Enhancing a Web Usage Mining based Tourism Website Adaptation with Content Information

Olatz Arbelaitz, Ibai Gurrutxaga, Aizea Lojo, Javier Muguerza, Jesús M. Pérez and Iñigo Perona
Dept. of Computer Architecture and Technology, University of Basque Country UPV-EHU, M. Lardizabal, 1, 20018 Donostia, Spain

Keywords: Adaptive Web, Link Prediction, User Profile, Collaborative Filtering, Machine Learning, Web Usage Mining, Web Content Mining, Semantics.

Abstract: Websites are important tools for tourism destinations. The adaptation of the websites to the users' preferences and requirements will turn the websites into more effective tools. Using machine learning techniques to build user profiles allows us to take into account their real preferences. This paper presents the first approach of a system that, based on a collaborative filtering approach, adapts a tourism website to improve the browsing experience of the users: it generates automatically interesting links for new users. In this work we first build a system based just on the usage information stored in web log files (common log format) and then combine it with the web content information to improve the performance of the system. The use of content information not only improves the results but it also offers very useful information about the users' interests to travel agents.

1 INTRODUCTION

Intelligent systems in the tourism sector are being studied recently (Gretzel, 2011). Intelligent systems are next generation information systems that might provide tourism consumers and service providers with the most relevant information, more decision support, greater mobility, and finally, the most enjoyable travel experiences. There is currently a wide range of technologies rellated to them such as recommender systems, context-aware systems, autonomous search agents, web mining tools, etc. Creating these systems requires a thorough knowledge of tourists' psychology, social structures where tourism is experienced, the ratio of tourists that use technology, the structure of the tourism industry, the language of tourism, etc. Moreover, travel agents are among service providers whom their internet adaption could be the best marketing device for their business and a tool for their competitive advantages (Abou-Shouk et al., 2012).

In this context the adaptation of tourism websites to the user requirements becomes specially important. That is, web personalization becomes essential. Web personalization (Pierrakos et al., 2003) can be defined as the set of actions that are useful to dynamically adapt the presentation, the navigation scheme and/or web content, based on preferences, abilities, or user requirements. Web personalization in tourism can positively affect both the feeling of the user and the business.

This paper presents a preliminary approach to adapt a tourism web page, *www.bidasoaturismo.com*, according to the browsing preferences of the users. The proposed adaptation is to automatically generate links to the users while they are navigating so that their objective is reached more easily. Furthermore, the system will provide useful information about the tourists to the service providers.

Adaptations of the web environments to specific users in navigation time require a previous phase of generating user profiles containing the most important facts about their navigation preferences in this case. The most widely used method for obtaining information about users is observing their actions (Schiaffino and Amandi, 2009). In adaptive systems, the user profile is used to behave differently for different users.

Our research is contextualized in the use of web mining (Mobasher, 2006) to build user profiles and then propose adaptations to the website based on the obtained profiles and extract semantic information from them. We could define web mining as the application of machine learning techniques to data from the Internet. This process requires a data acquisition and pre-processing stage which is not easy because it

Arbelaitz O., Gurrutxaga I., Lojo A., Muguerza J., M. Pérez J. and Perona I..
Enhancing a Web Usage Mining based Tourism Website Adaptation with Content Information.
DOI: 10.5220/0004171002870292
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2012), pages 287-292
ISBN: 978-989-8565-29-7

requires several steps such as cleaning and merging data from multiple log files, user identification, session identification, completion of route etc. 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. And finally, the patterns detected in the previous steps are used in the operational phase to adapt the system and make navigation more comfortable for new users. Our claim is that semantic information can improve the quality of profiles based on web usage information.

We have built a system based on the collaborative filtering approach that takes the content of the web and the minimum information stored in a web server as input: server log files stored in web Common Log Format (CLF) (W3C, 1995), i.e., the profiles are constructed without perturbing the user.

This paper uses our previous experience in automatically generating links in a local tourism web page and improves its behavior and gives it added value introducing semantic analysis of the content in the user profile. Although this is a preliminary work we claim that the combination of both usage and semantics can lead to more accurate and richer recommendations and moreover it gives to the travel agents greater insight about the real interests of the tourists.

