2011

User Profile Construction in the TWIN Personality-based Recommender System

Alexandra Roshchina
Technological University Dublin

John Cardiff
Technological University Dublin, john.cardiff@tudublin.ie

Paolo Rosso
Universidad Politecnica de Valencia

Recommended Citation

Roshchina A., Cardiff J., Rosso P., User Profile Construction in the TWIN Personality-based Recommender System, Workshop on Sentiment Analysis where AI meets Psychology SAAIP, 5th International Joint Conference on Natural Language Processing, IJCNLP, 2011

User Profile Construction in the TWIN Personality-based Recommender System

Alexandra Roshchina
ITT Dublin / Ireland
sasharo@itnet.ie

John Cardiff
ITT Dublin / Ireland
John.Cardiff@ittdublin.ie

Paolo Rosso
NLE Lab-ELiRF, Universidad Politécnica de Valencia / Spain
prosso@dsic.upv.es

Abstract

The information overload experienced by people who use online services and read user-generated content (e.g. product reviews and ratings) to make their decisions has led to the development of the so-called recommender systems. We address the problem of the large increase in the user-generated reviews, which are added to each day and consequently make it difficult for the user to obtain a clear picture of the quality of the facility in which they are interested.

In this paper, we describe the TWIN ("Tell me What I Need") personality-based recommender system, the aim of which is to select for the user reviews which have been written by like-minded individuals. We focus in particular on the task of User Profile construction. We apply the system in the travelling domain, to suggest hotels from the TripAdvisor[1] site by filtering out reviews produced by people with similar, or like-minded views, to those of the user. In order to establish the similarity between people we construct a user profile by modelling the user's personality (according to the Big Five model) based on linguistic cues collected from the user-generated text.

1 Introduction

With the transformation of the Web from the static data source into the interactive environment that allows users to actively communicate with each other and produce shared content, the amount of the information available online has grown tremendously. As a result the task of automatic data analysis has emerged to help people make better choices of the ever increasing number of products and services. In situations where the number of alternatives is very large, people tend to rely on the opinions of experts. Regarding the Web, so-called "recommender systems" (Ricci et al., 2010) have been constructed to serve this expert function following the user-oriented approach in the online world. There are many existing recommenders on the Internet nowadays serving different purposes such as recommending films (Movies2Go[2]), music and TV programs (Last.fm[3]), etc.

One of the domains in which the necessity of making a good choice is very important is travelling. People are faced with a high degree of uncertainty when choosing a place (hotel, restaurant) they have never been to, consequently they must rely on other travellers' reviews which sites such as TripAdvisor provide. However, when the number of such reviews becomes large, it is critical to provide a filtration system, whereby the reviews most likely to be valued by the reader are highlighted. In Viney (2008), it has been shown that people normally do not go further than the second or third page in search results.

In this paper, we propose the "Tell me What I Need" (TWIN) recommender system, the goal of which is to select reviews written by people with like-minded views to those of the reader to get the list of hotels that could be of interest for him.

[1] http://www.tripadvisor.com

[2] http://www.movies2go.net

[3] http://www.last.fm

A critical component of this task is to accurately construct the profile of the user. Traditionally, this has been constructed from the person's preferences or their explicit or implicit ratings (Ricci et al., 2010), however in our case we follow the emerging approach of personality-based user profile construction (Nunes, 2008) from linguistic cues retrieved from the text of the user reviews (Mairesse et al., 2007).

The paper is organized as follows. In Section 2 we provide an overview of online recommender systems and proceed to discuss the modelling of personality with the emphasis on the identification of a writer's personality from the text they write. We give an overview of the importance of the social sites in the online travelling domain. In Section 3 we present the prototype of the TWIN recommender system. In Section 4 we discuss the preliminary work on the evaluation of the TWIN system in regard to the User Profile construction. Finally, the conclusions are presented in Section 5.

2 Background

2.1 Recommender systems

The function of a recommender system is to assist a person to make the right decision about choosing a particular product or service, from the vast number that is available. This functionality is beneficial not only for customers but also for business providing the product or service, as positive recommendations will increase the volume of sales. Some of these systems are being built for commercial reasons (to sell more diverse goods, etc.), while others are purely for research needs (to improve recommendation algorithms, study users' needs more precisely, etc.) (Ricci et al., 2010).

Types of recommender systems

The main two types of recommender systems are *content-based* and *collaborative filtering* (Marmanis and Babenko, 2009). Content-based systems rely on the attributes of items and require users to provide their initial preferences in order to recommend items which match those preferences. One of the main advantages of this type of recommender system is that a user's unique taste is not smoothed by the preferences of others (Nageswara and Talwar, 2008) and people with extreme likes will still receive appropriate recommendations.

