Evaluating the Similarity Estimator Component of the TWIN Personality-based Recommender System

Alexandra Roshchina
Social Media Research Group,
ITT Dublin/Ireland
E-mail: sasharo@itnet.ie

John Cardiff
Social Media Research Group,
ITT Dublin/Ireland
E-mail: John.Cardiff@ittdublin.ie

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

Abstract

With the constant increase in the amount of information available in online communities, the task of building an appropriate Recommender System to support the user in her decision making process is becoming more and more challenging. In addition to the classical collaborative filtering and content based approaches, taking into account ratings, preferences and demographic characteristics of the users, a new type of Recommender System, based on personality parameters, has been emerging recently. In this paper we describe the TWIN (Tell Me What I Need) Personality Based Recommender System, and report on our experiments and experiences of utilizing techniques which allow the extraction of the personality type from text (following the Big Five model popular in the psychological research). We estimate the possibility of constructing the personality-based Recommender System that does not require users to fill in personality questionnaires. We are applying the proposed system in the online travelling domain to perform TripAdvisor hotels recommendation by analysing the text of user generated reviews, which are freely accessible from the community website.

Keywords: recommender system, personality, Web 2.0

1. Introduction

Recommender Systems have become an important part of everyday life in the online world (Schafer, 1999) as the ever increasing amount of information produces a serious challenge to the user searching for a particular piece of information. Recommender Systems have also become a part of successful marketing strategies of E-commerce firms (Bodapati, 2008) as a way of analyzing the history of product purchases that could help in the prediction of items that the user could find interesting in the future.

Traditionally, Recommender Systems collect the information from the user explicitly by asking the user to fill in the fields in a user profile (usually demographic data or products ratings) or implicitly by studying user behavior (logs of purchases, content analysis, etc.) (Tuzhilin, 2005). However there is increasing interest in the connection between the consumer personality and specific characteristics of the products (e.g. brands) the person is more likely to purchase (Mulyanegara et al., 2007). Accordingly, the challenging task of introducing the personality dimension into Recommender Systems has arisen.

However, existing Personality-based Recommender Systems tend to rely on questionnaires in order to estimate the personality of the user. While being sometimes an interesting activity on its own, questionnaires still require time and effort from the individual to accurately fill them in. Furthermore, people do not always provide honest answers and incorrect data can produce a negative impact on the quality of the recommendation.

One of the alternatives to questionnaires is the estimation of the personality from the user generated content that is freely available in many online communities. A lot of work has been done by psychology researchers to extract specific features from the text to establish the connection between the way the person writes and her personality (Tausczik & Pennebaker, 2009).

In this paper we report on experiments in which we exploit existing tools of personality from the text recognition (Mairesse, 2007) in order to estimate the possibility of building the TWIN Personality-based Recommender System to provide TripAdvisor [1] hotel recommendations based on the text of reviews that people contribute to the website.

The paper is organized as follows. In Section 2 we provide an overview of the basic data mining algorithms utilized, and give an overview of Personality-based Recommender Systems. In Section 3 we describe the TWIN Personality-based Recommender System. In Section 4 we describe our experiments in which we apply the personality from the text construction approach in the TWIN system and present the results. Finally, the conclusions are presented in Section 5.

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

2. Background

2.1 Data Mining and Recommender Systems

The increase in the amount of information available on the WWW requires the development of specific strategies to cope with it. It is possible to process such data automatically or semi-automatically by means of data mining techniques. The main purpose of practical data mining is to find hidden patterns in the training data (usually labeled with correct answers manually annotated by human experts) and describe them explicitly in a specific structural format, which will allow to assign previously unseen instances to a particular class (Witten & Frank, 2005).

Data mining algorithms can be broadly classified into two categories: *supervised* and *unsupervised*. Supervised algorithms at the learning stage make use of the data annotated with correctly assigned classes while unsupervised algorithms try to learn the structure from the unlabeled data by grouping similar objects together according to the specific distance function (Witten & Frank, 2005). *Decision trees*, *classification and association rules* are the types of supervised machine learning algorithms while techniques like *clustering* belong to the unsupervised algorithms category.

