UNIVERSITY OF BUCHAREST

Faculty of Mathematics and Computer Science

Department of Computer Science

Contributions in Recommender Systems

(Abstract)

Supervisor:

Prof. Ph.D. Denis Enăchescu

Ph.D. student:

Andreea Salinca

Bucharest, September 2017

1. Introduction

The huge amount of information available on World Wide Web and the number of Internet users had an explosive growth in the last ten years. Moreover, a consistent growth is registered year after year. For each person, it has become a challenge to access items of interest on the Internet. The ability of humans to dissect relevant information is being exceeded. Information retrieval systems have partially solved this problem, but without addressing the need for prioritization and personalization of information. Recommender systems research has roots in cognitive science, information retrieval, approximation theory and marketing modelling (Adomavicius & Tuzhilin, 2005).

In the last years, recommender systems have changed our lives from the way that we find information, products or we connect to other people. Recommender systems represent a set of advanced techniques that have the ability to predict the preferences of a particular user.

Recommender systems are used to suggest relevant items to a user in a decision-making process such as which product to buy, which movie to watch, what music to listen, what books or news to read, which restaurant to choose for eating, which hotel to choose for staying or travel recommendations based on his preferences, history, profile, individual traits or demographic information (Ricci, Rokach, & Shapira, 2011).

We first seek to ask for the opinion of our family or friends in a making a decision. Next, we choose to seek for other opinions to provide useful and effective suggestions because we consider that we may be biased by the first option. Recommender systems are intended for individuals who lack sufficient information, knowledge or competence in evaluating the majority of alternative options (Resnick & Varian, 1997).

Within the research community, the problem of having access to relevant information is addressed by developing algorithms and techniques that would provide high quality and personalised recommendations. It has become increasingly important for people to have access to various information according to their preferences and tastes especially in the context of large datasets. The same principle is valid for companies as well, they must provide their users relevant and accurate information to maximise their target and increase their revenue.

Although many approaches in building recommender systems have been developed in the last years, the interest in the domain remains still high due to the rapidly increasing amounts of information being produced nowadays.

2. Motivation

Significant research efforts have been put to provide relevant and desirable information by developing new algorithms and approaches in building efficient recommender systems.

Since the early 90s when collaborative filtering was introduced, algorithms have evolved along with researchers' understanding of the challenges and particularities of recommendation systems and application to ubiquitous domains. Around 2006 Netflix, a major online DVD rental and video streaming service, announced a competition of 1M $ for improving the accuracy of their systems with 10% that increased the interest on this research topic (Netflix, 2006). The wining approach was using an ensemble of collaborative filtering algorithms and blending a diverse set of predictors which surpassed Netflix's *Cinematch* algorithm for predicting ratings by 10.06% (Töscher, Jahrer, & Bell, 2009). In the same year, MyStrands organized a summer school on the recommender systems Recommenders06.

With the rapid commercialization of recommender systems on web, challenges in scale and value have arisen. These systems should handle more than previous research scale, millions of users and items in real time transactions dealing with thousands or more requests per second. Researchers understood the problem and addressed it along with the technological challenges. A set of new algorithms was developed including item-based correlation, dimensionality reduction approaches and evaluating a top-n list of recommendation items.

Since 2007, ACM has a dedicated conference: ACM Recommender Systems Conference to the topic. Other conferences have dedicated topics or even workshops on recommendations: IJCAI International Joint Conference on Artificial Intelligence, AAAI Conferences on Artificial Intelligence, ACM KDD Knowledge Discovery and Data Mining, European Conference on Information Retrieval, IEEE Conference on Data Mining, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Moreover, the research in recommender systems had a significant growth and influences from different domains: from artificial intelligence, data mining, information retrieval to security, privacy, and marketing research (Jannach, Zanker, Felfernig, & Friedrich, 2010).

The popularity of recommending systems has increased during the last ten years; both academia and industry have developed tools and techniques for reducing the complexity when searching for useful and relevant information. Recommender systems had proven to be a valuable solution and a powerful tool in electronic commerce or various services. Furthermore, some vendors, such as Google, Amazon, Netflix, LinkedIn, IMDb, Yahoo, YouTube, and Yelp have developed services in which incorporated recommendation into their platforms (Ricci, Rokach, & Shapira, 2011).

