AutoCardSorter: Designing the Information Architecture of a Web Site Using Latent Semantic Analysis

Christos Katsanos, Nikolaos Tselios, Nikolaos Avouris
Human-Computer Interaction Group, Electrical and Computer Engineering, University of Patras
Rio, Achaia, Greece
{ckatsanos, nitse, avouris}@ece.upatras.gr

ABSTRACT
In this paper, we describe an innovative tool that supports the design and evaluation of the information architecture of a Web site. The tool uses Latent Semantic Analysis and hierarchical clustering algorithms to provide optimal information navigation schemes in an automated manner. The proposed, tool-based, approach addresses the problem of reasonable content structuring, which established techniques such as card sorting also address. A real world case study depicted substantial effectiveness gain, without expense in the quality of results. We argue that such an approach could facilitate information-rich applications design, like most Web sites, by reducing time and resources required.

ACM Classification Keywords
H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces – Evaluation/Methodology, H5.4 Information interfaces and presentation (e.g., HCI): Hypertext/Hypermedia – Architectures, Navigation, User Issues.

Author Keywords
Information architecture, automated tools, card sorting, latent semantic analysis, clustering algorithms.

INTRODUCTION
The exponential growth of content within the Web has created an overabundance of information and a poverty of human attention. As a result, users often confront problems, while trying to navigate and locate desired pieces of information. In addition, Web design is a complex process characterized by an inherent dualism. The Web should be treated simultaneously both as a user interface as well as a hypertext system. However, despite the maturity of User Centered Design methods to address the first aspect, little has been done to assist the latter [4]. This is mainly due to the fact that the complexity and the unstructured nature of information interaction are not fully explained by typical models of human-computer interaction [11]. More to the point, even established techniques, such as card sorting, are often neglected, due to required resources and increased complexity to carry out the analysis.

Card sorting is one of the main methods used to design a Web site’s information architecture. While applying the method, representative users are given a stack of index cards, each containing one word or phrase representing the information or services provided on Web pages. They are asked to group the cards in piles that make sense to them and subsequently name the resulting groups. Card sorting “can provide insight into users’ mental models, illuminating the way that they often tacitly group, sort and label tasks and content within their own heads” [10]. Therefore, by understanding the users’ selections, we can increase findability of information [9]. There are two primary alternatives, open and closed card sorting. Open card sorting (with no pre-established set of groups) is used primarily in new Web sites, while the closed variation is used when adding content to an existing structure or extending the results of an open card sorting.

However, the method has a number of weaknesses. While card-sorting study results can be stable with 20 or even fewer participants [12], the process of finding representative users’ early in the lifecycle can be daunting, time consuming and costly. The problem of cost is further intensified given the fact that later in the iterative cycle users should be re-involved, in order to conduct much needed user testing sessions. In addition, thorough data pre-processing and statistical analysis is required, which diminishes the possibility of wider adoption.

There is a range of software tools that can support the gathering of card sorting data, and/or assist with their analysis (e.g. USort/EZCalc [2], CardZort1, OptimaSort2). These tools automate aspects of the data collection and/or analysis, and generally have three main components: an administration tool for setting up the experiment, a tool for participants to conduct the sort and, optionally, an analysis tool. Among these tools, MindCanvas3 is a remote research.


You can quote this paper as:

1 http://www.cardzort.com
2 http://www.optimasort.com
3 http://www.themindcanvas.com
tool that provides an interesting game-like elicitation approach to gather card sorting data and elaborate them to rich visualizations. However, even with these tools available, the effort for gathering and analyzing the data of a card sorting study should not be underestimated. As a result, the method is often neglected by practitioners, who opt for empirical and superficial approaches.

In this paper, an innovative, tool-based approach for the design and evaluation of a Web site’s information architecture is presented. This tool (AutoCardSorter) is inspired by the card sorting user method and employs Latent Semantic Analysis (LSA) [8] and clustering algorithms [1] to address the problem of reasonable content structuring. The tool is aimed at providing the necessary flexibility and efficiency to the practitioners, and can be used for both the initial design and redesign of web sites. In the rest of the paper, a typical scenario of using the tool is presented first, followed by a brief description of the statistical techniques it is based on. Subsequently, we present a study conducted to validate the quality of the results obtained using AutoCardSorter in comparison to the traditional card sorting approach. Finally, we discuss some directions for future work.