Data mining algorithms form the basis of a Recommender System and the choice of the appropriate one correlates with the system's performance. The widely accepted *k-nearest-neighbors approach* (kNN) (Almazro et al., 2010) provides a way of recommending groups of similar people by calculating the distances between them based on users preferences. But the above mentioned algorithm has some scalability problems as it implies the necessity of calculating nearest neighbors over the entire dataset in real-time. To overcome this obstacle the user data is usually pre-clustered offline (for example, using the simple and effective *k-means* algorithm (Witten & Frank, 2005; Ricci et al., 2010) and kNN is applied only within the appropriate cluster (Alag, 2009).

As k-means and k-nearest-neighbors approaches are among the most commonly used data mining algorithms (Wu et al., 2007) we have decided to utilize them for the construction of the TWIN system.

2.2 Personality-based Recommender Systems

Recent research shows that users tend to appreciate personality-based Recommender Systems more than classical ratings-based and return to sites that implement them more often (Hu & Pu, 2009). As the concept of such systems is still an emerging trend, the variety of proposed systems of this type is not extensive.

One of the first mentioned personality-based Recommender Systems is the system introduced by Nunes (2008), which follows the widely accepted Big Five model (Matthews, 2009). In her research Nunes (2008) proposes to provide a better personalized environment for the customer. She claims that one interesting outcome of introducing a psychological dimension into the recommender system could be the possibility of products categorization based not only on their attributes (price, physical parameters, etc.) but also on the effect they may have on the consumer.

Tkalčič et al. (2009) proposed a personality-based approach for the collaborative filtering systems that follows the Big Five model. The authors applied and tested two algorithms of calculating personality-based similarity measures (using Euclidian distance and Weighted Euclidian distance).

The example of the online personality-based system is the "What to rent" [2] movie Recommender System which utilizes the LaBarrie theory [3] in order to produce suggestions of the films to watch depending not only on the personality but also on the current mood of the user (Hu, 2010).

Personality-based music Recommender System was introduced by Hu and Pu (2010). The authors base their system on the correlations between musical preferences and personality types. Four preference groups were found according to various styles of the music compositions people are fond of. For example, the "reflective and complex" group (prefers jazz, blues and classical music) has correlations with openness to new experience Big Five dimension and "energetic and rhythmic" group (tends to appreciate rap, hip-hop, funk and electronic music) correlates positively with extraversion and agreeableness.

3. The TWIN System

3.1 Background

In previous research we have introduced the TWIN ("Tell me What I Need") Personality-based Recommender System (Roshchina et al., 2011). In this work, we proposed the hypothesis that "similarity" between people can be established by analyzing the context of the words they are using, in particular, that the occurrence of the particular words in the particular text reflects the personality of the author. This leads to the possibility of the text-based detection of a circle of "twin-minded" authors whose choices could be quite similar and thus could be recommended to each other.

We have decided to apply findings of the psychological research to introduce the personality dimension in Recommender Systems (Mairesse et al., 2007). One of the main advantages of the approach (comparing to the systems discussed in the previous section) is that the user is not required to perform any additional steps (fill in questionnaires, vote, provide descriptions of the content) to get appropriate recommendations. The personality is constructed automatically from the text of the users through the analysis of their natural styles of writing. Furthermore, our approach eliminates any element of subjectivity or interference that could be introduced by the user evaluating or describing content.

3.2 TWIN System architecture

TWIN system components are represented in Figure 1.

