Expert Systems with Applications 40 (2013) 7478–7491

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it

Olatz Arbelaitz *, Ibai Gurrutxaga, Aizea Lojo, Javier Muguerza, Jesús Maria Pérez, Iñigo Perona

University of the Basque Country UPV/EHU, Computer Architecture and Technology Department, Spain

ARTICLE INFO

Keywords:
Bidasoa tourism website
Web usage mining
Web content mining
Web user profiling
Clustering
Frequent pattern mining
Topic modelling

ABSTRACT

The tourism industry has experienced a shift from offline to online travellers and this has made the use of intelligent systems in the tourism sector crucial. These information systems should provide tourism consumers and service providers with the most relevant information, more decision support, greater mobility and the most enjoyable travel experiences. As a consequence, Destination Marketing Organizations (DMOs) not only have to respond by adopting new technologies, but also by interpreting and using the knowledge created by the use of these techniques. This work presents the design of a general and non-invasive web mining system, built using the minimum information stored in a web server (the content of the website and the information from the log files stored in Common Log Format (CLF)) and its application to the Bidasoa Turismo (BTw) website. The proposed system combines web usage and content mining techniques with the three following main objectives: generating user navigation profiles to be used for link prediction; enriching the profiles with semantic information to diversify them, which provides the DMO with a tool to introduce links that will match the users taste; and moreover, obtaining global and language-dependent user interest profiles, which provides the DMO staff with important information for future web designs, and allows them to design future marketing campaigns for specific targets. The system performed successfully, obtaining profiles which fit in more than 60% of cases with the real user navigation sequences and in more than 90% of cases with the user interests. Moreover the automatically extracted semantic structure of the website and the interest profiles were validated by the BTw DMO staff, who found the knowledge provided to be very useful for the future.

1. Introduction

For tourists, the decision on which destination or product to choose requires a considerable time and effort (De Ascaniis, Bischof, & Cantoni, 2013) because, as Ballantyne, Hughes, and Ritchie (2009) stated, tourism services are a class of product regarded as high risk and consumers are often led to engage in extensive information search.

Destination Marketing Organizations (DMO) are the tourism organizations responsible for managing the promotion of a destination. Previous works, (for example Gretzel, Yuan, & Fesenmaier, 2000) identified the importance of DMOs understanding new challenges and the meaningful use of new technologies to seek excellence in destination marketing. Marchiori, Milwood, and Zach (2013) affirmed that in order to maintain their share of the market tourism organizations have to respond not only by adopting new technologies, but also by interpreting and using the knowledge created by Internet users.

The tourism industry has experienced a shift from offline to online travellers. Experts underlined many years ago that the Internet is the main source of information in the tourist domain (Gretzel et al., 2000). An increasing number of travellers are no longer dependent on travel agencies to look for information for their next trip; they have replaced using agencies by the use of the Internet (ETC, 2012). Steinbauer and Werthner (2007) affirmed that the development of information communication technologies during the last decade has affected the tourism industry, as a growing number of travellers have begun to look for tourism information online. As the experts pointed out in the ENTER 2013 eTourism conference held in Innsbruck in January 2013, 'these systems have significantly changed the travel industry'. As a consequence, DMOs must use their official websites to interact with tourists in order to promote a destination and provide information on it and, furthermore, they should extract knowledge from this interaction. As e-Destinations serve as platforms where consumers can be inspired, get all the information they need about the desired destination and eventually book the holiday, the presence of destinations in the

* Corresponding author. Address: University of the Basque Country UPV/EHU, Computer Architecture and Technology, Manuel Lardizabal 1, 20018 Donostia, Gipuzkoa, Spain. Tel.: +34 943 018042; fax: +34 943 015590.

E-mail addresses: olatz.arbelaitz@ehu.es (O. Arbelaitz), i.gurrutxaga@ehu.es (I. Gurrutxaga), aizea.lojo@ehu.es (A. Lojo), j.muguerza@ehu.es (J. Muguerza), txus.perez@ehu.es (J.M. Pérez), inigo.perona@ehu.es (I. Perona).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.040

web is crucial (Pan & Fesenmaier, 2003). Moreover, as Hsu, Shih, Huang, Lin, and Lin (2009) conclude in their work, the web facilities provided to tourists affect their loyalty.

The success of electronic commerce, especially for the less well-known companies, is largely dependent on the appropriate design of their website (Turban & Gehrke, 2000). In the paper of Chaffey, Ellis-Chadwick, Johnston, and Mayer (2006) it is stated that a good website should begin with the users and understanding how they use the channel to shop. This confirms that understanding the needs and preferences of the website audience will help to answer questions about what the content of the website should be, how it should be organized and so on.

In the last decades, the same trends followed by the tourism industry have been noticed in many other industries. This evolution has led to a dramatic increase in the amount of information stored in the web, which often makes the information intractable for users. As a consequence, the general need for websites to be useful in an efficient way for users has become especially important. There is a need for easier access to the required information and adaptation to the users' preferences or needs. Web personalization thus becomes essential in industries such as tourism and it can be positive for both the user and the business. According to Pierrakos, Paliouras, Papatheodorou, and Spyropoulos (2003) web personalization can be defined as the set of actions to dynamically adapt the presentation, the navigation schema and the contents of the website, based on the preferences, abilities or requirements of the user. Nowadays, as Brusilovskys (2007) describe, many research projects focus on this area, mostly in the context of e-Commerce (Brusilovskys, 2007) and e-learning (García, Romero, Ventura, & Castro, 2009). Important websites such as Google and Amazon are clear examples of this trend.

Within this context, the use of intelligent systems in the tourism sector has become crucial (Gretzel, 2011). These information systems can provide tourism consumers and service providers with the most relevant information, more decision support, greater mobility and the most enjoyable travel experiences. As stated in the previous paragraphs, the Internet has become one of the most widely accepted technologies and there is currently a wide range of systems related to it such as recommender systems, context-aware systems, web mining tools, etc. Moreover, travel agents are among those service providers for whom adoption of the Internet could be the best marketing device for their business and a tool to give them a competitive advantage (Abou-Shouk, Lim, & Megicks, 2013).

In any web environment, the contribution of the knowledge extracted from the information acquired when the users navigate in a website is twofold: it can be used for web personalization (i.e., for the adaptation of the website according to the user requirements) and also to extract knowledge about the interests of the people browsing the website, which will then be useful for the service provider; in the case of tourism websites, the staff of the DMOs.

Although focused on a tourism website, the aim of the research presented in this paper is more ambitious. We aim to build a general web mining system (Mobasher, 2006) to work with any website in the two areas addressed in the previous paragraph. Web mining can be defined as the application of machine learning techniques to data from the Internet. This process requires a data acquisition and pre-processing stage. The machine learning techniques are mainly applied in the pattern discovery and analysis phase to find groups of web users with common characteristics related to the Internet and the corresponding patterns or user profiles. Finally, the patterns detected in the previous steps are used in the operational phase to adapt the system and make navigation more efficient for new users or to extract important information for the service providers.

In the first stage of our web mining approach we analysed the navigation of users (web usage mining) and built user navigation profiles that provide a tool to adapt the web to new users while they are navigating (through link prediction). We then automatically extracted thematic information from the content of the URLs (web content mining) and combined this with usage information to obtain information about the interests of the users browsing the website (i.e., we extracted user interest profiles). The system is general and non-invasive, because it has been built using the minimum information stored in a web server: the content of the web and server log files stored in web Common Log Format (CLF, 1995). The whole process is thus carried out without disturbing the user. Moreover, this makes the proposed system easily extensible to any other environment.

The system proposed in the paper was built for the Bidasoa Turismo (BTw) website, *www.bidasoaturismo.com*, based on usage data collected over 10 months and the corresponding content data.

The paper describes related work in Section 2. Section 3 introduces the data environment where the system was developed and Section 4 describes the generation of navigation profiles, their use for link prediction and the evaluation. Section 5 is devoted to describing the interest profiling process, its outcome and the expert validation. Section 6 explains the introduction of semantic structure in navigation profiles and is followed by Section 7, where semantics is used as a tool to diversify link proposals. Finally, Section 8 summarises the conclusions and future work.

2. Related work

User profiles are part of many electronic tourism applications and particularly recommender systems. The information needed to build a user profile can be obtained explicitly or by observing the actions of the user (Schiaffino & Amandi, 2009). In the area of tourism, user profiles are mainly generated by asking the users to fill in a questionnaire or by an interface (Luberg, Jrv, & Tammet, 2012). The users must usually complete some steps in order to create the profile; for instance, by selecting photos that they like (Cao et al., 2010; Berger, Denk, Dittenbach, Pesenhofer, & Merkl, 2007). Thus, the most common user profiling strategy in tourism is to use information provided explicitly by the user. The user profiles created are generally used to recommend some tourist plan or information based on a collaborative filtering approach.

Explicit acquisition of user information has several problems. First, users are generally not willing to provide information by filling in long forms. Second, they do not always tell or write the truth about themselves when completing forms. There is an alternative option that avoids this type of problem: to implicitly acquire user information by observing their actions. In the context of the web, this can be done using web server logs (Zanker, Fuchs, Hpken, Tuta, & Mller, 2008). It is also common to use Global Positioning Systems (GPS) or Smart phones to acquire user information such as location, time and language (Sarkaleh, Mahdavi, & Baniardalan, 2012). In these cases the user does not have to provide any personal information.