AUTOMATED CARD SORTING TOOL

The Automated Card Sorting Tool (AutoCardSorter) automates structuring of information, using a technique by calculating semantic similarity of different Web pages and clustering accordingly the information space. The tool addresses the problem of reasonable content structuring (structural navigation) and helps in creating semantic relationships between related pieces of content across levels of a hierarchy (associative navigation links – [5]). Such an automated approach could substantially accelerate the design and evaluation lifecycle, by providing valuable insight and increasing the possibility to explore alternative designs for the structure of a Web site. The proposed methodology can be used for both the initial design and redesign of web sites.

AutoCardSorter employs a novel algorithm that uses LSA and clustering algorithms to support the early steps of a web site’s architecture design or redesign, simulating an open card sorting experiment. This is achieved by creating a distance matrix of semantic dissimilarities among descriptions of content items, using LSA as a distance metric. Subsequently, clustering algorithms are used to construct the information space. A typical scenario of using the tool is the following: First, the designer provides descriptions of the pages that the Web site will contain (see Figure 1 - leftmost part of the user’s screen). The tool, then, runs an automated analysis, using the LSA algorithm and machine learning techniques. LSA is used to calculate the semantic similarity among all the Web pages descriptions, while clustering algorithms are applied to group together semantically similar pages.

LSA was developed to mimic human ability to detect deeper semantic associations among words, phrases or whole sentences. LSA builds a semantic space representing a given user population’s understanding of words sentences, and whole text segments, from documents that these users are likely to have read. The meaning of a word, sentence or any text is represented as a vector in a dimensional space, typically, with about 300 dimensions. The degree of semantic similarity between any pair of texts, such as the descriptions of the content of two Web pages, is measured by the cosine of the corresponding two vectors. Each cosine value lies between +1 (identical) and -1 (opposite). Near-zero values represent unrelated texts. Currently, three, complementary, hierarchical agglomerative clustering

![Figure 1. Using AutoCardSorter to identify the information architecture for a University Department’s Web.](image-url)
algorithms have been implemented to ensure the quality of the obtained results: (a) average linkage, (b) complete linkage and (c) single linkage.

The output of the tool is an interactive tree diagram, known as dendrogram, presenting the recommended clustering of the pages comprising the Web site’s navigational model. Specifically, the tool clusters the described information space, suggests how the web pages should be distributed and which pages should have links to each other, according to their semantic similarity. In addition, the designer is offered with the option to differentiate the number of the desired groups in a visual way, simply by dragging the line depicting the similarity strength among the grouped items (see Figure 1), and the tool reorganizes the results, showing the most effective item clustering in real time. Finally, the tool also provides a suggestion about the number of clusters that should be chosen, based on the elbow criterion [1].

A possible variation to the previous scenario could include the definition of the desired number and labels of the sections to be created. This variation implements an automated process of a closed card sorting technique, where the tool places each page to a section, according to their semantic similarity.

VALIDATING THE RESULTS PROVIDED BY THE TOOL

An appropriate study has been designed and implemented, in order to validate the robustness of the results obtained by using AutoCardSorter. The goal of the study was to compare results from a typical card sorting method involving representative users and from using AutoCardSorter to design the information architecture of the Qalibra project Web site. The aim of the site is to enable partners’ collaboration and communicate to the general public the project’s results on potential positive and negative effects of food consumption on human health.

Methodology and Procedure

AutoCardSorter was used to identify the structure that would meet efficiently the needs of the Web site’s representative users. The derived structure was compared to the results of an open card sorting study. For both cases, the same content descriptions have been used, which were selected from the available content inventory. First, the descriptions of the pages were enriched with the contextual information that they were referring to, in order to accommodate the differences in the ways humans and computers perceive the meaning of words. Inconsistencies in word-labels describing the same concept and the ambiguity created by use of pronouns can influence the results. To tackle this issue, we consistently used the same label for all semantically related terms. We, also, replaced pronouns with their implicitly related nouns or noun.