Collaborative filtering algorithms are the most popular nowadays (Ricci et al., 2010). They use various similarity measures to estimate the distance between items (item-item approach) or between people (user-based neighborhood construction methods). The widely used similarity measures are: *cosine similarity* (each item's attributes are seen as a multidimensional vector and to assess the similarity between two such vectors the cosine of the angle between them is considered). The *Pearson correlation similarity* is based on the correlation between two items, and *probability-based similarity*, where if the user purchased one item after another then the probability of the similarity of those items increases.

Other types of recommender systems which include *demographic* recommender systems (based on the age, country or language of the user), *knowledge-based* recommenders (specialize in recommending data from a particular domain of knowledge through estimating person's needs in that field), *community-based* (recommendations are based on the items that are favorable for user's friends) and *hybrid* recommender systems (utilize a combination of the above mentioned approaches) (Ricci et al., 2010).

2.2 Personality traits

One of the most the widely addressed philosophical questions (having its roots in works of Aristotle) has been the variance of personality traits between people, and how this variance influences people's behavior. The appearance of the scientific trait theory has become possible at the beginning of the 20th century through systematic data collection and the development of statistical methods like data correlation techniques and factor analysis (Matthews et al., 2009). A number of statistical approaches are used to find correlations between various traits and then factor analysis techniques are subsequently applied to group positively correlated traits into larger groups. Each dimension consists of a number of traits that are related to each other and thus if the person has one of the traits in a particular dimension he is likely to have other traits from the same group.

Big Five model

The Big Five personality trait classification is one of the most widely used and recognized models (Matthews et al., 2009) utilized for the research as well as for the staff recruitment purposes. It consists of the five major trait categories: *Extraversion* (the desire of active and ener-

getic participation in the world around), *Agreeableness* (the tendency to eagerly cooperate with others and generally be more helpful and generous), *Conscientiousness* (the ability to control the impulses and to hold to the long-term plans as well as being able to foresee the consequences of one's behavior), *Neuroticism* (which is positively correlated with the susceptibility to experiencing negative feelings such as anxiety, anger and depression) and *Openness to experience* (the tendency of the person to be sensitive to new ideas, non-conventional thinking and to being intellectually curious).

2.3 Personality from the text

Research has shown that there is a correlation between the "The Big Five" dimensions and linguistic features found in texts. In particular, Tausczik and Pennebaker (2009) have discovered that the use of first-person singular pronouns correlates with depression levels while the amount of positive emotions words reveals extraversion. Mairesse et al. (2007) has shown that emotional stability (as an opposite of neuroticism) is correlated with the amount of swearing and anger words used by the person while agreeableness is associated with back-channelling (personality types were estimated form self-reports and observers' reports). Some of the traits were studied more thoroughly (for example, extraversion) which could be caused by a higher level of representativeness of the particular trait-related linguistic cues (Mairesse et al., 2007).

In our research we utilize the Personality Recognizer which is one of the available tools that allow estimating Big Five personality scores (Mairesse et al., 2007).

2.4 Travelling and social media sites

One of the fast developing online domains is the travelling sector. Travellers trying to find a suitable accommodation tend to rely on a number of factors. In particular their choice depends on the hotel awareness (the place is somehow more familiar to the person, for example as a result of advertising) and hotel attitude based on the attributes of the hotel that are pivotal to the person (for example location, cleanliness, service, etc.) (Vermeulen and Seegers, 2008). Thus the choice the traveller makes can become highly influenced by the market games, advertising, popularity of some locations, etc. For this reason many people tend to trust more the opinions of other travellers when making a decision about a particular place to go (O'Connor, 2010).

Research shows that the role of social sites in online travelling domain that allows experience sharing is significant and a high percent of search engine results are links to the social media sites belonging to a number of major categories like virtual community sites, review sites, personal blog sites and social networking tools (Xiang and Gretzel, 2010).

Recently social sites such as TripAdvisor have started to emerge to allow their users to publish reviews of the places they travelled to. TripAdvisor provides the interface to search through the travel facilities (hotels, restaurants, etc.), check their availability for a specific date and read the reviews associated with them. Most of the users of TripAdvisor (97%) return to the site and utilize its content to plan their next trip (O'Connor, 2010). But as the volume of the available reviews is growing in size every day, it is impractical for users to manually retrieve and consider each review.

In our approach, we propose a recommender system based on the hypothesis that users are more likely to be interested in the views of others who have the same personality traits as themselves. Accordingly, in the TWIN System we aim to identify key personality characteristics of users (both readers and writers) based on their writings in order to identify texts written by reviewers who are similar to the reader.