Web usage mining can be used to extract knowledge from observed actions. A wide variety of techniques have been used with this aim, such as case-based reasoning (Godoy, Schiaffino, & Amandi, 2004), Bayesian networks (Garcia, Amandi, Schiaffino, & Campo, 2007), association rules (Schiaffino & Amandi, 2006), genetic algorithms (Yannibelli, Godoy, & Amandi, 2006), neural networks (Villaverde, Godoy, & Amandi, 2006), topic modelling (Fujimoto, Etoh, Kinno, & Akinaga, 2011) etc. Similar techniques have also been applied to the area of tourism but, to our knowledge, they have always been applied to extract knowledge from explicitly acquired user information. For example Hsu et al. (2009) presented a system that uses an integrated Bayesian network mechanism using a linear structural relation model (LISREL) to predict tourism loyalty

based on 425 valid answers to a poll (explicit information requirement) collected from tourists about their holiday experience at the Toyugi hot spring resort in Taiwan. In another work, Brejla and Gilbert (2012) use a data-driven approach to knowledge discovery. Their aim was to achieve a deeper understanding of guest-to-guest and guest-to- staff interactions on board cruise ships. They use holiday reviews retrieved through web content mining from CruiseCritic.com. Although the previous works describe the application of web mining techniques in the tourism context they are not based on user navigation logs but on user reviews or information explicitly acquired from users; i.e., they are not web usage mining techniques.

One of the most widely pursued objectives of web usage mining applications has been web access pattern discovery. Although web access pattern discovery has many applications (such as improving web cache performance, personalizing the browsing experience of the users, recommending related pages, etc.) the most widely explored application in the web research community has been web page prefetching. Many years ago, Kroeger, Long, and Mogul (1997) showed that the performance improvement achieved by combining prefetching and caching (i.e., downloading pages that are likely to be visited in the future and storing them in the cache) can be twice that of caching alone. Since then, many approaches have been published that, taking user click sequences as a starting point, concentrate on predicting the next page that will be accessed in order to prefetch it before the user requests it and thus reduce web access latency (Chen & Zhang, 2003; Anitha, 2010; Makkar, Gulati, & Sharma, 2010; Zukerman, Albrecht, & Nicholson, 1999). Common characteristics of these approaches are generally the use of clustering and/or Markov models to predict the next link to be accessed. In general, the results differ from paper to paper and they are difficult to compare.

Some different prefetching approaches can be found in Makkar et al. (2010) and Bhawsar, Pathak, and Patidar (2012). The former combines the user log information and the structure of the website, while the latter proposes a solution that can be used when the navigation logs are not large enough. The order in which the users access the pages is important in differentiating the usage patterns. In fact, this is probably the reason for the popularity of using sequence analysis methods to predict web access. Another approach would be to take into account the access sequence in a clustering process. The work of Chordia and Adhiya (2011) proposes an efficient implementation of sequence alignment methods for grouping web access sequences that combine global and local alignment techniques.

As previously mentioned, one of the objectives of this work is to discover navigation profiles and use these to personalize the browsing experience of the user, thus making it easier and more convenient. We consider that this can be done by adapting the navigation scheme by providing the users with a list of links of interest to them in the early stages of the navigation, so that they can achieve their objective faster. This would probably help them to have a shorter and more satisfactory navigation experience by skipping some of the intermediate pages on the way to their objective. In contrast, prefetching would take the users through the same path they would navigate without adaptation but faster.

Furthermore, we aim to combine the knowledge extracted from web usage information with knowledge extracted from the web content through web content mining techniques to provide the staff of the Bidasoa Txingudi Bay DMO with information about the types of users browsing their website, according to interests. This information will be useful to provide a better service or for future marketing campaigns.

Although not in the tourism area, and not very frequently, some researchers have published works that combine usage and content information for web page recommendations. For example Senkul and Salin (2012) investigated the effect of semantic information on the patterns generated through web usage mining in the form of frequent sequences. To do this, they developed a framework for integrating semantic information into the web navigation pattern generation process, where frequent navigational patterns are composed of ontology instances rather than web page addresses. They measured the quality of the generated patterns through an evaluation mechanism involving web page recommendation. Although this work combines web usage and content information, it requires having previously built an ontology representing the concepts in the website explored.

We can also find other descriptive works where web mining is used to generate suggestions for the web master. Carmona et al. (2012) present the methodology used in an e-commerce website for the sale of extra virgin olive oil. The data used were not access sequences but mechanical and use characteristics extracted from Google Analytics. The knowledge was extracted using unsupervised and supervised machine learning algorithms such as clustering, association and subgroup discovery. The results presented were mainly aimed at the website design team, providing some guidelines for improving its usability and user satisfaction. Hung, Chen, Yang, and Deng (2013) carried out a study with the main purpose of understanding the self-care behaviour of elderly participants in a self-care service system that provides self-care service and to analyse the daily self-care activities and health status of elderly people living at home alone. Each page visited by the elderly person belonged to a previously annotated topic that was used to generate interest-based representations of user sessions. The authors used the ART2-enhance *K*-means algorithm to mine cluster patterns, combined with sequence-based representation schemes in association with Markov models to capture sequential profiles. The authors provide qualitative results that can be used for research in medicine, public health, nursing and psychology and for policy-making in the health care domain.

As can be seen from the above, none of the works described included all the characteristics of the system described in this paper: a general, automatic and non-invasive system that combines usage and content information from a tourism website to make the optimum use of it. The works that had the aim of predicting or recommending links had no further aim and they did this based only on usage information. In the cases where usage information was combined with content information, the content information was not extracted automatically; some prior knowledge of the content structure of the website was required. Finally, focusing in the context of tourism, the web mining applications we found in bibliography were invasive; they required information obtained explicitly from the customers. In contrast, our work proposes a system that, based on usage and content information and without any prior knowledge of the content structure of the website, is able to predict links and to provide information about the types of users that navigate in a website according to their interests.

Fig. 1 presents a schema of the proposed system, including all the processes executed to automatically generate user interest and navigation profiles based on usage and content information. The different phases and their results will be described in the following sections.

3. Data environment

The application environment was the Bidasoa-Txingudi bay DMO, which is located at the western tip of the Pyrenees, straddling two countries (France and Spain) and linking the Basque provinces of Gipuzkoa and Lapurdi. The Bidasoa River has had the effect of socially and culturally linking the three towns surrounding the bay (Hendaye, Hondarribia and Irun). The area offers

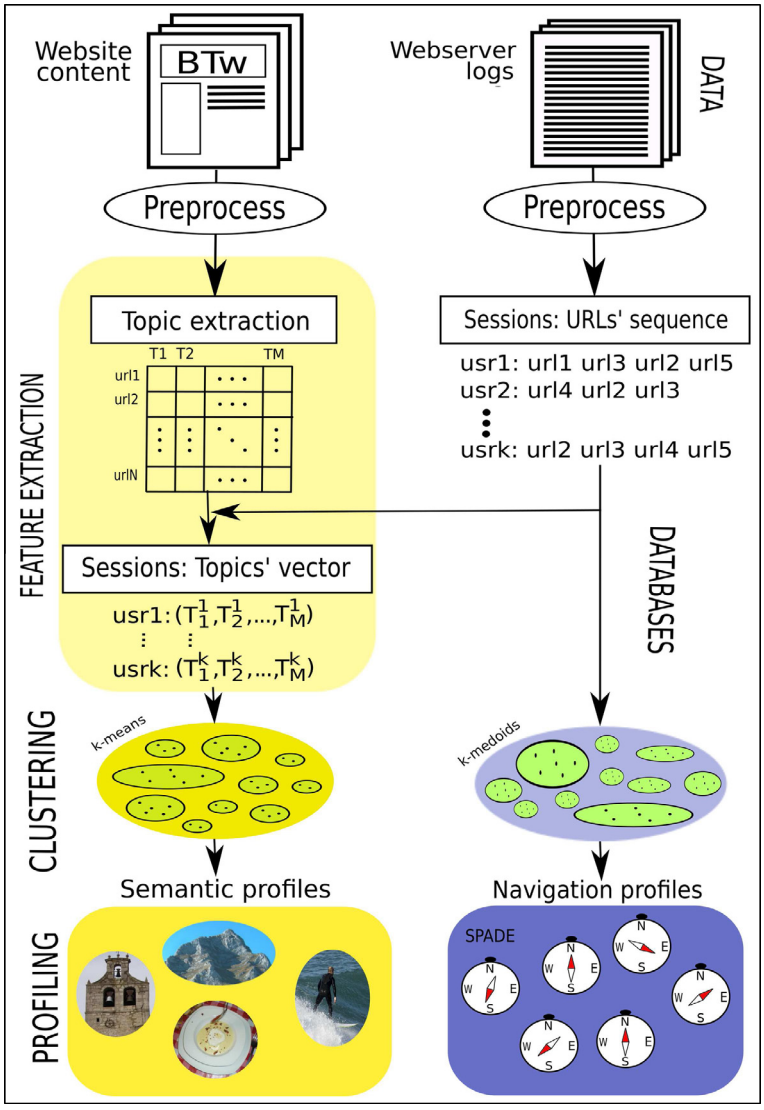

Fig. 1. Schema of the system architecture.

the opportunity of a wide range of tourism activities and the Bidasoa Turismo website (BTw) *www.bidasoaturismo.com* includes all sorts of practical tourist information on the area: thematic tourism, professional tourism, gourmet tourism, agenda, recommendations, etc. A screen shot of the website is shown in Fig. 2.

Nearly ten months of usage data for BTw were provided by the staff of the DMO (from January 2012 to October 2012). The information contained in this database is contained in the web server logs of requests (a total of 3,636,233) stored in Common Log Format (CLF, 1995). We also acquired and used the content information of the website; i.e., the text appearing in the website.

In addition, we used a third source of information: the language used when accessing the website. The web page can be accessed in four different languages (Basque, Spanish, French and English), which provides information about the origin of the users.