The interest in the research area of recommender systems remains high due to the need of users to deal with information overload and data diversity and the need for developing practical applications that will provide personalized recommendations and content.

In this thesis, we build upon previous recommendation approaches, by introducing several new methods and efficient algorithms in the design of recommender systems.

To achieve more effective recommendations, we focus on the research of designing and elaborating predictive models in building several recommender systems, using multiple large-scale databases provided by Yelp in a challenge for research purposes. Yelp is a company that has over 100 million user reviews on worldwide businesses and provides recommendations of top restaurants, shopping, entertainment, and various services.

We conduct our research in the area of sentiment analysis, opinion mining, sentiment mining and sentiment extraction to develop an algorithm for business reviews classification.

We focus on deep learning algorithms with significant scalability while using large real world datasets.

3. Thesis structure

The thesis is organized into seven chapters as follows:

1. In Chapter 1 we introduce and motivate the problem addressed in this thesis. This chapter is divided into 4 sections. We present an overview and the main contributions of this thesis.
2. Chapter 2 provides a review of the state of the art in recommender systems and a classification of the main types of different approaches and a literature survey on the performance evaluation of recommender systems. This chapter is divided into 3 sections:
    a) In the first section, we formally define the recommendation problem.
    b) In the second section, we present the main recommendation techniques in developing recommending systems which into three main categories: content-based, collaborative filtering and hybrid recommendation approaches. We describe the limitations of various recommendation techniques and present several hybrid methods that overcome these problems.
    c) In the third section, we discuss the evaluation of the performance of recommender systems and we present two main approaches in the evaluation experiment: offline and online. We introduce the most important metrics for the performance evaluation of recommender systems used in the literature (Shani & Gunawardana, 2011).
3. In Chapter 3, we develop a content-based recommender system, and we propose a hybrid method of feature extraction. We extract features from the Yelp dataset (Yelp, Yelp 2013 Dataset Challenge, 2013) using a hybrid technique, and we use content-based models to identify users' preferences. We evaluate the performance of our system and show that it is capable of making good rating predictions for businesses. This chapter is divided into 5 sections:
    a) In the first section, we illustrate our approach to developing a content-based recommendation system by building a model to predict the rating a user will give to a business.
    b) In the second section, we present an introduction to content based systems and previous work.

   c) In the third section, we describe the Yelp dataset and the predictive models we built for recommending businesses. We propose our approach to a feature extraction method based on clustering and two predictive models for the recommendation system.
   d) In the next section, we present the evaluation results of the proposed content-based recommendation system.
   e) In the last section, we present some conclusion remarks, comparison to other approaches and some ideas on future work.

4. Chapter 4 presents the research in developing an algorithm that processes users text reviews using information retrieval and deep learning techniques. We create a model for business classification using Restricted Boltzmann Machines to extract binary features. We propose multiple learning algorithms for the classification of business text reviews to predict if a business is good for children or not. We show that our experimental results using RBMs modelling outperform traditional classifiers. This chapter is structured as follows:
   a) In the first section, we show our approach in creating a model useful for business prediction using Restricted Boltzmann Machines to extract binary features.
   b) In the second section, we present a state of the art in developing recommending systems using Restricted Boltzmann Machines.
   c) In the third section, we use introduce a new algorithm using a binary version of RBMs as a feature extractor for modelling business text reviews which are processed in a Bag-of-words representation. Further, the features extracted using binary stochastic RBMs are used in a classification problem to predict if a business is good for children.
   d) Next, we present the results of the classification task using the proposed model comprising Restricted Boltzmann Machines feature extraction, using vectorized text reviews with tf-idf (Term Frequency-Inverse Document Frequency), and a Logistic Regression Classifier. We compare our results against the results obtained using traditional classifiers such Stochastic Gradient Descent Classifier.
   e) In the last section, we give some ideas on future work.

