Intelligent Promotions Recommendation System for Instaprom Platform

Marcos Martín Pozo[1], José Antonio Iglesias[2], and Agapito Ismael Ledezma[3]

Universidad Carlos III de Madrid, Leganés (Madrid), Spain
marcos.martin.pozo.delgado@gmail.com, {jiglesia,ledezma}@inf.uc3m.es

Abstract. The customized marketing is an increasing area where users are progressively demanding and saturated of massive advertising, which has a really low success rate and even discourage the purchase. Furthermore, another important issue is the smash hit of mobile applications in the most known platforms (Android and iPhone), with millions of downloads worldwide. Instaprom is a platform that joins both concepts in a mobile application available for Android and iPhone; it retrieves interesting instant promotions being close to the user but without invading the user's e-mail. Nowadays, the platform sends promotions based on the customized preferences by the user inside the application, although the intelligent system proposed in this paper will provide a new approach for creating intelligent recommendations using similar users promotions and the navigation in the application information.

Keywords: marketing, artificial intelligence, machine learning, recommendation systems, marketing segmentation.

1 Introduction

Over the last years, smartphones on the one hand and promotions applications on the other hand are increasing in size. The first expansion is due to the progress and decreasing costs of technology and the offers applications emerge to combat the effects of economic crisis. Instaprom is a platform that joins the necessity of offers (good for the shopper, which saves money, and for the seller, which frees stock) with the power of smartphones in a mobile application available for Android and iPhone that allows to the shops add offers easily and flexibly and permits to the shoppers receive the offers immediately.

In this paper CRISP-DM methodology [15] –the most used methodology for Data Mining tasks [6]– is applied to create an intelligent recommendation system for Instaprom.

This paper is structured in the following sections: Section 2 describes the CRISP-DM methodology, Section 3 analyzes the datasets and their preprocessing, Section 4 explains the proposed models, Section 5 evaluates these models and finally in Section 6 conclusions and future research are described.

E. Corchado et al. (Eds.): IDEAL 2014, LNCS 8669, pp. 231–238, 2014.

2 Methodology

The methodology applied in the approach proposed in this paper is CRISP-DM, the most used methodology in Data Mining [6]. This methodology has 6 steps in an iterative process which are detailed as follows.

First of all, the business has to be known the context of the process and the objectives that we want to achieve. Secondly, we need to know how many datasets we have and how these datasets are because they are the main point in Data Mining. Data preprocessing is the third step and the most time-consuming because real data are incomplete and inconsistent. Then, the models are designed using Data Mining techniques. The fifth step is to evaluate the generated models. Finally, the sixth step is to deploy the system in a production environment.

Instaprom platform is the business in our case so we want to get the most relevant promotions which can be interesting for each user. This is achieved comparing the available promotions with the previous favorite promotions and with the promotions that the most similar users like.

The available data are the user profile introduced in the mobile application, the most loved promotions for the user and similar users. Data are described with more details in the next section.

Finally, steps 3-5 are analyzed in greater details in next sections, and deployment details have less relevance to the Data Mining process.

3 Preprocessing

This project uses heterogeneous data from varied sources, and therefore each one has to be processed differently. These data have been obtained for two months by recording Android application information.

Firstly, promotions text is used to find similar promotions to the most valuated by the user. These promotions are stored in a database, but they must be processed in order to compare them correctly. This process consists in applying text mining to promotions fields (title, body and conditions). The text mining process applied is 1) *tokenize* the text (to split the text in words), 2) convert all characters to lower case, 3) remove stop words (words without relevance as prepositions and conjunctions with some specific of domain words as 'promotion'), 4) apply a stemmer (to convert words to their root, removing genre, number and conjugation ambiguity) and 5) applying TF-IDF (technique that assigns a relevance numeric value to each word in function of the frequency in the text and the inverse of the frequency in other documents).

Secondly, user navigation is recorded capturing events in Android application. For each event the following data are captured: name of the event, date, user, promotion, shop, commercial area and geolocalization. This task allows how to know the way in which the user uses the application (relevant to find users that use the applications for the same purpose) and also permits to know who is interested in each promotion. This interest is calculated assigning a score to each event and adding the event score to the promotions for the user every time

the user triggers the event. Events score has been assigned in function of the relevance of each event and it is shown in Table 1.

