Data-driven User Profiling to Support Web Adaptation through Cognitive Styles and Navigation Behavior

Panagiotis Germanakos, Efi Papatheocharous, Marios Belk, George Samaras

Department of Computer Science, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus
{pgerman, efi.papatheocharous, belk, cssamara}@cs.ucy.ac.cy

Abstract. This paper aims to analyze human navigation behavior and identify similarities of cognitive styles using measures obtained from psychometric tests. Specific navigation metrics are utilized 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 that entails a psychometric-based survey for extracting the users' cognitive styles, combined with a real usage scenario of users navigating in a controlled Web environment. A total of 84 participants of age between 17 and 25 participated in the study providing interesting insights with respect to cognitive styles and navigation behavior of users. Studies like the reported one can be useful for assisting adaptive interactive systems to organize and present information and functionalities in an adaptive format to diverse user groups.

Keywords: Data-driven Profiling/Personalization; Cognitive Styles; Human Behavior; Web Mining.

1 Introduction

The rapid increase of digital information and the widely diverse audience on the World Wide Web (Web) has generated research challenges in the area of adaptive interactive systems. Adaptive interactive systems [1][2] provide an alternative to the "one size fits all" approach of static user interfaces by adapting the interactive system's content and functionality to users' needs and preferences, with the aim to provide a positive user experience.

An important feature of an adaptive interactive system is its user model. The user model is a representation of static and dynamic information about an individual that is utilized by the adaptive interactive system aiming to provide adaptation effects (i.e., the same system can look different to users with different user models) [3][4]. For example, a personalized search engine may adaptively select and prioritize the most relevant results to the users' interests.

Apart from research-oriented system, major commercial Web service providers such as Google [5], Bing [6], and Amazon [7] have shown an increased interest in providing personalized services to its users. These service providers are nowadays offering personalized results and recommendations, by employing various user

modeling and adaptation techniques. Accordingly, the notion of personalization has found its way in users' everyday interactions in Web interactive systems. Nevertheless, some research issues are still open with regards to some influential factors of personalization, such as the behavioral drivers and navigation interaction of users in executing task-oriented reasoning processes. A useful example for investigation is the everyday interaction with online encyclopedia articles. In that case, the personalization process could be applied based on observations of human behavior. Since content and functionality of Web interactive systems is either presented visually or verbally, and users may have specific navigation behavior, e.g. linear or nonlinear navigation behavior, this work suggests that cognitive styles of users, which describe the way individuals, perceive and organize information [8], might be applied effectively on designing adaptive Web interactive systems.

In this respect, main objective of the paper is to study the humanistic behavior and interaction with Web interactive systems and identify patterns of similar cognitive style of users and navigation behavior. Subsequent objective is to investigate whether specific data analysis techniques can group users of particular cognitive style using measures obtained from psychometric tests.

The paper is organized as follows. In Section 2, we provide an overview of user modeling mechanisms in the context of adaptive interactive systems. In Section 3 we present a user study based on the proposed approach. Consequently, we conclude the paper and describe our directions of future work in Section 4.

2 User Modeling Mechanisms

Adaptation in interactive systems heavily depends on successful user modeling. 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 users' interface. More intelligent approaches for generating user models include cases 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 [9] and [10]. 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, Mobasher [11] 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 [12]. Nasraoui et al. [12] 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. In addition, data mining and machine learning techniques are used for modeling the interests and preferences of users towards specific items in e-commerce environments. For example, data analysis techniques are utilized for grouping users that visited, bought or rated similarly the same products [13][14]. Association rules are used in some 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 [10]. Accordingly, this information is further utilized by the system to offer recommendations based on the navigation behavior of users.

Apart from modeling the interests of users, adaptive Web interactive systems can be developed to accommodate a variety of individual differences, such as cognitive styles. Cognitive styles may correspond ideally to the structure of Web environments. Since cognitive styles describe the way users process and organize information, a personalized environment, that is supported by an automated cognitive style-based modeling mechanism, can be adapted at the levels of content selection and structure; the content is essentially either visual or verbal (or auditory), while the manipulation of links can lead to a more analytic and segmented structure, or to a more holistic and cohesive environment.

