ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Modeling consumption of contents and advertising in online newspapers

Iago Porto-Díaz [1] David Martínez-Rego [2]
Oscar Fontenla-Romero [3] Amparo Alonso-Betanzos [4] *

University of A Coruña - Department of Computer Science
Campus de Elviña s/n 15071 A Coruña - Spain

Abstract. This paper presents the design of a system for personalization of contents and advertising for readers of online newspapers. This software is conceived to work in a context of high network traffic with millions of URLs served each day. The model is divided into two subsystems. The first one takes care of the recommendation of news items. The mathematical model is based on the PageRank algorithm and considers several practical day-to-day scenarios. The second one, which is the subsystem of personalization of advertising, uses a Multinomial Logistic Regression model in order to predict categories of advertising for banners within the news content. The system has obtained practical satisfactory results using real data.

1 Introduction

Personalization technology enables the dynamic insertion, customization or suggestion of content in any relevant format. This technology is useful for online newspapers, with the objective of increasing revenue. This can be achieved in two ways: first, by recommending news items of interest to the readers and, second, by associating the right advertising to the right content. Both these task can be calculated from the information of previous navigations of users. These two are the topics covered by the system designed in this paper.

The key factor when trying to improve the user experience and to orientate the commercial contents to each user is being able to predict the contents that are going to be of the most interest to the user. Therefore, the approach to be followed is to try to make more accessible the contents that are susceptible of being of interest for the user. The system is designed for a mainstream online newspaper in a real environment. In consequence, there exist a series of limitations, inherent to the problem of personalizing news content and advertisements, that need to be addressed by the designed system:

- The only information available is the navigation of the users. Opinions, content ratings, purchases, etc. are not available. In consequence, data from users are binary (a user either clicks a link or not). This restriction limits classic approaches such as collaborative filtering, which would not very suitable if Hamming distance between binary patterns is used.

*This paper is supported by Spanish Ministerio de Ciencia e Innovación under project code TIN 2012-37954, partially supported by the European Union ERDF. Iago Porto-Díaz is supported by University of A Coruña pre-doctoral grant.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

- Due to the uncertain lifetime of cookies, the time period in which a user is known may be narrow. The model must take into account the fact that a known user will become unknown periodically. It should be noted that asking the user to register in the web site is not very adequate: virtually no online newspaper requires its users to be registered to read contents.
- News contents are short-lasting. A news item will lose its interest in a short period of time.
- Although the data volume for each user is reduced, the total data volume is very high (millions of URLs per day), which demands that the computational complexity of the algorithms is low or either the algorithm is parallelizable.
- An important amount of contents of interest are not reached by the user, because of the model of online publications.

There exist in the literature a considerable amount of recommenders for electronic commerce . However, there is less development in specific systems for online newspapers. Billsus & Pazzani designed a system that used the nearest neighbor algorithm and the Naive Bayes classifier for selecting news for a user based on feedback [1]. Chen & Shahabi utilized a genetic algorithm for user profiling [2]. Liu et al., from Google Inc., published in 2010 a hybrid recommender system based on a Bayesian framework that predicts news interests from the history of the users and the news trends [3]. The focus of the proposed system is similar to the latter approach, as it also uses the behavior of most users in order to provide recommendations, as will be seen. Moreover, the proposed algorithm for news items recommendation is based on the PageRank algorithm [4], which ranks web sites by analyzing the links among them.

The personalization of advertising began to develop in the nineties [5]. The usual schema of advertising in the Internet consists of displaying advertisements (usually banners) within the content of web sites. These banners are usually selected by a commercial agent through automated systems. These systems receive a category of advertising and provide a banner previously classified into that category. Therefore, the model designed in this paper has to output a category of advertising, which needs to be predicted from news content. This is achieved by training a multinomial logistic regression algorithm [6] in order to associate the tags of the news items to the category of advertising. This association is extracted from log files of users who click banners.

2 Design of the Subsystem of Personalization of Contents

In order to estimate a navigation model correctly, a series of assumptions about the user decision making have to be made:

- There exists a relationship between the contents the user has consumed and those the user is going to consume in the next visit to the newspaper. In other words, the user does not surf randomly through the pages, but the

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

contents that were of interest in the past drive him to consume contents related to those in the future.

