

UNIVERSITY OF CALIFORNIA

Los Angeles

**Tocorime Apicu: Design of an Experimental
Search Engine Using an Information Sharing
Model**

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Reginald Louis Walker

2003

UMI Number: 3089008

UMI[®]

UMI Microform 3089008

Copyright 2003 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346


PREVIEW

© Copyright by
Reginald Louis Walker
2003

The dissertation of Reginald Louis Walker is approved.



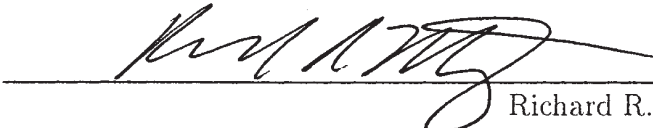
Gary B. Fogel



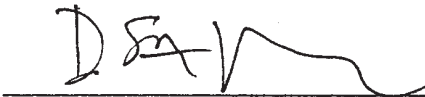
Christopher Lee



Jack W. Carlyle



Richard R. Muntz



D. Stott Parker, Committee Chair

University of California, Los Angeles

2003

*To my mother, father, grandparents, and my biological
lineage—people of Gullah descendant—
for the courage and fortitude needed to accomplish
this goal*

PREVIEW

TABLE OF CONTENTS

1	Historical Perspective on Search Engines	1
1.1	The Early Search Engines	1
1.2	Origin of Early Search Engines	2
1.3	Integrated Search Engines	4
1.3.1	WAIS Integrated Search Engine	4
1.3.2	Major Characteristics of the WAIS Search Engines	5
1.3.3	Major Characteristics of Current Search Engines	8
1.3.4	The Search Process	10
1.3.5	Use of the Biological Classification System	11
1.3.6	Naming Conventions	12
1.4	Improvements Needed by Current Search Engines	18
1.4.1	Precision and Recall	19
1.4.2	Search Engine Queries	20
1.4.3	Network Issues Associated with Search Engine Traffic	24
1.4.4	Information Retrieval Model for Reducing Useless Keyword Searches	26
1.5	Using an Information Sharing Model to Improve Current Search Engines	27
1.5.1	The Importance of Information Sharing	27
1.5.2	Replacing the Information Retrieval Model with an Information Sharing Model	28

1.5.3	Dissertation Roadmap	29
1.6	Related Work	30
1.7	Summary	32
2	Honeybee Information Sharing	34
2.1	The Social Hierarchy of Information Sharing among Honeybees . .	34
2.1.1	Information Sharing within a Bee Colony	34
2.1.2	Information Sharing within a Honeybee Ecosystem	36
2.2	Honeybee Information Sharing Methods	37
2.2.1	Tactile Communication	37
2.2.2	Auditory Communication	39
2.2.3	Chemical Communication Within the Colony	40
2.2.4	Chemical Communication Within the Colony Ecosystem .	43
2.3	Stochastic Information Fluctuations within the Honeybee Ecosystem	44
2.4	Background Information on the Characteristics of Honeybee Societies	45
2.4.1	Species of Bees	45
2.4.2	Foraging Honeybees	46
2.5	Navigational Skills of Foraging Honeybees	47
2.6	Hive to Hive Communication	48
2.7	Summary	49
3	The Tocatorime Apicu Information Sharing Model	50
3.1	Overview	51
3.2	A Model of Stored Information	52

3.2.1	Stochastic Fluctuations Within the Information Ecosystem	52
3.2.2	Design for Emulating the Honeybee Information Sharing Model and Ecosystem	55
3.2.3	Model Constraints for Simulating the Information Ecosystem	57
3.3	A Model for Transferring Stored Information	59
3.3.1	Hierarchical Task Topology for Retrieving Information from Within the Information Ecosystem	59
3.3.2	Hierarchical Communication Topology for Distributing Information	62
3.4	A Model for Mutating Stored Information	63
3.4.1	Stochastic Emulation of Communication between Nearest Neighbors	63
3.4.2	Stochastic Selection of Nearest Neighbors	64
3.5	A Model for Translating Stored Information for Value Judgements	65
3.5.1	Search Strategies for Finding and Translating Hidden Knowledge	65
3.5.2	Hierarchical Task Model for Translating Stored Information	67
3.6	Related Work	70
3.7	Summary	70
4	Adapting the Hierarchical Communication Topology of the Honeybee Information Discovery	72
4.1	Hierarchical Aspects of the Communication Topology	73

