Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 51–58
DOI: 10.15439/2014F464
ACSIS, Vol. 2

Global versus modular link prediction approach for discapnet: website focused to visually impaired people

O. Arbelaitz, A. Lojo, J. Muguerza, I. Perona
University of The Basque Country
Department of Computer Architecture and Technology
Donostia-San Sebastian, 20018, Spain
Email: {olatz.arbelaitz, aizea.lojo, j.muguerza, inigo.perona}@ehu.es.

***Abstract*—Web personalization becomes essential in industries and specially for the case of users with special needs such as visually impaired people. Adaptation may very much speed up the navigation of visually impaired people and contribute to diminish the existing technological gap. This work is the first stage of a web mining process carried out in discapnet: a website created to promote the social and work integration of people with disabilities where slow navigation has been detected. Based on observation in-use where behaviours emerge applying a web mining process to server log data, we designed a system to generate user navigation profiles and adapt to the web site through link prediction. Two approaches for user profiling were implemented: a global system built based on the complete database and a modular approach carried out discovering the navigation profiles within different zones. Although both approaches are effective, the modular approach outperforms. When 25% of the navigation of the new user has happened the designed system is able to propose a set of links where nearly 60% of them (2 out of 3) is among the ones the new user will be using in the future. This will definitely make the navigation easier saving a lot of time.**

I. Introduction

THE success of electronic commerce, especially for the less well-known companies, is largely dependent on the appropriate design of their website [1]. Chaffey et al. [2] stated in their work that a good website should begin with the users and understanding how they use the channel. 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. Organizations have to respond not only by adopting new technologies, but also by interpreting and using the knowledge created by Internet users.

In the last decades, the trends have 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 and specially for the case of users with special needs such as visually impaired people.

However, little is known about the navigation tactics employed by screen reader users when they face problematic situations on the Web. Modelling the navigation of users is of utmost importance as it allows not only to predict interactive behaviour, but also to assess the appropriateness of the content in a link, the information architecture of a site and the design of a web page [3].

Navigating through audio web interfaces is a challenging task mainly because content is serially rendered. Content serialisation has several negative implications: users cannot get an overview of the page, entailing that users can only catch a glimpse of the page as long as they scan through the document. Consequently navigation across different web pages is a time consuming task and web page exploration is a resource intensive activity that requires a dedicated attention span [3].

The sequential access of screen readers means that visually impaired users take up to five times longer than sighted users to explore a web page [4]; the screen reader itself requires an additional cognitive effort [5] and, moreover, inter-page and intra-page navigation problems are some of the problems that need to be faced when working with visually impaired users [3].

As a consequence, in the case of users with disabilities, adaptation becomes crucial and may very much speed up the navigation and contribute to diminish the existing technological gap. In order to be able to model the user, the modelling component must collect information about a number of observable parameters such as interest, characteristics, etc. This information can be requested to the user in a previous session, but this is annoying, disruptive and can produce false assumptions. Another option is to collect this information in-use while the user is accessing the web, and therefore, to build a non invasive system able to model the users in the wild. In this way the system can learn its interests, likes, etc.

According to Pierrakos et al. [6] 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

978-83-60810-58-3/$25.00

the user. Nowadays, as Brusilovsky et al. [7] describe, many research projects focus on this area, mostly in the context of e-Commerce [7] and e-learning [8]. Important websites such as Google and Amazon are clear examples of this trend.

In any web environment, the contribution of the knowledge extracted from the information acquired from observation in-use is twofold:

- It can be used for web personalization (i.e. for the adaptation of the website according to the user requirements).
- It can also be used to extract knowledge about the interests of the people browsing the website or about the possible design mistakes.

Data mining for web personalization has many advantages. It is not disruptive, it is based on statistical data obtained by real navigation data (decreasing the possibility of false assumptions) and is itself adaptive (when the characteristics of the user change, collected data allows the automatic change of the interaction schema). When the user is a person with physical, sensory or cognitive restrictions, data mining is the easiest (and frequently almost the only) way to obtain information about the uses of the person.

Data mining in this context has also some drawbacks. The most important one is its impact over privacy, due to the need of storing large quantities of data about the users. Diverse laws in different countries protect user rights for privacy. Even if it is difficult to reach a balance among privacy and personalization, some appealing proposals have been recently published.