4. Navigation profiling (link prediction)

The generation of user navigation profiles based on web usage information can be divided into two main steps: data acquisition and preprocessing (Cooley, Mobasher, & Srivastava, 1999) and pattern discovery and analysis.

4.1. Data preprocessing

The usage data in web mining refers to the data generated from the interaction between the user and the web server. The most basic records of these interactions are recorded in the log files of the web servers, which register all the web page requests received by the web server. These log files contain raw data that needs to be preprocessed before starting the user profiling phase.

Web server log files follow a standard format called Common Log Format (CLF, 1995). This standard specifies the fields all log files must have for each request received: remotehost, rfc931, authuser, date, request, status and bytes. The fields we used for this work are the remote host IP address, the time the request was recorded, the requested URL and the status field that informs about the success or failure when processing the request. Fig. 3 shows some sample lines of a log file.

The log files used in this work were the result of 10 months of recording of requests received by the web server. They contained 3,636,233 requests, which were reduced to 168,556 after the data preparation or data preprocessing phase described in the following lines. First of all we removed erroneous requests, those that had an erroneous status code (client error (4xx) and server error (5xx)).

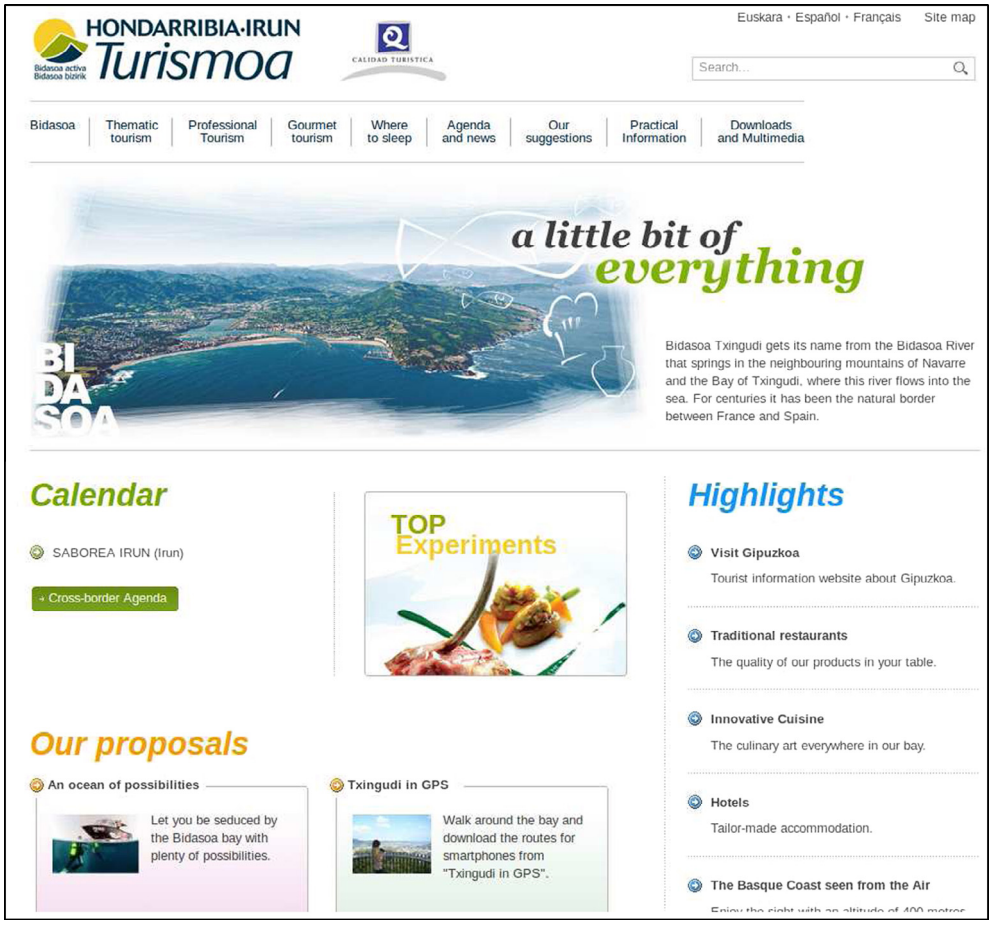

Fig. 2. Appearance of the home page of BTw website.

```
207.46.13.48 - - [22/Feb/2012:00:04:05 +0100] "GET /index.php?...&lang=es HTTP/1.1" 200 30055 "-" "Mozilla/5.0 (com
207.46.19.49 - - [22/Feb/2012:00:04:07 +0100] "GET /index.php?...&lang=en HTTP/1.1" 200 29646 "-" "Mozilla/5.0 (com
207.46.19.49 - - [22/Feb/2012:00:04:07 +0100] "GET /index.php?...&lang=es HTTP/1.1" 200 28088 "-" "Mozilla/5.0 (com
66.249.72.32 - - [22/Feb/2012:00:04:09 +0100] "GET /index.php?...&lang=es HTTP/1.1" 200 29440 "-" "Mozilla/5.0 (com
207.46.99.49 - - [22/Feb/2012:00:04:12 +0100] "GET /index.php?...&lang=fr HTTP/1.1" 200 28106 "-" "Mozilla/5.0 (com
207.46.19.49 - - [22/Feb/2012:00:04:13 +0100] "GET /index.php?...&lang=en HTTP/1.1" 200 29557 "-" "Mozilla/5.0 (com
207.46.19.49 - - [22/Feb/2012:00:06:06 +0100] "GET /index.php?...&lang=eu HTTP/1.1" 200 23380 "-" "Mozilla/5.0 (com
73.224.15.77 - - [17/Sep/2012:00:00:00 +0200] "POST /administ...index.php HTTP/1.1" 301 261   "-" "Mozilla/5.0 (com
13.4.215.228 - - [17/Sep/2012:10:21:58 +0200] "GET /templates/...logo.gif HTTP/1.1" 304 -     "-" "Mozilla/5.0 (com
194.69.224.7 - - [18/Sep/2012:09:16:31 +0200] "GET /templates/....uery.js HTTP/1.1" 200 55774 "-" "Mozilla/4.0 (com
194.69.224.7 - - [18/Sep/2012:09:16:33 +0200] "GET /images/....Button.gif HTTP/1.1" 200 368   "-" "Mozilla/4.0 (com
194.69.224.7 - - [18/Sep/2012:09:16:33 +0200] "GET /templates/...logo.gif HTTP/1.1" 200 12530 "-" "Mozilla/4.0 (com
194.69.224.7 - - [18/Sep/2012:09:16:33 +0200] "GET /templates/...ogo2.gif HTTP/1.1" 200 3451  "-" "Mozilla/4.0 (com
194.69.224.7 - - [18/Sep/2012:09:16:33 +0200] "GET /templates/...piti.gif HTTP/1.1" 200 45    "-" "Mozilla/4.0 (com
194.69.224.7 - - [18/Sep/2012:09:16:35 +0200] "GET /templates/...irun.png HTTP/1.1" 200 2450  "-" "Mozilla/4.0 (com
```

Fig. 3. Sample lines of a log file in CLF.

We therefore only took into account successfully processed requests. The next step consisted of selecting the requests directly related to the user activity. User clicks indirectly send many web browser requests to complete the requested web page with images, videos, style (css) or functionalities (scripts), for example. All these indirect requests were removed. We also removed requests related to the web administration activity, which could be detected in our case by analysing the URLs subfields. We then selected and normalized the format of the parameters of the URLs so that only those that really influence the web pages final appearance were taken into account, thus avoiding noisy parameters. As a result of this processing, different URLs with the same appearance are considered to be identical. We then carried out user identification based on IP addresses and session identification.

We completed the user identification process based on IP addresses and we fixed the expiry time of each session to 10 min of inactivity (He & Gker, 2000). We selected the most relevant of the sessions obtained (those with a minimum activity level; 3 or more clicks), and removed the longest sequences (those with more than 86 requests; out of 98% percentile), with the assumption that long sequences are outliers and might be caused by some kind of robot, such as, crawlers, spiders or web indexers. We represented each session as a sequence of URLs, where a session of length L could be represented as:

$$S_s = u_s^1 u_s^2 \dots u_s^L \tag{1}$$

where u_s^l corresponds to the lth URL in sequence s.

Table 1
Number of requests and sessions after the different preprocessing phases.

	Requests	Sessions
In log files	3,636,233	
Valid requests	2,002,827	
Sessions after sessionning and filtering	765,712	66,897
Sessions after removing cultural programming and news	168,556	21,916

Finally, the last stage in the preprocessing consisted of removing cultural programming and news. We noticed that nearly 35.9% of the requests belonged to this type of page but since the information they contain is volatile we decided to keep their analysis outside the scope of this work. Table 1 shows a summary of the number of requests and session after the different preprocessing phases we applied.

According to the numbers in Table 1 the final database contained 21,916 sessions with an average length of 7.7 clicks.

4.2. User navigation profile discovery

This is the stage that, taking the user click sequences as input, is in charge of modelling users and producing user navigation profiles. Most commercial tools perform statistical analysis on the data collected. They extract information about the most frequently accessed pages, average view times, average lengths of paths, etc., which are generally useful for marketing purposes. But the knowledge extracted from this kind of analysis is very limited. Machine learning techniques are generally able to extract more knowledge from data. In this context, unsupervised machine learning techniques have shown to be adequate to discover user profiles (Pierrakos et al., 2003).