4.1.1	The Hierarchical Retrieval Process of the HTML Resource Discovery System	73
4.1.2	Optimization of the Hierarchical Retrieval Process	74
4.2	Network Issues Associated with the Retrieval of Web Documents .	75
4.2.1	Effect of Multimedia Transmissions	75
4.2.2	Possible HTML Sources	76
4.3	HTML Resource Discovery System Web Dispatchers	77
4.3.1	Web Dispatcher Constraints	77
4.3.2	Role of the Web Dispatchers	77
4.3.3	Characteristics of the Web Probe Dispatchers	79
4.3.4	Characteristics of the Web Scout/Forager Dispatchers	80
4.4	Computational Methods for Measuring Web Dispatcher Performance	81
4.4.1	Stochastic Computational Methods for Probe Dispatchers	81
4.4.2	Network Congestion Detection and Avoidance Mechanisms using RS Statistics for Scout Dispatchers	83
4.5	HRD Experimental Results	86
4.5.1	Experimental Environment	86
4.5.2	Operational Web Probe Dispatchers	86
4.5.3	Operational Web Scout Dispatchers	102
4.5.4	Operational Web Forager Dispatchers	113
4.6	Related Work	118
4.7	Summary	119
5	Adapting Stochastic Honeybee Search Strategies	120

5.1	Stochastic Optimization Techniques for Adapting Honeybee Search Strategies	120
5.1.1	Solution Variations Among the Optimization Methodologies	120
5.1.2	Development Focus of the Optimization Methodologies . .	121
5.2	Optimization of Parallel Information Gathering	126
5.2.1	Hybrid Random Stochastic Search Strategies	126
5.2.2	Optimization of the Web Page Indexing System	126
5.2.3	Document Retrieval Algorithms for Computing Stochastic Web Page Measurements	130
5.3	Benefits of Incorporating Stochastic Optimization Techniques . . .	131
5.3.1	Stochastic Computational Measures	131
5.3.2	Reductions in the Computational Effort	133
5.4	Related Work	133
5.5	Summary	136
6	Applying the Stochastic Honeybee Search Strategies to Information Sharing Indexing System	137
6.1	The Web Page Indexing System	138
6.2	The Tocarime Apicu Web Page Indexer	138
6.2.1	The Web Document Parser	138
6.2.2	Generation of the Canonical Web Page	140
6.3	Information Sharing Indexing Web Page Dispatcher	143
6.3.1	The Stochastic Regulatory Mechanism	143

6.3.2	Applying the Search Strategies to Stochastic Information Fluctuations	144
6.3.3	Formulation of Nearest Neighbor Clusters (NNCs)	148
6.4	ISI Experimental Results	151
6.4.1	Experimental Environment	151
6.4.2	Operational ISI System	156
6.4.3	Operational ISI Web Page Indexers	159
6.4.4	Operational ISI Web Page Dispatcher	165
6.4.5	Timing Results	172
6.5	Related Work	175
6.6	Summary	179
7	The Tocarime Apicu Experimental Search Engine	180
7.1	Task Distribution Model	180
7.2	Centralized Hierarchical Topology	184
7.2.1	Event Manager Scheme	184
7.2.2	Event Manager Load-Balancing Model	185
7.3	Experimental Results for the ISI System	186
7.3.1	Search Engine Case Study Load-Balancing Results	186
7.3.2	Effects of File Server Saturation	189
7.4	Related Work	192
7.5	Summary	194
8	Conclusions	195

8.1	Summary	195
8.2	Future Work	198
8.2.1	The Browser Reporting Interface (BRI) System	198
8.2.2	Relevance Feedback	199
8.2.3	Distributive File System	200
	References	202