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 DATA

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 Bidasoa River has had the effect of linking socially and culturally the three towns surrounding the bay: Hendaye, Hondarribia and Irun. 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. Our work will make more pleasant and effective the navigation of the user and, as a consequence, it will also contribute to a more enjoyable travel experience for the tourist. Moreover the generated semantic profiles will be a good marketing device for service providers. 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, 1995). Furthermore, we also use the content information of the website, i.e., the text appearing in the website.

3 PROPOSED SYSTEM

The work presented in this paper enhances the performance of a web usage mining (Srivastava et al., 2005) application including a semantic analysis of the content information. As every web usage mining process, it can be divided into three main steps: data acquisition and preprocessing (Cooley et al., 1999), pattern discovery and analysis, and, exploitation.

3.1 Data Acquisition and Preprocessing

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

For the first type, nearly 4 months of usage data were collected from BTw and first of all, we preprocessed the used URLs so that further uses of the same URL were identified in the same way and reduced all agenda accesses to a single agenda page. We preprocessed the log files to obtain information from different users and sessions. Before identifying user sessions, we filtered erroneous requests, image requests, etc. so that the only requests taken into account for our experiments are the ones related to user clicks (the amount of requests was reduced nearly in a 50%: to 470,402 requests).

We performed the user identification process based on IP addresses and as an heuristic to identify sessions within a users' activity, we fixed the expire time of each session to 10 minutes of inactivity (He and Göker, 2000). Among the obtained sessions, we selected the most relevant ones; the ones with higher activity level (3 or more clicks), removed the outliers, i.e., the ones with more than 55 requests (out of 98% percentile), and, grouped consecutive requests of agenda in a single request because we detected that many of them are automatically generated by the server. After the whole preprocessing phase the database contains 55,454 user requests divided in

9,549 sessions, with an average length of 5.8 requests, where a total of 308 different URLs are visited.

To acquire content data we used the GNU Wget (GNU, 1996) computer program to retrieve content from the BTw web server. We downloaded the HTML files of the whole web site 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 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 having a total of 231 URLs. Note that there are some URLs, mainly private URLs related to web administration not accessible for a normal user, appearing in the usage information that will not have their equivalent in the content part.

3.2 Session Representation

Being the aim of this work to detect sets of users with similar navigation patterns and to use them to make the navigation of future users easier, and obtain semantic profiles, we represented the information corresponding to each of the sessions as a clickstream or sequence of clicks preformed in the URLs of BTw.

3.3 Pattern Discovery and Analysis

This is the stage that, taking as input the user click sequences, is in charge of modeling users and producing user profiles. Unsupervised machine learning techniques have shown to be adequate to discover user profiles (Pierrakos et al., 2003). We have used a crisp clustering algorithm to group users that show similar navigation patterns.

3.3.1 Clustering

We used *PAM (Partitioning Around Medoids)* (Kaufman and Rousseeuw, 1990) clustering algorithm and a Sequence Alignment Method, Edit Distance (Gusfield, 1997)(Chordia and Adhiya, 2011) as a metric to compare sequences to group into the same segment users that show similar navigation patterns. Although further analysis should be done, as a first approach and based on the analysis of the distribution of the different URLs in the sessions, we instantiated the maximum number of clusters, *K* parameter of *PAM* algorithm, to 50.

3.3.2 Profile Generation

The outcome of the clustering process is a set of groups of user sessions that show similar behavior. But we intend to model those users or to discover the associated navigation patterns or profiles for each one of the discovered groups. That is, to find the common click sequences appearing among the sessions in a cluster. We used SPADE (Sequential PAttern Discovery using Equivalence classes) (Zaki, 2001), an efficient algorithm for mining frequent sequences, to extract the most common click sequences of the cluster. In order to build the profiles of each cluster using SPADE, we mapped 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 for the sessions belonging to it. SPADE parameters such as minimum support and maximum allowed number of sequences per cluster regulate the system so that it finds an adequate number of URLs to propose and a balance between the precision and the recall of the system is achieved. We used the SPADE parameters to propose a similar amount of URLs per cluster in the two approaches, the one without semantic information and the other using semantic information.