We used the PAM (Partitioning Around Medoids) clustering algorithm (Kaufman & Rousseeuw, 1990) and the Edit Distance sequence alignment method (Gusfield, 1997; Chordia & Adhiya, 2011) as a metric to compare sequences. The aim was to group users that show similar navigation patterns. PAM requires the k parameter to be estimated. This parameter is related to the structure of the database and the specificity of the generated profiles, where the greater its value the more specific the profiles will be.

The outcome of the clustering process was a set of groups of user sequences that show similar behaviour. However, we intended to model those users or to discover the associated navigation patterns or profiles; i.e., to find the common click sequences appearing in the sessions in a cluster. For profiling we used SPADE (Sequential PAttern Discovery using Equivalence classes) (Zaki, 2001), an efficient algorithm for mining frequent sequences, which we used to extract the most common click sequences in the cluster. SPADE uses combinatorial properties to decompose the original problem into smaller sub-problems that can be independently solved in main-memory. All sequences are discovered in only three database scans. In order to adapt SPADE to the sequential nature of the data we matched each user session with a SPADE sequence, with events containing a single user click. The application of SPADE provides for each cluster a set of URLs that are likely to be visited in the sessions belonging to it. The number of proposed URLs depends on parameters related to the SPADE algorithm, such as minimum support and maximum allowed number of sequences per cluster.

The optimum value for the parameters depends on the characteristics of the clusters (such as size, structure, compactness, etc.), which will vary from one cluster to another. Hence, it is important to regulate the system so that it finds an adequate number of URLs to propose. In our system we have fixed minimum support to a small value (0.2) so that enough URLs are proposed to profile the sessions in a cluster. Nevertheless, some of the clusters seemed to be too diverse and SPADE was unable to find common patterns for them. This happened on average for 7% of the clusters containing 8% of the examples and we could not consider these clusters as user profiles.

As we described in the introduction, the generated navigation profiles will be used for link prediction. In this context, proposing a large number of links would stress the user and we therefore decided to fix the maximum number of URLs per profile to 4.

4.2.1. User navigation profiles: evaluation

Before the system is used in any real application the generated profiles need to be evaluated; i.e., we need to compare the generated profiles with the profiles of new users navigating the website and measure their similarity. We first generated user profiles by combining PAM with SPADE and compared these profiles to those for new users navigating the website. In order to carry out this evaluation we used a 10-fold-CV (Cross Validation) methodology, dividing each folder into a training set (15,341 examples), validation set (4380 examples) and test set (2196 examples). We used the validation set to select k (the number of clusters) and the test set to evaluate the performance of the system.

The internal structure of the data is completely unknown and we therefore tried a wide range of values for K to select the optimum number of clusters; we tried values ranging from 20 to 700. Although a usual exploration limit for the number of clusters is $\sqrt{n}$, which makes it advisable to explore values of K up to 124 for 15341 training examples, and having more profiles makes the phase of identifying the profiles of new users navigating the website more costly, the exploration area was so extensive because we prioritized finding the real structure of the data and obtaining good quality profiles.

The best way to select the optimum number of clusters would be to use a Cluster Validity Index (CVI). However, probably because most CVIs were designed and tried for a small number of clusters, the CVIs reported to be the best in Arbelaitz, Gurrutxaga,

Table 2
Profile evaluation: average validation set results for the 10-fold-CV.

K	avg(nURL/prof)	Pr	PrIncr	Re	ReIncr	F1	F1Incr
20	2.91	0.532		0.274		0.361	
40	2.93	0.556	0.00123	0.290	0.00081	0.381	0.00099
80	2.89	0.590	0.00085	0.309	0.00046	0.405	0.00060
120	2.93	0.603	0.00033	0.321	0.00032	0.419	0.00035
160	2.95	0.615	0.00029	0.331	0.00024	0.430	0.00027
200	2.89	0.633	0.00046	0.337	0.00016	0.440	0.00025
300	**2.90**	**0.648**	**0.00015**	**0.347**	**0.00010**	**0.452**	**0.00012**
400	2.90	0.656	0.00008	0.352	0.00005	0.458	0.00006
500	2.90	0.662	0.00006	0.356	0.00004	0.463	0.00005
700	2.96	0.663	0.00000	0.362	0.00003	0.468	0.00003
Test (300)	2.89	0.642		0.344		0.448	

Muguerza, Pérez, and Perona (2013) did not provide any coherent result. We therefore evaluated the quality of the partitions according to the precision, recall and F-measure performance metrics, and determined the trade-off between the value of the performance metrics and the number of clusters based on the analysis of the gradient of their level of improvement.

The generated profiles were evaluated by comparing them to new users navigating the website. With this aim, the system needs to select a profile for the new users which will then be compared to their click sequence. The selection is done according to a distance calculation. This can be done at any stage of the navigation process; i.e., from the first click of the new user to more advanced navigation points. The hypothesis is that the navigation pattern of the user will be similar to the user profiles of its nearest clusters. As a result, the system will propose to the new user the set of links that models the users nearest clusters.

In order to simulate a real situation we need to take into account that when a user starts navigating only the first few clicks will be available to be used for deciding the corresponding profile. We simulated this real situation using 25% (more or less 2 links, because the click sequences have on average 7.7 links) of the validation or test sequences to select the profile for the new user according to the built model.

According to previous works found in the bibliography (Arbelaitz et al., 2012), new users might not be identical to any of the profiles discovered in the training set; their profile might have similarities with more than one profile and, as a consequence, the diversification helps; it is better to build the profiles of the new users dynamically based on some of their nearest profiles. We propose the use of the *k*-Nearest Neighbour (Dasarathy, 1991) supervised learning approach to calculate the distance from the click sequence (Edit Distance to the medoid) of the new users to the clusters generated in the previous phase. We used 5-*NN* to select the nearest clusters and combined the profiles of the two nearest clusters with defined profiles, fixing URL selection probabilities according to their distance. More than two clusters have to be selected because, as we explained in the previous subsection, some of the clusters might not have their corresponding profile and we want to ensure that at least two of the clusters will have it defined. We combined these to propose profiles containing at most 4 URLs; those with the highest support values. If there are not enough URLs exceeding the minimum support value the profiles could have less than 4 URLs.

We computed performance metrics based on the results obtained for each of the new users. We compared the number of proposed links that are actually used in the test or validation examples (hits) and the number of proposals that are not used (misses) and calculated precision, recall and F-measure.

The greater the number of URLs proposed as profiles the smaller will be the significance of some of them and the risk taken by the system will thus be greater. As a consequence, the values for precision will probably drop. In contrast, by limiting the maximum number of URLs proposed for each profile to 4 the recall values will never reach 1. Since the average length of the sequences is 7.7, if we propose a profile (4 URLs) based on 25% of the navigation sequence (more or less 2 URLs), we would have sequences of 4 + 2 = 6 URLs, and the value of recall would be at most 0.78.

Table 2 shows the average results (precision, recall and F-measure) obtained for the validation set in the 10 folds for different values of *K*. The PrIncr, ReIncr and F1Incr columns show the increment of the performance metrics per extra cluster. The second column, avg (nURL/prof), shows the average number of links proposed in each profile. As we explained in the previous paragraph, this value limits the maximum recall value that can be obtained; in this case to 0.64.

The values of the performance metrics in Table 2 improve as the number of clusters increases, which means that the system users are very diverse and, as a consequence, smaller clusters better capture the different types of existing profiles. However, if we analyse the improvements in the performance metrics for each new cluster added, or normalized improvement differences, we detect an elbow, or change in order of magnitude of the improvement (marked in bold in Table 2), in $K = 300$. We therefore fixed the number of clusters to 300. Thus, from this point on, we set the system to 300 clusters, with a maximum of 4 URLs provided as profile. The last row in Table 2 shows the average results of the 10 folds, for the test set and $K = 300$.

The results show that the values obtained for the test samples are similar to those obtained for the validation samples, thus confirming that the behaviour of the system is stable. Moreover, if we focus on the values of the performance metrics we can say that the profiles proposed to the new users fit in more than 60% with the real navigation sequence and that the generated profiles therefore have good quality.

Nevertheless, a qualitative analysis of the generated profiles showed that many of the users visited the home page at intermediate points of the navigation. We consider this provides information about a users navigation behaviour; it could suggest situations such as being lost or changing their mind about their interests . . . so we still consider this as part of the profile. However, we consider that if the profiles are used to propose links to new users the proposal of the home page would not provide any help for them and we therefore decided to remove the home page from the generated profiles.

In this situation we generated new profiles in the same conditions in which we generated the previous ones ($K = 300$, maximum number of URLs = 4, minimum support = 0.2) but not admitting the home URL as a proposal. This will obviously affect the precision and recall values because having a forbidden URL reduces the highest achievable values. In order to obtain more realistic values, misses in the home page have not been counted when calculating recall. This process reduced the average number of URLs proposed per profile to 2.64, precision to 0.509 and recall and F-measure to 0.262 and 0.346, respectively. What this means is that half of the proposed URLs are used by the new users and the system is able to propose more than a quarter of the URLs that are actually used.

4.2.2. Use of the system for link prediction: evaluation

One of the uses of navigation profiles is link prediction during the access of new users, a task that has to be performed in real time. Up to this point, we have identified groups of users with similar navigation patterns and generated user profiles containing the URLs that are most likely to be visited, or the most common paths, for each of the groups. At this point we set out to use that information to improve the users' navigation experience by automatically proposing links to them during their navigation in the website.

In order to evaluate how good the generated profiles would be for link prediction we simulated the real situation by using 25% of each test sequence to select the profile of the new user according to the built model and compared the profile with the rest of the sequence (75%).