Taking into consideration the abovementioned works, the next section presents a user study for eliciting similar groups of users based on their navigation behavior and investigates how these groups may be related to cognitive styles. To the best of the author's knowledge, this is among the first attempts to study the relation between the cognitive style of users and their navigation behavior in an online encyclopedia system, apart from sporadic attempts, the first of 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 [15], and a second more recent research attempt which studied the connection between the way people moved in a museum and the way they preferred to approach and process information cognitively [16].

3 User Study

The objective of the study is to investigate the relation between cognitive styles and human navigation behavior, as well as investigate whether clustering techniques can group users of particular cognitive style using measures obtained from a cognitive-grabbing psychometric test.

3.1 Cognitive Style Theory used in the Study

Among the numerous proposed theories of individual styles [8][17][18], the proposed work utilizes Riding's Cognitive Style Analysis (CSA) that consists of two dimensions and classifies users to the cognitive typologies of Wholist-Intermediate-

Analyst and Verbal-Intermediate-Imager [8]. 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. 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 nonlinear approach in hypermedia navigation. Users that belong in between the two end points of the Wholist/Analyst scale (i.e., Intermediate) do not differ significantly with regards to information organization. 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.

In this regards, we consider that Riding's CSA [8] implications can be mapped on Web environments as described in Table 1, since they consist of distinct scales that respond directly to different aspects of the Web space. The CSA implications can provide clear guidelines in the context of Web design (i.e., selecting to present visual/verbal content and structuring information flow in a wholistic/analytic manner), and is probably one of the most inclusive theories, since it is actually derived from the common axis of a number of previous theories [19].

Table 1. Riding's CSA Scale Mapping to Web Environments

CSA Scale	Typology	Description	Web Implications
Verbal/Imager	Verbal	Process information in textual form more efficiently	Web content in text/auditory form
Verbal/Imager	Imager	Process images more efficiently	Web content in graphical, visual representation
Verbal/Imager	Intermediate	In between the Verbal/Imager scale	No significant difference in preference or information processing
Wholist/Analyst	Wholist	Views a situation and organizes information as a whole	Linear approach in Web-site navigation
Wholist/Analyst	Analyst	Views a situation as a collection of parts and only stresses one or two aspects at a time	Nonlinear approach in Web-site navigation
Wholist/Analyst	Intermediate	In between the Wholist/Analyst scale	No significant difference in information processing

3.2 Method of Study

A total of 84 individuals participated voluntarily in the study carried out within the last week of February 2012. All participants were undergraduate students and their age varied from 17 to 25. The students first completed a series of questions using a Web-based psychometric test based on Riding's CSA [8] that measured the response

time on two types of stimuli and computed the 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 navigated in a reproduced version of a Web application; Wikipedia [20].

Participants were assigned 10 problem-based tasks whose answers could be found inside the Wikipedia articles with the aim to increase the navigation activity of users and investigate their behavior in solving the problem-based tasks they had been assigned.

The navigation behavior of the participants was monitored at all times on the client-side. In particular, a browser-based logging facility was implemented with JQuery JavaScript Library [21] to collect the client-side usage data from the hosts accessing the Web-site used for the study. Furthermore, the following metrics were used to measure an individual's usage data: i) Absolute Distance of Links (ADL), ii) Average Sequential Links (ASL) and iii) Average non-sequential Groups of Links (AGL) visited. To better explain the metrics used, we provide an example navigation, 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 used, M is the number of sequential (linear) links used based on the Web-site content and B is the number of nonsequential groups of links derived from the links used. 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 nonsequential 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. A main hypothesis for these calculations was that the starting point of all users' navigations in a Web-page is the first link appearing in the menu (e.g., Home) since it is the first content the user is interacting with.

3.3 Results

This section presents and discusses the results and conclusions obtained from the study. The following analyses were performed: i) *k*-means clustering on the answers of users to the psychometric test, and ii) *k*-means clustering on the navigation pattern of users interacting with the online encyclopedia system while trying to answer to a set of specific questions. The aim is to discover knowledge from such complex humanistic data.

Since the data are derived from different users carrying out different navigations, they may be considered having the same probability distribution as the rest sequences of navigation, and thus all are mutually independent or generated independently. In addition, since the users were not directed in any way, the possible navigation patterns and user interactions with the user interface were close to a very large number. That is why *k*-means clustering was selected for the analysis provided in this section, to avoid

calculating all possible distances between all possible interactions. Other assumptions included that the structure of the Web-site's content was linear (i.e., the first section is introductory while the rest sections follow the first in a sequential manner) and that the number of classes is known in each case (i.e., k=3 and k=2 in each clustering case respectively). To better explain this, this study is used to differentiate users to three classes (k=3) based on their cognitive style dimension (i.e., Wholist-Intermediate-Analyst and Verbal-Intermediate-Imager) and to two classes (k=2) based on their navigation style/behavior (i.e., linear and non-linear). k was selected after k-means clustering was performed with various k (i.e., 2, 3, 4 and 5).

