Decomposing the Web Graph: Towards Information Architecture Mining

Matthias Keller
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
D-76128 Karlsruhe, Germany
matthias.keller@kit.edu

Antoine Blanchard
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
D-76128 Karlsruhe, Germany
uqcuf@student.kit.edu

Martin Nussbaumer
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
D-76128 Karlsruhe, Germany
martin.nussbaumer@kit.edu

ABSTRACT
A semantic gap exists between the Web Graph which is the standard model of current Web mining approaches and the Information Architecture as perceived by humans. We believe that this gap can be closed and present a poster of our current work on this topic. The idea is to segment the Web Graph into subgraphs that represent distinct navigation systems and to analyze the subgraphs individually. These subgraphs called Navigation Structure Graphs (NSGs) reveal the type and purpose of navigation systems. The novel method can open doors for a new kind of application that allows re-engineering the Information Architecture precisely and promises to solve known problems of Web structure mining as Site boundary detection and hierarchy extraction.

Categories and Subject Descriptors
H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia – Navigation; H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia – Theory

General Terms
Algorithms, Design, Human Factors, Theory

Keywords
Information architecture, Hypertext, Data mining, Navigation systems

1. INTRODUCTION
Today the “lost in hyperspace”-problem is a problem of poor engineering rather than a problem of lacking research. Problems of hypermedia navigation are addressed (and solved) by Information Architects, User Interface Designers and Usability Engineers. They shape a fine-grained information space by implementing a plethora of menus and navigation aids. These elements are based on an underlying concept of content organization that usually has a hierarchical, linear or networked structure.

Designing Web navigation is not about placing individual hyperlinks on pages but about organizing coherent systems of navigation that contain a number of hyperlinks and span a group of pages. Navigation systems do not contain a fixed set of links. Often the links displayed vary from page to page, e.g. in the case of a navigation system implementing the breadcrumb pattern [5] or if a sub menu is present on some pages.

Human users are able to decode the purpose and the semantics of navigation systems. For example they can distinguish the main hierarchical navigation system from a contextual navigation system showing related resources. The important aspect is that decoding navigation systems reveals the Information Architecture and the underlying content organization. Humans are able to conceive content hierarchies and other forms of organization while browsing.

Machines on the other hand are blind to the fine-grained information architecture. The standard model of Web structure mining is a graph whose vertices represent URLs and whose edges are defined by the hyperlinks on the pages identified by the URLs [1]. Link analysis based on this model has proven to be a working solution for crawling, indexing and ranking the resources of the WWW and a successful business model as well ([3],[4]). However these approaches do not consider the semantics of navigation systems and are not able to decompose the original Information Architecture the way humans do – we argue that there is a semantic gap that has not been closed yet.

Figure 1: Semantic gap between the Web Graph model and the Information Architecture as conceived by humans

We expect that closing this gap will solve central problems of Web structure mining as Web Site boundary detection and hierarchy extraction. It will allow analyzing the structure of Sites and the organization of the content with much more precision, providing very valuable information for data mining solutions, developers that work on re-engineering tasks and researchers interested analyzing the current Web.

2. APPROACH
In contrast to current approaches that are based on mining hyperlinks or groups of hyperlinks (e.g. [2]) we are working on a novel approach that is based on mining more complex navigation systems. We have implemted a prototype and conducted a first evaluation of its robustness. We also found that analyzing the

Copyright is held by the author/owner(s).
The 22nd ACM Hypertext Conference, Posters and Demos.
The detection was conducted without major errors for 91.5% of the Sites if the element
application was developed for inspecting NSG mining results. The grouping of navigation elements to NSGs succeeded
automatically classifying NSGs should be developed. This will further improve
the precision of navigation element detection and the discovering
of navigation system. The other direction is the research on the
interpretation of NSGs. The current version of the prototype
already provides very interesting data. The catalogue of
navigation systems should be extended and revised. A method for
automatically classifying NSGs should be developed. Evaluating
how a content model can be retrieve from the classified NSGs and
how it can be visualized is also a very promising and interesting
field of research.

3. ONGOING WORK

There are two basic directions of further developing the proposed
approach. One is the improvement of the NSG mining algorithms.
The metrics we used and derived from our experiments should be
replaced by machine-learning methods. This will further improve
the precision of navigation element detection and the discovering
of navigation system. The other direction is the research on the
interpretation of NSGs. The current version of the prototype
already provides very interesting data. The catalogue of
navigation systems should be extended and revised. A method for
automatically classifying NSGs should be developed. Evaluating
how a content model can be retrieve from the classified NSGs and
how it can be visualized is also a very promising and interesting
field of research.

4. REFERENCES

Advanced Techniques in Web Intelligence - I. Springer
Berlin / Heidelberg. 113-142.

Graphs for Representing and Analyzing Web Sites.


Bringing order to the Web. Proceedings of the 7th
International World Wide Web Conference (Brisbane,

[5] Yahoo! Design Pattern Library:
20.