5. Chapter 5 describes novel approaches capable of detecting the sentiment polarity of a business review. We conduct our research in the area of sentiment analysis

by introducing new algorithms in which we apply natural language processing techniques in the pre-processing steps to classify user sentiments with respect to the star ratings. We illustrate a comparative study on the effectiveness of the ensemble methods for sentiment the classification task. This chapter is organized into 5 sections:

   a) In the first section, we briefly describe our method for classifying the polarity of a high number of user-generated reviews for restaurants, businesses, and service providers with respect to the star rating scale.
   b) Next, we present previous work regarding sentiment analysis classification of text reviews using several machine learning techniques.
   c) In section three we propose two algorithms for feature extraction. The first approach is to build a custom dictionary from the training dataset. The second approach is to conduct a lexical analysis of business text reviews. We apply the following machine learning classifiers: Naive Bayes, Linear Support Vector Classification (SVC), Logistic Regression and Stochastic Gradient Descent (SGD) Classifier.
   d) In next section, we evaluate the accuracy of the proposes methods. Our approach explores several feature extraction methods and learning methods for classifying business text reviews using a large-scale dataset: *Yelp challenge dataset* that contains more than 1.6 million reviews (Yelp, Yelp 2015 Dataset Challenge, 2015).
   e) In the last section, we discuss new ideas that arise from this work.

6. Chapter 6 continues the research in the area of sentiment analysis and deep learning and the work described in the previous chapter by introducing a particular approach for business reviews classification using word-based Convolutional Neural Networks. We propose two neural network models approaches having different architectural size and multiple word vector representations: pre-trained word embeddings and end-to-end vector representations. We conduct an empirical study on the effect of hyperparameters on the overall performance in classifying the sentiment of text reviews. Furthermore, we implement several experiments on a large-scale dataset to capture the semantic relationship between reviews. We prove that our results are competitive with traditional methods approaches. This chapter is structured as follows:

a) In the first section, we present an introduction on Convolutional Neural Networks and we introduce our approach for business reviews classification.

b) Next, we discuss previous results in the area of sentiment analysis and text classification on large-scale databases using Convolutional Neural Networks.

c) In section three we introduce two neural network models approaches with different architectural size and several pre-trained or end-to-end learned word representations with different embedding sizes.

d) Next, we present a series of experiments made to explore the effect of architecture components on model performance along with the hyperparameters tuning, including filter region size, number of feature maps, and regularization parameters of the proposed convolutional neural networks. We discuss the design decisions for sentiment classification on Yelp 2017 challenge dataset (Yelp, Yelp 2017 Dataset Challenge, 2017) that comprises 4.1M user reviews and we offer a comparison between these models and report the obtained accuracy. Moreover, since as at the moment of this writing there are no results reported on Yelp 2017 challenge dataset on the sentiment classification task, we extend our experiments on the Yelp 2015 polarity dataset to compare with the other results reported in (Zhang, Zhao, & LeCun, 2015).

e) In section five we conclude the results present in this chapter and present future ideas for sentiment categorization of Yelp user text reviews.

7. In the last chapter, we conclude and summarize the main contributions of this thesis. We discuss open problems and possible future directions for the improvement of this research in recommendation systems.

4. Contribution

We discussed the motivation for our research in recommendation systems field. Next, we present the contributions of the work presented in this thesis. The most important results that have been published in (Salinca, A hybrid approach of feature extraction for content-based recommender system, 2014), (Salinca, Restricted Boltzmann Machines for modeling businesses, 2014), (Salinca, Business reviews classification using sentiment analysis, 2015) and (Salinca, Convolutional Neural Networks for Sentiment Classification on Business Reviews, 2017).

The principal objective of the current thesis is to develop effective approaches in building recommending systems using different machine learning techniques and focusing on deep learning algorithms with significant scalability while using large real-world datasets.

To achieve more effective recommendations, we focus on the research of designing and elaborating predictive models in building several recommender systems using Yelp databases Challenge (Yelp, Yelp 2013 Dataset Challenge, 2013), (Yelp, Yelp 2015 Dataset Challenge, 2015), (Yelp, Yelp 2017 Dataset Challenge, 2017).

We also participated in a competition organized by Yelp - a company that has over 100 million user reviews on businesses worldwide and provides recommendations of top restaurants, shopping, entertainment, and services: Yelp Dataset Challenge (Yelp, Yelp 2013 Dataset Challenge, 2013). The first contribution is discussed in Chapter 3. We introduce a hybrid approach on feature extraction on a recommender system using a large-scale dataset provided by Yelp, one of the most popular review websites. We extract features from the Yelp dataset using a hybrid technique, and we use content-based models to identify future users' preferences. We evaluate and compare our system performances using Root Metrics Mean Squared Error (RMSE) (Salinca, A hybrid approach of feature extraction for content-based recommender system, 2014).