The first clustering performed using the replies of the users in the psychometric test showed that the sample is balanced particularly with respect to the Verbal-Imager dimension, in the sense that on average a similar number of users was found to represent each cluster comparing the clustering results and CSA based on Riding's theory. The number of users clustered in the three groups and their respective cognitive style range are reported in Table 2.

Table 2. Ratio of Cognitive-based Profiles of Clustered Users from the Psychometric Test

Cluster	# Users	Wholist-Analyst Range	Users	Verbal-Imager Range
1	30	[0.543, 1.002]	41	[0.599, 0.980]
2	29	[1.046, 1.347]	22	[0.980, 1.084]
3	25	[1.377, 3.170]	21	[1.093, 1.489]
1	73	[0.543, 1.660]	30	[0.599, 0.898]
2	7	[1.534, 3.111]	35	[0.921, 1.102]
3	4	[1.619, 3.170]	19	[1.121, 1.489]

The upper part of the table represents users' profiling based on Riding's CSA and the lower part the cognitive ratios of the grouped users via k-means clustering. If we relate the figures of cognitive ratio in each cluster we might assume that the first cluster represents Wholist users, the second Intermediates and the third Analysts while the first cluster on the right side of the table seems to represent Verbals, the second Intermediates and the third Imagers. Another observation is that the ranges of users in each dimension are representative of the theoretical range scales identified by Riding in educational context of psychology [19] (upper part of Table 2). For example it is quite impressive how in the Wholist-Analyst dimension one of the clusters contains users with cognitive style ratio [0.543, 1.002] which is in line to Riding's Wholist scale (i.e., ≤1.02) and in the Verbal-Imager dimension the clustered users' cognitive style ratio [0.599, 0.980] is again in line to Riding's Verbal scale (i.e., ≤0.98). Accordingly, the clustering in the Verbal-Image case gave users with cognitive style ratio [0.599, 0.898], [0.921, 1.102] and [1.121, 1.489] (lower part of Table 2) which are again in line (or quite close) with (or to) the upper figures. This finding showed that the k-means clustering technique is promising and it can group users of particular cognitive style using measures obtained from psychometric tests, at least in the Verbal-Imager dimension and justifies further utilization.

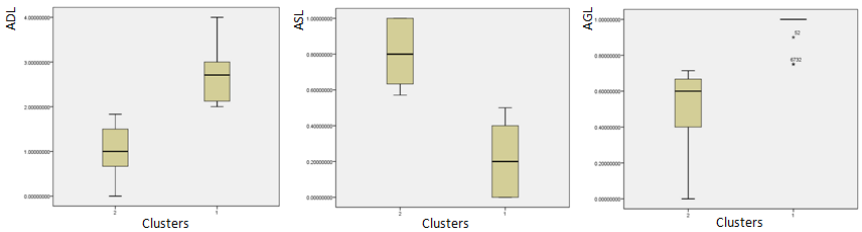
The results presented next include the cluster analysis from the content visit path browsed by the users and through three clustering metrics which measure the linearity of the users' navigation behavior. Table 3 presents the obtained ranges of the users'

cognitive-based profiles in each cluster. For example, the cognitive style of the users grouped based on their content navigation style in the first cluster is between the values 0.543 and 3.170 regarding the Wholist-Analyst dimension and between the values 0.599 and 1.242 regarding the Verbal-Imager dimension. Accordingly, the cognitive style of users in the second cluster present narrower ranges regarding the Wholist-Analyst dimension (from 0.684 to 3.111) and between 0.702 and 1.442 regarding the Verbal-Imager dimension (i.e., a shifted range). These results were obtained by using the combination of the content visit path metrics (i.e., ADL, ASL and AGL) whereas the ratios obtained for each of these metrics are provided in Table 3.