Web mining can be defined as the application of data mining techniques to data from the Internet. This process has three main stages:

- The data acquisition and pre-processing stage.
- 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 profile. Machine learning techniques are mainly applied in this phase.
- 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.

This work is the first stage of a web mining process carried out in discapnet website *www.discapnet.es* where 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). Being discapnet website addressed to people with disabilities, mainly to visually impaired people, link prediction will be specially important in the system. This is corroborated somehow because a preliminary analysis of the web logs showed that the time spent in link type or hub type pages is considerably longer than it would be expected to; it is longer than the one spent in pages devoted to content (content pages) and dynamic pages which are mainly related to news. This makes us suspect that the implementation of an efficient link prediction system will definitely help to make navigation easier, and as a consequence, diminish the time spent in link type pages.

So that the link prediction system is efficient, it is important for it to be based on observation in-use. This way behaviours emerge from the obtained data instead of looking for predefined models. Web logs are the most simple in-use information and the ones applicable to the wider set of users. Other more complete tools [9] capturing longitudinally low-level interaction unobtrusively limit the public to be used in the modelling process.

Summarizing, the aim of this paper is to design a link prediction system which contributes to make the navigation of users navigating in discapnet easier. The proposed system is based on observation in-use; behaviours emerge applying a web mining process to the obtained data, web server log data. The web mining process has required a thorough analysis of the environment and the data, a selection of the machine learning tools to be used and, finally, a design and evaluation of the system. After this process, we developed two approaches for user profiling: a global system built based on the complete website and a modular approach carried out discovering the navigation profiles within each zone. Both systems show to be useful for link prediction but the values of the evaluation measures are a bit higher for the modular approach. We consider that the inclusion of the described system in discapnet will contribute to improve inter-page navigation within the website and diminish the times spent in link type pages.

The paper describes the discapnet website and its main characteristics in Section II and the preprocess applied to the data in Section III. Section IV describes the machine learning techniques used for user navigation profile generation whereas Section V is devoted to describing the two profiling options implemented for discapnet. The evaluation of the two systems and their use for link prediction are presented in Section VI. Finally, Section VII summarises the conclusions and future work.

II. Discapnet website analysis

Discapnet is an initiative created to promote the social and work integration of the people with disabilities financed jointly by the Fundación ONCE [10] and Technosite. It contains two main action lines:

- An information service for organizations, professionals, people with disabilities and families.
- A platform to develop actions to promote the involvement of people with disabilities in the economic, social and cultural life.

Technosite provided us the server logs of two serves that store the activity generated in some areas of the web discapnet. The transferred data was basic anonimized server log data in Common Log Format [11] (see Figure 1). It contained all the requests served by two of the servers hosting discapnet website from the 2nd February 2012 to the 31st December 2012.

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

Fig. 1. Sample lines of a log file in CLF.

Fig. 2. Appearance of the front page of discapnet website.

In this context, the next stage of our research consisted on analysing the structure and content of the site. Figure 2 shows the appearance of the front page of discapnet.

The site is divided in different areas being main ones:

- *Areas Temáticas*
- *Comunidad*
- *Actualidad*

Some of these parts, such as, *Actualidad* (actuality) and *Noticias* (news) are very dynamic and can hardly be used for link prediction because it is impossible to build the models according to news that will be generated in the future. From the rest of the zones in the website, the experts in Technosite considered that *Areas Temáticas* (excluding *Salud*) and *Canal Senior* within *Comunidad* were the most interesting zones for modelling and introducing adaptation tools. And, as a consequence the provided data was limited to these zones.

The direct consequence of the previous assertion is that the built user models and link prediction system will be mainly limited to *Areas Temáticas* (see Figure 3).

Therefore, it shouldn't be forgotten that the provided sequences might not be complete user sessions what limits the data mining process and, as a consequence, the quality of the obtained profiles.

Finally, before starting with the modelling process we evaluated the accessibility of each of the pages of discapnet. We found this an important starting point because we considered that being discapnet a website addressed to people with special requirements, accessibility of the pages might become a source of problems. Therefore, accessibility was evaluated using EvalAccess [12]; the Automatic Accessibility Evaluator developed by EGOKITUZ according to the design guidelines published by WAI [13] and devoted to help designers to produce web sites that are accessible. The study showed that the accessibility rate was in average near 90% and this means that each individual page in discapnet was designed taking into account the accessibility guidelines.