PREVIEW

LIST OF FIGURES

1.1	The KDD architecture (courtesy of Adriaans and Zantinge 1996)	27
2.1	The noise of the direction indicator (the result of the degree of divergence versus distance in meters to feeder) declines with increasing distance in all species of honeybee (from J.L Gould and C.G. Gould). The dances are: a) round dance, b) tail-wagging dance, c) tail-wagging dance, and d) tail-wagging dance.	39
2.2	Inhibiting and stimulating effects of hive pheromones.	47
3.1	Web page connectivity as viewed by Google. The solid lines show the category connectivity of all Web pages that comprise the composite set of Web documents.	54
3.2	Web page connectivity in the Tocarime Apicu information ecosystem. The information ecosystem replaces this view of the Internet by viewing all Web page groupings as containing documents that connect all pages to and from the core pages.	58
3.3	The WWW as an information ecosystem as viewed by Web dispatchers.	60
3.4	Web page connectivity as viewed by the HRD system.	69
4.1	Weekly node access log for Web probes released (Version B). . . .	90
4.2	Weekly node access log for responding HTML ISPs (Version B). . .	90
4.3	Weekly node access log for DNS name resolutions (Version B). . .	91
4.4	Weekly node access log for Web probes released (Version C). . . .	97

4.5	Weekly node access log for responding HTML ISPs (Version C).	98
4.6	Weekly node access log for DNS name resolutions (Version C).	99
4.7	Self-similarity measures for Web scout dispatcher 0 (Version B).	105
4.8	Self-similarity measures for Web scout dispatcher 1 (Version B).	105
4.9	Self-similarity measures for Web scout dispatcher 2 (Version B).	106
4.10	Self-similarity measures for Web scout dispatcher 3 (Version B).	106
4.11	Self-similarity measures for Web scout dispatcher 0 (Version C Week 16).	109
4.12	Self-similarity measures for Web scout dispatcher 0 (Version C Week 17).	109
4.13	Self-similarity measures for Web scout dispatcher 1 (Version C Week 16).	110
4.14	Self-similarity measures for Web scout dispatcher 1 (Version C Week 17).	110
4.15	Self-similarity measures for Web scout dispatcher 2 (Version C Week 16).	111
4.16	Self-similarity measures for Web scout dispatcher 2 (Version C Week 17).	111
4.17	Self-similarity measures for Web scout dispatcher 3 (Version C Week 16).	112
4.18	Self-similarity measures for Web scout dispatcher 3 (Version C Week 17).	112
6.1	Distribution of Web Pages.	141

6.2	Supersedure emulation state diagram in the form of overlapping NNCs.	144
6.3	The flowchart for the Web page dispatcher.	146
6.4	A snapshot of the expandable search space for random NN.	147
6.5	A snapshot of the expandable search space for disjoint NNCs.	148
6.6	A snapshot of the expandable search space for overlapping NNCs.	149
6.7	Load-balancing model.	157
6.8	The collection mechanism for the BRI system.	158
6.9	Execution times for the workstations.	161
6.10	Execution times for the IBM SP2.	162
6.11	Speedup for the workstations.	163
6.12	Efficiency for the workstations.	163
6.13	Speedup for the IBM SP2.	164
6.14	Efficiency for the IBM SP2.	164
6.15	Version A through D—fitness measures for 1 indexer (sequential).	169
6.16	Version A—fitness measures for 2 through 8 indexers.	170
6.17	Version B—fitness measures for 2 through 8 indexers.	170
6.18	Version C—fitness measures for 2 through 8 indexers.	171
6.19	Version D—fitness measures for 2 through 8 indexers.	171
6.20	Timing results of the distinct versions.	173
6.21	Speedup of the distinct versions.	174
6.22	Efficiency of the distinct versions.	174

7.1	The architecture of the integrated search engine.	181
7.2	Communication topologies for the Tocorime Apicu project.	183

