Methods for Distributed Information Retrieval

Nicholas Eric Craswell

A thesis submitted for the degree of Doctor of Philosophy at The Australian National University

May 2000
© Nicholas Eric Craswell

Typeset in Palatino by \TeX{} and \LaTeX{}2ε.
Except where otherwise indicated, this thesis is my own original work.

Nicholas Eric Craswell
23 May 2000
This thesis is dedicated to my original PhD supervisor Paul Thistlewaite, who passed away in February 1999.
Acknowledgements

Thanks to the many people who contributed to this work, either directly by discussing the work or reviewing drafts, or indirectly by giving their time and friendship.

Thanks to my supervisor David Hawking for providing advice and support in so many areas. His research knowledge, technical understanding, common sense, good attitude and friendship improved my work and my PhD experience.

Thanks to the late Paul Thistlewaite for guiding me through my initial experiments in document filtering, my dabbling with aglets and on to the development of the work presented here. He guided me towards the principles of practicality, effectiveness and generality which I hope are embodied in this thesis.

Thanks to Peter Bailey, who joined our research group in late 1998. He has variously been my thesis advisor, research collaborator, temporary house-mate and friend. Thanks also to Chris Johnson for being on my panel and providing sound thesis advice.

Thanks to the postgrad discussion group attendees, including at various stages Steve Blackburn, Bill Clarke, Roland Goecke, Zhen He, Raj Nagappan, Rochelle O’Hagan, Sam Taylor, Linda Wallace and John Zigman.

Thanks to various other friends, collaborators and partners in crime: Francis Crimmins, Mark Grundy, Jason Haines, Steve Lawrence, Andy Macfarlane and Robert Umphelby.

Thanks to the Department of Computer Science for keeping me in offices, stationery and coffee, and for funding my Berkeley conference trip. Thanks to Richard Walker for TeX support and lending me useful books. For good Linux systems, the loan of a large monitor and miscellaneous advice, particularly during writeup, thanks go to Bob Edwards and the rest of the technical support group.

Thanks to the Cooperative Research Centre for Advanced Computational Systems (ACSys) for funding my travel to several conferences. Thanks to Mark Grundy and John O’Callaghan of ACSys for organising my AltaVista internship at Digital’s Gold Coast labs, and to Greg McCane for having me there. Thanks to Jan Bitmead for organising the ACSys Student Meetings, Professional Development Courses and everything else.

Thanks finally for support from family: Mum, Dad, Penny, the Craswells, the Gillespies and the Woltmanns.

This research was supported by an Australian Postgraduate Award PhD Scholarship. Additional financial support was provided in the form of a supplementary scholarship by the ACSys Cooperative Research Centre.
Presentational Conventions

A number of presentational conventions have been adopted in this thesis:

- Italics are used when a term is defined, but not thereafter.

- Spelling is according to the (Australian) Macquarie Dictionary, in particular the version which is searchable on the Web (http://www.dict.mq.edu.au/).

- References to archival publications are used in preference to Internet URLs.

- In cases when URL references are necessary, a URL is inserted in parentheses like this (http://pastime.anu.edu.au/nick/pubs/). Each URL refers to a Web page as it appeared on 23 May 2000.

- Equations from other studies are presented here using a standard notation, which is described in Table 2.2. While every effort has been made to avoid introducing errors, readers should consult the original studies for authoritative versions.
Abstract

Published methods for distributed information retrieval generally rely on cooperation from search servers. But most real servers, particularly the tens of thousands available on the Web, are not engineered for such cooperation. This means that the majority of methods proposed, and evaluated in simulated environments of homogeneous cooperating servers, are never applied in practice.

This thesis introduces new methods for server selection and results merging. The methods do not require search servers to cooperate, yet are as effective as the best methods which do. Two large experiments evaluate the new methods against many previously published methods. In contrast to previous experiments they simulate a Web-like environment, where servers employ varied retrieval algorithms and tend not to sub-partition documents from a single source.

The server selection experiment uses pages from 956 real Web servers, three different retrieval systems and TREC ad hoc topics. Results show that a broker using queries to sample servers’ documents can perform selection over non-cooperating servers without loss of effectiveness. However, using the same queries to estimate the effectiveness of servers, in order to favour servers with high quality retrieval systems, did not consistently improve selection effectiveness.

The results merging experiment uses documents from five TREC sub-collections, five different retrieval systems and TREC ad hoc topics. Results show that a broker using a reference set of collection statistics, rather than relying on cooperation to collate true statistics, can perform merging without loss of effectiveness. Since application of the reference statistics method requires that the broker download the documents to be merged, experiments were also conducted on effective merging based on partial documents. The new ranking method developed was not highly effective on partial documents, but showed some promise on fully downloaded documents.