Finally, application user profile is used to find similar users and recommend them their most valuated promotions. This information consists of gender, year of birth, province and interests. All the fields are optional, except interests, that must be at least three. This information is stored in a database, but it is necessary to extract and clean it. In addition, this profile information is merged with user events. A different dataset which has been considered in this research is the division of these events in sessions, considering a session as a sequence of same user events with less than 10 minutes between an event and the next one. Ultimately, profile information and sessions are merged in another dataset.

Table 1. Score of each user event in the application

Event	Description	Score
Promotion	User enters in a promotion	20
Conditions	User views the promotion conditions	10
QR	User views the promotion QR code (purchase is assumed)	80
Shop	User views store information	30
Like	User likes promotion	50
Not like	User unchecks that he likes promotion	-50
Preshare	User enters in share option	10
Share	User share promotion	100
Directory	User enters in shop directory	0
CA	User enters in commercial area	0
Preferences	User enters in preferences	0
About	User enters in about application section	0
List	User enters in promotions list	0
Back	User presses back button	0
E-mail	User enters in send e-mail to shop option	0
Call	User enters in call to shop option	0
Web	User enters in view shop web option	0
Contact	User enters in application contact option	0
Help	User enters in help option	0
Play	User starts promotions reception	0
Pausa	User pauses promotions reception	0
Reload	User reload promotions list	0

4 Models

This project uses different techniques to process the datasets. These techniques are renowned and massively used. For clustering tasks EM, K-means, Cobweb and Furthest First are used. For classification tasks C4.5 is used because decision trees permit to analyze results in an affordable manner. Finally, a less known adaptive technique is used to process the users events sequences. Many

other techniques have been studied, but they cannot be applied due to time restrictions: dynamic bayessian networks [12,13], ontologies [9,14], fuzzy logic, simulation [2], artificial neural networks [2], genetic algorithms [10,11] and user models [1,3,4,5,7,17].

On the one hand, text mining is used to preprocess promotions text and afterwards get a similarity measure among them. This process is applied to several datasets: promotion title, body, conditions and all fields merged in an unique dataset. To analyze the best dataset, and because analyzing TF-IDF results is intricate and complex, clustering and classification techniques are used. Firstly, EM is used to cluster the TF-IDF processed promotions because using Weka it gets automatically the number of clusters using cross validation [16] and because its mathematical base is suitable for the numerical nature of TF-IDF results. Next, C4.5 is applied to clusters to analyze the goodness of results.

On the other hand, user application and navigation profile are used to find similar users and to recommend their most valuated promotions. In a first iteration, clustering were used to both profiles. Thus, EM, K-means, Cobweb and Furthest First were applied on several datasets: profile, sessions, and sessions with profile. In a last iteration an adaptive clustering technique was used to user events sequences: this technique stores events sequences of a maximum length in a *trie* (a special type of tree suitable for information retrieval) and then it uses the cosine distance to cluster users. This technique uses also a special algorithm in which the behavior of the users can evolve, for more details see [8].

The before techniques constitute the intelligent recommendation system. After this processing a strict filter is used to remove the promotions that don't match the user gender, age, location or interests. This procedure is schematized in Figure 1.

5 Evaluation

Evaluation is essential in a data mining project in order to validate the created models: without a correct evaluation it is not possible to guarantee that the models work properly. In this research a complete evaluation couldn't be accomplished because data collected was not large enough. Nevertheless, bad models were rejected and remaining models seemed to be adequate, although a re-evaluation with more data is necessary to guarantee the fitness of the models.

Content-based recommendation models are validated analyzing the decision trees and the precision obtained by C4.5 on EM clusters. In a first iteration stop words were not removed from promotions fields, but despite getting more precision (due to overfitting) the models obtained are worse. For this reason this should be validated with more data. Precision obtained by C4.5 for each dataset is shown in Figure 2. In this chart the number of clusters that EM generated is shown in yellow color inside bars. In this chart we can notice that the dataset with more precision is when we use all fields of the promotions. The dataset that only contains conditions also has a high precision, but analyzing the decision tree we can see its incoherence, due to the short text of conditions (even there