3.3.3 Enriching the Profiles with Semantics

The profiles have been generated up to this point using only usage information but we propose to enrich the profiles with semantic information to improve performance. Obviously the navigation pattern of the users depends on their interests, and, as a consequence, URLs with similar or related content to the ones appearing in the user profile will also be interesting for the user. We have used two types of tools for finding similarity between URL contents: MG4J (Boldi and Vigna, 2006) search engine and a keyword extraction based approach (KYWD). MG4J is a full text indexer for large collection of documents written in Java, developed at the University of Milano. We used this tool with TfIdf distance to obtain similarity values between every possible pair of URLs in the website. This gives us the chance to obtain for each URL a list of URLs ordered by semantic similarity.

On the other hand, we have used Yahoo Term Extractor tool (Yahoo!, 2011) to extract keywords and compare semantic contents of different URLs. Keywords and key phrases (multi-word units) are widely used in large document collections. They describe the content of single documents and provide a kind of semantic meta data that is useful for a wide variety of purposes. Once the keywords of each URL have been extracted, we have used the cosine similarity distance (Madylova and gduc, 2009) to compare URLs.

$$similarity(URL_k, URL_l) = \frac{\sum_{i=1}^{n} w_{ki} \times w_{li}}{\sqrt{\sum_{i=1}^{n} (w_{ki})^2} \times \sqrt{\sum_{i=1}^{n} (w_{li})^2}} \quad (1)$$

For calculating the weight of keyword i in each document (URL_k), we used:

$$w_{ki} = (tf/tfmax) * log(N/n) \quad (2)$$

where tf is the keywords' frequency in URL_k; $tfmax$ is the maximum term frequency in URL_k; N is the number of URLs and n is the number of documents containing keyword i. The URLs with larger similarity value will be the semantically more similar ones.

The two previous approaches have been used in the same way to enrich usage information based user profiles: we added to the profiles generated using SPADE two extra links, the most semantically similar ones, for each proposed link. In the case those URLs already appeared in the profile we have not taken them into account.

Furthermore, the KYWD gives us the option to extract semantic information from the obtained profiles. With this aim we have analyzed which are the most important (frequent) keywords in each of the generated profiles.

3.4 Exploitation

This is the part that needs to be done in real time. Up to now, we have identified groups of users with similar navigation patterns and we have generated user profiles or most common paths for each of the groups. At this point we need to use that information to automatically propose links to new users navigating in the web. We propose the use of k-Nearest Neighbor (Dasarathy, 1991) learning approach to calculate the distance of the click sequence (average linkage distance based on Edit distance (Gusfield, 1997)) of the new users to the clusters generated in the previous phase.

Our hypothesis is that the navigation pattern of that user will be similar to the user profile of its nearest cluster. As a consequence the system will propose to the new user the set of links that models the users in the cluster.

4 EXPERIMENTS: RESULTS AND ANALYSIS

4.1 Experimental Setup

In order to evaluate the performance of the whole process, the best choice would be to carry out an experiment in a real environment. This real experiment should be carried out in a conversational manner with users and this interaction should provide the feedback required to validate our system following the guidelines of Conversational Case Base Reasoning (CCBR) (Aha et al., 2001). But in this work, as a first approach we carried out a standard validation method. In this case, we applied the hold-out method dividing the database into two parts. One for generating the clusters and extracting user profiles, and, another one for testing or using it in exploitation. To simulate a real situation we based the division of the database on temporal criteria: we used the oldest examples (66% of the database, 6366 user sessions) for training and the latest ones (33%, 3183 user sessions), for testing.

We validated the system in two different situations: when no content information is used (SP) and when profiles are enriched with semantic information extracted from content. For the latter, we have used two tools to compare URLs: MG4J and the Keyword based comparison (KYWD).

We validated the system from two points of view the user point of view and the service provider point of view. For the first part we used the test examples as described in the exploitation section and then we compared the automatically generated links with the real click sequences of the users. For the latter, we analyzed the description based on the most frequent keywords for each cluster.

We performed the evaluation taking into account that in real executions, when a user starts navigating, only its first few clicks will be available to be used for deciding the corresponding profile and proposing new links according to it. We have simulated this real situation using 10%, 25% and 50% of the user navigation sequence in the test examples to select the nearest cluster or profile.