[2] http://whattorent.com

[3] http://whattorent.com/theory.php

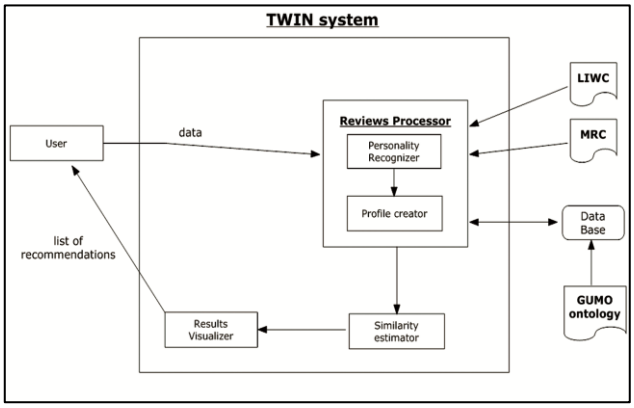

Figure 1: The TWIN System Architecture

3.2.1 Reviews Processor

The Reviews Processor component retrieves the textual data from the user (plain text written by the person) and performs the text preprocessing step (dealing with special characters, etc.). It consists of two components, the Personality Recognizer (Mairesse, 2007) and the Profile Creator. The Personality Recognizer allows the estimation of the personality from the text by calculating the overall percentage of words that belong to each of the Psycholinguistic database dictionary categories described by the Linguistic Inquiry and Word Count (LIWC[4]) and the Medical Research Council (MRC [5]). In order to establish the personality of the author, the Personality Recognizer applies Weka models (Hall et al., 2009) trained on the psychology essays corpora (Pennebaker & King, 1999), comprising texts and personality scores of the authors collected through the Big Five questionnaire). The personality is modeled according to the Big Five classification that consists of 5 categories: Openness to experience, Consciousness, Extraversion, Agreeableness and Neuroticism.

The Profile creator stores general information about the user (login, age group, etc.). In order to represent the user we have built the ontology based on Dublin Core[6]and FOAF [7] vocabularies and the GUMO - General User Model Ontology (Heckmann, 2005). The GUMO ontology provides a way of extensively describing the user and is a part of the framework that realizes the concept of ubiquitous user modelling. It includes demographic information, psychological state, among other aspects. It has appropriate classes to represent the Big Five model personality parameters as well as general user data (age, gender, etc.). The GUMO vocabulary defines two classes that we have utilized the purpose of the TWIN user profile construction: gumo:UnformattedText.100324 (to describe any text with no specific structure) and gumo:Person.110003 (to represent the general user).

The main classes introduced in the TWIN ontology are the Review class implemented as a subclass of gumo:UnformattedText.100324, the TWINUser class being a subclass of the gumo:Person.110003 and the corresponding GUMO classes to model the personality of the user.

The Profile creator is exporting the user data into the RDF format that follows the proposed ontology.

3.2.2 Similarity Estimator

The Similarity Estimator component performs the k-nearest neighbors algorithm to search for similarly typed people among the users' profiles within the system based on the assigned personality scores. Recommendations are calculated based on the items liked by the community of "similar" people.

3.2.3 Results Visualizer

The Results Visualizer represents the results of the recommendation for the user, i.e. the list of hotels. The resulting list of hotels is depicted on Google Maps[8].

3.3 TWIN System Development

The structural components of the TWIN system are shown on Figure 2. The system is designed to be a client-server web application. The Server part is written in Java under the Apache Tomcat server [9] and utilizes MySQL [10] database for data storage. The Client part utilizes Flash technology and is written in ActionScript3[11].

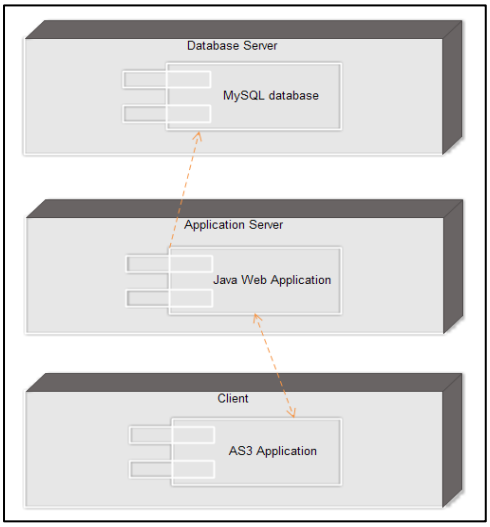

Figure 2: TWIN system: main structural components

4. Experiments and Evaluation

4.1 The TripAdvisor Dataset

For the purposes of our experiment we have created a Java crawler and collected a reviews dataset from the TripAdvisor site. TripAdvisor provides a large amount of the user-generated content including reviews of the