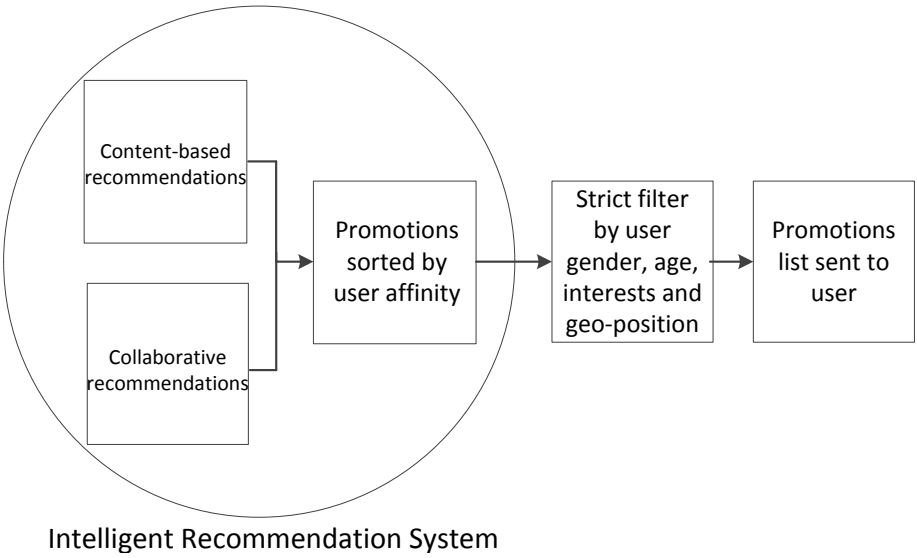

Fig. 1. Intelligent recommendation model

are promotions without conditions). Then, we can say that the best model for content-based recommendation is to use all fields of the promotion removing stop words.

Collaborative recommendation using traditional clustering techniques are validated similarly: clusters are classified by C4.5 and precision and trees coherence are analyzed. In a first iteration data types were incorrect, dealing numeric values as nominal with several possible values. Again, in some cases the precision of this iteration is greater than the precision with correct types due to overfitting, but obviously with correct data types the models are more robust (it can be seen analyzing decision trees) and with more data this would be validated. Precision gotten for each dataset is shown in Figure 3 (hierarchical techniques results are not shown because they were very low), and again number of the clusters that EM generated is shown in yellow color inside bars. In this chart we can notice that precision is high in all datasets, though if we consider sessions the precision is worse. This is because there are few sessions, and they are short. Between profile and sessions with profile the difference is not large, but analyzing sessions with profile decision trees we noticed that user events are almost not present, and they appear in the bottom of trees, where the relevance is low. Finally, in all models K-means is a bit better than EM, but EM calculates the number of clusters automatically, therefore, we determine that the best model for profile clustering is EM using the dataset in which the user profiles are stored.

Finally, since traditional clustering techniques were not adequate for user events profile, it was analyzed using a different technique. This technique generates event sequences of a maximum length (for instance with maximum length 3 it generates sequences of length 1, 2 and 3), then it generates a *trie* with these

