

# Automatic Persistent Personalization of Ads in Tourism Websites

Alberto Rezola, Aitor Gutierrez, and Maria Teresa Linaza

**Abstract** Information and Communication Technologies (ICT) have dramatically increased the ability of advertisers to target advertising campaigns and make sure that ads are shown to only certain targeted groups of people. Usage of appropriate ads to each visitor may increase Click Through Rates (CTR) and chances of conversion. This paper presents a novel online advertising approach for automatic “persistent personalization” of Web ads on the basis of Web-mining techniques that combine representative parameters for advertising in a unique platform. The functionality of the approach as well as the problems that arose during the implementation are posed and discussed. Finally, the recommendation system has been successfully validated in a travel blog Website. The implemented prototype made it possible to serve the appropriate ads to the targeted audience on the basis of the classification of user profiles. The obtained CTR was the double of the expected common CTR rates in online advertising campaigns.

**Keywords** Ad personalization • Persistent personalization • Supervised learning • Implicit user profiling • Hybrid recommender system

## 1 Introduction

Many studies have generally found a negative public attitude towards advertising, with consumers often annoyed due to intrusive advertising messages (Donnell & Cramer, 2015; Tsang, Su-Chun, & Ting-Peng, 2004; Wang, Zhang, Choi, & Eredita, 2002; Watson, McCarthy, & Rowley, 2013; Zanot, 1984). However, this consumer attitude can change when advertising is personalized and relevant to the lifestyle of the consumer (Goldfarb & Tucker, 2011). Information and Communication Technologies (ICT) have dramatically increased the ability of advertisers to target advertising and make sure that ads are shown to only certain targeted groups of people (Tucker, 2012).

Web advertising personalization controls the visualization of campaigns to appropriate customers at the proper time based on the user preferences. For

---

A. Rezola (✉) • A. Gutierrez • M.T. Linaza  
eTourism and Cultural Heritage Department, Vicomtech-IK4, Donostia-San Sebastian, Spain  
e-mail: [Arezola@vicomtech.org](mailto:Arezola@vicomtech.org); [Agutierrez@vicomtech.org](mailto:Agutierrez@vicomtech.org); [Mtlinaza@vicomtech.org](mailto:Mtlinaza@vicomtech.org)

example, users reading articles about traveling regularly may be interested in flight booking. Thus, travel Websites should display ads of airlines offers and discounts. In such way, the usage of appropriate ads to each visitor may increase Click Through Rates (CTR) and chances of conversion. Obviously, the more detailed the representation of the context of the user is, the more effective these approaches become.

This paper presents a novel online advertising approach for automatic “persistent personalization” of Web ads on the basis of Web-mining techniques that combine representative parameters for advertising in a unique platform: most suitable content of the Website of the advertiser; probability of the click-through; advertising targets arising from contracts with advertisers; and fuzzy mechanisms. This approach has been named “*Sistema Publicitario Integrado de Gestión Automática*” (eSPIGA).

The rest of this paper is structured as follows. Section 2 describes the research background related to ad personalization and recommendation. Section 3 describes the research hypothesis that will be validated. Section 4 describes the eSPIGA system which has been implemented to validate such hypothesis, while Section 5 summarizes the validation of the system in a travel blog Website. Finally, Section 6 contains the conclusions.

## 2 State of the Art

While traditional media (e.g. TV, radio station, newspaper, billboards) advertise products and services non-selectively to people, sponsors are more interested in personalized ads that present products and services for their prospective customers. Sponsor interest in personalized advertising lies on ad personalization strategies so as to minimize the costs of marketing campaigns while maximizing their impact. Personalized advertising also has focused the attention from researchers in multimedia, e-commerce and Artificial Intelligence because contents and presentations can be tailored to the user preferences in order to maximize their attention.

Personalized recommendation systems aim at presenting online content tailored to the specific interests of each user (Tucker, 2012). The personalization of ad recommendation has received particular interest from the research community in recent years, since the Internet business model heavily relies on advertising. However, personalizing ad recommendation is challenging, since the lack of the metadata associated to ads makes it very difficult for recommender systems to adequately filter ads.