[4] http://www.liwc.net
[5] http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm
[6] http://dublincore.org
[7] http://www.foaf-project.org
[8] http://code.google.com/apis/maps/documentation/flash/
[9] http://tomcat.apache.org
[10] http://www.mysql.com
[11] http://www.adobe.com/devnet/actionscript.html

restaurants, holiday rentals, etc. In our research we were focusing on the hotels reviews. Table 1 shows the description of the dataset that includes the texts of the reviews, hotels ratings (including detailed ratings: value, rooms, location, cleanliness, service and sleep quality) and the information about the authors (name, city, age, gender and the number of contributions to the TripAdvisor site).

For the purposes of the analysis we have applied the Personality Recognizer tool to produce the Big Five scores for each of the reviews texts and filtered out the small percentage of outliers (approximately 12%) for which the scores were incorrectly calculated.

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.

4.2 Experiment 1

In order to evaluate the possibility of the Similarity Estimator component construction we have hypothesized that reviews of the same author should have approximately the same personality scores and should appear in the group of the nearest neighbors (the number of the neighbors to search for equals to the total number of the reviews of the particular author) found by the kNN algorithm initialized by one of the reviews of the current person. We have repeated the same procedure starting from different reviews of the particular author (10 entry point reviews per each person).

For the purposes of this experiment we have chosen 26 people from the TripAdvisor dataset who have contributed more than 35 reviews. As we have found (Roshchina et al., 2011) that different traits of the Big Five have different levels of estimation complexity we have experimented with the various combinations of the Big Five parameters to feed the Weka (Hall et al., 2009) kNN algorithm. The results are summarized in Table 2. As the distance function for the kNN we have chosen the most commonly used Euclidian distance (Witten & Frank, 2005):

$$\sqrt{\left(a_1^{(1)} - a_1^{(2)}\right)^2 + \left(a_2^{(1)} - a_2^{(2)}\right)^2 + \ldots + \left(a_k^{(1)} - a_k^{(2)}\right)^2},$$

where k is the number of attributes (in our case the maximum was 5) and a_k are attribute values.

Big Five parameters	Correctly found reviews (%)
All 5 parameters	10.2
Consciousness&Openness	8.4
Agreeableness& Consciousness&Openness	8.9
All without Neuroticism	9.6
Extraversion& Neuroticism&Openness	9.9
All without Agreeableness	10

Table 2: Percentage of correctly found reviews (from the kNN algorithm output).

As can be seen from the Table 2 the results of the classification are not very optimistic but still promising considering the difficulty of the personality from the text estimation on real-world data. It can be concluded also that the kNN algorithm performs optimally when considering all the Big Five dimensions and other combinations of various dimensions do not improve the personality construction.

Figure 3 shows the actual percentage of the correctly found reviews per each of the 26 people (considering all the Big Five dimensions).

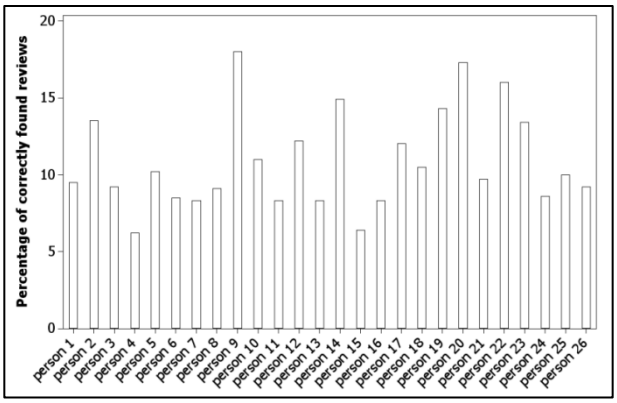

Figure 3: Percentage of the correctly found reviews (from the kNN algorithm output)

4.3 Experiment 2

In this experiment, we investigated the impact that the application of clustering produces on the procedure described in the previous experiment.