III. Data Preprocessing

After the preliminary analysis the logs must be preprocessed to extract the useful information. Web server log files follow a standard format called Common Log Format [11]. 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

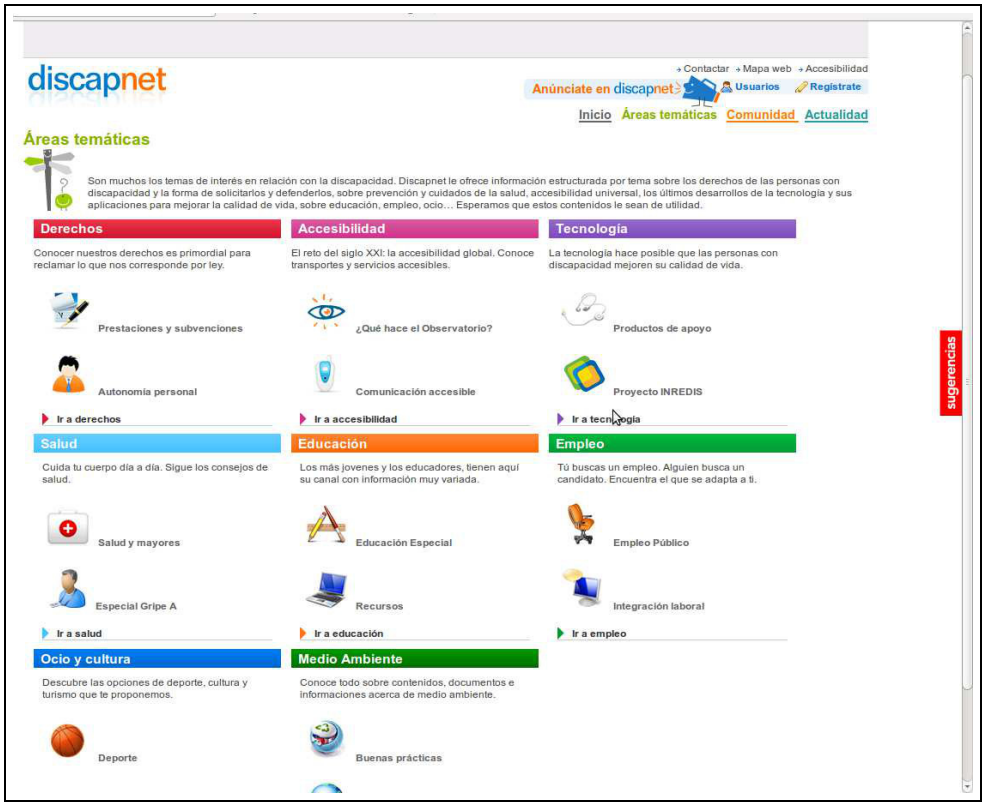

Fig. 3. Appearance of *Areas tematcias* within discapnet website.

code given by the anonymization process to remote host IP addresses, the time the request was recorded, the requested URL and the status field that informs about the success or failure when processing the request.

The log files used in this work contained 157,527,312 requests, which were reduced to 13,352,801 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)). Therefore, we 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 then carried out session identification: we fixed the expiry time of each session to 10 minutes of inactivity [14] obtaining a total amount of 907,404 sessions.

A. Session Representation

Once the database to be used for the process has been selected in the filtering and preprocessing steps, we need to decide how to represent the information to be used in the machine learning algorithms. 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, we represented the information corresponding to each of the sessions as a clickstream or sequence of clicks preformed in different URLs. Since we want to build user navigation profiles the order of the visited URLs will be important.

We selected the most relevant sessions (those with a minimum activity level; 3 or more clicks) and removed the longest sequences (those with their length out of the 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.

IV. Pattern Discovery

This is the stage that, taking as input the user click sequences, is in charge of modelling users and producing user profiles. Most commercial tools perform statistical analysis over the collected data. They extract information about most frequently accessed pages, average view times, average lengths of paths, etc. that are generally useful for marketing purposes. But the knowledge extracted from this kind of analysis is very limited. Machine learning techniques are in general able to extract more knowledge from data. In this context unsupervised machine learning techniques have shown to be adequate to discover user profiles [6]. We have used a crisp clustering algorithm to group users that show similar navigation patterns.

