

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

May 1, 2003

1 Abstract

Pages returned by Web search engines are often used as starting points for further navigation. The links 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.

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, which can together 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 relationship - 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.

2 Thesis Contents

I The Navigation Problem and Related Issues

1 Introduction

1.1 Motivation

1.2 Outline of the Thesis

1.3 Acknowledgements

2 Problems and Solutions

2.1 Introduction

2.2 A Brief History of Hypertext

2.3 The World Wide Web

2.3.1 Structure of the Web

2.4 Resource Discovery

2.5 Information Retrieval Techniques

2.5.1 Document Representation

2.5.2 Scoring Metrics

2.5.3 Evaluation Metrics

2.5.4 The Text REtrieval Conference

2.6 Web Page Metrics

- 2.6.1 HTML Tag Weighting
- 2.6.2 Landmark Nodes
- 2.6.3 Hubs and Authorities
- 2.6.4 PageRank
- 2.6.5 Combining Metrics
- 2.7 Non-Linear Search
 - 2.7.1 Question Answering
 - 2.7.2 Category-based Clustering
 - 2.7.3 Link-based Clustering
- 2.8 The Navigation Problem
- 2.9 Navigation Aids
 - 2.9.1 Link Suggestion
 - 2.9.2 Site Maps
 - 2.9.3 Trees and Graphs
- 2.10 Trail Recording
- 2.11 Trail Finding
- 2.12 Summary

II Implementation of a Trail-Based Navigation Engine

- 3 The Best Trail Algorithm
 - 3.1 Introduction
 - 3.2 Graph Traversal and Path Finding
 - 3.3 The Best Trail Algorithm
 - 3.4 Auxillary Functions
 - 3.5 Scoring Trails
 - 3.6 Filtering and Sorting
 - 3.7 Implementation
 - 3.8 Complexity
 - 3.9 Performance Evaluation
 - 3.10 Concluding Remarks and Future Work
- 4 Navigability and Starting Point Selection
 - 4.1 Introduction
 - 4.2 Potential Gain and Related Metrics
 - 4.3 Computing Potential Gain
 - 4.4 Experiments
 - 4.5 Correlations between Ranking Metrics
 - 4.5.1 Experimental Methods

- 4.5.2 Discussion
- 4.6 Improving Starting Point Selection
- 4.7 Concluding Remarks and Future Work
 - 4.7.1 Query Specific Potential Gain
- 5 Architecture of a Navigation Engine
 - 5.1 Introduction
 - 5.2 Top level Overview
 - 5.3 Webcases
 - 5.4 An Extensible Component Architecture
 - 5.4.1 TrailAlgorithm subclasses
 - 5.4.2 Post-Processing and Index Creation
 - 5.5 Advanced Features
 - 5.6 Improving File-Type Recognition
 - 5.7 Web Page Summaries
 - 5.7.1 A Summarization Algorithm
 - 5.7.2 Implementation
 - 5.7.3 Examples
 - 5.7.4 Titles and Short Titles
 - 5.8 Concluding Remarks and Future Work
 - 5.8.1 Multipage Search
 - 5.8.2 Partial Collection Ranking
 - 5.8.3 Incremental Crawling and Merging Webcases
- III Applications for Trail-Discovery
 - 6 Navigating the Web
 - 6.1 Introduction
 - 6.2 Navigation Interfaces
 - 6.2.1 NavSearch
 - 6.2.2 TrailSearch
 - 6.2.3 GraphSearch
 - 6.3 Web Site Examples
 - 6.3.1 SleepyCat
 - 6.3.2 Department of Trade and Industry
 - 6.3.3 University College London
 - 6.4 Evaluation
 - 6.5 Scaling to the Web
 - 6.5.1 Graph Partitioning

- 6.6 Merging Results
- 6.7 Concluding Remarks and Future Work
 - 6.7.1 User Studies and Evaluation
- 7 Search and Navigation in Database Systems
 - 7.1 Introduction
 - 7.2 Indexing Relational Databases
 - 7.2.1 Translating a Relation to a Full-Text Index
 - 7.2.2 Generating the Link Graph
 - 7.2.3 Computing Joins with Trails
 - 7.3 Extending the Navigation System
 - 7.4 Semi-Structured Data and XML
 - 7.5 Query Expressiveness
 - 7.6 Examples
 - 7.7 Evaluation
 - 7.8 Hierarchies, Taxonomies and Ontologies
 - 7.8.1 Generating the Link Graph
 - 7.8.2 Graph Construction Algorithms
 - 7.9 Related Work
 - 7.10 Future Work and Concluding Remarks
 - 7.10.1 Queries
 - 7.10.2 Presentation
 - 7.10.3 Security
 - 7.10.4 Closing the Loop
 - 7.10.5 Concluding Remarks
- 8 Trails and Program Comprehension
 - 8.1 Introduction
 - 8.2 OOP, Java and Object Coupling
 - 8.2.1 Object Oriented Programming
 - 8.2.2 Java
 - 8.2.3 Coupling
 - 8.2.4 The Jakarta Project
 - 8.2.5 Java Documentation Systems
 - 8.3 Autodoc
 - 8.4 AutoCode
 - 8.4.1 Architecture
 - 8.4.2 Source Code Display
 - 8.4.3 Examples

- 8.5 Power Law Distributions in Class Relationships
 - 8.5.1 Analysis Techniques
 - 8.5.2 Methods, Fields and Constructors
 - 8.5.3 Graph Structure
- 8.6 Evaluation
- 8.7 Concluding Remarks and Future Work
- 9 Future Work and Concluding Remarks
 - 9.1 Summary of the Thesis
 - 9.2 Personalization
 - 9.3 Meta Search
 - 9.4 The Software Navigation Problem
 - 9.5 Navigation in Virtual Environments
 - 9.6 Concluding Remarks