Using the new methods, an effective search broker can be built, capable of addressing any given set of available search servers, without their cooperation.
Contents

Acknowledgements vii

Presentational Conventions ix

Abstract xi

1 Introduction 1
   1.1 Research approach ................................. 2

2 Distributed Information Retrieval 3
   2.1 Relationship with previous research ................ 3
   2.2 Problem description ................................. 5
      2.2.1 Clients, servers, brokers and users .......... 5
      2.2.2 Practical considerations ....................... 11
   2.3 Solutions .......................................... 14
      2.3.1 Server selection methods ...................... 14
      2.3.2 Results merging methods ....................... 22
      2.3.3 Search brokers ................................ 25
   2.4 Published evaluation experiments .................. 27
      2.4.1 Overview of information retrieval evaluation ... 27
      2.4.2 Selection evaluation ........................... 30
      2.4.3 Merging evaluation ............................. 32
   2.5 Summary and conclusion ............................. 33

3 New Methods and Hypotheses 35
   3.1 Using downloaded documents ....................... 35
   3.2 New server selection methods ...................... 37
   3.3 New merging methods ................................ 41
   3.4 Hypotheses ......................................... 43
   3.5 Summary and conclusion ............................. 43

4 Evaluation Methodology 45
   4.1 Selection evaluation ............................... 45
      4.1.1 Merit definitions may be incorrect ............ 46
      4.1.2 Server merit depends on other selected servers 47
      4.1.3 System level evaluation ....................... 49
   4.2 Merging evaluation ................................. 50
      4.2.1 Simulating input rankings ...................... 51
4.3 Summary and conclusion ............................................. 52

5 Server Selection Experiments ......................................... 53
  5.1 Method ................................................................. 53
  5.1.1 Search servers: documents and retrieval .................... 53
  5.1.2 The broker: selection and merging ......................... 57
  5.1.3 User model ......................................................... 59
  5.2 Results ................................................................. 59
  5.3 Discussion .......................................................... 62
  5.4 Further experiments ................................................ 65
  5.5 Conclusion .......................................................... 68

6 Merging Experiments ..................................................... 71
  6.1 Method ................................................................. 71
  6.1.1 Search servers: documents and retrieval .................... 71
  6.1.2 The broker: selection and merging ......................... 72
  6.1.3 User model ......................................................... 74
  6.2 Results ................................................................. 75
  6.3 Discussion .......................................................... 78
  6.4 Further experiments ................................................ 79
  6.5 Conclusion .......................................................... 84

7 Conclusions .............................................................. 85
  7.1 Methods .............................................................. 85
  7.2 Other contributions ............................................... 86
  7.3 The future .......................................................... 87
  7.4 Overall conclusion ................................................ 88

A Variation in Terminology .............................................. 89

Bibliography ............................................................... 91
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Document request</td>
<td>5</td>
</tr>
<tr>
<td>2.2</td>
<td>Simple search</td>
<td>6</td>
</tr>
<tr>
<td>2.3</td>
<td>Search broker network communication</td>
<td>7</td>
</tr>
<tr>
<td>2.4</td>
<td>Search broker information flow</td>
<td>9</td>
</tr>
<tr>
<td>2.5</td>
<td>vGLOSS scenarios</td>
<td>18</td>
</tr>
<tr>
<td>2.6</td>
<td>Propagating statistics for merging</td>
<td>23</td>
</tr>
<tr>
<td>2.7</td>
<td>Test collection example</td>
<td>28</td>
</tr>
<tr>
<td>3.1</td>
<td>Search broker with document download</td>
<td>36</td>
</tr>
<tr>
<td>3.2</td>
<td>Probe queries and test queries</td>
<td>38</td>
</tr>
<tr>
<td>3.3</td>
<td>Predicting effectiveness</td>
<td>39</td>
</tr>
<tr>
<td>4.1</td>
<td>Retrieving relevant documents may reduce broker effectiveness</td>
<td>48</td>
</tr>
<tr>
<td>5.1</td>
<td>WT2g server sizes</td>
<td>54</td>
</tr>
<tr>
<td>5.2</td>
<td>WT2g topic skew</td>
<td>54</td>
</tr>
<tr>
<td>5.3</td>
<td>Probe query results (heterogeneous servers)</td>
<td>60</td>
</tr>
<tr>
<td>5.4</td>
<td>Effectiveness estimation results (heterogeneous servers)</td>
<td>61</td>
</tr>
<tr>
<td>5.5</td>
<td>Probe query results (homogeneous servers)</td>
<td>66</td>
</tr>
<tr>
<td>5.6</td>
<td>Effectiveness estimation results (homogeneous servers)</td>
<td>67</td>
</tr>
<tr>
<td>6.1</td>
<td>Feature distance vs BM25 results (TREC-6)</td>
<td>77</td>
</tr>
<tr>
<td>6.2</td>
<td>BM25 (ref stats) vs BM25 (true stats) (TREC-6)</td>
<td>83</td>
</tr>
</tbody>
</table>