CONTENTS

Acknowledgement iii
Summary iv
List of Figures v
List of Tables vii
List of Abbreviations viii

1. INTRODUCTION 1
1.1 Introduction 1
1.2 Web information retrieval 2
1.3 Web crawlers and coherence problem 3
1.4 Search process and relevance problem 4
1.5 Objectives 6
1.6 Thesis organization 6

2. STUDY OF SEARCH ENGINE REPOSITORY 10
2.1 INTRODUCTION 10
2.2 SEARCH ENGINE REPOSITORY 10
  2.2.1 Structures 11
  2.2.2 Edifice methods 14
  2.2.3 Web crawlers 16
2.3 PROBLEM OF COHERANCE 26
  2.3.1 Needless requests 26
  2.3.2 Freshness metrics 28
  2.3.3 Factors involved in crawler based updation 29
2.4 Discussion 34

3. STUDY OF SEARCH ENGINE SEARCH PROCESS 36
3.1 INTRODUCTION 36
3.2 SEARCH PROCESS 36
  3.2.1 Phases of search process 37
  3.2.2 Role of phases on relevance of search results 38
3.3 PROBLEM OF RELEVANCE 53
  3.3.1 Context in search process 53
  3.3.2 Personalization of search 56
3.4 Discussion

4. TRAFFIC ADAPTIVE OPTIMUM UPDATING SCHEME
4.1 INTRODUCTION 60
4.2 SYSTEM DESIGN CHARACTERISTICS 60
  4.2.1 Autonomic characteristics of updating scheme 61
  4.2.2 Delta encoding for updating scheme 62
  4.2.3 Top down approach 69
4.3 ARCHITECTURE 70
  4.3.1 Conceptual view 72
  4.3.2 Detailed architecture 76
  4.3.3 Data structures 76
4.4 SUMMARY 79

5. TAOS PROTOTYPE AND TESTING
5.1 INTRODUCTION 80
5.2 FINITE STATE MACHINES 80
5.3 TAOS ALGORITHMS 82
  5.3.1 Policy management for AC node 83
  5.3.2 Policy management for AC server 85
5.4 TAOS RESULTS 86
  5.4.1 Testbed 86
  5.4.2 Freshness test 87
  5.4.3 Load test on servers and network bandwidth 92
5.5 CONCLUSION 102

6. SEMANTIC SEARCH PROCESS
6.1 INTRODUCTION 103
6.2 PROCESS DESIGN CHARACTERISTICS 103
  6.2.1 Semantic entity headers using HTTP extension framework 104
  6.2.2 Conceptual diagram 105
6.3 PROCESS ARCHITECTURE 106
  6.3.1 Phases of semantic search process 109
  6.3.2 Data structures 109
6.4 RESULTS

6.4.1 Relevance test 112

6.4.2 Impact of additional headers 119

6.5 Discussion 120

7. PERFORMANCE EVALUATION

7.1 INTRODUCTION 122

7.2 TRAFFIC ADAPTIVE OPTIMUM UPDATING SCHEME 122

7.2.1 Repository freshness 123

7.2.2 Resource utilization 126

7.3 SEMANTIC SEARCH PROCESS 136

7.3.1 Conventional search results 136

7.3.2 Semantic search results 138

7.4 Discussion 140

8. CONCLUSIONS AND FURTHER RESEARCH

8.1 SUMMARY OF CONTRIBUTIONS 143

8.2 CONCLUSIONS 145

8.3 LIMITATIONS 147

8.4 RECOMMENDATIONS FOR FURTHER RESEARCH 147

References 149

Appendices

A – Glossary of terms 155

B – UML diagrams 160

C - Test data 164

D – Sample screens 167