Encrypted Search: Enabling Standard Information Retrieval Techniques for Several New Secure Index Types While Preserving Confidentiality Against an Adversary With Access to Query Histories and Secure Index Contents

By Alexander R. Towell, Bachelor of Science

A Thesis Submitted in Partial
Fulfillment of the Requirements
for the Degree of
Master of Science
in the field of Computer Science

Advisory Committee:

Hiroshi Fujinoki, Chair

Gunes Ercal

Tim Jacks

Graduate School
Southern Illinois University Edwardsville
September, 2015

ABSTRACT

ENCRYPTED SEARCH: ENABLING STANDARD INFORMATION RETRIEVAL TECHNIQUES FOR SEVERAL NEW SECURE INDEX TYPES WHILE PRESERVING CONFIDENTIALITY AGAINST AN ADVERSARY WITH ACCESS TO QUERY HISTORIES AND SECURE INDEX CONTENTS

by

ALEXANDER R. TOWELL

Chairperson: Professor Hiroshi Fujinoki

Encrypted Search is a way for a client to store searchable documents on untrusted systems such that the untrusted system can *obliviously* search the documents on the client's behalf, i.e., the untrusted system does not know what the client is searching for nor what the documents contain. Several new secure index types are designed, analyzed, and implemented. We analyze them with respect to several performance measures: confidentiality, time complexity, space complexity, and search retrieval accuracy. In order to support rank-ordered search, the secure indexes store frequency and proximity information. We investigate the risk this additional information poses to confidentiality and explore ways to mitigate said risk. Separately, we also simulate an adversary who has access to a history of encrypted queries and design techniques that mitigate the risk posed by this adversary.

KEYWORDS: (Encrypted Search, Information Leaks, Perfect Hash Filter, Query Obfuscation, Secure Indexes)

ACKNOWLEDGEMENTS

I would like to recognize my better half, Kimberly Wirts, for her patience and encouragement.

I would like to thank my parents, Bill and Sharon Towell, for their constant support throughout the entire process.

I would like to express my gratitude to my thesis advisor, Prof. Hiroshi Fujinoki, for introducing me to my thesis topic and lending me his time and support.

I would like to thank the rest of my thesis committee, Prof. Gunes Ercal and Prof. Tim Jacks, for their insights and challenging questions.

TABLE OF CONTENTS

ABSTRACT ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES vi

LIST OF TABLES ix

Chapter

I. INTRODUCTION 1

II. REVIEW OF LITERATURE 3

Confidentiality 3
Information Leaks 6
Online and Offline Searching 7
Mapping Queries to Documents 9

III. RESEARCH OBJECTIVES 15

Secure Indexes 15
Relevancy Metrics 30
Information Leaks 34

IV. EXPERIMENTS 41

Inputs 41
Outputs 42
Platforms 43
Global Parameters 43
Adversary Simulation Results 44
Secure Index Results 50

V. EXTENSIONS 84

Set-Theoretic Queries 84
Fuzzy Set-Theoretic Search 84
Boolean Proximity Searching 86
Caching Results 87

VI. FUTURE WORK 88

Simulating an Adversary with Secure Index Access 88
Mitigating Access Pattern Leaks 88

Semantic Search..89
Topic Search (Classification)..90
Letter N-Grams and Word N-Grams 91
Learning Optimal Parameters 91

VII. CONCLUSIONS...92

REFERENCES 94

LIST OF FIGURES

Figure Page

1. Overview of Encrypted Search on a Secure Index Database 16
2. Overview of Secure Index Construction 17
3. Perfect Hash Filter (Using A Minimal Perfect Hash) 20
4. The Perfect Hash Filter Secure Index (PSI) 22
5. The PSI Block-Based Secure Index (PSIB) 24
6. The PSI Frequency Secure Index (PSIF) 25
7. The PSI Postings List Secure Index (PSIP) 26
8. The PSI Minimum-Pairwise Distance Secure Index (PSIM) 28
9. Experiment #1 44
10. Unique Obfuscations vs Accuracy of Adversary Using MLE Attack 45
11. Obfuscation Rate vs Accuracy of Adversary Using MLE Attack 46
12. Experiment #2 46
13. Secrets vs Accuracy of Adversary Using MLE Attack 47
14. Experiment #3 47
15. History with Secrets vs Accuracy of Adversary Using MLE Attack 48
16. History vs Accuracy of Adversary Using MLE Attack 48
17. Number of Secrets vs Accuracy of Adversary Using MLE Attack 49
18. Experiment #4 49
19. Vocabulary Size vs Accuracy of Adversary Using MLE Attack 50
20. Experiment #5 50
21. Location Uncertainty vs Accuracy of Top 10 BM25 MAP Search Results 51
22. Mean Average Precision of Random Results 52
23. Experiment #6 52
24. Location Uncertainty vs Accuracy of Top 10 Mindist* Search Results..........53
25. Experiment #7 53
26. Number of Secrets vs Compression Ratio 54