We computed statistics based on results for each one of the new users. We compared the number of proposed links that are really used in the test examples (hits), the number of proposals that are not used (misses), and the amount of links used by the test users. An ideal system would maintain precision and recall as high as possible. But, in order to compare the two options, we keep the amount of proposed links not too high (around 50% of the average sequence length) and focus on recall because it gives us an idea of the achieved coverage, i.e., the amount of links really used in the test examples that our system proposes. In order to ensure that precision values do not suffer a sudden drop we present values for two statistics: recall (Re) and F-measure (Fm). Note that the obtained values could be seen as a lower bound because, although not appearing in the user navigation sequence, the proposed links could be useful and interesting for her/him. Unluckily their usefulness could only be evaluated in a controlled experiment us-

ing the user feedback. Besides, since some links appearing in the usage analysis (mainly those related to web adminstration issues) do not appear in the stored content information, recall and F-measure values are limited (could never be 100%).

We calculated two values for the used statistics: an upper bound (ReUp, FmUp) that takes into consideration the whole test sequence, and the values calculated using only the clicks in the test sequence that have not been used to select the nearest profile (Re, Fm); that is, taking into account the remaining 90%, 75% or 50% (for the cases 10%, 25% and 50% respectively).

4.2 Results and Analysis

A first analysis of the results shows that most of the users start navigating from the initial page of the website and visit the agenda. Although this is interesting information for the travel agent, the proposal of those two URLs as part of the profiles inflates the values of the calculated statistics, and, as a consequence we have removed them form the generated profiles.

The first conclusion we can draw from the results is that even if the values of the measured metrics vary depending on the selected option, all of them are able to predict a certain percentage of the links a new user will be visiting.

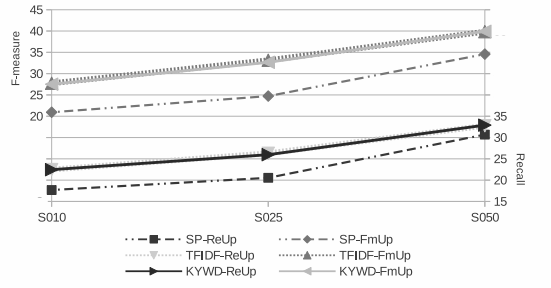

Figure 1: SP vs Semantics Enriching. Upper bound.

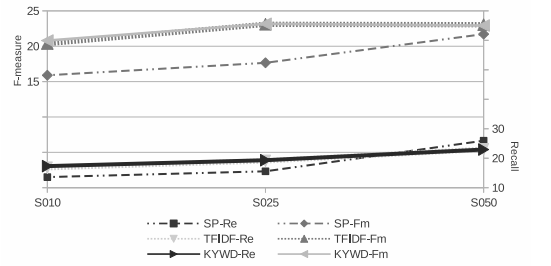

Figure 2: SP vs Semantics Enriching.

We present in figures 1 and 2 recall and F-measure values for the usage based system (dashed lines) and the system combining usage and content information (continued lines). We adjusted the parameters so that both system proposed a similar amount of links to the new users: 3.5 in average for SP option and 3.8 in average for MG4J and KWYD options. Note that when larger the proposed amount of links is, smaller is the support of some of them, so the system is risking more and, as a consequence, a drop in the F-measure value is very probable as a consequence of a drop in precision.

Table 1: Semantics of profiles.

Topic	Keywords
Mountain	cycle trails, mountain paths, path cycle
Sea	bay, river, ocean bay, river boat trips, natural treasures
Accommodation	hotel, accommodation, youth hostels
Cuisine	sugar, eggs, recipes typical product, innovative cuisine
History	walled city, borda, palace
Events	events activities, markets fairs, rural sports

Graphics in figures 1 and 2 show that when enriching the system with content information recall and the F-measure values increase, that is, the system guesses more links among the ones really used by the users clicks (Re). The improvement is larger in the case of the upper bound but even in the real case content information seems to be important. On the other hand, this improvement is more evident at early stages of the navigation when the usage information is very limited. Those are the moments when the prediction can probably contribute more to the user navigation experience becoming more satisfactory.