First, we clustered all the data in the above mentioned TripAdvisor dataset using Weka's SimpleKMeans algorithm to construct 600 clusters. We then repeated all the steps from the Experiment 1 but this time the kNN algorithm was applied only to the instances within the same cluster that was found for the test instance. Finally we manually constructed 243 clusters reflecting the low (less than 30%), normal (between 30% and 70%) and high (more than 70%) scores for each of the Big Five traits. The results are summarized in Table 3.

Type of the experiment	Mean number of correctly found reviews
Without clustering	3.93
With SimpleKMeans	3.74
With manually constructed clusters	3.85

Table 3: Mean numbers of correctly found reviews (from the kNN algorithm output).

The ANOVA test has not shown significant difference ($p > 0.8$) between the results of the three approaches. This allows us to conclude that we can use the manually constructed clustering approach to speed up the calculation of the kNN.

5. Conclusions and future work

In this paper we have presented our approach to estimating personality from the text in the TWIN personality-based Recommender System. We have also shown the progress on the ongoing work of the TWIN system construction.

The results that we obtained experimenting with the TripAdvisor dataset reflect the difficulty of the task but are still promising. We have shown that the combination of all the Big Five parameters produces better results for the kNN algorithm utilized by the TWIN system. Finally we found that the application of clustering does not change the results significantly and thus can be used in order to increase the speed of the nearest neighbors algorithm calculation.

Our future work will include the modification of the personality from the text recognition algorithm. It will also involve contacting existing TripAdvisor users, whose reviews we have used, in order to fill in the Big Five questionnaire in order to evaluate the performance of the recommendation algorithm.

6. References

Alag, S. (2009). Collective Intelligence in Action, Greenwich, CT: Manning Publications, p.365.

Almazro, D. et al. (2010). A Survey Paper on Recommender Systems. Arxiv preprint arXiv, abs/1006.5, pp.12-151.

Bodapati, A.V. (2008). Recommendation Systems with Purchase Data. Journal of Marketing Research, 45(1), pp.77-93.

Hall, M. et al. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1), pp.10-18.

Heckmann D. (2005). Ubiquitous User Modeling. IOS Press.

Hu, R. (2010). Design and User Issues in Personality-based Recommender Systems. Perception, 36(3), pp. 357-360

Hu, R., Pu, P. (2009). Acceptance issues of personality-based recommender systems. Proceedings of the third ACM conference on Recommender systems RecSys 09, p.221.

Hu, R., Pu, P. (2010). A Study on User Perception of Personality-Based Recommender Systems. In: P. De Bra, A. Kobsa, and D. Chin (Eds.): UMAP 2010, LNCS 6075, pp. 291-302.

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, pp. 457-500.

Matthews G., Deary I. J., Whiteman M. C. (2009). Personality Traits. Cambridge, UK: Cambridge University Press, pp.23-26.

Meloun M., Militky J. (2011). Statistical data analysis: A practical guide. India: Woodhead Publishing, pp. 40-423.

Mulyanegara, R.C., Tsarenko, Y., Anderson, A. (2007). The Big Five and brand personality: Investigating the impact of consumer personality on preferences towards particular brand personality. Journal of Brand Management, 16(4), pp.234-247.

Nunes, M.A.S.N. (2008). Recommender Systems based on Personality Traits. Thèse de Doctorat Doctorat en Informatique. Université Montpellier 2.

Ricci F., Rikach L., Shapira B., Kantor P. (2010). Recommender Systems Handbook. Springer, US. p. 62.

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

Schafer, J.B., Konstan, J. & Riedi, J. (1999). Recommender systems in e-commerce. Proceedings of the 1st ACM conference on Electronic commerce EC 99, pp.158-166.

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), pp.24-54.

Tkalčič, M. et al. (2009). Personality Based User Similarity Measure for a Collaborative Recommender System. 5th Workshop on Emotion in HumanComputer InteractionReal World Challenges, p.30.

Tuzhilin, A. (2005). 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), pp.734-749.

Witten I. H., Frank E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Machine Learning. San Francisco, CA: Morgan Kaufmann, pp. 3-8, 83-141.

Wu, X. et al. (2007). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), pp.1-37.