12.4, 2002, pp. 331-370.

[5] Ankerst, Mihael, et al. "Optics: Ordering points to identify the clustering structure." ACM Sigmod Record. Vol. 28. No. 2. ACM, 1999.

[6] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.

[7] Jahrer, Michael, Andreas Töscher, and Robert Legenstein. "Combining predictions for accurate recommender systems." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.

[8] Shani, Guy, and Asela Gunawardana. "Evaluating recommendation systems." In Recommender systems handbook, Springer US, 2011, pp. 257-297.

[9] L. Breiman, and A. Cutler, "Random Forests", http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm . Accessed September, 2014.

[10] https://www.kaggle.com/c/yelp-recsys-2013

[11] N. Vladimir. "Hybrid Recommender System for Prediction of the Yelp Users Preferences." *Advances in Data Mining. Applications and Theoretical Aspects*. Springer International Publishing, 2014, pp. 85-99.

[12] N. Carrillo et al."Recommender Systems Designed for Yelp.com", http://www.math.uci.edu/icamp/summer/research/student_research/recommender_systems.pdf