As the home page of BTw can be reached from any point in the website, it did not make any sense to predict or propose the home URL. The performance metrics were thus calculated for the option that does not admit the home URL as a proposal, where the average number of URLs proposed per profile was 2.68 and the values obtained were 0.267 for precision, 0.155 for recall, and, 0.196 for F-measure. These values showed that, even limiting the system to not proposing the home URL, our system is able to predict some of the links the new users are using and thus help users to have a more pleasant navigation experience.

Note that these results should be seen as lower bounds because, although not appearing in the user navigation sequence, the proposed links could be useful for them. Unfortunately, their usefulness/relevance could only be evaluated in a controlled experiment, by using user feedback.

5. Interest profiling

Web usage information can be combined with content information to profile users interests. In order to be able to profile the interests of users accessing BTw, we first need to link the information appearing in each URL to interests; i.e., discover the semantic structure of the website. This could be provided by the website designers, stored in an ontology or extracted automatically using Natural Language Processing (NLP) techniques. We combined this information with usage information to deduce the interests of the users accessing BTw.

5.1. Automatic extraction of the semantic structure of BTw

It is a common situation where websites do not have an underlying ontology and the communication with the service provider is not fluent. We therefore decided that the best strategy was to use NLP techniques to automatically extract the semantic structure of a website and we specifically chose topic modelling techniques.

Topic modelling is a text mining technique, a way of identifying patterns in a corpus. Several statistical models have recently been developed for automatically extracting the topical structure of large document collections, Blei (2011). For this work, we used the Latent Dirichlet Allocation (LDA) model (Blei, Ng, & Jordan, 2003). LDA allows each document to exhibit multiple topics in different proportions and can thus capture the heterogeneity in the grouped data, showing various latent patterns.

Topic modelling algorithms take as input a set of documents (a set of URL contents in our case) and provide as output a list of topics represented by the words belonging to them and a vector with the probability that each text item or document has of belonging to each of the topics.

The document-topic probability vector can be used to deduce the bias of the content of each document. This information could be used to compare documents with each other and also to group them by thematic structure.

In order to obtain the thematic structure of the website, we downloaded the HTML files of the whole website using recursive downloading. We then applied an HTML parser to obtain the content of each page and filtered the menus of the web pages so that we worked only with the real content. In order to limit our work, we only performed the analysis of the most stable part of the website; we removed the parts of the website that vary daily, such as news and agenda, due to their heterogeneity and we worked with a total of 231 URLs. Then, in order to get information about the topics hidden in the collection of URLs belonging to BTw, we used the Stanford Topic Modelling Toolbox (STMT) (Ramage & Rosen, 2009). We gave as input to STMT the dataset containing all the content information of each URL. After running STMT we obtained a list of topics represented by the keywords related to them (topic-keyword list) and a vector for each of the URLs in the database, containing the probability or affinity to each of the topics (document-topic probability vector). For BTw we obtained a URL-topic vector for each of the 231 URLs, which could be represented as:

$$UT_d = \left(A_{u_d}^{T_0}, A_{u_d}^{T_1}, \ldots, A_{u_d}^{T_N}\right), \quad (2)$$

where N represents the number of topics and UT_d represents the vector corresponding to URL_d and $A_{u_d}^{T_n}$ represents the affinity of URL_d with Topic T_n.

We performed several experiments to determine the optimum number of topics. The decision was made by analysing the coherence of the keywords proposed by the STMT tool for each of the topics and trying to find a trade-off between the number of topics and the coherence of the keywords proposed for each topic; i.e., we selected the minimum number of topics with a coherent set of keywords: 10 main topics or abstract themes. Once the STMT tool had extracted the different topics from the URL collection we named them manually, inferring a topic title based on the keywords grouped under each topic. In this way the presented results will be more readable. Table 3 shows the titles we selected for the 10 topics proposed by STMT and some of the related keywords. These are the topics used to generate profiles of users' preferences.

Slightly changing the number of topics would change the structure, but not too much. For example, if 9 topics were extracted instead of 10 the system would group the Sea& Sports and Sports topics together. Furthermore, we discussed the topics obtained with the staff of the DMO and they confirmed that the proposed structure was coherent with their aims when designing the website.

The other output that the STMT tool provides, a document-topic probability vector per URL, could be used to automatically analyse the theme of the content associated with each of the URLs. Although there could be URLs containing various topics, in this work we assigned a single topic per URL: the one with the greatest probability in the document-topic vector.

5.2. User semantic profile discovery

In this phase we combined web usage and content information to extract knowledge about the interests of the users accessing BTw. The process consisted of detecting sets of users with similar interests and extracting their semantic profiles.

Before starting the profile discovery process we discussed the information provided by the topics accessed with the staff of the DMO. Although topics related to accommodation might be very important for the tourist, they are nearer to a requirement than to a preference or interest. We therefore decided to delete them from the interest profiling phase. However, the analysis carried out on sessions accessing only accommodation-related URLs provided information of interest to the staff of the DMO: 17% of the BTw user sessions were completely devoted to accommodation. We removed these sessions from the database for discovering the semantic profiles of the users.

The web page can be accessed in four different languages (Basque, Spanish, French and English) and we used the access language as an indicator of the origin of the users. The access language will help us to differentiate between local people (those accessing the site in Basque), Spanish people or people from Spanish-speaking countries (those accessing the site in in Spanish), French people or people from French-speaking countries (those accessing the site in French) and, finally, people from the rest of the world (those

Table 3
Topic-keyword list proposed by STMT and titles for each topic.

Topic title	Topic-keyword List
Nature (Na)	mountain, river, beach, bay, ...
Historical Monuments (HM)	church, chapel, castle, history, ...
Cuisine (Cu)	cuisine, restaurant,cider-house, ...
Accommodation Camping (AC)	accommodation, pilgrim, camp, sleep, ...
Accommodation Hotel (AH)	room, accommodation, countryside, ...
Events (Ev)	festival, exhibition, theater, event, ...
Culture (Cl)	culture, organization, artist, visitor, ...
Sea & Sports (SS)	sea, surf, sport, kayak, ...
Sports (Sp)	sport, golf, tennis, pelota, ride, ...
Tradition (Tr)	tradition, celebration, typical, activity, ...

accessing the site in English). Taking this into account, we carried out two analyses of the users' interests: a global analysis and an analysis that took into account the access language. The comparison of the two analyses can provide extra information to the staff of the DMO to be used in future marketing campaigns or redesigns of the website and it can also be used to propose specific adaptations depending on the origin of the new user.

In order to perform origin-dependent analysis, we separated the examples in the database depending on the language used. We first identified the language of each link based on the *lang* parameter of the URL. We then assigned a language to each session or sequence of URLs, according to the language with the highest proportion in the sequence. Sequences or sessions with at least 70% of their navigation in the same language were labelled with that language; otherwise they were labelled as multilingual sessions. Thus, bearing in mind that the first access of the page might not be in the desired language, even in the sessions labelled with one of the languages, we allowed a certain degree of mixture of languages. After dividing the database into languages we obtained 5 sub-databases, the sizes of which are shown in Table 4. For further analysis, as we could not presuppose anything about the origin of people generating multilingual sessions, we ignored their sessions.

The data shows that, as expected, the number of accesses in every language is not the same; accesses in Spanish are more frequent than accesses in other languages. In order to avoid imbalances and to avoid obtaining results biased by the interests of the users navigating in Spanish, we obtained global profiles with a stratified sample of more or less 2650 examples for each language. To obtain the sample we randomly discarded some of the sessions in Spanish.

5.2.1. Session representation

As a first step to discover user semantic profiles we analysed the session-topic relationship and represented the sessions as session-topic vectors. As described in Section 4, the usage data has been organized such that each session is represented as a URL sequence. In addition, as described in Section 5, by applying topic modelling to the BTw content information we obtained for each URL a document-topic vector that represents its degree of affinity to each of the 10 topics extracted from the whole website. In order to model the user interests according to their navigation we combined both sets of information: we added the probabilities of the topics for every URL appearing in the session and obtained a vector representation of the user sessions: session-topic vectors.

For a session s of length L, the session-topic vector can be represented as:

$$ST_s = \left(ST_s^{T_0}, ST_s^{T_1}, \ldots, ST_s^{T_N}\right), \tag{3}$$

where $ST_s^{T_n} = \sum_{l=0}^{L} A_{u_s^l}^{T_n}$ and $A_{u_s^l}^{T_n}$ represents the affinity of URL_l of session s with Topic T_n.

Each session-topic vector represents the degree of affinity of that session to each of the topics (it has been normalized so that the sum of the vector elements is 1). Topics with higher values will denote a higher interest of the user in that topic.

5.2.2. Semantic profile generation

This stage is in charge of modelling users and producing user profiles, taking as input the session-topic vectors (i.e., a vector representation of user sessions). Unsupervised machine learning techniques have proved to be adequate to discover user profiles (Pierrakos et al., 2003). We used a crisp clustering algorithm (*K*-means) (Lloyd, 1982) to group users with similar navigation patterns and Euclidean Distance to compare two sessions. Using these techniques, we grouped users showing similar interests into the same segment.

The outcome of the clustering process was a set of groups of session-topic vectors. We used this information to deduce the probability of each topic for the cluster. Where S_j represents the number of sessions in a cluster j, the cluster-topic vector would be calculated by adding (for each topic) the topic-probability of the S_j sessions in the cluster, as can be seen in Eq. (4).

The cluster-topic vector for cluster j with S_j sessions would be:

$$CT_j = \left(CT_j^{T_0}, CT_j^{T_1}, \ldots, CT_j^{T_N}\right), \tag{4}$$

where $CT_j^{T_n} = \sum_{s=0}^{S_j} ST_s^{T_n}$.