In a conventional approach for the provision of personalized contents, the ad platform either contains a server that keeps track of the user preferences and interactions (e.g., Webpage visits, search keywords, shopping transactions) or requires that the interaction device of the user (e.g., mobile devices, Web browsers, IPTV) sends his personal information over the network in order to select pertinent ads.

However, these approaches have potential threats to the privacy of the customers. For instance, his personal information could be disclosed in the process of collecting and managing his profile in a system based on a server storing user profiles. If the system is violated, the privacy of thousands of customers will be threatened.

Ad personalization is a challenging research topic for current advertisers, which aims at assigning a suitable ad to a single Web user rather than a group of individuals. To achieve this challenge, personalization systems need to have some input about the user. One of the current approaches to create such profiles takes advantage of the information gained during the registration process or just asks questions to the users about their preferences (explicit user profiles). Many personalized advertising methods have been proposed on the basis of explicit user profiles, which are gathered, maintained, and analysed by the system (Bilenko & Richardson, 2011; Bleier & Eisenbeiss, 2015). However, this process can discourage many users.

On a research level, AdRosa is an advertising tool that works through remote open site agents. It deals with the automatic personalization of Web ads without using explicit user personal information, but their navigation patterns in order to maintain the privacy of the users. The functionality of the system is based on deriving the needed knowledge from Webpage content, previous sessions and current behaviour of users (Kazienko & Adamski, 2007).

Furthermore, Bae, Park, and Ha (2003) developed a system based on Web usage mining to cluster navigation paths to create usage patterns. Pages were manually classified from both the Website of the publisher and the target sites of the advertisers into thematic categories. Appropriate ads were assigned to each active user according to the pages and categories visited during the current session. Such matching was based on fuzzy rules stored in the system.

The commercial online advertising system AdSense (Google Ads) delivers a targeted ad to the Website of a publisher and consists of two options: “AdSense for content” and “AdSense for search.” While the former delivers appropriate text or image ads based on the content of a site of the publisher, the latter encourages publishers to add the Google search box to their pages, so a set of targeted text-based ads are attached to the search result pages in the form of “sponsored links”. The complementary Google program, AdWords, is targeted to advertisers, who define keywords associated with their ad, so that Google matches the available ad subset with all activities in which given keywords occur (Davis, 2006).

Similarly, Rusmevichientong and Williamson (2006) studied algorithms for the selection of profitable search keywords that are especially useful for fixed advertising budgets. Since the AdSense system can access only data available for the Google search engine and the content of Websites, it is able to provide only “ephemeral personalization” of advertising. The ephemeral approach can deliver a different item on every page of a Website but be the same for all users (Schafer, Konstan, & Riedl, 2001). The more adaptive method—“persistent personalization”—uses the history of a user’s behaviour and generates a different item for each user in each context. However, Barford, Canadi, Krushevskaja, Ma, and

Muthukrishnan (2014) analysed the personalization over 175 K distinct display ads from a variety of Websites and they found that while targeting was widely used, there were many remaining instances in which delivered ads did not depend on the profile of the user.

### 3 Research Hypothesis

Technological advances mean that consumer information can be used to personalize the actual content of the ads shown in order to match the interests of the user.

#### **H1: Web ads can be automatically personalized to different user profiles**

Advertising campaigns are generated for specific audience targets, which are based on different demographic parameters such as gender, age range, user preferences and localization of the users. It is assumed that online activities and behaviours of the users are diverse enough to allow the characterization of different user profiles and preferences. Thus, it is possible to serve the appropriate ads to the targeted audience on the basis of the classification of user profiles in an automatic way.

The main objective of the user modelling is to collect as much data as required about the user and then customize the advertising content to fulfil the preferences of the users. In order to build user profiles, only explicit information of the customers will be used.

#### **H2: User preferences can be obtained implicitly from the Web usage data**

As a user navigates the Web, his navigation activity can be obtained and stored. This information, along with the content analysis of each visited Webpage, serves to determine the major topics of interest of the user (i.e. user preferences). Creating an accurate user profile is challenging, since the appropriate set of data about users should be collected in a usable way.