Finally, if we analyze the semantics of the generated profiles based on the extracted keywords, we realize that most of the clusters group users with similar interests. Table 1 shows an example of the semantics of some of the clusters and some of its related keywords. The names of the topics appearing in the left hand column in Table 1 have been assigned manually whereas the keywords on the right hand column have been obtained automatically. The results in the table clearly show that the users clustered in different groups based on their usage patterns, besides navigating in a different way, have clearly different interests and our system is able to extract information about them. Calculating statistics of the amount of users in each cluster the service providers could obtain very useful information about the main interests of the people accessing BTw and use it in the future for marketing campaigns or modifications in the website.

5 CONCLUSIONS

We designed a system that, without disturbing the users, based just on server log information, content information and machine learning techniques, identifies different groups of users, builds the corresponding

profiles, automatically generates useful link proposals for new users, and moreover, it gives insight about the users' preferences to the tourism agents. This work has been done for Bidasoa Turismo, a tourism website in our environment, but it could be extended to any other environment since it uses the minimum information stored in any web server (in common log format). We preprocessed the data, prepared it so that it could be used with machine learning algorithms, we divided the database into two parts training and test, applied *PAM* to the training data to discover groups of users with similar navigation patterns and SPADE to discover the profiles associated to each of the clusters. We further enriched those profiles adding semantic information using two options MG4J and KYWD. In the exploitation phase we related each test example to just one of the built profiles (1-NN).

We evaluated, based on a a hold-out strategy, different configurations of the system and how it performs at different stages of the user navigation: 10%, 25% and 50%. We calculated recall and F-measure statistics and analyzed the semantic profiles.

Results showed that the use of the semantic knowledge extracted from the website content information improves the performance, recall and F-measure values, of the system proposed, and, moreover, this improvement is greater at early stages of the navigation so, the system deals better with the zero day or cold start problem. Furthermore, using content information gives the option to enrich the generated profiles with semantic information that can be very useful for service providers.

This work opens the door to many future tasks. A deeper analysis of the differences of the two options implemented for URL content comparison MG4J and KYWD should be done. Moreover, the Topic Modeling option could be another option to extract semantic knowledge from the websites' content. More sophisticated strategies to build semantic could also be explored.

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 (TIN2010-15549 project); by the Diputación Foral de Gipuzkoa, Zer4You (DG10/5); and by the Basque Government's SAIOTEK program, Datacc (S-PE11UN097).

REFERENCES

Abou-Shouk, M., Lim, W. M., and Megicks, P. (2012). Internet adoption by travel agents: a case of egypt. *International Journal of Tourism Research*, pages n/a–n/a.

Aha, D. W., Breslow, L., and Muñoz-Avila, H. (2001). Conversational case-based reasoning. *Appl. Intell.*, 14(1):9–32.

Boldi, P. and Vigna, S. (2006). Mg4j at trec 2006. In Voorhees, E. M. and Buckland, L. P., editors, *TREC*, volume Special Publication 500-272. National Institute of Standards and Technology (NIST).

Chordia, B. S. and Adhiya, K. P. (2011). Grouping web access sequences using sequence alignment method. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(3):308–314.

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

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

GNU (1996). Gnu wget.

Gretzel, U. (2011). Intelligent systems in tourism: A social science perspective. *Annals of Tourism Research*, 38(3):757–779.

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

He, D. and Göker, A. (2000). Detecting session boundaries from web user logs. *Proceedings of the 22nd Annual Colloquium on Information Retrieval Research*.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data An Introduction to Cluster Analysis*. Wiley Interscience, New York.

Madylova, A. and gduc, S. G. (2009). A taxonomy based semantic similarity of documents using the cosine measure. In *ISCIS*, pages 129–134. IEEE.

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

Pierrakos, D., Paliouras, G., Papatheodorou, C., and Spyropoulos, C. D. (2003). Web usage mining as a tool for personalization: A survey. *User Modeling and User-Adapted Interaction*, 13(4):311–372.

Schiaffino, S. and Amandi, A. (2009). Artificial intelligence. chapter Intelligent user profiling, pages 193–216. Springer-Verlag, Berlin, Heidelberg.

Srivastava, T., Desikan, P., and Kumar, V. (2005). Web mining – concepts, applications and research directions. pages 275–307.

W3C (1995). The world wide web consortium: The common log format.

Yahoo! (June 15 2011). Term extraction documentation for yahoo! search.

Zaki, J. M. (2001). Spade: An efficient algorithm for mining frequent sequences. *Mach. Learn.*, 42(1-2):31–60.