Automatic Web Site Authoring with SiteGuide[1]

V. de Boer [a] V. Hollink [b] M. W. van Someren [a]

[a] *Human-Computer Studies Laboratory, Informatics Institute, University of Amsterdam*
[b] *Centre for Mathematics and Computer Science (CWI)*

1 Introduction

An important step in the design process for a web site is to determine which information is to be included and how the information should be organized on the web site's pages. Usually, web designers are not experts on the content or domain of a new site and the domain experts are no designers. The goal of SiteGuide is to assist both groups by creating a first description of the content topics with a tentative structure for the site. In this paper we present 'SiteGuide', a system that helps both amateur and professional web designers to create a setup for a new web site by presenting an initial information architecture.

In the early phases of web site design, reviewing web sites from the same domain as the target site are often used as source of inspiration for the new site. For instance, a person who wants to build a site for a small soccer club will look at web sites of some other small soccer clubs. However, comparing sites manually is very time-consuming and error-prone, especially when the sites consist of many pages. SiteGuide takes as input a set of user-selected web sites of the same type as the target website (typically 3 to 10). The system creates an initial information architecture for a new site by efficiently and systematically comparing a set of example sites identified by the user. SiteGuide automatically searches the sites for *topics* and structures that the sites have in common. For example, in the soccer club domain, it may find that most example sites contain information about youth teams or that pages about membership always link to pages about subscription fees. The common topics are brought together in a model of the example sites.

The tool then presents this found common information architecture to the user in both textual and visual form. SiteGuide can be used as a standalone tool or its output can serve as a starting point for further design refinement. SiteGuide can also be used in in a critiquing scenario for a first draft of a new web site. The draft is compared with the model, so that missing topics or unusual information structures are revealed.

2 Constructing the web site model

To construct the web site model, SiteGuide identifies common topics that occur on most example sites. For this, SiteGuide identifies pages of different example sites that handle on the same topic and forms clusters of these pages. The clustering method must allow that pages appear in more than one cluster (a single page includes text about different topics). Also, a cluster may not contain a page from each site, because a topic may not be included in each site. Pages occur in more than one cluster, when they contain content about more than one topic.

The total quality of a clustering is determined by the similarity between the pages in the clusters. We use five page similarity measures: For **text similarity** uses the $tf \cdot idf$ weighted cosine similarity between the stemmed words on the pages. **Anchor text similarity** is defined as the cosine similarity between the anchor texts of the links that point to the pages. **URL similarity** and **page title similarity** are defined as the inverse of the Levenshtein distance between the pages' URLS and page titles respectively. cosine similarity between the anchor texts of the links that point to the pages. The final similarity measure is **link structure similarity** which assigns higher scores to clusters who's pages link to and from pages in the same clusters.

[1]The full version of this paper is accepted at the 17th International Joint Conference Intelligent Information Systems (IIS 2009). Krakw, Poland, June 15-18, 2009.

These similarity measures are weighted with an additional cluster size parameter combined to determine the cluster similarities. Because we are interested in *inter-site similarity* rather than *intra-site similarity*, we only consider similarities between pages of different web sites to determine the final cluster quality. We use heuristic hill-climbing search to find a clustering with a high quality score.

Each cluster of the final clustering becomes a topic in the model. Topics are presented to the user in the form of *characterizing features* and *structural features*. Characterizing features are the most descriptive keywords extracted from the contents, page titles, URLs and link anchor texts of the pages in the clusters. An example page is also determined for each topic. Structural features show how topics are embedded in the model. They include the average number of pages in a site on the topic, the average number of in- and outgoing links and the linked topics.

The model, consisting of the topics described by their features is presented to the user in the SiteGuide tool. This can either be in the form of a textual list of topics or in the form of a visual graph. SiteGuide also offers the option to output the model to other (XML) formats so that it can be used in other web-authoring tools. One example of this is the DENIM web site sketching and prototyping tool for which the SiteGuide-produced model can serve as an initial setup, which can be further refined.

In the critiquing scenario, the user's draft web site is mapped onto the model constructed from the example sites. SiteGuide then informs the user which topics in the model do not have corresponding pages in the draft and vice versa. It also compares the structural features of the topics in the draft site to the model.

3 Experiments and Results

To evaluate whether SiteGuide provides useful assistance to users who are building a web site, we set up an evaluation study that answers two questions: 1) Do the discovered clusters and topics represent the subjects that are really addressed at the example sites and are the textual topic descriptions understandable for humans? and 2) Do people actually build better web sites when using the SiteGuide tool?

To answer the first question we test whether the generated model topics correspond to a gold standard. For this, we used web sites from three domains: windsurf clubs, primary schools and small hotels. For each domain 5 sites were selected as example sites with sizes ranging from 8 to 61 pages. For each site, we manually identified the topics and annotated them with short descriptions.

For each of these domains, SiteGuide generated a web site model based on the example sites and presented the SiteGuide output to 5 evaluators. The evaluators were asked to produce a short description of what they thought each generated topic was about. Next, an expert coder matched the generated topics' description to the gold standard topics' description, with partial matches counting as half a match. The expert coder assigned a full match to 73% of the generated topics while an additional 21% of the topics were classified as a partial match. This shows that for most topics SiteGuide is capable of generating an understandable description and that the found topics are the important topics that should be found according to the gold standard. We also found that more than half of the topics that should be detected according to the gold standard have indeed been found by SiteGuide.

To show that the use of SiteGuide's output actually improves the early phases of the web design process we asked 12 participants to sketch web site setups for a primary school. The group was split up in a control group, who were presented links to example web sites and a test group who were presented with both the links and the SiteGuide generated information architecture for inspiration.

A double-blind evaluation of the sketches was performed by a web design professional. He first ranked the 12 sketches on overall quality and then assigned a rating on a five point scale to each sketch for five quality criteria (completeness, relevancy, detailedness, content structure and link structure). The results show a significantly higher ranking for the test group sketches and significantly higher scores on completeness, content structure and link structure (at $\alpha = 0.05$). From this we can conclude that the setups made by participants presented with the SiteGuide output are considered to be better than those made without SiteGuide.

To analyze the objective acceptance of the SiteGuide setup for the school task, we compared the topics in the presented SiteGuide setup to the resulting sketches from the test group. For each page in a participant's sketch, we checked if a corresponding topic occurred in the SiteGuide setup. This average acceptance score for the test group is significantly higher than that of the control group. This indicates that, as was also perceived by the participants, the SiteGuide information is actually used in the final setup.