As part of our research, we develop a content-based recommendation system and a hybrid method of feature extraction. Feature extraction has a major role in building accurate recommendation systems. Feature extraction is essential for the reduction of computational cost and data dimensionality reduction. We identify several variables relating to features extracted from business and user modelling, and we evaluate the impact of the variables on the user-modelling effectiveness with an offline evaluation based on 200 000 recommendations displayed to 40 000 users. The proposed content-based business recommender system has proven a RMSE score of 1.24 when using Decision Regression Trees as a predictive model. In the ACM Yelp RecSys challenge (Kaggle, 2013) the best RMSE score obtained was 1.21251 using a hybrid recommender system combining multiple filtering approaches as collaborative filtering and content-based filtering (Salinca, A hybrid approach of feature extraction for content-based recommender system, 2014).

Next, the contribution from Chapter 4 presents the research in developing the second recommender system focusing on modelling the user-generated business reviews from Yelp dataset to predict if a business is good for children using only text reviews. In our approach, we extract nonlinear features from raw user's text reviews using a probabilistic model: Restricted Boltzmann Machines (RBMs), and we create a model for business prediction. We combine information retrieval methods for processing business text reviews along with a deep learning technique (RBMs) to extract binary features, and we consider the business prediction problem as a classification problem using different learning algorithms (Salinca, Restricted Boltzmann Machines for modeling businesses, 2014).

In our experimental results, we prove that the RBMs feature extraction using vectorized text reviews with tf-idf and a Logistic Regression Classifier outperforms results using traditional classifiers such as Stochastic Gradient Descent Classifier (SGD) and a Logistic Regression Classifier (Salinca, Restricted Boltzmann Machines for modeling businesses, 2014).

Furthermore, in Chapter 5, we aim our research to develop an algorithm for business reviews classification using another large-scale review dataset provided by Yelp: Yelp Challenge dataset (consisting of more than 1.6 million user reviews) in the area of sentiment analysis, opinion mining, sentiment mining and sentiment extraction. We build several *ensemble learning* methods for *sentiment classification* of business text *reviews* using two feature extraction methods and four machine learning models. On top, we propose new algorithms in which we apply natural language processing techniques in the pre-processing steps to achieve high accuracy in the classification task. We classify users' sentiments as either positive or negative with respect to the star ratings using only the text review for restaurants, businesses and service providers (Salinca, Business reviews classification using sentiment analysis, 2015).

This is the first approach to classifying sentiment of text reviews using the latest reviews from Yelp challenge dataset (Yelp, Yelp 2015 Dataset Challenge, 2015). We demonstrate that the sentiment analysis algorithm built on top of machine learning algorithms like Naïve Bayes and Linear Support Vector Classification (SVC) have accuracy above 90% and the best results were obtained using Linear SVC and Stochastic Gradient Descent Classifier (SGD) classifiers with an accuracy of 94.4%. However, in the research of Wilson et al. human analysis of sentiments expression, feelings, emotions, and opinions represents a difficult task. A rate of 82% agreement has been measured for automatically identify subjective expression between two persons, proving that the algorithm we have proposed is efficient in classifying sentiments (Salinca, Business reviews classification using sentiment analysis, 2015).

Convolutional Neural Networks (CNNs) models have proven remarkable results for text classification and sentiment analysis. We further continue our research on sentiment analysis described in Chapter 6, and we present our approach to the task of classifying text reviews using word embeddings. We compare word-based CNNs using several pre-trained word embeddings and end-to-end vector representations for text reviews classification.

We conduct several experiments on a large-scale dataset, Yelp 2017 challenge dataset provided by Yelp (Yelp, Yelp 2017 Dataset Challenge, 2017), to capture the semantic relationship between reviews and we use deep learning techniques that prove that the obtained results are competitive with traditional methods. We use several pre-trained models of word embeddings built with an unsupervised learning algorithm for obtaining vector representations of words: GloVe (Karpathy & Fei-Fei, 2015), word2vec along with pre-trained vectors trained on 100 billion words of Google News dataset. The models contain 100-dimensional vectors for 3 million words and phrases (Mikolov, Chen, Corrado, & Dean, 2013). We also experiment fastText pre-trained word vectors for the English language that are an extension of word2vec. These vectors in dimension 300 were trained on Wikipedia using the skip-gram model described in (Bojanowski, Grave, Joulin, & Mikolov, 2017) with default parameters. Moreover, we use in the embedding layer of both proposed CNNs models a 100-dimensional word2vec embedding vectors that we have trained using the text reviews in the training dataset.

