Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Improving Centruflow Using Semantic Web Technologies

A thesis presented in partial fulfillment of the requirements for
the degree of Master of Science in Computer Science at
Massey University, Palmerston North, New Zealand.

Jonathan Andrew Giles
2007

Abstract

Centruflow is an application that can be used to visualise structured data. It does this by drawing graphs, allowing for users to explore information relationships that may not be visible or easily understood otherwise. This helps users to gain a better understanding of their organisation and to communicate more effectively. In earlier versions of Centruflow, it was difficult to develop new functionality as it was built using a relatively unsupported and proprietary visualisation toolkit. In addition, there were major issues surrounding information currency and trust. Something had to be done, and this was a sub-project of this thesis.

The main purpose of this thesis however was to research and develop a set of mathematical algorithms to infer implicit relationships in Centruflow data sources. Once these implicit relationships were found, we could make them explicit by showing them within Centruflow. To enable this, relationships were to be calculated based on providing users with the ability to 'tag' resources with metadata. We believed that by using this tagging metadata, Centruflow could offer users far more insight into their own data.

Implementing this was not a straight-forward task, as it required a considerable amount of research and development to be undertaken to understand and appreciate technologies that could help us in our goal. Our focus was primarily on technologies and approaches common in the semantic web and 'Web 2.0' areas. By pursuing semantic web technologies, we ensured that Centruflow would be considerably more standards-compliant than it was previously. At the conclusion of our development period, Centruflow had been rather substantially 'retrofitted', with all proprietary technologies replaced with equivalent semantic web technologies. The result of this is that Centruflow is now positioned on the forefront of the semantic web wave, allowing for far more comprehensive and rapid visualisation of a far larger set of readily-available data than what was possible previously.

Having implemented all necessary functionality, we validated our approach and were pleased to find that our improvements led to a considerably more intelligent and useful Centruflow application than was previously available. This functionality is now available as part of 'Centruflow 3.0', which will be publicly released in March 2008. Finally, we conclude this thesis with a discussion on the future work that should be undertaken to improve on the current release.

Acknowledgements

> "I would have written a shorter letter, but I did not have the time."
>
> \- Blaise Pascal

It is my pleasure to thank the people and institutions that made this thesis possible.

Firstly I want to thank my wife Julia, who has supported me throughout the year. Undertaking this thesis required considerable time, and I thank her for allowing me to work as hard as I needed to to complete this thesis.

Thanks to my parents Ian and Sue who have always supported me in whatever endeavour I set out on, and for buying me that Apple IIe back in the late 80's for $5000. It introduced me to the wonderful world of computing, and got me my first (unpaid) job teaching my friends and teachers at my local kindergarten (at the age of four).

I wish to acknowledge my supervisor Dr. Jens Dietrich from Massey University for his time, support, and ideas during this project. Without Jens, this project would not have been as successful as it was, particularly in terms of the formulation of the necessary mathematical functions. I am by no means a mathematician.

The Foundation for Research Science & Technology financially supported this project through the provision of a Technology in Industry Fellowship (TIF). Additionally, I appreciate the support of Massey University, through their provision of both the Massey Scholar scholarship and the Massey Masterate scholarship.

This project could not have been implemented without the support of a large number of open source projects, of which there are too many to mention. The key projects I want to thank include:

- The Jena semantic web framework by Hewlett-Packard Laboratories Bristol
- Chris Bizer for his D2R Server, and D2RQ mapping language.

Finally, thank you for taking the time to read this thesis. I appreciate any thoughts you may have on this research, and invite you to contact me via email at Jo@JoGiles.co.nz.