- The user, at the time of deciding which contents are of interest, can follow one of the following three patterns:
  - Once the user begins surfing the newspaper, a subset of all the published content lies within reach (the contents directly linked from the front page). The user will decide which are of interest among them. To put it in another way, a user will never know what is the complete set of contents published in the day, only the ones which are linked in the visited pages.
  - Once finished reading a page, the user decides to keep reading other news items of interest. It is considered that the contents of interest for a user are not totally independent, although the relationship among them may be hard to explain.
  - The user may consume a determined content arriving to it randomly from external sources like social networks, news aggregators or front pages related with the information.

  It is important to emphasize that a user does not reach a large amount of the published contents: some of the present contents may have been considered of interest, had he been aware of its existence.
- There exist sets of users with similar consumption patterns: the interests of a user align with those of a subset of all the readers of the online newspaper.

2.1 Estimation of the Model of Consumption

The event of a user becoming interested in an editorial content might occur in two ways: (a) the user considers it interesting independently of the other contents of the newspaper and (b) the user becomes interested in the content after reading another content of the newspaper. The second scenario is observable, because the navigation of users is logged. However, the first scenario is not, as the user may have been influenced by a content outside of the newspaper or even the user may have not been aware of this specific content. Therefore, the probability that a user is interested in an editorial content i is defined as:

$$P(interest\ i) = \frac{1-s}{N} + s \sum P(interest\ i \mid interest\ j) P(interest\ j) \qquad (1)$$

where N is the number of editorial contents and s is a weighting coefficient. The first addend represents the uncertainty associated to the probability of scenario (a). The second addend corresponds with the probability associated with scenario (b). Actually, in order to estimate the probability of (a), it is considered that each content has a uniform probability $1/N$. This model is similar to PageRank [4], which is a well established criterium to sort the pages returned in a search engine. The difference is that the probability of (b) is going to be estimated via the analysis of the access logs, whereas the search engine has no access to the equivalent of this information and has to use other criteria of connectivity between pages.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

2.2 Calculation of the Model of Consumption

The total probability of interest of a content can be expressed as the limit of a Markov chain of order 1. Each element of the matrix of transition of the Markov chain is defined (from (1)) as:

$$\mathbf{M}_{ij} = (1-s)\mathbf{A}_{ij} + s * \mathbf{T}_{ij} \tag{2}$$

where $\mathbf{M}_{ij}$ is the total probability of considering interesting the content i having considered interesting the content j, which includes the uncertainty of the model $\mathbf{A}_{ij} = 1/N$. $\mathbf{T}_{ij} = P(interest\ i \mid interest\ j)$ is the observed probability of deciding that the content i is interesting because it is related to the content j, which has been found interesting before. It should be noted that the transition matrix $\mathbf{M}$ is ergodic.

2.3 Implementation of the Model of Consumption

The computation of the model needs to estimate the parameters $\mathbf{T}$ and s. $\mathbf{T}_{ij}$ is calculated in the following way: every time a user consumes a content j and, next, a content i, 1 is added to the component $\mathbf{T}_{ij}$. The parameter s is fixed to 0.85, which is a value recommended in previous experiences [4].

Once the matrix $\mathbf{M}$ is built (after normalizing each column), the vector μ of probability of interests can be calculated by power iteration, which is an efficient and parallelizable method that only implies algebraic operations with matrices. The initial probability of the interests of the user is represented through μ^0, which is a vector of the absolute frequencies of the contents consumed by the user. The probability that the user, in subsequent clicks, is interested in each of the contents is estimated by:

$$\mu^k = \mathbf{M}\mu^{k-1} \tag{3}$$

being μ^k_j the probability that the user is interested by the content j in the future click k.

3 Design of the Subsystem of Personalization of Advertising

The subsystem of personalization of advertising proposed predicts categories of advertising using a statistical approach. Banners are displayed within the content of web sites and are usually selected by an automated system from an advertising category. Therefore, the model designed outputs a category of advertising, predicted from news content. It consists of a Multinomial Logistic Regression model (MLR) [6], which is trained with data obtained from the logs of the online newspaper.

MLR is a generalization of logistic regression that allows more than two classes at the output. It is utilized to estimate the probabilities of the different values of a categorical variable, given a set of features:

$$p(c = i|\mathbf{x}) = \frac{e^{\mathbf{w}_i^T\mathbf{x}+b_i}}{\sum_{j=1}^{C} e^{\mathbf{w}_j^T\mathbf{x}+b_j}} \tag{4}$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

where C is the number of classes, $\mathbf{x}$ is the vector of features describing a given observation, $\mathbf{w}$ is a vector of weights and $\mathbf{b}$ is the vector of biases. In the subsystem, the categorical variable to be estimated is the category of the advertising. The set of features are, for each click on a banner present in the logs, the city and country of the user and the tags of the news item in which the banner was. The tags represent the content of news item. MLR is usually used as an alternative to Naive Bayes when statistical independence between the features can not be assumed. Since the tags of the news items of the online newspaper are hierarchical, independence can not be assumed.