We report an accuracy score of 95.6% using 3-fold cross validation using pre-trained fastText vector embeddings and a CNN having three filter regions sizes and 128 feature maps. (Salinca, Convolutional Neural Networks for Sentiment Classification on Business Reviews, 2017).

Moreover, since at the moment of the as at the moment of this writing there are no results reported on Yelp 2017 challenge dataset, we evaluate our model approach on Yelp 2015 polarity dataset to compare our approach against other works. We report a score of 94.27% using the word-based CNN with pre-trained word2vec with 300-dimension word embeddings.

In (Zhang, Zhao, & LeCun, 2015) the authors propose character-level CNNs that achieve an accuracy of 94.11% for the large-featured architecture and 93.47% for the small-featured architecture on Yelp 2015 polarity dataset. The authors also present two baseline implementations for word-based CNNs with pre-trained word2vec and achieve a score of 95.40% for a large-featured architecture and 94.44% for a small-featured architecture on the Yelp 2015 polarity dataset (Zhang, Zhao, & LeCun, 2015). We show that our CNNs obtain comparable results to the ones in the literature for the task of classifying business reviews (Zhang, Zhao, & LeCun, 2015), (Tang, Qin, & Liu, 2015).

In this thesis, we will provide an in-depth description of machine learning challenges that arise in the context of developing recommender systems and will describe novel approaches, ideas, concepts, and techniques.

We have presented the work from this thesis at leading artificial intelligence conferences, such as International Joint Conference on Artificial Intelligence (IJCAI) (Salinca, Convolutional Neural Networks for Sentiment Classification on Business Reviews, 2017).

We hope that our innovative approaches will help recommender systems research to move forward to new rises.

5. Bibliography

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering, 17*(6), 734-749.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics 5, 2*, 135-146.

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: an introduction.* Cambridge University Press.

Kaggle. (2013). *ACM Yelp RecSys challenge*. Retrieved September 2014, from Kaggle: https://www.kaggle.com/c/yelp-recsys-2013

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 3128–3137).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Netflix. (2006). *The Netflix Prize*. Retrieved from Netflix: http://www.netflixprize.com/

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM, 40*(3), 56-58.

Ricci, F., Rokach, L., & Shapira, B. (2011). *Recommender Systems Handbook.* Springer US.

Salinca, A. (2014). A hybrid approach of feature extraction for content-based recommender system. *4th International Conference on Hybrid Intelligent Systems, HIS* (pp. 54-57). Trends in Innovative Computing.

Salinca, A. (2014). Restricted Boltzmann Machines for modeling businesses. *International Conference on Artificial Intelligence and Pattern Recognition, AIPR* (pp. 24-28). SDIWC.

Salinca, A. (2015). Business reviews classification using sentiment analysis. *Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2015 17th International Symposium on.* Timisoara: IEEE.

Salinca, A. (2017). Convolutional Neural Networks for Sentiment Classification on Business Reviews. *IJCAI Workshop on Semantic Machine Learning.* Melbourne: CEUR.

Shani, G., & Gunawardana, A. (2011). *Evaluating Recommendation Systems.* Boston, MA: Springer US.

Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. *EMNLP*, (pp. 1422-1432).

Töscher, A., Jahrer, M., & Bell, R. M. (2009). The bigchaos solution to the Netflix grand prize. *Netflix prize documentation*, 1-52.

Yelp. (2013). *Yelp 2013 Dataset Challenge*. Retrieved from Yelp Dataset Challenge: http://www.yelp.com/dataset_challenge

Yelp. (2015). *Yelp 2015 Dataset Challenge*. Retrieved from Yelp 2015 Dataset Challenge: https://www.yelp.com/dataset_challenge

Yelp. (2017). *Yelp 2017 Dataset Challenge*. Retrieved from Yelp Dataset Challenge: https://www.yelp.com/dataset_challenge

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, (pp. 649-657).