A Navigation-log based Web Mining Application to Profile the Interests of Users Accessing the Web of Bidasoa Turismo

Olatz Arbelaitz, Ibai Gurrutxaga, Aizea Lojo, Javier Muguerza, Jesús M. Pérez and Iñigo Perona

Department of Computer Architecture and Technology
University of the Basque Country UPV/EHU, Donostia, Spain
{olatz.arbelaitz, i.gurrutxaga, aizea.lojo, j.muguerza, txus.perez, inigo.perona}@ehu.es

Abstract

Websites are important tools for tourism destinations. The information acquired from the use of tourism websites can be very useful for the travel agents. It will provide insight about the users' preferences, requirements and habits that are very useful for marketing campaigns or website redesign. Using machine learning techniques to build user profiles allows taking into account their real preferences. This paper presents a navigation-log based web application to profile users accessing the web of Bidasoa Turismo. The profiles are built based on the combination of web usage information stored in web log files and web content information to obtain more comprehensible profiles. The experiments show that we are able to find specific user profiles.

Keywords: User Profiling, Web Usage Mining, Web Content Mining, Semantics.

1 Introduction

Intelligent systems in the tourism sector are being studied recently (Gretzel, 2001). They are next generation information systems that might provide tourism consumers and service providers with the most relevant information, more decision support, greater mobility and the most enjoyable travel experiences. There is currently a wide range of technologies related to them such as recommender systems, context-aware systems, web mining tools, etc. Moreover, travel agents are among the service providers whom Internet adaption could be the best marketing device for their business and a tool for their competitive advantages (Abou-Shouk, 2012).

This paper presents an approach to discover the interests of the people navigating in the web of Bidasoa Turismo according to their browsing habits. Our research is contextualized in the use of web mining (Mobasher, 2006) to build user profiles. The first step of this project is to analyze the navigation of users (web usage mining), combine it with information extracted form the content of the URLs (web content mining) and generate semantic user profiles. Those profiles will be useful in the future to propose web adaptations on the one hand, and, on the other hand, they will be a good marketing device for service providers.

The paper tries to answer the next research question: are there different types of users in Bidasoa Turismo website according to their interests or are the interests of most users similar? In other words, is the clustering able to find structure in the web data?

The article summarizes in Section 2 the data acquisition environment. Section 3 is devoted to describing the characteristics of the system we have developed. Then, Section 4 presents some of the results obtained in the performed experiments. Finally, we summarize in Section 5 the conclusions and future work.

2 Application Environment: Bidasoa Turismo

In this work we have used a database from our environment: Bidasoa-Txingudi bay, which is located at the western tip of the Pyrenees mountains and, straddling two countries, France and Spain, links the Basque provinces of Gipuzkoa and Lapurdi. The area offers the opportunity of a wide range of tourism activities and, Bidasoa Turismo website (BTw), *www.bidasoaturismo.com*, includes all sorts of practical tourist information to visit the area: thematic tourism, professional tourism, gourmet tourism, agenda, suggestions, etc. We acquired nearly four months of usage data of BTw: from January 9, 2012 to April 30, 2012. The information contained in this database belongs to web server logs of requests (a total of 897,301) stored in common log format (W3C). Moreover, we also used the content information of the website, i.e., the text appearing in the website.

3 Proposed system

The work presented in this paper introduces a web mining application that combines web usage mining (Mobasher, 2006) and web content mining (Srivastava, 2005) strategies. That is, the usage information is combined with the semantic analysis of the content information, so that semantic profiles of the users can be obtained.

3.1 Data Acquisition and Preprocessing

We acquired two types of data. On the one hand, we acquired usage information, and, on the other hand, we acquired the content information for BTw.

Usage Data. Nearly 4 months of usage data were collected from BTw. First of all, we preprocessed the data so that further uses of the same URL were identified in the same way and removed erroneous requests. The only requests taken into account for our experiments are the ones related directly to user clicks. Then we preprocessed the log files to obtain information from different users and sessions. Among the obtained sessions, we selected the most relevant ones; the ones with higher activity level (3 or more clicks) and removed the outliers, i.e., the ones with more than 55 requests (out of 98% percentile). After the whole preprocessing phase the database contains 55,454 user requests divided in 8,678 sessions, with an average length of 5.8 requests.