In general, user profiling for advertising purposes is related to the knowledge generation about the gender, age range and user preferences on the basis of the Web navigation. In order to add gender and age range profiling to user model generated from the Web usage data, the following hypothesis will be considered.

#### **H3: Available surveys and statistics about demographics and Internet services can provide a clear insight of the preferences of the users**

The main challenge of the proposed approach is the use of only implicit information. Thus, no personal information about users is known, i.e. all users are considered as non-registered. In this scenario, gender and age range characterization of online users is not obvious. In particular, merging Web usage data of the users with available surveys and statistics that relates demographics and Internet services, aids to solve the cold start problem for non-registered users, i.e. this

system can be used to train and enhance profiling algorithms to infer user profiles when no previous data is available.

4 The Concept of the eSPIGA System

The proposed method for automatic “persistent personalization” of Web ads is based on Web-mining techniques that combine representative parameters for advertising in a unique platform: the most suitable content of the Website of the advertiser; the probability of the click-through; the advertising targets arising from contracts with advertisers; and fuzzy mechanisms.

Figure 1 displays the general scheme of the eSPIGA system, which includes both, authoring tool to manage the advertising campaigns and the Websites; and the analytics dashboard to analyse the results of the ad campaigns and users’ profiles. First, Web content data of a publisher is processed using Natural Language Processing (NLP) algorithms to index the content of all pages of the Websites. Terms obtained from the HTML content are filtered on the basis of several categories to classify such pages.

Secondly, the acquisition of HTTP requests and the extraction of the sessions of the users are necessary to determine the usage mining data, which is the set of historical user sessions together with information about the ads clicked during these sessions. The current user behavior consists of data about visited pages as well as

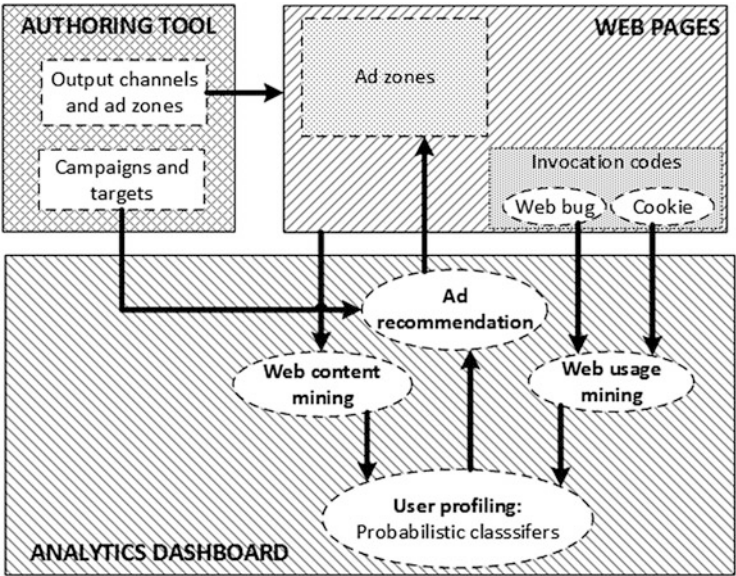


Fig. 1 Overview of the persistent personalization approach of eSPIGA system

the viewed and clicked ads during the active session. Such data is obtained using Web-bugs and cookies to categorize users into user profiles.

Finally, the advertising personalization algorithm selects the most adequate ads for the user on the basis of the Web content mining and usage mining data by means of classification and clustering techniques.

This section details the corresponding functionality of developed modules as well as the problems that arouse during the implementation.

## 4.1 Web Content Mining

Web content mining extracts useful information and Web knowledge from Web sources or Web contents such as text, image, audio, video and structured records. The eSPIGA system captures, pre-process, cleans and categorizes the text content of the Webpages of the publishers into several predefined categories using some Open Source libraries such as *ascrawler4j* (Ganjisaffar, 2012), *boilerpipe* (Kohlschütter, Fankhauser, & Nejd, 2010) and *OpenNLP* (Morton, Kottmann, Baldrige, & Bierner, 2005).