PREVIEW

LIST OF TABLES

1.1	Characteristics of WAIS incorporated in selected (crawler-based and human editor-based) search engines.	5
1.2	Major characteristics incorporated in selected (crawler-based and human-editor-based) search engines.	9
1.3	Comparison of current search engine indexer technology—Part 1.	13
1.4	Comparison of current search engine indexer technology—Part 2.	14
1.5	Comparison of current search engine indexer technology—Part 3.	15
1.6	Comparison of current search engine indexer technology—Part 4.	16
1.7	Comparison of current search engine indexer technology—Part 5.	17
1.8	Comparison of current search engine indexer technology—Part 6.	18
1.9	Advantages and disadvantages of various types of queries—Part 1.	21
1.10	Advantages and disadvantages of various types of queries—Part 2.	22
1.11	Advantages and disadvantages of various types of queries—Part 3.	23
2.1	Methods of information sharing within the colony.	35
2.2	Dances of the honeybee.	38
2.3	Pheromone production and its significance to other honeybees. . .	41
2.4	Environmental factors that affect the concentration of hive pheromones.	42
2.5	Characteristics of honeybee species.	45
2.6	The types and purpose of forage.	46
3.1	Comparison of the requirements of the information sharing and search engine models.	56

3.2	Benefits of emulating the purposive behavior of honeybees.	57
3.3	Benefits of simultaneously locating multiple information (forage) sources.	61
3.4	Comparison of the honeybee IS model, the KDD model, and the Tocarime Apicu IS model.	68
4.1	Sources of multimedia transmissions.	75
4.2	IP addresses.	76
4.3	Web probe/scout/forager translations of the honeybee tactile (dance) information sharing strategies within a bee colony.	79
4.4	Navigational model used by honeybee and Web foragers.	81
4.5	Cumulative access log summaries for Web probe dispatchers.	88
4.6	Weekly access log summaries for all Web probe dispatchers per week starting 25 Jan 2001 and terminating on 05 Mar 2001 (Versions A and B).	89
4.7	Cumulative access log summary for all Web probe dispatchers for one week, starting on 25 Jan 2001 and terminating 01 Feb 2001 (Version A).	92
4.8	Cumulative access log summary for all Web probe dispatchers per week, starting on 05 Feb 2001 and terminating 05 Mar 2001 (Version B).	93
4.9	Cumulative access log summary for all Web probe dispatchers per week, starting on 15 Oct 2001 and terminating 28 Jan 2002 (Version C)—Part 1.	94

4.10	Cumulative access log summary for all Web probe dispatchers per week, starting on 15 Oct 2001 and terminating 28 Jan 2002 (Version C)—Part 2.	95
4.11	Stochastic measures associated with Web probe dispatchers.	100
4.12	Web scout dispatcher results for customized routing (Version B).	103
4.13	Web scout dispatcher results for customized routing (Version C).	107
4.14	ISP responses to Web foragers inquiries based on customized routing (Version B).	116
4.15	ISP responses to Web foragers inquiries based on customized routing (Version C).	117
5.1	Comparison of the basic attributes of optimization methodologies.	122
5.2	EC methodologies that form the stochastic search strategies of the Tocarime Apicu engine.	127
5.3	Component optimization of the Tocarime Apicu ISI system.	128
6.1	The ISI system input parameters.	152
6.2	Best-case operational rates of the stochastic regulatory system for 1024 Web pages.	166
7.1	Load-balancing results for 512 Web pages using a network of workstations.	187
7.2	Load-balancing results for 512 Web pages using the IBM SP2.	188
7.3	Version A—Best-case sequential and parallel load balancing results for 1024 Web pages at iteration 200 with the time format hh:mm:ss.	189

7.4	Version B—Best-case sequential and parallel load balancing results for 1024 Web pages at iteration 200 with the time format hh:mm:ss.	190
7.5	Version C—Best-case sequential and parallel load-balancing results for 1024 Web pages at iteration 200 with the time format hh:mm:ss.	191
7.6	Version D—Best-case sequential and parallel load-balancing results for 1024 Web pages at iteration 200 with the time format hh:mm:ss.	192

PREVIEW

ACKNOWLEDGMENTS

The author wishes to thank D. Stott Parker, Walter Karplus, Gary B. Fogel, Richard R. Muntz, Jack W. Carlyle, and Christopher Lee for their direction and suggestions. The author acknowledges Elias Houstis, Ahmed K. Elmagarmid, Apostolos Hadjidimos, and Ann Catlin for their support, encouragement, and access to the computer systems at Purdue University. The author would like to acknowledge Janice Martin-Wheeler, Sandra Nadazdin, Lee Gillie, and numerous other proofreaders (and worker bees) for their support and encouragement. Partial support for this work came from the Raytheon Fellowship Program, Tapicu, Inc., and Honeybee Technologies.