3 TWIN system

The proposed personality-based recommender system follows a user-based collaborative filtering approach. We make an assumption that the "similarity" between people can be established by analyzing the context of the words they are using. Accordingly, the occurrence of the particular words in the particular text reflects the personality of the author. This suggestion leads to the possibility of the text-based detection of a circle of "twin-minded" authors whose choices of particular places to stay (hotels reviewed in TripAdvisor, in our case) could be quite similar and thus could be recommended to each other. This approach provides recommendations that rely on the factors independent in many ways from the user's preexisting attitudes in the hotels' market and also avoids the subjective step of specifying explicit preferences.

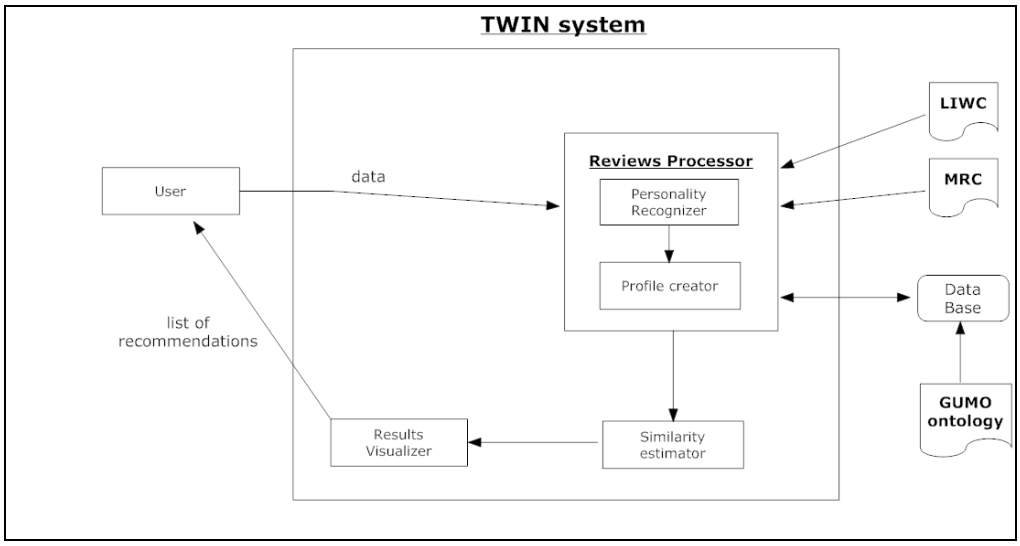

Figure 1. TWIN system architecture

3.1 Architecture

The diagram in Figure 1 represents the main components of the proposed TWIN recommender system described below.

Reviews Processor

The Reviews Processor component retrieves the textual data from the user (plain text written by person) and does the text preprocessing step (dealing with special characters, etc.).

Personality Recognizer

The Personality Recognizer tool is utilized for the estimation of personality scores (ranging from 1 to 7). It maps words found in the text to LIWC [1] (Linguistic Inquiry and Word Count) (Tausczik and Pennebaker, 2009) and MRC (Medical Research Council) Psycholinguistic database[2] (Coltheart, 1981) categories and calculates the number of words in each one. Then it applies the pre-constructed WEKA[3] (data analysis tool) (Hall et al., 2009) models of each of the Big Five dimensions to calculate the corresponding scores based on the found correlations between above the mentioned categories and each of the traits. There are 4 different models that are currently supported by the Personality Recognizer: Linear Regression, M5' Model Tree, M5' Regression Tree and SVM with Linear Kernel.

Profile creator

The Profile Creator stores the general information about the user (login, age group, etc.) as well as personality scores in the user profile that follows the GUMO ontology (General User Model Ontology) (Heckmann, 2005). This model provides a way of extensive description of the user and is a part of the framework that realizes the concept of ubiquitous user modelling. It includes demographic information, psychological state and a lot of other aspects. It has appropriate classes to represent the Big Five model personality parameters as well as general user data (age, gender, etc.).

Similarity Estimator

The Similarity Estimator component utilizes the Weka clustering model built using the K-Means algorithm (Witten and Frank, 2005). During the recommendation process the above-mentioned model is assigning the person to the appropriate cluster based on his profile information. Recommendations are calculated considering the items liked by people in this estimated cluster.

Results Visualizer

The Results Visualizer is constructed as a web-based Flash application to represent the results of the recommendation for the user, i.e. the list of hotels.