The cluster-topic vector will allow us to identify the most significant topics for each cluster. We could thus use the titles related to these topics to label each cluster; in this work we assigned a single label per cluster. In order to consider a topic to be representative of a cluster, and thus to be able to select it as a label, we required at least 40% cluster-topic affinity; i.e., $CT_j^{T_n} \geqslant 0.4$.

5.3. Global profiling

In order to evaluate the interest profiling carried out, we compared the topic preferences of all the users calculated without clustering to the topic preferences extracted from the profiling process. The comparison was made by calculating the degrees of topic affinities in two different ways: using the whole dataset and using the output of the global clustering (the clustering performed with the whole dataset). For the first option we computed the topic distribution for all the sessions in the database. For the second option we extracted interest profiles as described above; we then grouped the sessions using the *K*-means clustering algorithm and assigned the topic with the highest affinity to each cluster, only labelling them if the affinity was greater than 40%. Hence, for clusters labelled with a topic A, we could say that all the users in the cluster are interested in topic A with its corresponding degree of affinity.

Fig. 4 shows the affinity levels obtained for the two options (whole database, WD.a; and, after the profiling, CL.a). The figure shows in the *Y* axis the eight topics selected as meaningful for the staff of the DMO and in the *X* axis the level of affinity to each of these topics. The figure includes for the profiling option the number of sessions covered by each topic; i.e., the number of users linked to each topic and what percentage of the whole database this represents.

As can be observed, the topic preference rates obtained for the whole database are very low; the one with the highest rate is around 25%. However, these values rise when profiling has been used. For example, there is a group of 464 users whose affinity to the Sea&Sports topic reaches nearly 80% and a group of 2662 users whose affinity to the Nature topic is 60%. Moreover, although some of the clusters were discarded in the CL.a option because we considered them not to be representative, 75.55% of the users are covered by those represented in Fig. 4.

If we compare the distribution of topics obtained after profiling to the distribution obtained for the whole database, we can easily infer that the use of clustering is a good option because it finds a clear structure in the data; i.e., it finds profiles with a high level of interest in specific topics and thus opens the door to future personalization strategies.

Table 4
Sizes of the databases divided by language.

English (en)	Spanish (es)	Basque (eu)	French (fr)	multilingual (mul)
2630	10198	2616	2784	4557

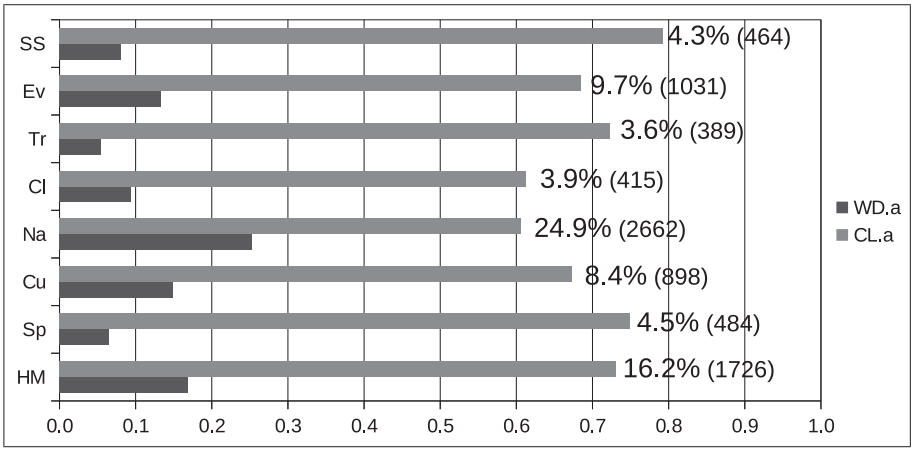

Fig. 4. Comparison of affinity levels for the whole database and the affinity levels of the profiles obtained with clustering.

5.4. Language-dependent profiling

We divided the database into four segments according to the access language and repeated the profiling phase described in the previous section for each of them. This profiling is important in order to determine the existence or otherwise of any common pattern among users with the same origin and whether or not this pattern differs from the profiles obtained for other languages. If a clear structure existed in the data for each access language the process would discover the interests of users depending on their origin, which would be interesting from two points of view: it would allow specific adaptations to be proposed depending on the origin of the new user and it would provide very useful information to the service provider for use in future marketing campaigns or redesigns of the website.

Table 5 shows the results of this experiment. The table shows, for the whole sample and for each of the languages, the popularity or number of sessions profiled as belonging to each of the topics and the average degree of affinity shown by these users. In order to make the results more visual we marked the four topics with most users interested as: 1 – (**bold + underlined**), 2 – (underlined), 3 – (***bold*** + ***italics***), 4 – (*italics*). The values in the table show that by dividing the database according to the language used our profiling methodology was able to find groups of users with high levels of affinity to specific topics. The degree of affinity is on average about 70% and the number of users covered by the significant clusters is on average 78% of the total number of examples (row Covered % int Table 5). It can thus be stated that the profiles obtained for each language are robust, in the sense that they have a high degree of affinity with one of the topics and also that they cover the majority of the users who have used the website. We cannot forget that, as shown in Fig. 4, when no clustering is applied the topic with highest affinity is Nature (with values of around 25%) and for the rest of topics the affinity values are under 15%.

Table 5 shows the results of this experiment. The table shows for the whole sample and each of the languages, the popularity or number of sessions profiled as belonging to each of the topics, and the average degree of affinity shown by these users. In order to make the results more visual we marked the 4 topics with more users interested as: The values in the table show that dividing the database according to languages, our profiling methodology was able to find groups of users with high affinity levels to specific topics. The degree of affinity is, in average, about 70%, and the number of users covered by the significant clusters is in average 78% of the total number of examples (row Covered % int Table). Thus, it can be stated that the profiles obtained for each language are robust, in the sense that they have a high degree of affinity with one of the topics, and also cover the majority of the users who have used the web. We can not forget that as shown in Figure, when no clustering is applied the topic with highest affinity is Nature with values around 25% but for the rest of topics affinity values are under 15%.

We can compare the ranking of topic preferences for the global database with the rankings in each language according either to the number of users covered or the degree of affinity obtained for each of the 8 topics. If we compare the results with those obtained with the global sample, deviations can be found for both criteria.

Table 5 shows that for the global database the most popular topics (those with the most users) were, in descending order: Nature (Na), Historical Monuments (HM), Events (Ev) and Cuisine (Cu). The two most popular topics, Nature (Na) and Historical Monu-

Table 5
Comparison of average affinity values and popularity of the different topics for global and language-dependent profiles.

Topic	Global		En		Es		Eu		Fr	
	Size	Affi.	Size	Affi.	Size	Affi.	Size	Affi.	Size	Affi.
SS	464	0.79	109	0.83	228	0.76	144	0.77	143	0.79
Ev	***1031***	0.69	*186*	0.76	*879*	0.65	*261*	0.66	***402***	0.64
Tr	389	0.72	75	0.89	194	0.77	126	0.67	120	0.69
Cl	415	0.61	98	0.68	231	0.62	74	0.72	77	0.73
Na	**2662**	0.61	**802**	0.53	**3109**	0.58	536	0.67	416	0.69
Cu	*898*	0.67	***337***	0.60	***1022***	0.61	***279***	0.62	*224*	0.68
Sp	484	0.75	164	0.67	312	0.65	137	0.78	135	0.77
HM	1726	0.73	522	0.69	1310	0. 68	**563**	0.66	**452**	0.77
Covered (%)	75.55		87.19		71.44		81.04		70.73	

ments (HM), were the same for the users connecting in English and Spanish but were reversed (being almost equal) for users connecting in Basque or French. Moreover, the topics appearing in third and fourth positions, Events (Ev) and Cuisine (Cu), only matched the global sample in the case of users accessing in French, who also show a higher interest in Events (Ev). For users accessing in the other languages the order of these two topics is reversed compared to the global data.

Fig. 5 shows the comparison of the interest profiles obtained for the whole database with those obtained for each language. In order to show the results we obtained a single value per topic by combining its affinity with its popularity. The combination was performed in each of the databases by normalizing coverage and affinity values so that the sum of all the coverage values and the sum of all the affinity values were both 1, and then calculating the average of the two values obtained for each topic. To compare the global profiles to the language-dependent ones the differences between the values obtained were calculated and these are shown in Fig. 5.

The first conclusion we can draw from Fig. 5 is that for the topic preferences, there are differences between the users accessing in different languages. The figure shows that the results for tourists accessing the site in English are not very different from the global profile but that they seem to be more interested in Cuisine (Cu) and Tradition (Tr) and less interested in Events (Ev) than the generic tourists. However, the results for Spanish tourists are very different from the generic ones, as they seem to be more interested in Nature (Na) and Cuisine (Cu) and less interested in the other topics. In the case of Basque tourists, the main interest appears to be in Historical Monuments (HM) and Sea and Sports (SS and Sp) and they seem to have very little interest in Nature (Na). Finally, French tourists seem to focus on Events (Ev) when visiting the area although they also show an interest in Historical Monuments (HM) and Sea&Sports (SS). However, their interest in nature (Na) is clearly lower than it is for the global profiles.

5.5. Expert validation

The results of the semantic profiling were presented to the BTw DMO staff, who confirmed that the semantic structure of the web obtained automatically with the STMT topic modelling tool resembles the actual structure of the website with minor variations.

The BTw DMO staff consider boating to be a very important activity that is very much in demand in their environment. This is also reflected in our results, where specific topics for Sports (Sp), Sea&Sports (SS) and Nature (Na) appeared.

The DMO staff were aware that users from different origins, and thus accessing the site in different languages, are likely to have different interests, and this was confirmed by the results we obtained. In fact, in their opinion repeating the web structure for every access language is not the right decision, since the different interests shown by users of different origin (and therefore using different languages) would require different structures.