A content-targeted system relies on matching ads and their associated keywords to the text of a Webpage. In contrast to content-targeted systems, the eSPIGA system categorizes each Webpage into a category label on the basis of a categorization algorithm which requires previous training of a set of predefined categories. The set of categories used in eSPIGA are defined in Table 1 and have been chosen from the main categories of Google Ads.

One of the main difficulties for the method is the multilingualism of Web contents. In particular, the categorization algorithm needs to be trained in all the languages considered to be analysed. Thus, only two languages (Basque and Spanish) have been included for this pilot.

The Web categorization algorithm has been trained with texts from Wikipedia. Based on its categories and lists, the training process has been automated in order to simplify the inclusion of new languages. 14,730 Wikipedia pages have been considered to train the selected 20 categories in Spanish.

**Table 1** Web content categories considered in eSPIGA

Web content categories						
Sports	Finances	Politics	Travel	Arts	Gastronomy	Animals
Cars and vehicles	House and garden	Health and wellness	Science	Online communities	Jobs and Education	Real state
Software and hardware	Companies and industry	Books and literature	Games	Food and beverages	Internet and communications	

However, as there are not many Wikipedia pages in Basque, news from Basque online media has been used to train the categorization algorithm (14,525 pieces of news from 732 Webpages).

Once categories have been trained, the categorization algorithm creates the probabilistic model that calculates the probabilities of a Webpage belonging to each category based on its content. For this purpose, a maximum entropy probabilistic model is generated for each language using the Open Source module Apache openNLP, which selects the most representative words for each category, assigning a specific weight to its ability to classify a trained category. When the classification model is built, the content of the Webpage of the publisher is classified into the best rated category based on its keywords, main text and language. To automate this process, a specific crawler has been implemented so that all the Webpages of a main URL are obtained and classified once a day.

## **4.2 Web Usage Mining**

Web usage mining is described as the application of Data Mining techniques on Web access logs to optimize a Website for the preferences of the user, exploring new ways to navigate and perform during a visit to a Website. It is mainly based on the sequential analysis of pages visited during a given session, analysing Web clicks.

The first step of Web usage mining is the acquisition of the HTTP requests and the definition of user sessions, which is a list of URLs requested by the user during a navigation period through the Website of the publisher. Once a new Website is registered, an invocation code is generated. When the invocation code is placed on the main Webpage of the registered Website, a cookie is inserted the first time a user visits the Website. Cookies generate specific identifiers and parameters uniquely linked to individual users. However, this user identification technique has several limitations. Particularly, the assigned cookie is different in each browser for the same user. Furthermore, if the user deletes cookies, user will be considered as a new user.

Moreover, a Web-bug is also inserted to follow the activities of the users in any Webpage where this invocation code is inserted. It is a small file embedded in a Webpage that acts like a “spying agent”. It is a tool used to monitor who is visiting a Website, collecting information about the IP address, the type of Web browser used, the URL of the visited Webpages and the time and duration when the Webpage was viewed. Thus, Web-bugs added to the Website cookie, allow tracking the navigation and actions of a specific user within Websites over time.

Once a customer visits a Webpage with a Web-bug embedded, the eSPIGA server receives a request to check whether it is a new user or not. On the one hand, if the user is a recurrent visitor, he has already a cookie assigned and the corresponding navigation information is sent to the server. On the other hand, if the user has no cookie assigned, a new cookie is generated for such user before his

navigation information is sent to the server. Hence, the internet browsing history of the users are stored in the database. Furthermore, each user session is linked up to the set of ads viewed and clicked by the user during the session.

Finally, though stored and processed data does not comprise private information about the users, the system alerts and notifies users about the use of cookies and the purpose of these cookies on each of the registered Websites.