A. Clustering

In this work, in order to grouping into the same segment users that show similar navigation patterns, we need a clustering algorithm that is able to deal with sequences and an adequate distance to compare sequences. Based in our experience in previous works, as clustering algorithm, we selected *PAM (Partitioning Around Medoids)* [15] [16] which is similar to k-means but uses medoids or examples as centres instead of centroids what makes it suitable for cases where the examples are represented as sequences. Furthermore, we selected a Sequence Alignment Method, Edit Distance [17][18], as a metric to compare sequences. As it happens with most clustering algorithms, *PAM* requires the *K* parameter to be estimated. This parameter is related to the specificity of the generated profiles, when greater its value is more specific the profiles will be. We didn't have prior knowledge of the structure of the data, that is, we have no idea of the number of different user profiles. Therefore we performed an analysis to try to find the value of *K* that is enough to group the sessions with common characteristics but does not force to group examples with not similar navigation patterns in the same cluster (the range of values tested for *K* will be described in Section VI).

B. Profile Generation

The outcome of the clustering process is a set of groups of user sessions that show similar behaviour. But we intend to model those users or to discover the associated navigation patterns or profiles for each one of the discovered groups. The model will be composed by the common click sequences appearing among the sessions in a cluster. We used SPADE (Sequential PAttern Discovery using Equivalence classes) [19], an efficient algorithm for mining frequent sequences, used to extract the most common click sequences of 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 build the profiles of each cluster using SPADE 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 for the sessions belonging to it. The number of proposed URLs depends on parameters related to SPADE algorithm such as minimum support and maximum allowed number of sequences per cluster. We fixed the value for the minimum support to 0.2 and limited the amount of proposed URLs to 3 because proposing too many could disturb the user.

V. DISCAPNET USER NAVIGATION PROFILE DISCOVERY

We are aware that nowadays navigation in a website can be difficult for any type of users but this is still harder for users with special requirements. As a consequence, an adaptive system able to propose the adequate links to the user during her navigation would be specially helpful for them.

TABLE I
SIZES AND CHARACTERISTICS OF DATABASES USED FOR THE MODULAR SYSTEM AND THE GLOBAL SYSTEM.

Website zone	User Sessions	Average length	*K*
Accesibilidad	10,259	5.08	90
Derechos	22,561	4.78	80
Educación	1,773	4.27	27
Empleo	4,720	3.89	60
Medioambiente	852	5.05	20
Ocio y cultura	3,603	4.86	50
Tecnología	3,954	4.54	50
canal senior	338	4.60	13
Global	48,060	4.63	130

The structure of each of the subtopics within *Areas Temáticas* is very different and this will probably affect to the navigation the users do within them. Moreover a preliminary analysis of the sequences showed that nearly 50% of the user sessions extracted from the database belonged to navigations in a single zone. We considered those sessions representative of the navigation within each zone and decided to build the link prediction system based on them. After selecting the most relevant sessions and removing the longest sequences from it, the database contains 48,060 user sessions. We used two approaches to face the problem:

- The design of a global system using the data of all the analysed zones (see Figure 4).
- A modular system which builds profiles independently for each of the navigation zones (see Figure 5).

A. Global approach

The global system consists on applying the clustering and profiling processes as described in Section IV to the complete database; the 48,060 user sessions; the patterns are grouped using PAM clustering algorithm and the profile for each of the clusters discovered based on SPADE. The schema of the system is represented in Figure 4.

B. Modular approach

Being the structure of each zone different, we decided to build a modular approach to the user navigation profiling within discapnet. This means to build the profiles focusing on each of the possible analysis zones for user navigation profile discovery. With this aim, instead of working with the whole database, we worked with the user sessions located in a single web zone. That is we divided the database according to navigation zones and we worked with 8 different subsets; one for each of the zones where user navigation profile discovery will be carried out. Table I summarizes the sizes of each subset and Figure 5 shows the schema of the system where it can be observed that the profile discovery process within each zone was carried out as described in Section IV.

The set of profiles in the modular approach will be composed by the set of profiles generated for each one of the 8 zones.