[1] http://www.liwc.net
[2] http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm
[3] http://www.cs.waikato.ac.nz/ml/weka/

4 Evaluation

This section provides an overview of the work undertaken to date on the TWIN system construction. In particular we focus here on the structure of the personality-based user profile. For the purposes of the experiments we describe, a dataset of hotel reviews was constructed, as described in the following section.

4.1 Dataset description

We built a Java crawler and constructed a dataset based on reviews submitted to the TripAdvisor website. The dataset consists of hotels reviews (texts and numerical ratings of the particular hotel) and the information about their authors (username, age group and gender) crawled from the TripAdvisor user profiles. For evaluation purposes, we have considered only authors who have more than 5 reviews. The description of key characteristics of the dataset is shown in Table 1.

Dataset parameter	Value
Num of reviews	14 000
Num of people	1030
Total amount of words	1.9 million
Avg num of reviews per person	13.8
Min reviews per person	5
Max reviews per person	40
Num of all words	2.9 million
Avg num of words per review	210.8
Avg num per sentence	16.6
Min words per sentence	3
Max words per sentence	39.7

Table 1. TripAdvisor dataset description

4.2 User profile

The common way to store information about people and model their identity within the recommender system is to create User Profiles. These profiles can be knowledge-based (if person's details are acquired through the questionnaire) or behavior-based (extracted by means of various natural language processing techniques) (Nunes, 2008). Here we follow the behavior-based approach, retrieving the profile data implicitly through the analysis of the reviews written by the particular person.

To model the personality we store the mean score of each of the Big Five parameters calculated from the text of each of the reviews written by the particular person. We selected 15 people from our dataset who have contributed more than 35 reviews. Using the Personality Recognizer (with a linear regression algorithm) we have obtained personality scores for each of the texts written by each individual. As each score is calculated from the text of the review independently we have analyzed them separately. The visualized scores per each of the Big Five dimensions are presented in Figures 2 – 6.

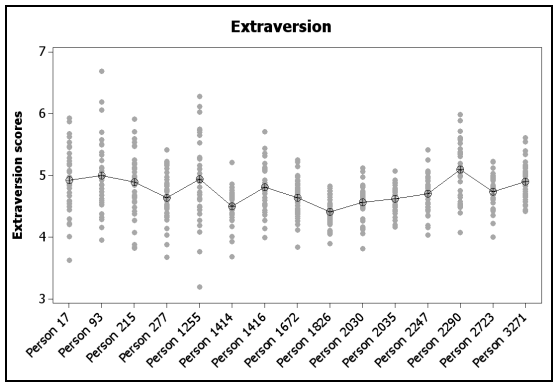

Figure 2. Extraversion scores distribution with means per each set of 15 people's reviews

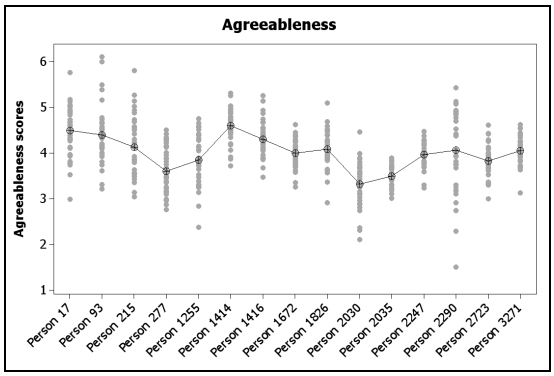

Figure 3. Agreeableness scores distribution with means per each set of 15 people's reviews

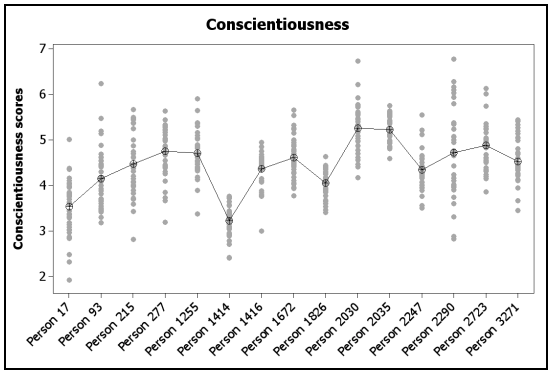

Figure 4. Conscientiousness scores distribution with means per each set of 15 people's reviews

The ANOVA test (Meloun and Militky, 2011) has shown significant differences ($p < 0.001$) between persons in each of the Big Five categories. Thus it could be concluded that the mean scores vary sufficiently from one person to another which enables us to use the mean score as the estimation of the personality in each of the 5 dimensions.