Content Data. To acquire content data we downloaded the HTML files of the whole web site. We then applied an HTML parser to obtain the content of each page and

filtered the menus of the web pages so that in further steps we work only with the real content. In order to limit our work, we only performed the analysis of the static 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.

3.2 Extraction of Semantic Information from Content

Several statistical models have recently been developed for automatically extracting the topical structure of large document collections (Blei, 2011). For this work, we used latent Dirichlet allocation (LDA) (Blei, 2003) model. LDA allows each document to exhibit multiple topics with different proportions, and therefore can capture the heterogeneity in the grouped data showing various patterns latent. We used the Stanford Topic Modeling Toolbox (STMT) in order to get information about the topics hidden in the collection of URLs belonging to BTw. We gave as input to STMT a dataset containing all the content information about each URL. After running STMT we obtained a list of topics where each topic is represented by the keywords related to it (topic-keyword list) and a matrix containing the probabilities of each topic in each of the URLs in the database (document-topic probability matrix). Some examples of the list of words assigned to each topic could be: Nature for mountain, river, beach, bay, etc., Historical monuments for church, chapel, castle, history, and Cuisine for cuisine, restaurant, cider-house, etc.

After several experiments to test different values and analyzing the coherence of the keywords related to each of the topics proposed by the STMT tool, we have determined that the website has 10 main topics or abstract themes. Once the STMT tool has extracted the different topics from the URL collection, we have named them manually, inferring a title for the topic based on the keywords grouped under each topic. This way the presented results will be more readable. The titles we selected for the 10 topics proposed by STMT are: Nature, Historical monuments (HistMon), Cuisine, Accommodation Camping (AccCamp), Accommodation Hotel (AccHotel), Events, Culture, Sea and Sports (Sea&Spo), Sports and Tradition.

3.3 Combining Usage and Content Information for Session Representation

The aim of this work is to use web usage and content information to detect sets of users with similar interests and to use them to obtain semantic profiles. First of all, we represented the information corresponding to each of the sessions as a clickstream or sequence of clicks preformed in the URLs of BTw. To build semantic profiles, we decided to analyze the session-topic relationship and obtain session-topic vectors. We obtained a vector representation for each session with mainly semantic information: the affinity of the user to each of the topics by adding the probabilities of the topics for every URL appearing in the session. We normalized the values between 0 and 1. Topics with higher affinity will denote higher interest of the user in that topic.

3.4 Pattern Discovery and Analysis

This stage is in charge of modeling users and producing user profiles taking as input the vector representation of user sessions. Unsupervised machine learning techniques have shown to be adequate to discover user profiles (Pierrakos, 2003). We have used

a crisp clustering algorithm *k-means* (Lloyd, 1982) to group users with similar navigation patterns and Euclidean Distance to compare two sessions. Using those techniques, we grouped into the same segment users that show similar interests.

The outcome of the clustering process is a set of groups of session-topic vectors and we used this information to deduce the probability of each topic for the cluster. For doing this work, we repeated the procedure used for representing sessions: we added the probability of the topics of every session in the cluster. We normalized the results in order to obtain probability values. After the calculation, we could identify the most significant topics for each cluster. Moreover, we could use the titles related to those topics to label each cluster. We required at least 40% of cluster-topic affinity for considering the topic a leader topic of the cluster and, as a consequence, for selecting it as a label. In this first approach, we assigned a single label for each cluster.

4 Experiments: results and analysis

In the following section we will describe the experiments performed to try to answer the research question posed in the introduction. We analyzed if the clustering process is able to find structure in the database or not. That is, we analyzed if using clustering is worth it to profile the different types of users of BTw or extracting the general distribution of affinities in different topics is enough.