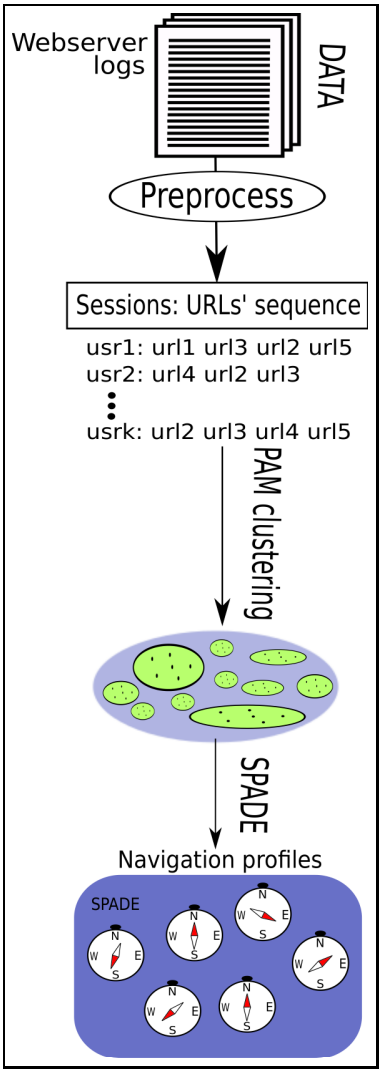

Fig. 4. Global approach to user profile discovery.

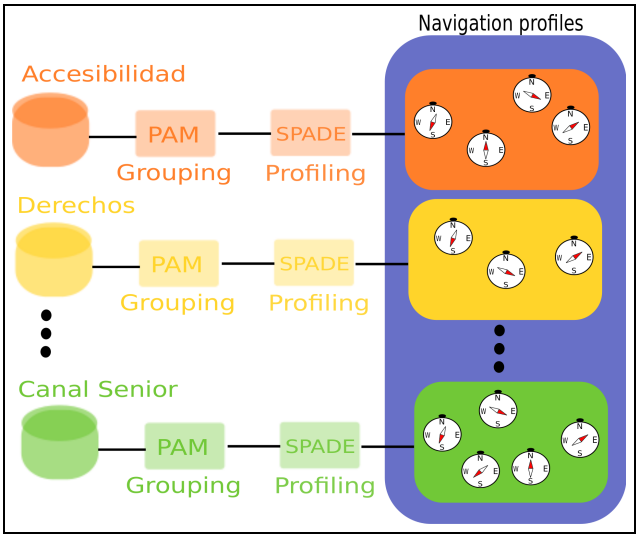

Fig. 5. Modular approach to user profile discovery.

VI. Evaluation and Link Prediction

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. The evaluation procedure was exactly the same for the two approaches used to build the model: the modular approach and the global approach.

In order to carry out this evaluation we used a hold-out methodology, dividing each folder into a training set (70% of the examples), validation set (20% of the examples) and test set (10% of the 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. Based on the usual exploration limit for the number of clusters, $\sqrt{n}$, in each of the databases we tried the following 5 values: $\sqrt{n}/4$, $\sqrt{n}/3$, $\sqrt{n}/2$, $2 * \sqrt{n}/3$ and $\sqrt{n}$. Obviously, being the sizes of the subsets very different, the number of clusters or user profiles generated in each of them has been different; Table II shows in column K the number of profiles generated in each of the modules according to the evaluation carried out using the validation set. This number has been selected using the validation set to evaluate results the same way the test set has been used to evaluate the final system. The procedure is explained in the following lines (although the explanation is given for the test the procedure used with the validation set is exactly the same).

The generated profiles were evaluated by comparing them to new users navigating the website (test set). 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% of the test sequences to select the profile for the new user according to the built model (between 1 and 2 links, because, as it is shown in column average length of Table I, the click sequences have in average near 5 links).

According to previous works found in the bibliography [20], new users might not be identical to any of the profiles

TABLE II
SUMMARY OF THE RESULTS.