Table 3. Ratio of Cognitive-based Profiles of Clustered Users

Metric	Range of Cluster 1		Range of Cluster 2	
	Wholist-Analyst	Verbal-Imager	Wholist-Analyst	Verbal-Imager
ADL	[0.731, 3.170]	[0.599, 1.242]	[0.543, 3.111]	[0.702, 1.442]
ASL	[0.543, 3.170]	[0.599, 1.242]	[0.684, 3.111]	[0.702, 1.442]
AGL	[0.684, 3.111]	[0.702, 1.442]	[0.543, 3.170]	[0.599, 1.242]
All	[0.543, 3.170]	[0.599, 1.242]	[0.684, 3.111]	[0.702, 1.442]

The statistics of the actual values regarding each variable (ADL, ASL and AGL) which measure the linearity of the users' content visit path in each formed cluster are illustrated in Figure 1. We observe that two clusters are formed, which are well balanced and each two clusters formed (i.e., Cluster 1 and 2) are significantly different in terms of values and mean value. Thus, the box plots clearly show that Cluster 1 and Cluster 2 in all metrics contain users following linear and nonlinear navigation behavior respectively.



(a) (b) (c)

Fig. 1. Box plots of clusters formed using (a) ADL, (b) ASL and (c) AGL metrics

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

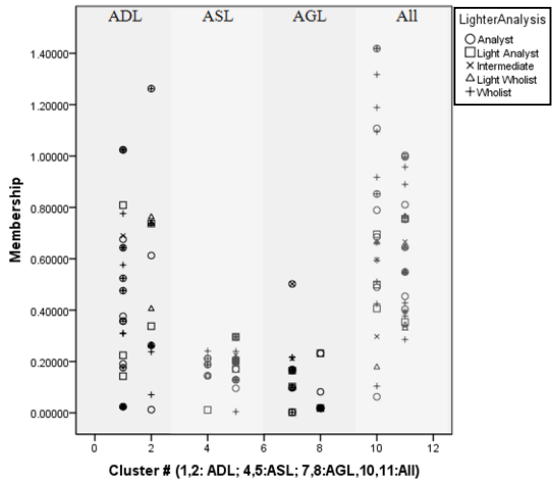


Fig. 2. Degree of Membership in each Cluster (in columns of the X axis) per metric ADL, ASL and AGL (in grey and dark grey columns) and cognitive style identification in the scale of Analyst, Light Analyst, Intermediate, Light Wholist and Wholist.

Based on the values of ADL, ASL and AGL the users grouped in the first cluster (46% of the subjects) indicate highly linear navigation behaviour. Particularly, the users indicate low values in the ADL and AGL metrics and high values in the ASL metric which measure the non-linear and linear aspect of navigation respectively. The opposite behaviour is observed for the users in the second cluster.

Table 4. Users in Cognitive Style Analysis per Metric

Metric / CS	ADL		ASL		AGL		ALL	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Analyst	13	4	6	11	6	11	9	7
Light Analyst	3	2	1	4	3	2	2	3
Intermediate	5	0	0	5	3	2	2	2
Light Wholist	2	7	1	8	2	7	7	2
Wholist	17	5	7	15	11	11	10	12

The users grouped in each cluster present however variability in terms of their cognitive style, something clearly visible in Figure 2 and also in Table 4. This shows a weak correlation of navigation patterns with specific cognitive-based profiles. However, the navigation metrics proposed seem to successfully distinguish clusters of users that according to their respective cognitive-based profile range (Table 4) belong to two overlapping groups whose range covers a lower and a higher fragment in the Riding CSA scale. Finally, the resulting membership degree to each cluster (shown in Figure 2) can be used to characterize the degree of linearity in the interaction of the users in fuzzy terms to optimally capture their navigational behaviour.

4 Conclusions

Processing of humanistic data, and particularly implicit data captured during the interaction of users with Web systems, is particularly intriguing since the human aspect is hardly understood or studied by researchers. Accordingly, the overarching aim of this work was to increase our understanding on mining humanistic behavior through navigation interaction of users with Web environments. Specific navigation metrics have been proposed and utilized by a clustering technique, with the aim to identify patterns, trends, similarities in navigation behavior, and investigate the relation to their cognitive style. Such a finding could provide a promising direction towards the identification of adaptation rules, based on the abovementioned relationship, for a more user-centric interface design. Furthermore, the paper also investigated whether *k*-means clustering can group users of particular cognitive style using measures obtained from psychometric tests.

