A Web of Trails

Richard Wheeldon
School of Computer Science and Information Systems
Birkbeck College, University of London
London WC1E 7HX, U.K.
Supervisor: Mark Levene
E-mail: {richard,mark}@dcs.bbk.ac.uk

January 28, 2004

1 Abstract

Pages returned by Web search engines are often used as starting points for further navigation. The hyperlinks which users follow from the start page form a trail. Despite this, navigation possibilities are not considered by conventional search engines. Nor do search engines provide any support in suggesting trails for users to follow.

Users experience the "Navigation Problem", where they are said to be "Lost in Hyperspace", whenever they are navigating the Web (or some other hypertext) and are either unsure of where they are relative to another page, unsure of which link to follow in order to find what they're looking for, unsure of where they've been or unsure of where they will get to when they follow any given link.

This thesis describes a potential solution to the navigation problem – a "navigation engine", developed to provide memex-like information trails in response to a user's query. This unifies ideas from the information retrieval and hypertext communities. The main body of the thesis is presented in two parts, covering two major contributions – the implementation of this system and its application to the fields of web navigation, database search and program comprehension.

An existing algorithm, the Best Trail, has been refined and enhanced. The described implementation – the first effective, publically-accessible implementation of this approach to trail-finding – has been made possible by changes to the selection functions, the addition of new methods for removing redundant information and the introduction of a new link-based metric.

The Potential Gain metric improves the selection of starting points from which the Best Trail algorithm will find the trails by evaluating the potential of a page to provide future navigation opportunities. It is defined, for a given page, in terms of the fraction of trails of various lengths which start from that page. Algorithms for computing Potential Gain are shown along with techniques for using it to improve node selection. Experiments have been performed which show the effectiveness of the metric in increasing the likelihood of finding high scoring trails.

A comprehensive description of the architecture of the navigation system covers not only the basic components and the algorithm for index creation but also methods for handling extended query syntax, indexing content from non-standard file types and summarizing Web documents. It is shown how this system provides useful trails and enhances user navigation experiences on Web sites. The use of trails alleviates the navigation problem in two ways. Firstly, by semi-automating the navigation process, the user is able to follow a pre-determined path which can be assumed to be revelant. Secondly, by providing contextual information, the trails allow users to make more informed navigation decisions and hence avoid getting "lost".

Building on this work, a tool called DbSurfer has been developed which provides an interface to relational databases. Data is extracted in the form of an inverted index and a graph of foreign key dependencies. Together, these can be used to construct trails of information, solving the join discovery problem and allowing free text search on the contents. The free text search and database navigation facilities can be used directly, or can be used as the foundation for a customised interface.

The navigation problem also exists in automatically-generated program documentation and in the source code from which such documentation is typically generated. The final contribution in this thesis is a pair of tools, AutoDoc and AutoCode, for indexing Javadocs and Java source code, respectively. AutoCode shows trails according to graphs of coupling relationships – graphs which are shown to be Web-like in their scale-free topology.

The development of the navigation engine called for solutions to several interesting problems and leaves a potential solution for many more. This represents an important step on the path to Bush's *Web of Trails*.