		Validation			Test		
	k	Pr	Re	F.5	Pr	Re	F.5
Global system	130	0.55	0.40	0.51	0.55	0.40	0.51
Modular system	49	0.58	0.44	0.54	0.58	0.44	0.54
Improvement%		5.65	9.06	6.10	6.49	8.97	6.70

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 (k-*NN*) [21] 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. Due to its characteristics, the k-*NN* algorithm allows to select naturally the set of profiles generated in the training phase with higher similarity to the new user, and moreover, it showed to have good performance in previous works [20]. We used 2-*NN* to select the nearest clusters and combined the profiles of the two nearest clusters with defined profiles, weighting URL selection probabilities according to their distance. We combined these to propose profiles containing at most 3 URLs; those with the highest support values. If there are not enough URLs exceeding the minimum support value the profiles could have less than 3 URLs.

We computed performance metrics based on the results obtained for each of the new users of the test set. We compared the number of proposed links that are actually used in the test examples (hits) and the number of proposals that are not used (misses) and calculated precision (percentage of clicks used among the proposed ones), recall (percentage of clicks proposed among the used ones) and F.5-measure (a relationship between precision and recall giving more importance to precision).

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. Furthermore, by limiting the maximum number of URLs proposed for each profile to 3 the recall values will never reach 1. Since the average length of the sequences is near 5, if we propose a profile (3 URLs) based on 25% of the navigation sequence (between 1 and 2 URLs), we would be proposing less links than the really used ones, what makes impossible for the recall to achieve the highest values. We consider that in the concrete environment we are working it is really important to propose links that the user finds interesting because other proposed links would probably disturb the user. As a consequence, it is more important for the proposed links to be of good quality (precision) than guessing more of the used links (recall). This is why we used F.5-measure.

Table II shows the average results (precision, recall and F.5-measure) obtained for the test and validation sets in both cases: with the global system, and the modular system. The numbers show that the modular approach achieves better results than the global one, obtaining improvements of around 9% in recall and around 6.5% in precision and F.5-measure. Furthermore, results are similar for both, the validation set and the test set what means that the concrete data used to evaluate the system does not severely affect to the obtained performance.

The values obtained for the modular system show that if we would use the profiles for link prediction, nearly 60% (precision=0.58) of the proposed links (tending to 2 out of 3) would be among the ones used by the new user. This could probably make the user navigation easier.

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

Moreover, taking into account that the preliminary analysis showed that the time spent in hub pages is longer than usual we could assert that using those profiles for link prediction would save a big part of the time spent by users in their navigations.

The designed system seems to obtain near balanced values for precision and recall. Therefore analysing the recall we could state that nearly 45% of the links used by the new users would be among the ones proposed by the system (recall=0.44).

We need to take into account that this is a very strict evaluation of the models because, in a real situation, although not used during the navigation, some of the proposed links might also interest to the new users.

The main use of navigation profiles is link prediction and our system could be directly used for link prediction following the methodology described in the evaluation procedure.

VII. CONCLUSIONS

Web personalization becomes essential in industries and specially for the case of users with special needs such as visually impaired people. Adaptation may very much speed up the navigation of visually impaired people and contribute to diminish the existing technological gap. This work is the first stage of a web mining process carried out in discapnet: a website created to promote the social and work integration of people with disabilities. Based on observation in-use where behaviours emerge applying a web mining process to server log data, we designed a system to generate user navigation profiles and propose adaptations to the site through link prediction. The work was limited to the most static zones of the website.

We used *PAM (Partitioning Around Medoids)* clustering algorithm and Edit Distance to group into the same segment users with similar navigation patterns and SPADE (Sequential PAttern Discovery using Equivalence classes) to extract the user profiles from the cluster. These techniques were used to implemented two approaches: a global system built based on the complete website and a modular approach carried out discovering the navigation profiles within different zones of

the website. We then used a k-*NN (k-Nearest Neighbour)* based heuristic for link prediction.

Using a hold-out strategy and precision, recall and F.5-measure as performance measures for evaluation, we could deduce that both approaches showed to be effective for link prediction but the modular approach outperforms obtaining values of nearly 60% for precision and 45% for recall. This means that when 25% of the navigation of the new user has happened the designed system is able to propose a set of links where nearly 60% of them (2 out of 3) is among the ones the new user will be using in the future and this will definitely make the navigation easier saving a lot of time.

Being this a preliminary work, the system is open and many new ideas to be implemented in the future appeared during its development. First of all, the introduction of the designed link prediction system in the website and its evaluation in a real experiment would be the best way to discover the efficiency of the system. On the other hand, based on the web server log data provided by discapnet, other types of characteristics of the user sessions could be extracted which would allow to analyse the use of the website from another point of view mainly for problem detection.

