



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.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

For tourists, the decision on which destination or product to choose requires a considerable time and effort (De Ascaniis, Bisc-hof, & 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).

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, Jr, & 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, & Miller, 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

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات