Investigating the Relation between Users' Cognitive Style and Web Navigation Behavior with K-means Clustering

Marios Belk, Efi Papatheocharous, Panagiotis Germanakos, George Samaras

Department of Computer Science, University of Cyprus, Nicosia, Cyprus
{belk,efi.papatheocharous,pgerman,cssamara}@cs.ucy.ac.cy

Abstract. This paper focuses on modeling users' cognitive style based on a set of Web usage mining techniques on navigation patterns and clickstream data. Main aim is to investigate whether k-means clustering can group users of particular cognitive style using measures obtained from a series of psychometric tests and content navigation behavior. Three navigation metrics are proposed and used to find identifiable groups of users that have similar navigation patterns in relation to their cognitive style. The proposed work has been evaluated with a user study which entailed a psychometric-based method for extracting the users' cognitive styles, combined with a real usage scenario of users navigating in a controlled Web environment. A total of 22 participants of age between 20 and 25 participated in the reported study providing interesting insights with respect to cognitive styles and navigation behavior of users.

1 Introduction

The World Wide Web today has expanded to serve millions of different users for a multitude of purposes in all parts of the world. Naturally, Web content nowadays needs to be filtered and personalized based on the particular needs of individual users. The users' interests, expectations and expertise, cognitive style and perception are some of the factors that need to be considered when creating personalized interactive systems. Therefore, the first step towards Web personalization is specifying the type of users and creating the user model which reflects the user's intrinsic needs and preferences which ultimately influence the adaptation of Web interactive systems.

The user model is a representation of static and dynamic information about an individual, and it represents an essential entity for an adaptive interactive system aiming to provide adaptation effects (i.e., a set of tasks or content of a system can be presented differently between users with different user models) [1]. For example, an adaptive information retrieval system may recommend the top most relevant items based on the user's interests. An adaptive educational hypermedia system may provide adjusted educational material and navigation support to users that have particular level of knowledge on a subject. An adaptive e-commerce system may enhance the security and privacy-preserving measures and can present an adapted content to users that have a specific level of knowledge and experience towards security terms (e.g., provide novice users with personalized security awareness information by using simpli-

fied security terms and additional explanations). The mechanism utilized for user modeling can be based on explicit or implicit information gathering approaches. Explicit information is provided directly by the user, usually through Web registration forms, questionnaires, or specially designed psychometric instruments. On the other hand, implicit information is extracted by the system automatically to infer characteristics about the user and is usually obtained by tracking the user's navigation behavior throughout the system. For example, such implicit information can be extracted from the time spent on a particular Web-page by a user, which can be used to infer the interest of the user towards the main subject of that Web-page.

To this end, the work presented studies the relation between users' cognitive styles and navigation behavior with explicit and implicit user information gathering approaches. Main objectives of the paper are to: i) investigate whether a specific clustering technique (i.e., *k*-means clustering) can group users of particular cognitive style using measures obtained from psychometric tests, ii) propose navigation content metrics to find identifiable groups of users that have similar navigation patterns in relation to their cognitive style, and iii) investigate whether there is a possible relationship between users' cognitive style and their navigation behavior. The identification of users with specific cognitive and navigation style will ultimately help in defining an adaptation mechanism that will target a different user interface experience in Web-based environments for various cognitive typologies of users.

2 User Modeling based on Data Analysis Techniques

The ability of adaptation in interactive systems heavily depends on successful user modeling. A user model is created through a user modeling mechanism in which unobservable information about a user is inferred from observable information from that user [2]; for example, using the interactions with the system (i.e., time being active on a Web-page, buying history, ratings of products, bookmarked or saved content, etc.).

The simplest approach of user model generation is in the case where the information collected by the user is used as-is and remains unprocessed. For example, users might explicitly express their interest on specific topics of a news publishing system which will be further used by simple rule-based mechanisms to adapt the interface by displaying the selected topics on the top of the user's interface. More intelligent approaches for generating user models is in the case where the browsing activities of users may be utilized by data mining and machine learning techniques to recognize regularities in user paths and integrate them in a user model. A thorough literature review on how data mining techniques can be applied to user modeling in the context of personalization systems can be found in [3]. The data mining techniques mentioned enable pattern discovery through clustering and classification, association rules (or association discovery) and sequence mining (or sequential pattern discovery). They represent popular approaches appearing in the data mining literature. In addition, [4] describes data mining algorithms based on clustering, association rule discovery, sequential pattern mining, Markov models and probabilistic mixture and hidden (latent) variable models for Web personalization purposes.

Nowadays, the process of Web user modeling has become attached to automated data mining or knowledge discovery techniques due to the large volumes of available user data on the Web [5]. Nasraoui et al. [5] perform clustering on user sessions to place users in homogeneous groups based on the similar activities performed and then extract specific user profiles from each cluster. Clustering techniques are also used in order to divide users into segments containing users with similar navigation behavior. Using a similarity metric, a clustering algorithm groups the most similar users together to form clusters. Some algorithms classify users into multiple segments and describe the strength of each relationship [6]. The same concept is found within fuzzy clustering techniques, examples of which include the work of Castellano and Torsello [7] that categorized users based on the evaluation of similarity between fuzzy sets using a relational fuzzy clustering algorithm and Castellano et al. [8] that derived user profiles by analyzing user interests. Variations of fuzzy clustering methods include Fuzzy c-medoids, Fuzzy c-trimmed-medoids, relational Fuzzy Clustering-Maximal Density Estimator (RFC-MDE) algorithm and hierarchical clustering approaches.

The abovementioned works primarily focus on applying data mining and machine learning techniques for modeling the interests and preferences of users towards specific items of Web environments. For example, clustering techniques are utilized for grouping users that visited, bought or rated similarly the same products. Association rules are used in many cases to relate different products based on their viewing history, e.g., when users view product A and afterwards view product B, then an association rule is created between product A and B indicating a high relationship between the two products. Accordingly, this information is further utilized by the system to offer recommendations based on the navigation behavior of users.

Taking into consideration these works, the next section presents a user modeling approach for eliciting similar groups of users based on their navigation behavior in the context of adaptive interactive systems and relates these groups to cognitive styles. To the best of the authors' knowledge, this is among the first works to study the relation between the cognitive style of users and their navigation behavior in an online encyclopedia system, apart from sporadic attempts which utilized a number of clustering techniques to understand human behavior and perception in relation with cognitive style, expertise and gender differences of digital library users [9], and recent research attempts which studied the connection between the way people move in a museum and the way they prefer to approach and process information cognitively [10].

3 Cognitive-based User Modeling for Web Adaptation

The proposed approach, as shown in Figure 1, focuses on the user modeling part of an adaptive interactive system. User modeling is associated to information regarding the users of a system, the users' interactions as well as the context in which communication or data transaction takes place. It is mainly responsible for gathering information regarding the user, building the user model and feeding this information to the adaptation mechanism which in turn will modify the user interface accordingly.

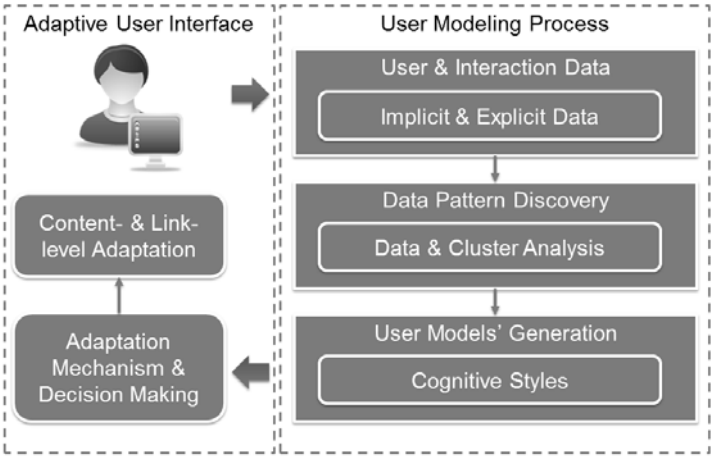

Fig. 1. Cognitive-based User Modeling Approach

Based on Figure 1, the first step of the proposed approach starts with collecting the user's interaction data with the system. Specifically, the browsing history is used as the main source of information about the user's interaction data which contains the URLs visited by the user and the date/time of the visits. Accordingly, meaningful information is derived based on this information, e.g., the number of visits to a particular URL, the time spent and the specific sequence of visits. In the next step, specific data analysis and clustering techniques are applied on the raw data in order to classify users to groups with similar navigation behavior and extract other important information about the users. Finally, the user models are generated containing information about the cognitive style of users which is further provided to the adaptation mechanism to apply the adaptation effects. The next section makes a brief introduction of the theory behind the cognitive styles utilized in this work.

3.1 Cognitive Styles

Cognitive styles represent an individually preferred and habitual approach to organizing and representing information [11]. Among numerous proposed theories of individual styles [11-13], this study utilizes Riding's Cognitive Style Analysis (CSA) [11]. In particular, Riding's CSA consists of two dimensions and classifies users to the cognitive typologies of Wholist-Intermediate-Analyst and Verbal-Intermediate-Imager. The Wholist-Analyst dimension refers to how individuals organize information. Specifically, users that belong to the Wholist class view a situation and organize information as a whole and are supposed to take a linear approach in hypermedia navigation (i.e., users read the material in a specific order based on the context). On the other hand, users that belong to the Analyst class view a situation as a collection of parts, stress one or two aspects at a time and are supposed to take a non-linear approach in hypermedia navigation. Users that belong in between the two end points (i.e., Intermediate) do not differ significantly with regards to information organiza-

tion. The Verbal-Imager dimension refers to how individuals process information. Users that belong to the Verbal class can proportionally process textual and/or auditory content more efficiently than images, whereas users that belong to the Imager class the opposite. Users that belong in between the two end points (i.e., Intermediate) do not differ significantly with regards to information processing.

Riding's CSA might be applied effectively on designing adaptive hypermedia systems, since it consists of distinct scales that correspond directly to different aspects of information systems, i.e., content and functionality is either presented visually or verbally, and users may have specific navigation behavior, i.e. linear vs. non-linear, based on their cognitive style.

4 User Study

The objective of the study is threefold; firstly, investigate whether clustering techniques can group users of particular cognitive style using measures obtained from Riding's CSA test, secondly, evaluate the use of content navigation metrics to find identifiable groups of users that have similar navigation patterns within the group of users that participated in the study, and finally investigate whether a relation exists between cognitive style and navigation behavior of users.

4.1 Method of Study

A total of 22 individuals participated voluntarily in the study carried out within the first week of November 2011. All participants were undergraduate Computer Science students in their third and fourth year of study, and their age varied from 20 to 25. The students first completed a series of questions using a Web-based psychometric test (http://adaptiveweb.cs.ucy.ac.cy/profileConstruction) based on Riding's CSA [11] that measures the response time on two types of stimuli and computes a ratio between the response times for each stimuli type in order to highlight differences in cognitive style. The stimuli types are: a) statements (i.e., identify whether a statement is true or false), and b) pictures (i.e., compare whether two pictures are identical, and whether one picture is included in the other). Then, the users were asked to read various articles of a Web environment and navigate freely through its hyperlinks. Main aim was to track the navigation sequence of users within the Web environment. Accordingly, an appropriate environment for tracking the navigation sequence of users is a reproduced version of Wikipedia (http://wikipedia.org) since it consists of content and hyperlinks that are placed in a context-dependent order and thus enables users either navigate sequentially, or in an unordered form. Furthermore, the articles were enriched to include verbal-based content, i.e., content in textual form without images (Figure 2A), or image-based content, i.e., content represented with images and diagrams (Figure 2B). The same content was presented to all users, while verbal-based and imager-based content was presented to users that belonged to the Verbal class and the Imager class, respectively.

Fig. 2. Verbal-based (A) and Image-based (B) User Interface of the Web-site used in the Study

The navigation behavior of the students was monitored at all times on the client-side. In particular, a browser-based logging facility was implemented with JQuery JavaScript Library (http://jquery.com) to collect the client-side usage data from the hosts accessing the Web-site used for the study.

4.2 Definition of Metrics

The reproduced version of Wikipedia consists of articles that are interconnected through hyperlinks based on a context-dependent hierarchy (i.e., articles of similar context are interconnected). We consider that sequential links are connected based on the articles' content and the distance between each point is equal to 1. Thus, a linear navigation behavior is represented with a minimum distance covered considering the links visited by a user whereas a higher distance describes a non-linear navigation behavior. Accordingly, we measure the distance between the links visited by users utilizing the following metrics: i) Absolute Distance of Links (ADL), the total absolute distance between the links visited, ii) Average Sequential Links (ASL), the average number of sequential links visited by a user, and iii) Average non-Sequential Groups of Links (AGL), the average number of non-sequential groups of links visited by a user, if all sequential links are considered to represent one group. To better explain the metrics used, we provide a navigation example, e.g., the clickstream navigation pattern "Node: 8, Pat: 4 | Node: 9, Pat: 3-2 |" which means that the user visited the eighth content-page and then read the content of the fourth link and so on. For this particular navigation the metrics, as defined above, are calculated as: ADL=(|4-1|+|3-1|+|2-3|)/N=2, ASL=M/N=0.333 and AGL=B/N=0.667, where N is the number of total links visited, M is the number of sequential links visited based on the Web-site content and B is the number of non-sequential groups of links derived from the links visited. In our example, M is equal to 1 and B is equal to 2 as for pattern "3-2" the only sequential link clicked was the second and the two non-sequential groups of links were patterns "4" and "3-2". The user's interaction with the Web-site content was captured through these metrics which were also normalized based on the number of user interactions by dividing each variable to the total number of clicks.

4.3 Results

This section presents and analyzes the results obtained from the study. A non-hierarchical method based on the Euclidean distance (*k*-means clustering) was used [14]. The following methods were applied: i) *k*-means clustering on the responses of users to the psychometric test, and ii) *k*-means clustering on the navigation pattern of users in the online encyclopedia system.

Since the data obtained was from different users, and thus generated independently, we may assume that the i.i.d. assumption holds. Moreover, since the possible navigation patterns and user interactions with the user interface were close to a very large number, *k*-means clustering was selected for the analysis to avoid calculating all possible distances between all possible interactions. Other assumptions made was that the structure of the Web-site's is linear based on the content (i.e., it contains an introduction and sections that follow the introduction in a sequential manner) and that the number of clusters is known in each case (i.e., k=3 and k=2 in each clustering case respectively). Thus, using *k*-means clustering we are trying to differentiate users based on their Cognitive Style (CS) typology (i.e., Wholist-Intermediate-Analyst and Verbal-Intermediate-Imager) and navigation style (i.e., linear and non-linear).

Table 1. Ratio of Cognitive-based Profiles of Clustered Users from the Psychometric Test

Cluster	Users	Wholist-Analyst Range	Users	Verbal-Imager Range
1	8	[0.786, 1.030]	6	[1.121, 1.248]
2	12	[1.099, 1.424]	7	[0.958, 1.040]
3	2	[1.776, 1.853]	9	[0.832, 0.941]

Table 1 presents the number of clustered users in each cluster using *k*-means and the range of ratios obtained from the psychometric tests. From the clustering performed using the users' responses in the psychometric test, we observe that the users are clustered in three groups. The figures show that the users are clearly distinguished based on their cognitive profile and that they cover the whole range of the scale as suggested by Riding [11]. For example, in the Wholist-Analyst dimension one of the clusters contains users with CS ratio [0.786, 1.030] which is in line to Riding's Wholist typology (i.e., $\leq$1.02) and in the Verbal-Imager dimension the clustered users' CS ratio [0.832, 0.941] is again in line to Riding's Verbal typology (i.e., $\leq$0.98). This finding shows that the *k*-means clustering technique can group users of particular CS using measures obtained from psychometric tests; it has provided encouraging results and justifies further utilization.

The next clustering applied involves content visit path analysis by using the three metrics proposed which measure the linearity of the users' navigation behavior. Table 2 presents the ranges of the users' cognitive-based profiles appearing in each cluster. For example, the CS of the users grouped based on their content navigation style in the first cluster is within the range [0.819, 1.776] regarding the Wholist-Analyst dimension and within the range [0.861, 1.248] regarding the Verbal-Imager dimension.

Table 2. Ratio of Cognitive-based Profiles of Clustered Users from Content Navigation Style

Metric	Ranges of Cluster 1		Ranges of Cluster 2	
	Wholist-Analyst	Verbal-Imager	Wholist-Analyst	Verbal-Imager
ADL	[0.819, 1.776]	[0.861, 1.248]	[0.786, 1.853]	[0.832, 1.154]
ASL	[0.819, 1.248]	[0.861, 1.121]	[0.786, 1.853]	[0.832, 1.248]
AGL	[0.819, 1.776]	[0.885, 1.248]	[0.786, 1.853]	[0.832, 1.128]
All	[0.819, 1.776]	[0.861, 1.248]	[0.786, 1.853]	[0.832, 1.128]

The normalized values of the metrics were used to perform clustering with the combination of all three clustering metrics (results of which are shown in the last row of Table 2) and to also visualize the degree of membership of each user per metric in each cluster (Figure 3).

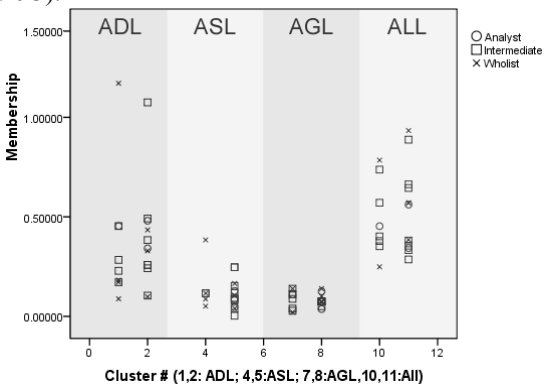

Fig. 3. Degree of Membership in each Cluster (in columns) per Metric ADL, ASL and AGL and CS Identification regarding the Wholist-Analyst Dimension

Table 3. Mann-Whitney Rank-sum Statistical Test per Clustering Metric

Metric	Cluster 1			Cluster 2		
	U	z	p	U	z	p
ADL	63	0.6	0.274	47	0.46	0.322
ASL	66	-1.6	0.054	54	-0.66	0.254
AGL	43	0.62	0.267	28	1.7	0.044

From Figure 3 we observe that the users grouped in each cluster present variability in terms of their CS. In addition, the results of the rank-sum statistical test (Mann-Whitney [15]) performed between the two clusters (Table 3), has shown that in most cases the two clusters did not differ significantly. However, we observed that using the ASL and AGL metrics, statistically significant differences were identified between the first and the second clusters' medians for the Wholist-Analyst and Verbal-Imager ratios respectively (U=66, p=0.05, and U=28, p=0.04). This means that the ASL and AGL metric proposed can be used to identify users of particular navigation behaviour that differ in their Wholist-Analyst and Verbal-Imager ratio styles respectively. Conclusively, users (in Cluster 1) with linear navigation behavior have statisti-

cally significant different cognitive style ratio concentration than non-linear users (in Cluster 2), which is an important result, since it can be further concluded that some relation has been found between navigation and cognitive styles in these particular cases.

4.4 Final Remarks

Based on the results obtained, currently, no safe conclusion can be drawn, whether there is a cohesive correlation between the cognitive style and the navigation pattern followed by each user, and further experimentation needs to be carried out. In particular, most users in the same cluster although had similar navigation behavior (i.e., linear/non-linear), their respective cognitive style was variant. However, a statistical comparison of the CS ratios of users between the clusters showed that the users' cognitive style differed significantly ($p \leq 0.05$) indicating that clustering users based on their navigation behavior is likely to cause separation of users into distinctive groups that differ significantly in terms of cognitive style. In addition, the navigation metrics proposed seem to successfully distinguish clusters of users that according to their respective cognitive-based profile range belong to two overlapping groups – range that covers a lower and a higher fragment in the Riding CSA scale. Finally, the resulting membership degree to each cluster can be used to characterize the degree of linearity in the interaction of users in fuzzy terms to optimally capture navigation behavior. Such findings could provide a promising future direction towards modeling cognitive styles of users by tracking their navigation behavior with implicit information gathering approaches that are transparent to the users, as well as the identification of adaptation rules for a more user-centric interface design.

5 Conclusions and Future Work

This paper investigated the relation between cognitive style and navigation behavior of users. Specific navigation metrics have been proposed and utilized by a clustering technique, with the aim to identify groups of users that have similar navigation behavior and investigate the relation to their cognitive style.

The limitations of the current work are related to the small sample of users participating in the study, the number of clusters used in the analysis (for example in larger samples perhaps a higher number of clusters should be used), the selection of clustering method, the effect of outliers and the order of cases analyzed. We have addressed some of these threats by assuming to know how many clusters we wanted to extract based on the cognitive profiles of the users and since we also had a moderately sized dataset this meant that the selection of *k*-means clustering was a reasonable choice. The solution of *k*-means clustering may depend on the order of the users using the online environment and thus we have arranged the samples in a random order to address this threat. In addition, the fact that the *k*-means clustering algorithm is sensitive to the initial randomly selected cluster centers, we have eliminated this threat by repeating the algorithm execution several times.

The relevant research is in its infancy and further empirical studies are needed to investigate such issues. A future research prospect is to evaluate other Web environments, techniques (e.g., fuzzy clustering and neural networks) and metrics that might also assess human behavior in order to shed more light on this complex phenomenon.

Acknowledgements. This work is co-funded by the EU project CONET (INFSO-ICT-224053), SocialRobot (285870) and the project smarTag (University of Cyprus).

References

1. Brusilovsky, P., Millán, E.: User Models for Adaptive Hypermedia and Adaptive Educational Systems. In: Kobsa, A., Nejdl, W. (eds.) The Adaptive Web. LNCS, vol. 4321, pp. 3-53. Springer, Heidelberg (2007)
2. Frias-Martinez, E. Magoulas, G., Chen, S., Macredie, R.: Modeling Human Behavior in User-Adaptive Systems: Recent Advances Using Soft Computing Technique. J. Expert Systems with Applications. 29(2), 320-329 (2005)
3. Eirinaki, M., Vazirgiannis, M.: Web Mining for Web Personalization. J. ACM Transactions on Internet Technology. 3(1), 1-27 (2003)
4. Mobasher, B.: Data Mining for Web Personalization. In: Kobsa, A., Nejdl, W. (eds.) The Adaptive Web. LNCS, vol. 4321, pp. 90-135. Springer, Heidelberg (2007)
5. Nasraoui, O., Soliman, M., Saka, E., Badia, A., Germain, R.: A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites. J. IEEE Transactions on Knowledge and Data Engineering. 20(2), 202-215 (2008)
6. Perkowitz, M., Etzioni, O.: Adaptive Web Sites. Communications of the ACM. 43(8), 152-158 (2000)
7. Castellano, G., Torsello, M.: Categorization of Web Users by Fuzzy Clustering. In: 12th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, pp. 222-229. Springer, Heidelberg (2008)
8. Castellano, G., Fanelli, A., Mencar, C., Torsello, M.: Similarity-Based Fuzzy Clustering for User Profiling. In: Conferences on Web Intelligence and Intelligent Agent Technology Workshops, pp. 75-78. IEEE Computer Society, Washington, USA (2007)
9. Frias-Martinez, E., Chen, S., Macredie R., Liu, X.: The Role of Human Factors in Stereotyping Behavior and Perception of Digital Library Users: A Robust Clustering Approach. J. User Modeling and User-Adapted Interaction. 17(3), 305-337 (2007)
10. Antoniou, A., Lepouras, G.: Modeling Visitors' Profiles: A Study to Investigate Adaptation Aspects for Museum Learning Technologies. J. Computing Cultural Heritage. 3(2), 1-19 (2010)
11. Riding, R., Cheema, I.: Cognitive styles - An Overview and Integration. J. Educational Psychology. 11(3/4), 193-215 (1991)
12. Felder, R., Silverman, L.: Learning and Teaching Styles in Engineering Education. J. Engineering Education. 78(7), 674-681 (1988)
13. Witkin, H., Moore, C., Goodenough, D., Cox, P.: Field-dependent and Field-independent Cognitive Styles and their Educational Implications. Review of Educational Research. 47(1), 1-64 (1977)
14. Aldenderfer, M., Blashfield, R.: Cluster Analysis. Sage Publications, Newbury Park, California (1984)
15. Mann, H., Whitney, D.: On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. J. Annals of Mathematical Statistics. 1(8), 50-60 (1947)