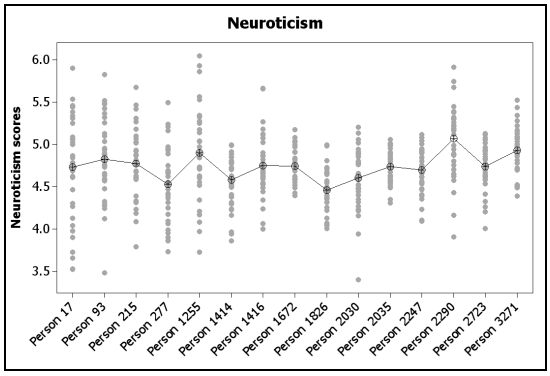

Figure 5. Neuroticism scores distribution with means per each set of 15 people's reviews

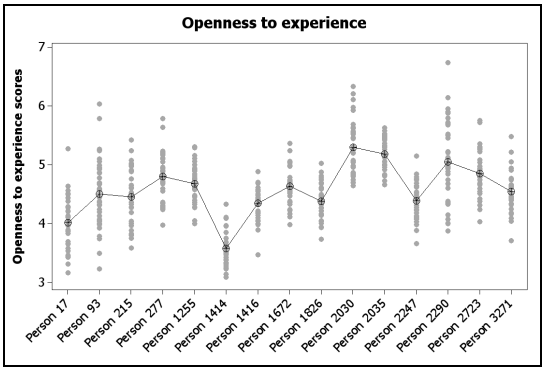

Figure 6. Openness to experience scores distribution with mean scores per each set of 15 people's reviews

It can be seen that openness to experience scores have the highest variability in means which suggests that this trait may be the easiest to detect. This result is in agreement with Mairesse et al. (2007) who had found that openness to experience is the easiest trait to model.

5 Conclusion and future work

In this paper, we have described the architecture of the TWIN Personality-based Recommender System. A fundamental tenet of our approach is that users will value reviews of like-minded people more highly. A critical factor in the success of our approach is the ability to determine personality characteristics (i.e. User Profiles) of reviewers using only the texts they write.

In this paper, we have determined that using a set of texts written by individuals who have contributed a large number of reviews, it is possible to differentiate personality types, and consequently to match a user with reviews written by like-minded people.

In the future work we are going to compare the performance of the other 3 algorithms available in the Personality Recognizer to choose the one that will provide better results. In order to evaluate the TWIN system we are planning the experiment that involves clustering (K-Means algorithm) of reviews in the collected dataset to estimate the percentage of rightly grouped ones.

Acknowledgments

This work has been done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and it has been partially funded by the European Commission as part of the WIQEI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework, by MICINN as part of the Text-Enterprise 2.0 project (TIN2009-13391-C04-03) within the Plan I+D+i.

References

Coltheart, M. 1981. *The MRC Psycholinguistic Database.* Quarterly Journal of Experimental Psychology, 33A(4):497-505.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H. 2009. *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, 11(1).

Heckmann D. 2005. *Ubiquitous User Modeling*. IOS Press.

Mairesse F., Walker M. A., Mehl M., Moore R. 2007. *Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text*. Journal of Artificial Intelligence Research, 457-500.

Marmanis H., Babenko D. 2009. *Algorithms of the intelligent web.* Manning Publications. USA.

Matthews G., Deary I. J., Whiteman M. C. 2009. *Personality Traits*. Cambridge University Press, Cambridge, UK.

Meloun M., Militky J. 2011. *Statistical data analysis: A practical guide.* Woodhead Publishing India.

Nageswara Rao K., Talwar V. G. 2008. *Application Domain and Functional Classification of Recommender Systems – A Servey*. DESIDOC Journal of Library & Information Technology, 28(3):17-35.

Nunes M. A. S. N. 2008. *Recommender Systems based on Personality Traits*, PhD thesis. Université Montpellier 2.

O'Connor P. 2010. *Managing a Hotel's Image on TripAdvisor*. Journal of Hospitality Marketing & Management, *19*:754-772.

Ricci F., Rikach L., Shapira B., Kantor P. 2010. *Recommender Systems Handbook.* Springer, US.

Tausczik Y. R., Pennebaker J. W. 2009. *The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods*. Journal of Language and Social Psychology, *29*(1):24-54.

Vermeulen I. E., Seegers D. 2008. *Tried and tested: The impact of online hotel reviews on consumer consideration.* Tourism Management, 30(1):123-127.

Viney D. 2008. *Get to the Top on Google: Tips and Techniques to Get Your Site to the Top of the Search Engine Rankings -- and Stay There*. Nicholas Brealey Publishing.

Witten I. H., Frank E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques. Machine Learning*. Morgan Kaufmann.

Xiang Z., Gretzel U. 2010. *Role of social media in online travel information search.* Tourism Management, 31(2):179-188.