### 4.3 User Profiling

There are several supervised classification techniques for user profiling. A common approach to develop efficient algorithms includes the following steps: (1) obtaining a reasonable set of training and testing data (i.e. actual users gender and age range as well as their navigation patterns or preferences); (2) generating an appropriate supervised classification model trained with gathered data; (3) testing the classification model; (4) measuring the error; and (5) iteratively improving generated model with the last two steps so as to minimize the error (avoiding overfitting). However, as the proposed approach does not consider any information about registered users, neither real users training nor testing data can be considered. Hence, required training data is obtained from a novel approach based on official statistics about the relation between demographics and the usage of Internet.

Developed approach uses Web content mining and Web usage mining to infer users' preferences and link them with official statistics about the relation between demographics and the usage of Internet in order to assign most likely gender and age range to each user based on his/her preferences. Web content mining added to user Web usage mining allows inferring user preferences from visited Webpages topics. The first step is to store the Webpages visited by the users and the corresponding dates into the database. Then, users' preferences are ranked over time and corresponding gender and age range are inferred.

As no data about gender and age range from real users is known, a novel technique based on probabilistic classifiers has been implemented. The relation between population demographics and the usage preferences of the Internet services has been used as training data for supervised learning algorithms. In this scenario, only one sample representing all the population is available for each of the variables considered in the model. Probabilistic classifiers are used for this type of input data as they are usually useful to make simple but effective approaches.

A probabilistic classifier is a classifier that predicts, given a sample input, a probability distribution over a set of categories, rather than only outputting the most likely category that the sample should belong to. In a common classifier, a category is assigned to a sample by a modelled function, i.e.  $\hat{y} = f(x)$ , being  $x$  a sample and  $\hat{y}$  its assigned category label. This function is fitted using the training dataset  $X$  while the category labels form a finite set  $Y$  defined prior to training. Probabilistic classifiers generalize this notion of classifiers: instead of functions, they consider conditional distributions, i.e.  $\Pr(Y|X)$ , meaning that for a given  $x \in X$ , they assign



probabilities to all  $y \in Y$ , being the sum of these probabilities equal to one. Once these probabilities are known, “hard” classification can be done by the selection of the predicted category as the one with the highest probability.

A Bayesian probabilistic classifier has been used to create a probabilistic model of user preferences that predicts the probability of users belonging to each category by means of the training data. In particular, two probabilistic models have been built. While the first model computes the probability of each user of being male or female, the second model calculates users’ probability of belonging to each of the six age ranges (in years old) considered (16–24, 25–34, 35–44, 45–54, 55–64, 65–74).

Furthermore, Bayesian classifiers can be simplified to a Naïve approximation if model characteristics, e.g. user preferences, can be considered as independent, i.e. the probability of a user to have a specific preference does not depend on other preferences. It can be assumed that the correlation between the 20 selected preferences is not significantly high in this case. Thus, independence among preferences is assumed and Naïve Bayesian classifier is considered.

The Naïve Bayesian classifier defined predicts the profile (gender or age range) of each user by means of the probability that each preference belongs to each category, e.g.  $\Pr(\text{preference for sports}|\text{male})$ . Thus, assuming that the preferences of users are independent, these probabilities can be calculated as shown in Eq. (1).

$$P(C_i | P_{j|j \in (1,n)}) = \frac{P(C_i) \cdot \prod_{j=1}^n P(P_j | C_i)}{P(P_{j|j \in (1,n)})} \quad (1)$$

Being  $C_i$  the predicted model category  $i$  with  $i \in (1, 2)$  for gender model and  $i \in (3, 8)$  for age range model; and being  $n$  the number of preferences considered ( $n = 20$ ) and  $P_j$  the user preference  $j$  with  $j \in (1, n)$ . For example, to determine if a specific user tends to be male or female according to his main preferences, e.g. “sports” ( $j = 1$ ) and “house and garden” ( $j = 9$ ), the greatest probability from Eqs. (2) to (3) will be chosen.