4 Testing of the System

In order to test the contents subsystem, logs of connections of users have been used. Almost 5 million lines of log of the online newspaper were available, which amounted for 1.3 GB of text. Each line of log includes the cookie identifier for the user and the visited URL. Half these lines were used for constructing the graph, and the other half were used for testing. Some preprocessing was performed: the URLs were grouped by the cookie identifier and the users with less than 5 (not traceable users) or more than 100 URLs (crawlers) were discarded. The constructed graph contained almost 20000 unique URLs. During testing, for each cookie identifier, navigation vectors were built using at least three news items, in order to have reasonable evidence to generate a recommendation. This lower limit was selected because the prototype developed for the online newspaper, due to space limitations in the newspaper, displays three recommendations. For each of these cases, the recommendations were generated using the previously built graph. If the next URL visited by the user matches one of the recommendations, it is considered a correct recommendation, and an incorrect one otherwise.

The obtained results show an accuracy of 10.92% over the test set. This result, despite being low, greatly outperforms a randomly generated recommendation, which obtains an accuracy of 0.23% and is slightly better than always recommending the three most viewed news items of the day, which obtains an accuracy of 8.77%. There are very few studies performed in environments of similar features. However, these results can be indirectly compared to others, i.e. [7], where a field study was carried out, in which a number of users were presented recommendations in a real environment. The percentage of users that clicked a recommendation was lower than 8.18% in all nine categories considered.

Regarding the advertising subsystem, in order to check its validity, the model was tested over a data set constructed from logs that contain information about the clicks on the banners. For each click, it is available the city and country of the user, the tags of the news article the user was reading at the moment of clicking the advertisement and the category of the advertisement. This data set consists of approximately 4000 samples. The tags have been codified using binary variables. For each of the approximately 2000 unique tags, there is a feature of the data set which may take value 1 if the tag is present in that sample or 0 otherwise.

Since the data set has around 2000 features, feature selection was performed.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

20 attributes were selected using the algorithm Information Gain [8]. A 10-fold cross validation was performed to evaluate the performance. Since a typical page may present several banners, the three most likely categories given by the model where considered. If the category of a banner in the test fold matches one of the three predicted by the model, it is considered a correct recommendation; an incorrect one otherwise. The average test accuracy through the 10 iterations was 70%. A baseline to compare this result could be to always choose the three majority classes. In this case, the performance would be 64%, which shows that the obtained results are slightly better. To the best knowledge of the authors, there do not exist papers which carry out a similar experiment.

5 Conclusions

In this paper, a system of personalization of contents and advertising for readers of online newspapers is designed. The system works efficiently with a big volume of traffic and is reliable even if cookies are not very persistent. The mathematical model of the contents subsystem considers a model of navigation where the interest for a news item may have generated (a) randomly, from external sources; (b) through the consumption of other contents within the online publication; (c) by the own interest of the consumer, even though he might have not found it. The system has been tested empirically and obtains better results than the two baselines considered: random recommendations and the most viewed news items as recommendations.

The advertising subsystem predicts a category of advertising from tags of news items and information of users, taking into account statistical dependence between features. The experimental study performed generates a satisfactory result.

References

[1] Daniel Billsus and Michael J Pazzani. A hybrid user model for news story classification. *Courses and Lectures - International Centre for Mechanical Sciences*, pages 99–108, 1999.

[2] Yi-Shin Chen and Cyrus Shahabi. Automatically improving the accuracy of user profiles with genetic algorithm. In *Proceedings of IASTED International Conference on Artificial Intelligence and Soft Computing*, pages 283–288, 2001.

[3] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 31–40. ACM, 2010.

[4] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[5] Raymond R Burke, Arvind Rangaswamy, Jerry Wind, and Jehoshua Eliashberg. A knowledge-based system for advertising design. *Marketing Science*, 9(3):212–229, 1990.

[6] D. Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[7] Dan Cosley, Steve Lawrence, and David M Pennock. Referee: An open framework for practical testing of recommender systems using researchindex. In *Proc. 28th Int. Conf. on Very Large Data Bases*, pages 35–46. VLDB Endowment, 2002.

[8] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.