Summarizing, no safe conclusion can be drawn, whether there is a cohesive correlation between the cognitive styles and the human navigation behavior. In particular, the users of the same cluster group although had similar navigation behavior (i.e., linear/nonlinear), their respective cognitive style was variant. This is maybe due to potential weaknesses of the reproduced version of Wikipedia at a conceptual level which needs to be further studied. Nevertheless, results have shown that the majority of users in the same cluster obtained from using all the metrics and applying the clustering mechanism were worth further analysis.

Furthermore, results have shown that *k*-means clustering could be applied effectively on the data extracted from cognitive-based psychometric tests for highlighting differences in users' cognitive styles by assigning the users into a cognitive-based cluster. Clustering mechanisms can be used as an alternative method for forming groups of users with particular cognitive style, handling the uncertainty and fuzziness of the information available (e.g., ratio) to extract optimum groups of users, rather than using a strict rule-based mechanism that highlight differences in cognitive style based on specific thresholds.

A future research prospect is to conduct further studies investigating the relation of cognitive styles of users with other types of navigation behavior than the ones investigated in this work (i.e., linear/nonlinear) as well as capture and study the users' interactions with other Web objects (e.g., drop down lists, drag and drop tools, Web forms). Furthermore, investigate the effect of users' navigation behavior during collaborative and sharing activities as well as evaluate the proposed approach and metrics in other Web environments (e.g., social networks, collaboration platforms).

Acknowledgements. The work is co-funded by the EU project Co-LIVING (60-61700-98-009), smarTag (University of Cyprus) and PersonaWeb (Cyprus Research Promotion Foundation).

References

1. Schneider-Hufschmidt, M., Kühme, T., Malinowski, U.: Adaptive user interfaces: Principles and practice. Human Factors in Information Technology. North-Holland, Amsterdam (1993)
2. Brusilovsky, P.: Adaptive Hypermedia. J. User Modeling and User-Adapted Interaction, vol. 11, no. 1,2, pp. 87--110 (2001)
3. 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)
4. Frias-Martinez, E., Magoulas, G.D., Chen, S.Y., Macredie, R.D.: Modeling Human Behavior in User-Adaptive Systems: Recent Advances Using Soft Computing Technique. J. Expert Systems with Applications. vol. 29, no. 2, 320--329 (2005)
5. Google Inc., http://www.google.com
6. Bing Inc., http://www.bing.com
7. Amazon Inc., http://www.amazon.com
8. Riding, R.: Cognitive Style Analysis - Research Administration. Learning and Training Technology (2001)
9. Eirinaki, M., Vazirgiannis, M.: Web Mining for Web Personalization. J. ACM Transactions on Internet Technology, vol. 3, no. 1, pp. 1--27 (2003)
10. Pierrakos, D., Paliouras G., Papatheodorou, C., Spyropoulos, C.D.: Web Usage Mining as a Tool for Personalization: A Survey. J. User Modeling and User-Adapted Interaction, vol. 13, no. 4, pp. 311--372 (2003)
11. 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)
12. 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, vol. 20, no. 2, pp. 202--215 (2008)
13. Linden, G., Smith, B., York, J.: Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Internet Computing, vol. 7, no. 1, pp. 76--80 (2003)
14. Su, X., Khoshgoftaar, T.: A survey of collaborative filtering techniques. Advances in Artificial Intelligence. Article 4, 19 pages (2009)
15. Frias-Martinez, E., Chen, S.Y., Macredie R.D., 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, vol. 17, no. 3, pp. 305--337 (2007)
16. Antoniou, A., Lepouras, G.: Modeling Visitors' Profiles: A Study to Investigate Adaptation Aspects for Museum Learning Technologies. J. Computing Cultural Heritage, vol. 3, no. 2, pp. 1--19 (2010)
17. Felder, R., Silverman, L.: Learning and Teaching Styles in Engineering Education. Engineering Education, vol. 78, pp. 674--681 (1988)
18. Witkin, H.A., Moore, C.A., Goodenough, D.R., Cox, P.W.: Field-dependent and field-independent cognitive styles and their educational implications. Review of Educational Research, vol. 47, no. 1, pp. 1--64 (1977)
19. Riding, R.J., Cheema, I.: Cognitive styles - An Overview and Integration. J. Educational Psychology, vol. 11, no. 3/4, pp. 193--215 (1991)
20. Wikipedia, http://wikipedia.org
21. JQuery Javascript Library, http://jquery.com