The profiles obtained for different languages are similar to what the experts perceive in the tourist offices. British tourists tend to be very interested in tradition (Tradition Tr). The Spanish tourist generally adheres to what is being promoted by the DMO: Nature and Cuisine (Na and Cu). Access by Basque tourists generally involves local people interested in their roots and also sports activities, which seems consistent with their interest in Historical Monuments (HM) and Sea and Sports (SS and Sp). Finally, the staff affirmed that the French tourist generally comes to "faire la fête". They are very interested in festivals and this is reflected in their specific interest in the Events topic (Ev).

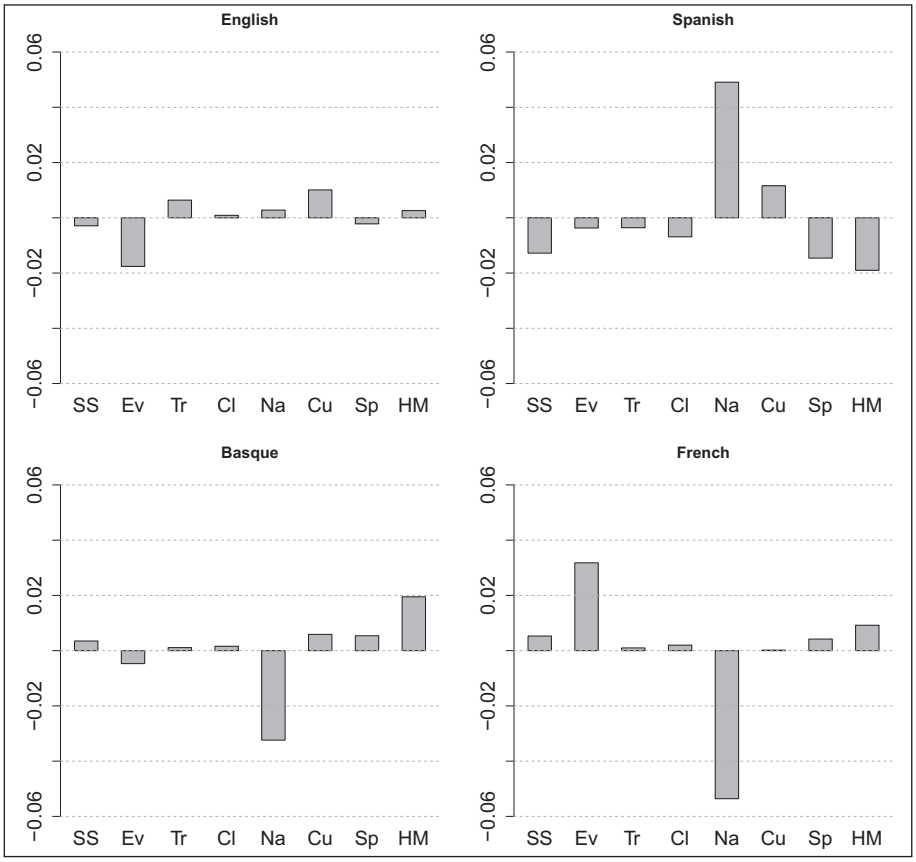

Fig. 5. Profile differences for the language-dependent profiles compared with global profiles.

In general, the experts validated the results and noted that some of the findings presented were useful for use in future web designs, specific marketing operations, etc.

6. Combining semantic structure with navigation profiles

As we mentioned in the evaluation of the generated navigation profiles, the values obtained should be seen as a lower bound, since the user could find the proposed links to be of interest even if they do not appear in the user navigation sequence. The semantic structure of the website could undoubtedly provide hints about whether or not the links proposed are of interest to the new users and it could also be used to explore more generic navigation profiles that represent the navigation of the users through different interest areas, instead of specific URLs. With this aim we combined the semantic structure of the website, specifically the main topic of each URL, with the navigation profiles obtained by the system; i.e., the sets of URLs that the user is likely to visit.

This combination will allow us to evaluate the navigation profiles according to interests and will also allow a further evaluation of the quality of the link prediction tool. It will allow us to measure to what extent the links proposed to the new user are of interest for her.

With this aim, we first used the document-topic probability vector provided by the STMT tool (see Section 5.1) to automatically analyse the theme of each of the URLs. Although there could be URLs with content for various topics, in this work we propose to assign a single topic per URL: the topic with the greatest probability in the document-topic vector. In the generated navigation profiles (see Section 4.2.1) we replaced the URLs by their main topic and calculated precision, recall and F-measure in the same way as we did in Sections 4.2.1 and 4.2.2, but using to the new profiles. Table 6 shows the results obtained and compares them to the performance metrics calculated using the URLs.

As can be observed, our intuition was correct. Precision and recall values, for example, soar up to 90.9% and 73% when evaluating profiles and up to 73.6% and 59.8% in the case of link prediction, which means that the generated navigation profiles are accurate according to interests and, moreover, that a high percentage of the links proposed by our system is of interest to the new users, which will make their browsing experience more pleasant and will help them to achieve their objectives faster.

7. Diversifying link proposals

The DMO staff suggested that although they found the outcome of the link prediction part of our system to be very satisfactory, their marketing campaigns might require diversification in the proposed links; i.e., to propose other links that might be of interest to the users even if they are not present in the profiles discovered. This could be a common practice in any website with commercial objectives.

The system we designed offers the option to diversify the link proposals by combining the profiles generated using clustering + SPADE with links proposed based on semantic information; i.e., the system would enrich profiles generated based on usage information with semantic information.

We implemented and evaluated this new option where, in order to locate the system in the same point of the learning curve, we limited the usage profiles to two URLs and enriched them with a single URL; the URL that is semantically most similar to that with the greatest support among those obtained with usage information. When the proposed URL already appeared in the profile, we selected the one with the next highest similarity. We calculated semantic similarity between URLs using the Hellinger distance (Blei & Lafferty, 2007; Deza & Deza, 2006) between the document-topic (URL-topic) vectors obtained with the STMT tool.

Table 6
Evaluation of profiles and link prediction according to interests.

Evaluation	Criteria.	N. prop	Pr	Re	F1
Profile	URL	2.89	0.642	0.344	0.448
	Topic		0.909	0.730	0.810
Link Prediction	URL	2.68	0.267	0.155	0.196
	Topic		0.736	0.598	0.660

Table 7
Evaluation of semantically-enriched link proposals according to interests.

System	N. prop	Pr	Re	F1
Usage	2.680	0.736	0.598	0.660
Usage + semantic	2.902	0.736	0.603	0.663

This variant of the system for link prediction was evaluated by calculating the same performance metrics we calculated to evaluate the previous system. Table 7 shows the values of the performance metrics when the link proposal is diversified using semantic information and when link proposals are made based only on the navigation profiles. The number of proposed URLs is in the same range for both options although it is slightly greater for semantically enriched link proposals. The values of the three performance metrics calculated are also in the same range for both cases, although they are slightly better for the new system.

We could therefore conclude that when using semantic information to diversify the links proposed to new users and introduce other products that might be of interest to the service provider, the system is still able to propose links to the new users that seem to be of great interest to them. In this case the system is achieving a double objective: it is helping the DMO staff in their campaigns and it is also providing a tool to enable the users to achieve their objectives faster and in a more pleasant way.

8. Conclusions and further work

In order to continue being successful destinations, Tourism organizations and DMOs not only have to respond by adopting new technologies but also by interpreting and using the knowledge created by Internet users and understanding how people use the channel to shop. In this context, the paper contributes with the application of a general automatic and non-invasive system to the Bidasoa Turismo website, which, taking the standard web usage information (log files from the web server) and the website content as input, automatically extracts its semantic structure and combines both information sources to extract knowledge for many applications.

User navigation profiles obtained from usage information will be useful for link prediction. These profiles can be enriched with different URLs using semantic information, with the result that the set of proposed links will be diversified. This could have a direct application for the DMOs; they can introduce links that suit the taste of the users, according to their interests. Finally, global and language-dependent user interest profiles obtained by combining usage and content information will enable the DMO staff into design future marketing campaigns for specific targets and will provide important information for future web designs. The results presented in this paper show that all the applications could be carried out successfully.

The user navigation profiles obtained were stable; the system obtained good quality profiles that match in more than 60% of cases the real user navigation sequences. When they were used for link prediction, even limiting the system by not allowing the

home URL to be proposed, nearly 30% of the proposed URLs matched the navigation of new users. Since the previous evaluation was too rigid we evaluated the link prediction capacity of our system according to user interests instead of specific URLs. The precision and recall values soar up to 90.9% and 73% when evaluating profiles and up to 73.6% and 59.8% in the case of link prediction, which means that the generated navigation profiles were accurate in terms of interests, which will make the browsing experience of the new user more pleasant and faster. Moreover, link proposals diversified using semantic information obtained similar values for precision and recall.

The BTw DMO staff corroborated that the automatically extracted semantic website structure captured the ideas behind the website. Moreover, some of the information provided by the interest profiles obtained by combining content and usage information was also validated by the BTw DMO staff. They also found the knowledge provided to be very useful for future web design and marketing campaigns.

Although we have presented a complete work, the system is not closed and many new ideas to be implemented in the future appeared during its development. In this work we used a crisp-based approach to assign interest profiles to clusters; we selected a single topic as representative. However, a fuzzy approach, where more than one topic is extracted as interest profile for each cluster, would probably be more realistic. We have developed a general system that we intend to apply to other websites. We are already working on a website for people with special requirements, where we believe the knowledge extracted with our system will be very useful for adapting the system and facilitating easier navigation in the site.