27. Number of Secrets vs Secure Index Build Time 54
28. Number of Secrets vs Secure Index Load Time 55
29. Experiment #8 55
30. False Positive Rate vs BM25 MAP 56
31. False Positive Rate vs Precision 56
32. Experiment #9 57
33. Obfuscations/Query vs BM25 MAP with 6 Terms/Query, 6 Words/Term 57
34. Obfuscations/Query vs BM25 Lag Time with 6 Terms/Query, 6 Words/Term 58
35. Obfuscations/Query vs BM25 MAP with 1 Term/Query, 2 Words/Term 58
36. Experiment #10 59
37. Obfuscations/Query vs Mindist* Mean Average Precision 59
38. Obfuscations/Query vs Mindist* Lag Time 60
39. Experiment #11 60
40. Page Count vs Secure Index Size 61
41. PSIB and BSIB Intersection for Pages/Document vs Secure Index Size 61
42. Experiment #12 61
43. Blocks per PSIB/PSIB vs Secure Index Size 62
44. Blocks per PSIB/PSIB vs Compression Ratio 63
45. Experiment #13 63
46. Documents per Corpus vs Corpus Size 64
47. Experiment #14 64
48. Pages per Document vs Build Time 65
49. A Closer Look a Pages per Document vs Build Time 65
50. Experiment #15 65
51. Documents per Corpus vs Total Build Time per Corpus 66
52. Experiment #16 66
53. Pages per Secure Index vs Secure Index Load Time 67
54. Experiment #17 67

55. Documents per Corpus vs Corpus Load Time 68

56. Experiment #18 68

57. Pages per Secure Index vs Mindist* Lag Time 70

58. Experiment #19 70

59. Location Uncertainty vs Absolute Location Error 71

60. Experiment #20 72

61. Location Uncertainty vs Mindist* MAP 72

62. A Closer Look at Location Uncertainty vs Mindist* MAP 73

63. Experiment #21 73

64. Pages per Secure Index vs BM25 Lag Time 74

65. A Closer Look at Pages per Secure Index vs BM25 Lag Time 74

66. Experiment #22 75

67. Location Uncertainty vs BM25 MAP with 2 Words/Term, 3 Terms/Query 75

68. Location Uncertainty vs BM25 MAP with 2 Words/Term, 2 Terms/Query 76

69. Experiment #23 76

70. Pages per Secure Index vs Boolean Search Lag Time 77

71. A Closer Look at Pages per Secure Index vs Boolean Search Lag Time 77

72. Experiment #24 78

73. Percentage of Junk Terms vs Secure Index Compression Ratio 78

74. Percentage of Junk Terms vs BM25 MAP 79

75. Experiment #25 79

76. Relative Frequency Error vs BM25 MAP with 2 Terms/Query 80

77. Relative Frequency Error vs BM25 MAP with 1 Term/Query 80

78. Experiment #26 81

79. Compression Ratio vs Mindist* MAP 81

80. Experiment #27 82

81. Words per Term vs Precision and Recall 83

LIST OF TABLES

Table Page

1. Comparison of Confidentiality Techniques 3
2. Online Searching vs Offline Searching...7
3. Boolean Search vs Rank-ordered Search 9
4. BNF Query Grammar 9
5. BNF Hidden Query Grammar...16
6. Testbed System for Experiments 43
7. BNF Set-Theoretic Query Grammar..84
8. BNF Fuzzy Set-Theoretic Grammar 86