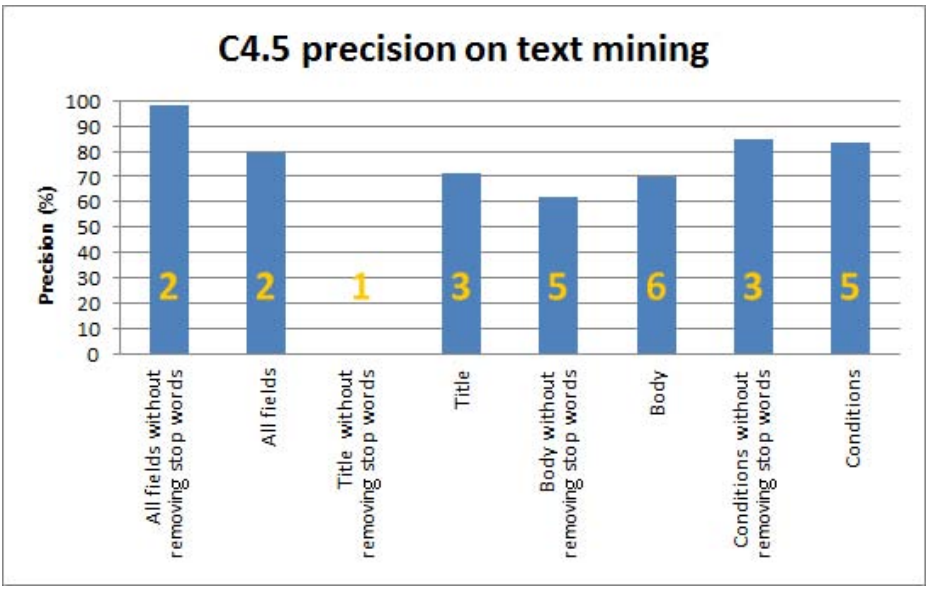

Fig. 2. Content-based recommendation evaluation

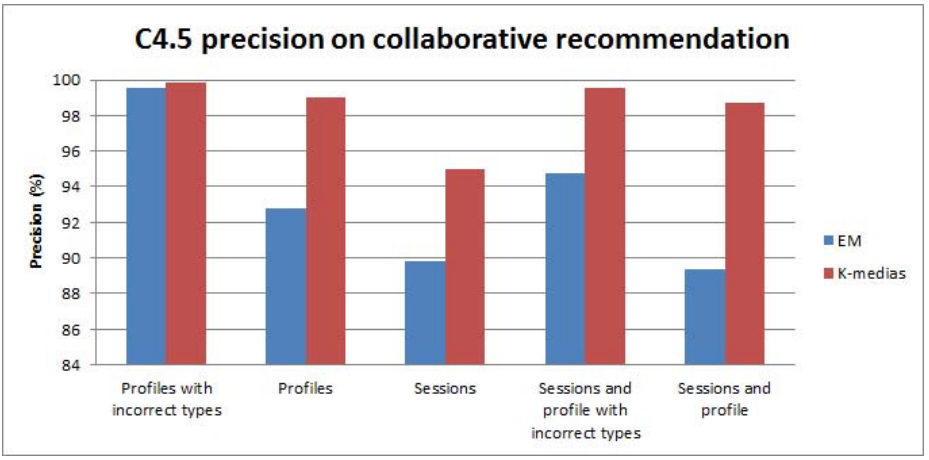

Fig. 3. Collaborative recommendation evaluation

sequences and finally it uses cosine distance to generate clusters. For more details see [8]. This technique is evaluated extracting the sequence of highest χ^2 of each user and then analyzing the average χ^2 and the occurrences of each sequence. This analysis was done with sequences of maximum length 3, 4 and 5; higher maximum lengths were not analyzed because results are similar with these maximum lengths and higher maximum length won't change them. Average χ^2 is