$$P(C_{i=1} | P_{j=1}, P_{j=9}) = \frac{P(C_{i=1}) \cdot P(P_{j=1} | C_{i=1}) \cdot P(P_{j=9} | C_{i=1})}{P(P_{j=1}, P_{j=9})} \quad (2)$$

$$P(C_{i=2} | P_{j=1}, P_{j=9}) = \frac{P(C_{i=2}) \cdot P(P_{j=1} | C_{i=2}) \cdot P(P_{j=9} | C_{i=2})}{P(P_{j=1}, P_{j=9})} \quad (3)$$

Due to the consideration that the highest probability is finally chosen, i.e.  $\max\{P(C_1 | P_1, P_9), P(C_2 | P_1, P_9)\}$ , and both probabilities have equal denominator, the later can be neglected in order increase computation efficiency. This process is repeated for each age range. Once these conditional probabilities have been computed, most probable gender and age range are selected, i.e.  $\max\{P(C_i | P_1, P_9) | i \in (1, 2)\}$  and  $\max\{P(C_i | P_1, P_9) | i \in (3, 8)\}$ , respectively.

**Table 2** Comparison table between official statistics of population gender and age range and resulting gender and age range statistics from random user profiles with  $k = 1$  and  $k = 6$  for gender and age range models, respectively

	Gender ( $k = 1$ )		Age range ( $k = 6$ )					
	Male ( $i = 1$ )	Female ( $i = 2$ )	16–24 ( $i = 3$ )	25–34 ( $i = 4$ )	35–44 ( $i = 5$ )	45–54 ( $i = 6$ )	55–64 ( $i = 7$ )	65–74 ( $i = 8$ )
Expected	49.2 %	50.8 %	9.8 %	13.7 %	16.9 %	14.9 %	11.4 %	8.7 %
Obtained	57 %	43 %	3 %	30 %	21 %	29 %	4 %	13 %
MSE	$6.084 \times 10^{-3}$		$1.001333 \times 10^{-2}$					

In order to improve the fitting of actual results with official statistics, the  $k$  most relevant preferences must be defined. The number of user preferences  $k$  to be considered in the probabilistic model is determined by generating random user behaviours and minimizing the difference between official and random users' aggregated statistics. In particular, from the 16 preferences evaluated in the official statistics, 65.536 ( $2^{16}$ ) behaviour combinations were generated with random values.

It is important to avoid considering all ( $k = n$ ) preferences as representative as this consideration leads to the same profile for any user.

Thus, best  $k|k < n$ , must be selected so that user profile depends on the most relevant  $k$  preferences. Thus, the most appropriate  $k$  value was selected by minimizing the Mean Square Error (MSE) between the actual official statistics and the statistics inferred from the set of random profiles generated (Table 2). Obtained results were originated by the use of the most relevant preference ( $k = 1$ ) for gender profiling and the six most relevant preferences ( $k = 6$ ) for age range profiling.

Finally, as a user navigates the Web, his profile must be frequently updated. In particular, a user profile is updated every 10 page views. User profiling (gender and age range) algorithms are coded in R statistical software while other automated functions are developed in Java.

#### 4.4 Advertisement Recommendation

The objective of the eSPIGA system is to maximize the utility of advertising campaigns, which can be defined as the value perceived by the users, representing how much users like each ad. Specific ad utility for each user is measured by a positive value if the user clicks on the ad or negative when there is no click. The recommender system considered to maximize the ads utility is as follows.

Consider  $W$  the set of users and  $A$  the set of ads available. Being  $u$  the utility function that measures the utility of an ad  $a$  perceived by a user  $w$ . The recommendation objective is to measure the ad  $a' \in A$  that maximizes the utility for each user  $w \in W$ :

$$\forall w \in W, \quad a'_w = \underset{a \in A}{\operatorname{argmax}} u(w, a)$$

A hybrid recommendation model has been implemented, combining both the data of similar advertising campaigns (content-based filtering) and the utility of the ads within such campaigns for similar users (collaborative filtering). The hybrid model is executed in cascade. In the first step, appropriate campaigns are pre-selected based on content-based filtering, i.e. a subset of similar campaigns whose target properties fulfil user profile (gender, age range, preferences, location). Once this subset of campaigns is obtained, a collaborative filter is applied to select those ads from the pre-subset of campaigns which utility is higher for the subset of similar users defined.