PREVIEW

VITA

- 1958 Born, Youngstown, Ohio
- May 1981 BS Mathematics, Morris Brown College, Atlanta, GA
- June 1986 First place in the 58th National Technical Association's Student Symposium Competition, NASA Goddard Space Flight Center, Washington DC
- June 1986 MS Applied Mathematics, Atlanta University, Atlanta, GA
- 1991-1994 Recipient of NASA Graduate Researchers Fellowships
- August 1994 MS Computer Sciences, Purdue University, West Lafayette, IN
- 1996 - 2000 Recipient of Hughes Doctoral Fellowships
- April 2000 Third place in the 2000 Annual UCLA Student Research Poster Competition, UCLA Computer Department Research Review, University of California, Los Angeles, CA

PUBLICATIONS

R.L. Walker. "Simulating an Information Ecosystem within the WWW," in *Soft Computing Systems: Design, Management and Applications*, IOS Press. December 2002.

R.L. Walker. “Applying Evolutionary Computation Methodologies for Search Engine Development,” in *SEAL’02: Proceedings of the 2002 Asia-Pacific Conference on Simulated Evolution and Learning*, November 2002.

R.L. Walker. “Using Nearest Neighbors to Discover Web Page Similarities,” in *PDPTA’02: Proceedings of the 2002 International Conference on Parallel and Distributed Processing Techniques and Applications*, June 2002.

R.L. Walker. “Tocorime Apicu: Design and Validation of an Experimental Search Engine,” in *ITCom 2001: Commercial Applications for High-Performance Computing, Proceedings of SPIE—The International Society of Optical Engineering*, Vol. 4528, August 2001.

R.L. Walker. “Preliminary Study of Web Scouts/Foragers for a Bioinformatic Application: A Parallel Approach,” in *Computational Methods and Experimental Measurements X*, WIT Press. June 2001.

R.L. Walker. “Parallel Clustering System Using the Methodologies of Evolutionary Computations,” in *CEC2001: Proceedings of the 2001 IEEE Congress on Evolutionary Computation*, May 2001.

R.L. Walker. “Search Engine Case Study: Searching the Web using Genetic Programming and MPI.” *Parallel Computing*, Vol. 27(1/2):71-89, Elsevier Press, March 2001.

R.L. Walker. “Preliminary Development of a Genetically Enhanced High Perfor-

mance KDD System for a Biologically Inspired Application,” in *ICA3PP 2000: Proceedings of the 4th International Conference on Algorithms and Architectures for Parallel Computing*, December 2000.

R.L. Walker. “Preliminary Development of a Parallel Search Engine Indexer Simulator using the Methodologies of Evolutionary Computations and KDD,” in *CSMA 2000: Proceedings of the 2nd Conference on Simulation Methods and Applications*, October 2000.

R.L. Walker. “Dynamic Load Balancing Model: Preliminary Results for Parallel Pseudo-Search Engine Indexers/Crawler Mechanisms Using MPI and Genetic Programming,” in *VECPAR’2000. Lecture Notes in Computer Science 1981*, June 2000.

R.L. Walker. “Dynamic Load Balancing Model: Preliminary Assessment of a Biological Model for a Pseudo-Search Engine,” in *IPDPS 2000 Workshops, Lecture Notes in Computer Science 1800*, May 2000.

R.L. Walker. “Preliminary Development of an Indexer Simulator for a Parallel Pseudo-Search Engine,” in *ASTC 2000: Proceedings of the 2000 Advanced Simulation Technologies Conference*, April 2000.

R.L. Walker, M.Y. Ivory, S. Asodia, and L. Wright-Peg. “Study of Search Engine Indexing and Update Mechanisms: Usability Implications,” in *CATA-2000: Proceedings of the ISCA 15th International Conference on Computers and Their Applications*, March 2000.