Using AutoCardSorter

We provided as input to the tool the descriptions of the pages to be clustered, selected a general semantic space as the representative user profile to better reflect the broad audience of the Web site, and employed the average linkage clustering method, which is a good compromise between the extremes of the other two [3]. The whole process required approximately 1 hour.

Open Card Sorting

18 participants took part in the open card sorting session [7], 10 male, 8 female, aged 22-55, with a mean of 28. Research [12] has shown that for card sorting studies 15 to 20 users are sufficient for stable results. The participants shared a high level of education and reported their internet experience as high. First, participants were given information regarding the general nature of the study, and the overall scope of the Web site. Next, a pile of cards representing the Web site’s pages was provided to them, and they were asked to sort the cards into groups, which were meaningful to them. The facilitator kept notes of participants’ comments. Each session lasted approximately 1 hour. For the analysis of the data, we used EZCalc [2], which facilitates analysis of computer-based card sorting studies. A pre-processing of the physical sorts was conducted in order to allow usage of EZCalc for the

---

**Figure 2.** Part of the average linkage dendrograms produced for the Qalibra project site by AutoCardSorter (a) and by analyzing card sorting data (b). Number of clusters is specified based on the elbow criterion.
analysis. The pre-processing and analysis of the data required approximately 9 hours.

**Analysis and Results**
First, a comparison of the similarity matrices produced by each approach was conducted. In the context of our analysis, a similarity matrix is a symmetric, square matrix containing a measurement of the semantic similarity for each pair of page descriptions. In our tool-based approach, this metric is the LSA index, while for the card sorting case we used the normalized frequency of card-pairs appearing in the same pile. The correlation analysis indicated a very high degree of correlation \((r = 0.80, p < 0.01)\).

To further investigate the accuracy of our tool-based method, a comparison of the dendrogram produced by each approach was conducted. Since perfect agreement has not been achieved, the resulting navigation scheme depends on where the designer decides to draw the cutting line. In order to be able to compare two possible navigation schemes proposed by the two approaches in an objective manner, we used the elbow criterion to determine the number of clusters for each case. The resulting navigation schemes were identical (see Figure 2). Therefore, our approach proved to provide similar results with a typical card sorting study.

Furthermore, the study demonstrated that the tool-based approach is substantially more effective than the traditional card sorting approach. Specifically, the site’s design with AutoCardSorter required approximately 1 hour overall, whereas the card sorting approach required 27 hours.

**CONCLUSION**
In this paper, an innovative tool to support design and evaluation of information architecture has been described. This tool uses LSA and three different hierarchical clustering algorithms to identify optimal information navigation schemes in an automated manner, thus minimizing the resources and effort required by established user-based techniques, such as card sorting. A validation study was conducted, based on a real world example, in which the automated approach was compared to a typical user-based approach. The study demonstrated substantial effectiveness gain in the use of the automated approach – proved approximately 27 times faster - without expense in the quality of results.

Automated tools such as the one described here, confirm the usefulness of traditional user methods, which can be applied in a more focused and formalized manner. As a result, the presented tool, coupled with other automated tools, e.g. a tool to evaluate semantic appropriateness of hyperlink’s descriptions [6], can inform both designers and evaluators while designing or redesigning a Web site, allowing deeper exploration of alternative scenarios.

A possible disadvantage of the proposed tool-based approach is the lack of the qualitative feedback obtained from representative users, while conducting the study. In addition, lack of an appropriate LSA semantic space to accurately represent intended users of specific characteristics in terms of knowledge, could deteriorate the quality of the results obtained. The latter is under investigation, with scheduled large scale studies for Web sites in various domains. Future work, also, includes investigating other techniques for providing alternative views of the results, taking also into account other factors.

**ACKNOWLEDGMENTS**
We thank European Social Fund (ESF), Operational Program for Educational and Vocational Training II (EPEAEK II) and particularly the Program PYTHAGORAS, for funding the above work. We, also, acknowledge the support of the EU-funded project QALIBRA (Quality of life - Integrated Benefit and Risk Analysis). Finally, we thank Dimitra Ioannou for her valuable help with the statistical analysis.

**REFERENCES**