Efficient Index Maintenance for Text Databases

A thesis submitted for the degree of
Doctor of Philosophy

Nicholas Lester B.E. (Hons.), B.Sc,
School of Computer Science and Information Technology,
Science, Engineering, and Technology Portfolio,
RMIT University,
Melbourne, Victoria, Australia.

1st August, 2006

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; and, any editorial work, paid or unpaid, carried out by a third party is acknowledged.

Nicholas Maxwell Lester
School of Computer Science and Information Technology
RMIT University
1st August, 2006

Acknowledgments

There are many people who deserve thanks for their help, leading to the publication of this thesis.

During the course of my candidature, I have been truly privileged to work with three excellent supervisors. Firstly, Hugh Williams, who introduced me to research and without whom I would not have undertaken this doctorate. Justin Zobel, my primary primary supervisor, whose friendly conversation and enthusiasm guided me through good times and bad. Finally, Alistair Moffat, who demonstrates a standard of research to aspire to and doesn't mind sharing.

The Search Engine Group at RMIT also deserve my gratitude for the friendship and occasionally studious atmosphere they provided. I would like to particularly thank Simon Wilkinson, William Webber, Bodo von Billerbeck, Steven Garcia, Adam Cannane, Michael Cameron, and Yaniv Bernstein for our association. In addition, Bodo and Simon kindly volunteered to proof-read this thesis.

Naturally, none of this would have been possible without my family. My mother, Rosemary, is an amazingly capable woman and was my first intellectual role model. My father, Chris, who fostered my interest in computing and whose common-sense advice can be relied upon. I would also like to thank my sisters, Rebecca and Josie, for providing a sense of perspective whenever it was required, and my grandparents, Elaine and Ray, for offering the sanctuary of their home more often than I was able to escape there.

Finally, I would like to thank my girlfriend, Katherine Chong. Her patience and love have humbled me for four years, and I hope that they continue to do so.

Credits

Preliminary versions of material from this thesis have appeared in the following publications. Chapters 3, 4, 5 and 6 contain material previously previously published by Lester, Moffat, and Zobel [2005b]. Chapters 3, 4, and 5 also contain material previously published by Lester, Zobel, and Williams [2004], and by Lester, Zobel, and Williams [2005c], Chapter 8 contains material previously published by Lester, Moffat, Webber, and Zobel [2005a]. William Webber, co-author of [Lester, Moffat, Webber, and Zobel, 2005a], is a research student at the University of Melbourne. Material produced in collaboration with William is individually attributed in Chapter 8.

This work has been supported by grants from the Australian Research Council and the RMIT Virtual Research and Innovation Institutes. We are grateful to MSN Search for the provision of a large (anonimised!) query log containing government-related queries. The thesis was written in the `Vim` editor on Debian Linux, and typeset using the LaTeX 2_ε document preparation system. All trademarks are the property of their respective owners.

Contents

List of Figures

List of Tables

List of Algorithms

Abstract

All practical text search systems use inverted indexes to quickly resolve user queries. Offline index construction algorithms, where queries are not accepted during construction, have been the subject of much prior research. As a result, current techniques can invert virtually unlimited amounts of text in limited main memory, making efficient use of both time and disk space. However, these algorithms assume that the collection does not change during the use of the index.

This thesis examines the task of index maintenance, the problem of adapting an inverted index to reflect changes in the collection it describes. Existing approaches to index maintenance are discussed, including proposed optimisations. We present analysis and empirical evidence suggesting that existing maintenance algorithms either scale poorly to large collections, or significantly degrade query resolution speed. In addition, we propose a new strategy for index maintenance that trades a strictly controlled amount of querying efficiency for greatly increased maintenance speed and scalability. Analysis and empirical results are presented that show that this new algorithm is a useful trade-off between indexing and querying efficiency. In scenarios described in Chapter 7, the use of the new maintenance algorithm reduces the time required to construct an index to under one sixth of the time taken by algorithms that maintain contiguous inverted lists.

In addition to work on index maintenance, we present a new technique for accumulator pruning during ranked query evaluation, as well as providing evidence that existing approaches are unsatisfactory for collections of large size. Accumulator pruning is a key problem in both querying efficiency and overall text search system efficiency. Existing approaches either fail to bound the memory footprint required for query evaluation, or suffer loss of retrieval accuracy. In contrast, the new pruning algorithm can be used to limit the memory footprint of ranked query evaluation, and in our experiments gives retrieval accuracy not worse than previous alternatives.

The results presented in this thesis are validated with robust experiments, which utilise collections of significant size, containing real data, and tested using appropriate numbers of real queries. The techniques presented in this thesis allow information retrieval applications to efficiently index and search changing collections, a task that has been historically problematic.