Hence, the objective of the recommendation system is to serve the most valuable ad for each user, taking into account the utility of the ad within the campaigns that better fit the user and campaign targets.

Finally, some additional customizable filters have been added to the recommendation system in order to reduce user rejection to the ads recommended. For example, the same ad is never shown more than five times to the same user, and an ad that is clicked is not shown again.

5 Demonstration of the eSPIGA System

The eSPIGA system has been validated on Travel And Twitts Website ([www.travelandtwitts.com](http://www.travelandtwitts.com) [Sept. 10, 2015]). This Website is a travel blog where the Basque Regional Tourism Organization Basquetour ([www.basquetour.net](http://www.basquetour.net) [Sept. 10, 2015]) wanted to insert some marketing campaigns to promote several tourism activities in the Basque Country (Fig. 2).

The validation process was evaluated as follows. First, Basquetour was registered as a publisher in the eSPIGA system to create advertising campaigns and uploading their corresponding creative elements. Several marketing campaigns



Fig. 2 Travel and Twitts blog and eSPIGAs' analytics and authoring tools

were created related to family, gastronomy or culture events. Each campaign had several creative elements customized for specific user profiles. For instance, while the “Visit Euskadi” campaign was oriented to the general public (it was a general ad about the Basque Country), the remaining campaigns were segmented by age, gender, location and user preferences. Secondly, the administrator of the Travel and Twitts Website registered the main URL of the blog into the eSPIGA system.

Once the Website was registered, all the Webpages of the Website and their contents were processed to assign one category (from the 20 possible) to each Webpage. Note that this process is run each time the classification model is changed. In addition, in order to keep updated the classification of all the Webpages, new pages are daily detected and classified. The validation pilot included three ad zones: two square zones for the right hand menu of  $250 \times 208$  pixels, and one rectangular zone for the bottom end of  $630 \times 90$  pixels. Once the invocation codes have been added to the Website pages, personalized ads are shown to Travel And Twitts Web visitors.

During the 2 months validation process, a 0.44 % Click Through Ratio (CTR) was gained, which represents 4.4 clicks per 1000 visualizations. This CTR may seem too low, but it was higher than expected, as common CTR rates in online advertising campaigns are decreasing steadily (Idemudia, 2014) and usually located between 0.06 % and 0.5 % (Cole, 2008; Luna-Nevarez & Hyman, 2012).

## 6 Conclusions and Further Work

This work presents a novel approach to serve personalized ads to Websites users using their navigation stream. The novel online advertising approach for automatic “persistent personalization” of Web ads is based on Web-mining techniques that combine representative parameters for advertising in a unique platform: the most suitable content of the Website of the advertiser; the probability of the click-through; the advertising targets arising from contracts with advertisers; and fuzzy mechanisms.

Automatic ad personalization is based on user preferences. In particular, online activities and behaviours of the users are diverse enough to characterize different user profiles and preferences, which confirms hypothesis H1. As a user navigates the Web, his navigation activity can be obtained and stored. This information, along with the content analysis of each visited Webpage, serves to determine the major topics of interest of the user (i.e. user preferences), which supports hypothesis H2.

Each Website content is related to its main category. A set of 20 predefined topics were considered. Once user preferences are calculated, presented methodology infers the user gender and age range from his preferences and the relation between demographic and Internet usage surveys and statistics. This last statement serves to confirm hypothesis H3. With the users profiled, a hybrid recommendation system has been proposed to serve ads within the campaigns that better fit the user and other similar users.

Finally, the recommendation system has been successfully validated in a travel blog Website. The implemented prototype made it possible to serve the appropriate ads to the targeted audience on the basis of the classification of user profiles. The obtained CTR (0.44 %) was high compared with common CTR rates in online advertising campaigns, usually located between 0.06 % and 0.5 %.