Acknowledgements

This work was funded by the University of the Basque Country, general funding for research groups (Aldapa, GIU10/02), by the Science and Education Department of the Spanish Government (ModelAccess project,TIN2010-15549), by the Basque Governments SAIOTEK program (Datacc2 project, S-PE12UN064) and by the Diputación Foral de Gipuzkoa. Finally, we would like to thank The staff in Bidasoa activa for providing the data and their support and feedback.

References

Abou-Shouk, M., Lim, W. M., & Megicks, P. (2013). Internet adoption by travel agents: a case of egypt. In *International Journal of Tourism Research, 15*, 298–213.

Anitha, A. (2010). A new web usage mining approach for next page access prediction. *International Journal of Computer Applications, 8*, 7–9.

Arbelaitz, O., Gurrutxaga, I., Lojo, A., Muguerza, J., Pérez, J., & Perona, I. (2012). Adaptation of the user navigation scheme using clustering and frequent pattern mining techniques for profiling. *4th International Conference on Knowledge Discovery and Information Retrieval (KDIR)*, 187–192.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. n. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition, 46*, 243–256.

Ballantyne, R., Hughes, K., & Ritchie, W. B. (2009). Meeting the needs of tourists: The role and function of Australian visitor information centers. *Journal of Travel and Tourism Marketing, 26*, 778–794.

Berger, H., Denk, M., Dittenbach, M., Pesenhofer, A., & Merkl, D. (2007). Photo-based user profiling for tourism recommender systems. In *Proceedings of the 8th international conference on electronic commerce and web technologies* (pp. 46–55).

Bhawsar, S., Pathak, K., & Patidar, V. (2012). Article:new framework for web access prediction. *International Journal of Computer Technology and Electronics Engineering (IJCTEE), 2*, 48–53.

Blei, D. M. (2011). Introduction to probabilistic topic models. *Communications of the ACM.*.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied Statistics, 1*, 17–35.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Brejla, P., & Gilbert, D. (2012). An exploratory use of web content analysis to understand cruise tourism services. *International Journal of Tourism Research*, http://dx.doi.org/10.1002/jtr.1910.

Brusilovsky, P., Kobsa, A., & Nejdl, W. (Eds.). (2007). *The adaptive Web: Methods and strategies of web personalization. Lecture notes in computer science* (Vol. 4321). Berlin: Springer.

Cao, L., Luo, J., Gallagher, A. C., Jin, X., Han, J., & Huang, T. S. (2010). A worldwide tourism recommendation system based on geotagged web photos. *ICASSP*, 2274–2277.

Carmona, C., Ramrez-Gallego, S., Torres, F., Bernal, E., del Jesus, M., & Garca, S. (2012). Web usage mining to improve the design of an e-commerce website: Orolivesur.com. *Expert Systems with Applications, 39*, 11243–11249.

Chaffey, D., Ellis-Chadwick, F., Johnston, K., & Mayer, R. (2006). *Internet marketing* (3rd ed.). Prentice Hall/Financial Times.

Chen, X., & Zhang, X. (2003). A popularity-based prediction model for web prefetching. *Computer, 36*, 63–70.

Chordia, B., & Adhiya, K. (2011). Grouping web access sequences using sequence alignment method. *Indian Journal of Computer Science and Engineering (IJCSE), 2*, 308–314.

CLF. (1995). Common log format (clf). The World Wide Web Consortium (W3C): http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems, 1*, 5–32.

Dasarathy, S. (1991). *Nearest neighbor (NN) norms: NN pattern classification techniques*. IEEE Computer Society Press.

De Ascaniis, S., Bischof, N., & Cantoni, L. (2013). Building destination image through online opinionated discourses. the case of swiss mountain destinations. *Enter*, 94–106.

Deza, E., & Deza, M. M. (2006). *Dictionary of distances*. Amsterdam: Elsevier.

ETC. (2012). New media trend watch – online travel market. The European Travel Commission (ETC): http://www.newmediatrendwatch.com/world-overview/91-online-travel-market?showall=1.

Fujimoto, H., Etoh, M., Kinno, A., & Akinaga, Y. (2011). Web user profiling on proxy logs and its evaluation in personalization. In X. Du, W. Fan, J. Wang, Z. Peng, & M. Sharaf (Eds.), *Web technologies and applications. Lecture notes in computer science* (vol. 6612, pp. 107–118). Berlin Heidelberg: Springer.

Garcia, P., Amandi, A., Schiaffino, S., & Campo, M. (2007). Evaluating bayesian networks precision for detecting students learning styles. *Computers and Education, 49*, 794–808.

García, E., Romero, C., Ventura, S., & Castro, C. D. (2009). An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. *User Modeling and User-Adapted Interaction, 19*, 99–132.

Godoy, D., Schiaffino, S., & Amandi, A. (2004). Interface agents personalizing web-based tasks. *Cognitive Systems Research Journal*, 207–222.

Gretzel, U. (2011). Intelligent systems in tourism: A social science perspective. *Annals of Tourism Research, 38*, 757–779.

Gretzel, U., Yuan, Y., & Fesenmaier, D. (2000). Preparing for the new economy: Advertising strategies and change in destination marketing organizations. *Journal of Travel Research, 39*, 146–156.

Gusfield, D. (1997). *Algorithms on strings, trees, and sequences – Computer science and computational biology*. New York, NY, USA: Cambridge University Press.

He, D., & Gker, A. (2000). Detecting session boundaries from web user logs. In *Proceedings of of the BCS-IRSG 22nd annual colloquium on information retrieval research* (pp. 57–66).

Hsu, C.-I., Shih, M.-L., Huang, B.-W., Lin, B.-Y., & Lin, C.-N. (2009). Predicting tourism loyalty using an integrated bayesian network mechanism. *Expert Systems with Applications, 36*, 11760–11763.

Hung, Y., Chen, K., Yang, C., & Deng, G. (2013). Web usage mining for analysing elder self-care behavior patterns. *Expert Systems with Applications, 40*, 775–783.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis* (9th ed.). Wiley-Interscience.

Kroeger, T. M., Long, D. D. E., & Mogul, J. C. (1997). Exploring the bounds of web latency reduction from caching and prefetching. In *Proceedings of the USENIX symposium on internet technologies and systems on USENIX symposium on internet technologies and systems USITS'97* (pp. 2–2). Berkeley, CA, USA: USENIX Association.

Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory, 28*, 129–137.

Luberg, A., Jrv, P., & Tammet, T. (2012). Information extraction for a tourist recommender system. *Enter*, 332–343.

Makkar, P., Gulati, P., & Sharma, A. (2010). A novel approach for predicting user behavior for improving web performance. *International Journal on Computer Science and Engineering (IJCSE), 2*, 1233–1236.

Marchiori, E., Milwood, P., & Zach, F. (2013). Drivers and benefits of analysing dmos' e wom activities. *Enter, 1*, 107–118.

Mobasher, B. B. (2006). 12 web usage mining. *Encyclopedia of Data Warehousing and Data Mining Idea Group Publishing*, 449–483.

Pan, B., & Fesenmaier, D. R. (2003). Travel information search on the internet: A preliminary analysis. In *Proceedings of the International Conference in Helsinki, Finland* (pp. 242–251).

Pierrakos, D., Paliouras, G., Papatheodorou, C., & Spyropoulos, C. D. (2003). Web usage mining as a tool for personalization: A survey. *User Modeling and User-Adapted Interaction, 13*, 311–372.

Ramage, D., & Rosen, E. (2009). Stanford topic modeling toolbox(stmt). Stanford Topic Modeling Toolbox http://nlp.stanford.edu/software/tmt/tmt-0.2/.

Sarkaleh, M. K., Mahdavi, M., & Baniardalan, M. (2012). Designin a tourism reccomender system based on location, mobile device and user features in museum. *International Journal of Managing Information Technology, 4*, 13–21.

Schiaffino, S., & Amandi, A. (2006). Polite personal agent. *IEEE Intelligent Systems, 21*, 12–19.

Schiaffino, S., & Amandi, A. (2009). Artificial intelligence. In M. Bramer (Ed.), *LNAI 5640 chapter Intelligent user profiling* (pp. 193–216). Berlin, Heidelberg: Springer.

Senkul, P., & Salin, S. (2012). Improving pattern quality in web usage mining by using semantic information. *Knowledge and Information Systems, 30*, 527–541.

Steinbauer, A., & Werthner, H. (2007). Consumer behaviour in e-tourism. In M. Sigala, L. Mich, & J. Murphy (Eds.), *Information and communication technologies in tourism, ENTER 2007. Proceedings of the international conference in Ljubljana, Slovenia* (pp. 65–76). Springer.

Turban, E., & Gehrke, D. (2000). Determinants of e-commerce website. *Human Systems Management, 19*, 111–120.

Villaverde, J., Godoy, D., & Amandi, A. (2006). Learning styles' recognition in e-learning environments with feed-forward neural networks. *Journal of Computer Assisted Learning, 22*, 197–206.

Yannibelli, V., Godoy, D., & Amandi, A. (2006). A genetic algorithm approach to recognize students learning styles. *Interactive Learning Environments, 14*, 55–78.

Zaki, M. J. (2001). Spade: An efficient algorithm for mining frequent sequences. *Machine Learning, 42*, 31–60.

Zanker, M., Fuchs, M., Hpken, W., Tuta, M., & Mller, N. (2008). Evaluating recommender systems in tourism – a case study from austria. *Enter*, 24–34.

Zukerman, I., Albrecht, D., & Nicholson, A. E. (1999). Predicting users' requests on the www. In *Procedings of the seventh international conference on user Modeling (UM99)* (pp. 275–284).