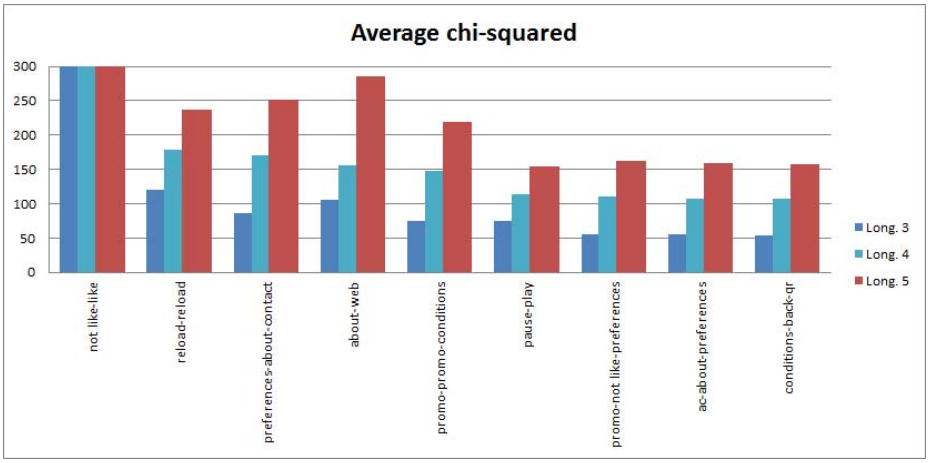

Fig. 4. Average Chi-squared for best sequences of each user

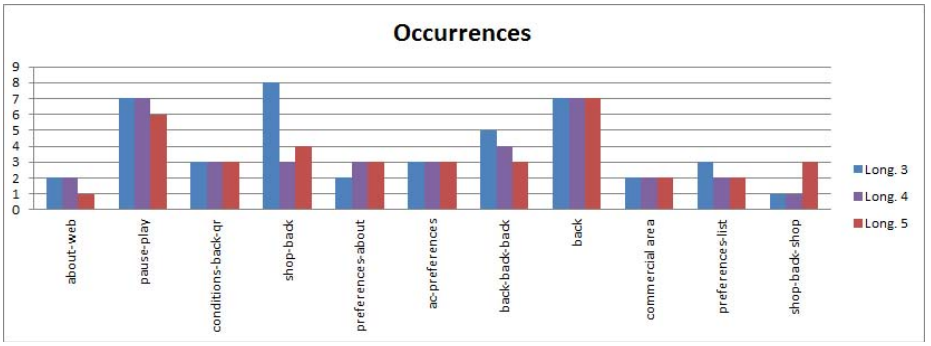

Fig. 5. Occurrences of best sequences of each user

shown in Figure 4 and occurrences are shown in Figure 5. We can appreciate that results hardly change varying maximum length. The best sequences are those with higher average χ^2 and more occurrences because they are the cluster prototypes. These sequences are *pause-play*, *conditions-back-qr* and *about-web*.

In conclusion, the best model for content-based recommendations is text mining using all fields of the promotions and the best model for collaborative recommendations are combining EM for profile recommendations with *tries* technique for navigation profile recommendations.

6 Conclusions and Future Research

In this project an intelligent recommendation system for Instaprom platform has been designed and evaluated. CRISP-DM methodology has been used for this purpose, and as result several techniques and datasets have been proposed.

In spite of these models are promising, more data are required to completely validate them. A set of validation users would be helpful too in order to guarantee the validity of the recommendation system.

Future Research will be about this validation with more data and cross validation, though other techniques can also be evaluated. According to the research detailed in Section 3, fuzzy logic, artificial neural networks, dynamic bayesian networks, simulation and genetic algorithms are the most promising.

References

1. Bouneffouf, D.: Mobile Recommender Systems Methods: An Overview. CoRR (2013)
2. Çağil, G., Erdem, M.B.: An intelligent simulation model of online consumer behavior. J. Intell. Manuf. 23(4), 1015–1022 (2012)
3. Davidsson, C.: Mobile Application Recommender System (2010)
4. Gavalas, D., et al.: Mobile recommender systems in tourism. Journal of Network and Computer Applications (2013)
5. Godoy, D., Schiaffino, S.N., Amandi, A.: Integrating user modeling approaches into a framework for recommender agents (2010)
6. KdNuggets: Data Mining Methodology (2007), http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm
7. Lakiotaki, K., et al.: Multicriteria User Modeling in Recommender Systems. IEEE Intelligent Systems 26(2), 64–76 (2011)
8. Iglesias, J.A., Angelov, P., Ledezma, A., Sanchis, A.: Evolving classification of agents behaviors: a general approach. In: Evolving Systems. Springer (2010)
9. Liu, W., et al.: Ontology-Based User Modeling for E-Commerce System. In: Third International Conference on Pervasive Computing and Applications, ICPCA 2008, pp. 260–263 (2008)
10. Martínez-López, F.J., Casillas, J.: Marketing Intelligent Systems for consumer behaviour modelling by a descriptive induction approach based on Genetic Fuzzy Systems. Industrial Marketing Management (2009)
11. Martínez-López, F.J., Casillas, J.: Mining uncertain data with multiobjective genetic fuzzy systems to be applied in consumer behaviour modelling. Expert Syst. Appl. 36(2), 1645–1659 (2009)
12. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference (1998)
13. Prinzie, A., Poel, D.: Modeling complex longitudinal consumer behavior with Dynamic Bayesian networks: an Acquisition Pattern Analysis application. J. Intell. Inf. Syst., 283–304 (2011)
14. Rodríguez Rodríguez, A., Iglesias García, N., Quinteiro-González, J.M.: Modelling the psychographic behaviour of users using ontologies in web marketing services. In: Moreno-Díaz, R., Pichler, F., Quesada-Arencibia, A. (eds.) EUROCAST 2011, Part I. LNCS, vol. 6927, pp. 121–128. Springer, Heidelberg (2012)
15. Turban, E., Sharda, R., Delen, D.: Decision Support and Business Intelligence Systems (2010)
16. Weka doc.: http://weka.sourceforge.net/doc.dev/weka/clusterers/EM.html
17. Zhang, L., Chen, S., Hu, Q.: Dynamic Shape Modeling of Consumers' Daily Load Based on Data Mining. In: Li, X., Wang, S., Dong, Z.Y. (eds.) ADMA 2005. LNCS (LNAI), vol. 3584, pp. 712–719. Springer, Heidelberg (2005)