Fig. 1 shows the comparison of degrees of topic affinities calculated in two different ways: taking into account the whole data set (BL) and using the output of the clustering (CL). The figure also includes for the case when clustering has been used, the number of sessions covered by that topic, that is, the amount of users linked to each topic. For the first option we computed the topic distribution taking into account all the sessions in the database. As it can be observed in the figure, the obtained topic preference rates are very low; the one with the highest rate does not reach 20%. On the other hand, we grouped the sessions using *k-means* clustering algorithm and assigned the topic with higher affinity to each cluster. We only selected the representative clusters. That is, when the affinity value of the topic with highest affinity was smaller than 40% we considered the cluster not specialized enough and we did not take it into account. Hence, for clusters labeled with the main topic, we are able to say that the N users in the cluster are interested in topic A with its corresponding degree of affinity.

The values in Fig. 1 (CL) are average affinity values of the topics that are representative for the clusters obtained. This means, for example, that there is a group of 120 users whose affinity to Sea&Spo topic reaches nearly a 90%, or, a group of 68 users whose affinity to Culture topic is of 80%. Moreover, although some of the clusters were discarded in CL option, 83.19% of the users are covered by the ones represented in Fig. 1. If we compare the distribution of topics obtained taking into account groups of users to the ones 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, that is, it finds profiles highly interested in concrete topics and, as a consequence, it opens the doors to future personalization strategies.

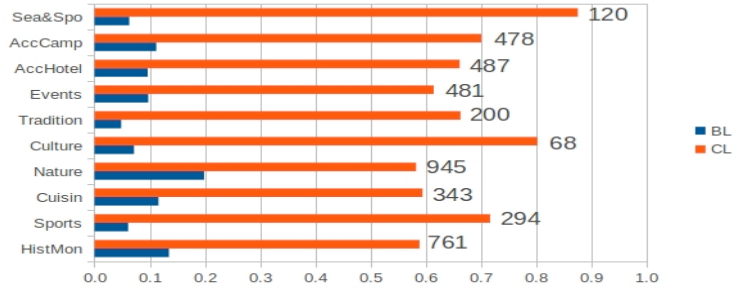

Fig. 1. Topic affinities for the whole DB (BL) vs topic affinities after clustering (CL).

5 Conclusions and Further Work

The work presented in this paper introduces a navigation-log based web mining application to profile users from different origins that combines two of the web mining approaches: web usage mining and web content mining. The application is divided into many steps: data acquisition and preprocessing, extraction of semantic information from the content based on topic modeling, combining the usage and the content information and clustering based pattern discovery and analysis. The obtained results show that the users accessing BTw have different types of interests and, the clustering process is able to find that structure in the database; it finds profiles of users highly interested in concrete topics.

In the future more specific profiles such as origin dependent profiles can be analyzed, the obtained profiles can be used to adapt the web page to the requirements of the different types of tourists and the extracted knowledge should be shared with the service providers.

References

Abou-Shouk, M., Lim, W. M. & Megicks, P.:. (2012). Internet Adoption by Travel Agents: a Case of Egypt. *In International Journal of Tourism Research* .

Blei, D. M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993--1022.

Blei, D. M. (2011). Introduction to Probabilistic Topic Models. *Communications of the ACM*.

Gretzel, U. (2001). Intelligent systems in tourism: A Social Science Perspective. *In Annals of Tourism Research, 38*, 757--779.

Lloyd, S. P. (1982). Least squares quantization in PCM. Technical Note, Bell Laboratories. *In IEEE Transactions on Information Theory, 28*, 129--137.

Mobasher, B. (2006). 12 Web Usage Mining. In Encyclopedia of Data Warehousing and Data Mining Idea Group Publishing. *Liu, B. (ed.) Springer Berlin Heidelberg*, 449--483.

Pierrakos, D., Paliouras, G., Papatheodorou, C. & Spyropoulos, C. D.: (2003). Web Usage Mining as a Tool for Personalization: A Survey. *In User Modeling and User-Adapted Interaction, 13*, 311--372.

Srivastava, T., Desikan, P. & Kumar, V. (2005). Web Mining -- Concepts, Applications and Research Directions. *In Foundations and Advances in Data Mining*, 275--307.

STMT. Stanford Topic Modeling Toolbox: http://nlp.stanford.edu/software/tmt/tmt-0.2/

W3C. The World Wide Web Consortium. The Common Log Format: http://www.w3.org/Daemon/User/Config/Logging.html\#common-logfile-format