Ontology Based Semantic File Search Assistant

J H P C D Jayasekara

(Reg.No: MS19817590)
M.Sc. in Information Technology

Supervisor: Mr. Samantha Rajapaksha

April 2021

Department of Information Technology
Faculty of Graduate Studies and Research
Sri Lanka Institute of Information Technology

Contents

1 Introduction **4**
1.1 Background . . . 4
1.2 The problems with traditional file search . . . 5
1.2.1 File Organization . . . 5
1.2.2 File Search . . . 7
1.3 Proposed System . . . 7
1.3.1 Significance of the Study . . . 8
1.3.2 Applications . . . 8
1.3.2.1 Mathematics . . . 8
1.3.2.2 History and Social Studies . . . 10
1.3.2.3 English . . . 11
1.3.2.4 Science . . . 17
1.3.2.5 Religious and Cultural Coexistence . . . 19
1.3.3 Arts and Crafts . . . 19
1.3.3.1 Children Stories . . . 20
1.3.3.2 Entertainment . . . 20
1.3.3.3 Physical Education . . . 20
1.3.3.4 Art . . . 20
1.4 Aims and Objectives . . . 21
1.4.1 Main Objective . . . 21
1.4.2 Sub Objectives . . . 21
1.4.3 Research Questions . . . 21
1.5 Outline of the Thesis . . . 22

2 Literature Review **23**
- 2.1 Related Work 23
- 2.2 What is an Ontology 24
  - 2.2.1 RDF and OWL 25
- 2.3 Ontology Learning 25
  - 2.3.1 The Process of Ontology Learning 25
  - 2.3.2 Techniques used in Ontology Learning 26
    - 2.3.2.1 Techniques that are Linguistic 28
    - 2.3.2.2 Techniques that are Statistical 28
    - 2.3.2.3 Inductive Logic Programming 29
    - 2.3.2.4 Natural Language Processing for Ontology Learning 30
  - 2.3.3 Challenges in Ontology Learning 30
  - 2.3.4 Ontology Learning Tools 30
    - 2.3.4.1 ASIUM 30
    - 2.3.4.2 Text-to-Onto 31
    - 2.3.4.3 TextStorm/Clouds 32
    - 2.3.4.4 CRCTOL 32
- 2.4 SPARQL 32
- 2.5 Natural Language to SPARQL 32
  - 2.5.1 Natural Langauge to SPARQL Tools 33
    - 2.5.1.1 PAROT 33
    - 2.5.1.2 PANTO 34
    - 2.5.1.3 DEANNA 34
    - 2.5.1.4 AutoSPARQL 34
    - 2.5.1.5 AquaLog 34

3 Design of the Research **36**
- 3.1 Type of study 36
- 3.2 Data collection method and research instruments 36
- 3.3 Intended data analysis techniques 37

4 Project Timeline **38**

5 **Methodology** **39**
- 5.1 Ontology Learning Engine . . . 39
  - 5.1.1 Support and Confidence . . . 41
    - 5.1.1.1 Support . . . 41
    - 5.1.1.2 Confidence . . . 42
- 5.2 System Frontend . . . 42
- 5.3 NL to SPARQL Engine . . . 42
  - 5.3.1 How to Implement Negation . . . 44
  - 5.3.2 How to Implement OR . . . 45
- 5.4 File Metadata Database . . . 45
- 5.5 Ontology Search Module . . . 45
- 5.6 Ranking Algorithm . . . 47
- 5.7 Prototype Development . . . 48

6 **Evaluation** **49**
- 6.1 Accuracy . . . 49
- 6.2 Performance . . . 50
- 6.3 System Evaluation . . . 51
- 6.4 Discussion . . . 52

7 **Conclusion** **54**

Bibliography **59**

List of Figures

1.1 Ontology Based File Search 5

2.1 Ontology Learning Layer Cake 26
2.2 Ontology Learning Techniques 27
2.3 Comparison of Ontology Learning Tools 31
2.4 Comparison of Natural Language to Structured Query Tools 33

5.1 System Architecture 40
5.2 Sample Ontology 41
5.3 Support Example 41
5.4 Confidence Example 42
5.5 System Frontend 43
5.6 Formulas used in NL to SPARQL conversion 44
5.7 Ontology Search Algorithm 46
5.8 Ranking Algorithm 47

6.1 Accuracy of Results Based the complexity of Queries 50
6.2 Performance of Different Types of Queries 51
6.3 Evaluation of the System 53

List of Tables

4.1 38
6.1 52

Declaration

I hereby declare this submission is my own work and nothing has been copied.

Certified by

Signature : ....

Date : ....

The above candidate has carried out research for the M.Sc. thesis under my supervision.

Name of Supervisor : Mr. Samantha Rajapaksha

Signature : ....

Date : ....

Abstract

Many computer users do not give much attention to the organization of files into proper folders when saving them. Even if they do, with the increase of the number of files it becomes very difficult to locate them quickly without extensive search. If the files are saved with proper organization, it will be easier to search for a file. But the users still have to remember in which folders the files are saved. This research takes a different approach to saving and searching for files. It focuses on creating a bot which is integrated into the operating system that takes care of saving, searching and opening files for the user. The bot accepts natural language queries and return file search results by querying an automatically created ontology. Further, the results are ranked by applying a ranking algorithm before presenting to the users. Though this concept can be applied to any domain the research will apply this to a set of Request for Comments (RFC) text files. The experiments done on the results show that the system has an accuracy level better than that of an ordinary file search in an operating system.

Express thanks for

I express my thanks for my Supervisor Mr.Samantha Rajapaksha sir who helped me to choose the proper research field and advising me through out the research project. Also, I thank everybody who helped me.

April 2021, Charith Jayasekara.