Future work will be oriented towards most accurate Web content mining techniques based on NLP such as Named Entity Recognition. Such techniques will allow the extraction of additional characteristics to complete the preferences information about the users. Other field of extension is the adaptation of the approaches presented here to mobile applications, where user monitoring is still a challenge.

**Acknowledgements** Authors would like to thank the Basque Government for partially funding this project. Authors would also like to thank the staff of Goiena, Basquetour and Grupo Turiskopio for their valuable help and participation on the validation of the project.

## References

- Bae, S. M., Park, S. C., & Ha, S. H. (2003). Fuzzy web ad selector based on web usage mining. *IEEE Intelligent Systems*, 18(6), 62–69.
- Barford, P., Canadi, I., Krushevskaja, D., Ma, Q., & Muthukrishnan, S. (2014). Adscape: Harvesting and analyzing online display ads. *IW3C2*, 597–608.
- Bilenko, M., & Richardson, M. (2011). Predictive client-side profiles for personalized advertising. *Proceedings of the 17th ACM SIGKDD – KDD '11*, 413.
- Bleier, A., & Eisenbeiss, M. (2015). Personalized online advertising effectiveness: The interplay of what, when, and where. *MARKET SCI*, (0).
- Cole, S. (2008). Creative insights on rich media (Tech Rep). *DoubleClick Research*
- Davis, H. (2006). Google advertising tools: Cashing in with adsense. In *Adwords, and the Google APIs*. O'Reilly Media.
- Donnell, K. O., & Cramer, H. (2015). People's perceptions of personalized ads. In *IW3C2* (pp. 1293–1298).
- Ganjisaffar, Y. (2012). *Crawler4j*–Open Source Web Crawler for Java.
- Goldfarb, A., & Tucker, C. (2011). Rejoinder – Implications of “online display advertising: Targeting and obtrusiveness”. *Marketing Sci*, 30(3), 413–415.
- Idemudia, E. C. (2014). The visual-cognitive model for internet advertising in online market places. *International Journal of Online Marketing*, 4(3), 31–50.
- Kazienko, P., & Adamski, M. (2007). AdROSA-Adaptive personalization of web advertising. *Information Sciences*, 177(11), 2269–2295.
- Kohlschütter, C., Fankhauser, P., & Nejd, W. (2010). Boilerplate detection using shallow text features. In *WSDM 2010, New York City* (pp. 441–450).
- Luna-Nevarez, C., & Hyman, M. R. (2012). Common practices in destination website design. *Journal of Destination Marketing and Management*, 1(1–2), 94–106.
- Morton, T., Kottmann, J., Baldrige, J., & Bierner, G. (2005). *OpenNlp*: A java-based nlp toolkit
- Rusmevichientong, P., & Williamson, D. P. (2006). An adaptive algorithm for selecting profitable keywords for search-based advertising services. *ACM-EC*, pp. 260–269.
- Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. In *Applications of Data Mining to Electronic Commerce* (pp. 115–153). Springer US.

- Tsang, M. M., Su-Chun, H., & Ting-Peng, L. (2004). Consumer attitudes towards mobile advertising: An empirical study. *International Journal of Electronics and Communications*, 8(3), 65–78.
- Tucker, C. E. (2012). The economics of advertising and privacy. *International Journal of Industrial Organization*, 30(3), 326–329.
- Wang, C., Zhang, P., Choi, R., & Eredita, M. D. (2002). Understanding consumers attitude toward advertising. *AMCIS*, 2002, 1143–1148.
- Watson, C., McCarthy, J., & Rowley, J. (2013). Consumer attitudes towards mobile marketing in the smart phone era. *International Journal of Information Management*, 33(5), 840–849.
- Zanot, E. J. (1984). Public attitudes towards advertising. *International Journal of Advertising*, 3, 3–15.

Information and Communication Technologies in  
Tourism 2016

Proceedings of the International Conference in Bilbao,  
Spain, February 2-5, 2016

Inversini, A.; Schegg, R. (Eds.)

2016, XV, 792 p. 103 illus., 35 illus. in color., Softcover

ISBN: 978-3-319-28230-5