ACKNOWLEDGMENT

This work was funded by the Department of Education, Universities and Research of the Basque Government (Eusko Jaurlaritza/Gobierno Vasco) through Grant IT-395-10, by the Science and Education Department of the Spanish Government (ModelAccess project,TIN2010-15549), by the Basque Governments SAIOTEK program (Datacc2 project, S-PE12UN064)

REFERENCES

[1] E. Turban and D. Gehrke, 2000. “Determinants of e-commerce website”. *Human Systems Management*, vol. 19, pp.111-120.

[2] D. Chaffey, F. Ellis-Chadwick, K. Johnston and R. Mayer, 2006. “Internet Marketing”. *Prentice Hall/Financial Times*.

[3] M. Vigo, S. Harper, 2013. “Challenging information foraging theory: screen reader users are not always driven by information scent”. *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pp.60-68, http://dx.doi.org/10.1145/2481492.2481499.

[4] J. Craven, P. Brophy, 2013. “Nonvisual access to the digital library: The use of digital library interfaces by blind and visually impaired people”. *In Technical report No. 145. Manchester, United Kingdom: Centre for Research in Library and Information Management*.

[5] S. Chandrashekar, T. Stockman, D. Fels, R. Benedyk, 2006. “Using Think Aloud Protocol with Blind Users: A Case for Inclusive Usability Evaluation Methods”. *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pp.251-252, http://dx.doi.org/10.1145/1168987.1169040.

[6] D. Pierrakos, G. Paliouras, C. Papatheodorou and C.D. Spyropoulos, 2003 “Web usage mining as a tool for personalization: A survey”. *User Modeling and User-Adapted Interaction*, vol. 13, pp.311-372, http://dx.doi.org/10.1023/A:1026238916441 .

[7] P. Brusilovsky, A. Kobsa and W. Nejdl, 2007. “The Adaptive Web: Methods and Strategies of Web Personalization”. *Lecture Notes in Computer Science (Springer)*, Berlin, http://dx.doi.org/10.1007/978-3-540-72079-9 .

[8] E. García, C. Romero, S. Ventura and C.D. Castro, 2009.“An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering”. *User Modeling and User-Adapted Interaction*, vol. 19, pp.99-132, http://dx.doi.org/10.1007/s11257-008-9047-z.

[9] A. Apaolaza, S. Harper, C. Jay, 2013.“Understanding Users in the Wild”. *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pp.1-4, http://dx.doi.org/10.1145/2461121.2461133.

[10] “Fundación ONCE for cooperation and social inclusion of people with disabilities”. Available: http://www.fundaciononce.es/EN/Pages/Portada.aspx, accessed 04-05-2014.

[11] “Common log format (clf)” 1995.*The World Wide Web Consortium (W3C)*. Available: http://www.w3.org/Daemon/User/Config/Logging.html, accessed 04-05-2014.

[12] “EvalAccess: Web Service tool for evaluating web accessibility”. Available: http://www.adm.aau.dk/rektor/aalborgeksperiment/engelsk/preface.html, accessed 04-05-2014.

[13] “WAI Guidelines and Techniques”. Available:http://www.w3.org/WAI/guid-tech.html, accessed 04-05-2014.

[14] D. He, A.Gker, 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.

[15] L. Kaufman and P. J. Rousseeuw, 1990. “Finding Groups in Data: An Introduction to Cluster Analysis”. *Wiley-Interscience*, .

[16] L. Liu and M. T. Özsu, 2009. “Encyclopedia of Database Systems. In: PAM (Partitioning Around Medoids).“ *Springer US*, .

[17] D. Gusfield, 1997 “Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology”. *Cambridge University Press*, New York, NY, USA, .

[18] B. Chordia and K. Adhiya, 2011. “Grouping web access sequences using sequence alignment method”. *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 2, pp.308-314.

[19] M. J. Zaki, 2001. “Spade: An efficient algorithm for mining frequent sequences”. *Machine Learning*, vol. 42, pp.31-60, http://dx.doi.org/10.1023/A:1007652502315.

[20] O. Arbelaitz, I. Gurrutxaga, A. Lojo, J. Muguerza, J. Pérez and I. Perona, 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)*, pp.187-192.

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