Contents

1 Introduction **1**

1.1 The Big Picture 1

1.2 Research Overview 2

1.2.1 The Internet 2

1.2.2 Web 2.0 And Social Networking 4

1.2.3 The Semantic Web 5

1.2.4 Visualisation Technology 5

1.3 Centruflow 6

1.3.1 Templates 7

1.3.2 Centruflow Release Timeline 7

1.3.3 Configuring Centruflow 8

1.4 Project Objective and Scope 9

1.5 Business Requirements 10

1.6 Overview of Thesis 11

2 Background **13**

2.1 Semantic Web 13

2.1.1 XML 15

2.1.2 RDF 16

2.1.3 Ontology Languages 16

2.1.4 SPARQL 17

2.1.5 SWRL 21

2.1.6 What Is The Semantic Web? 22

2.1.7 Open and Closed-World Assumptions 23

2.2 Trust Systems 24

2.3 Semantic Similarity 25

2.4 Representation State Transfer (REST) 26

2.5 Tagging Systems 26

2.5.1 Tagging 26

2.5.2 Folksonomy 27

2.5.3 Tagging Architectures 29

2.5.4 Tagging User Interface 31

2.5.5 Inferred Relationships 35

2.5.6 Public and Private Tags 35

2.6 Reflections 37

3 Algorithm Theory And Implementation 39

3.1 Introduction 39

3.2 User Trust 39

3.2.1 User Profiles 40

3.2.2 User Profile Findings 43

3.2.3 Algorithm Implementation 43

3.2.4 Algorithm Summary 51

3.3 Resource Distance 52

3.3.1 Metrics 53

3.3.2 Distance 54

3.3.3 Metric Algorithm Basics 54

3.3.4 Trust Similarity and Distance 56

3.3.5 Tag Similarity and Distance 56

3.3.6 Resource Distance Algorithm Conclusion 59

3.4 Relation Disagreement 59

4 Software Implementation 61

4.1 Implementation of Similarity and Trust Functions 61

4.1.1 Brute-Force Computation 61

4.1.2 On-Demand Calculation 62

4.1.3 Our Approach 62

4.2 Tagging Architecture 63

4.2.1 Taggings Relational Database 63

4.2.2 Tags as OWL Classes 65

4.2.3 The Tagging Pipeline 66

4.2.4 Querying The Tagging System 68

4.3 Centruflow Server 69

4.3.1 Requirements of a Centruflow Server 69

4.3.2 Alternative Software 70

4.3.3 Handling Graph-Based Data 73

4.3.4 Handling Non-Graph-Based Data 77

4.3.5 Features Overview 78

4.4 Centruflow Client 79

4.4.1 Important Centruflow Concepts 79

4.4.2 Data Retrieval Using SPARQL 80

4.4.3 Data Translation 81

4.4.4 Data Integration 82

4.4.5 Search 83

4.4.6 Tagging User Interface 84

4.5 Centruflow Standalone Version 88

4.6 Centruflow Boot Loader 90

4.6.1 Short-Term Solution 90

5 Validation 93

5.1 Validation Process 93

5.2 Validation Results 96

5.3 Thoughts 97

5.4 Stress and Performance Test 98

6 Conclusion 103

6.1 Introduction 103

6.2 Conclusion 103

6.3 Future Work 104

6.3.1 Further Research 104

6.3.2 Algorithm Improvements 104

6.3.3 Software Improvements 105

6.4 Summary 106

References 107

Glossary 113

A Use Cases **115**

A.1 Use Cases 115

A.1.1 Tagging Entities (Bookmarking) 115

A.1.2 Tag Searching 117

A.1.3 Tag Untagging 118

A.1.4 Browsing Related Entities 119

B Resource Distance Examples **121**

B.1 Tagging Distance Examples 121

B.1.1 Example Set One 121

B.1.2 Example Set Two 124

B.1.3 Example Set Three 126

B.1.4 Example Set Four 129

List of Figures

1.1 A screenshot of Centruflow 3.0's main visualisation perspective. 6

2.1 The semantic web layer cake [29]. 14

2.2 The 'MySQLicious' tag database schema. 29

2.3 The 'Scuttle' tag database schema. 29

2.4 The 'Toxi' tag database schema. 29

2.5 The Del.icio.us 'add bookmark' screen. 32

2.6 The main Del.icio.us page 33

2.7 Viewing tags as a tag cloud. 33

2.8 Del.icio.us can infer relationships between tags based on peoples taggings. . 34

3.1 An example resource with the tags that belong to it (as well as the associated users and the timestamps at which they were applied). 45

3.2 An example of three resources with some tagging overlap (but with the timestamps stripped to reduce the number of taggings shown). 55

4.1 The tag transformation pipeline. 66

4.2 The tag preferences dialog. 68

4.3 The Centruflow Server SPARQL Architecture. 74

4.4 The Centruflow Server implementation using open standards and open source components. 75

4.5 The Centruflow Server REST Architecture. 78

4.6 How the Centruflow client and server communicate. 81

4.7 The search dialog. 84

4.8 The add tags dialog. 85

4.9 The dialog shown to users when they select a tag they wish to untag. A similar dialog is shown when a user disagrees or wishes to delete a tag. . . 86

4.10 A tag as it is shown to the user inside Centruflow 3.0. It has the text 'SemanticWeb', and has been applied to the 'Semantic Web' topic, and to a person called 'Sonia'. 88

4.11 A number of inferred relationships shown inside Centruflow 3.0 (all dashed lines are inferred). The inferred relationships are the visible result of the algorithms developed in chapter 3. 89

4.12 Centruflow Boot Loader Dialog shown to users who are not presently running Centruflow, and who do not already have an already installed Centruflow client. 92

5.1 Results of the performance/stress tests. 100

List of Tables

3.1 Summary of prototypical users. 44

3.2 Examples of the user consensus function. Note that the maximum user consensus value is 20. 47

3.3 Algorithm evaluation table for user consensus metric 48

3.4 Examples of the tagging activity algorithm. 51

3.5 Algorithm evaluation table for user tagging activity metric 51

3.6 The result of calling *taggings(r)* on the resources shown in figure 3.2. 55

3.7 Example *tagCount* based on figure 3.2 and table 3.6. 57

5.1 User groups and the number of users in each 94

5.2 Tags input by our prototypical users. 95

5.3 Untags input by our prototypical users. 96

5.4 The calculated trust values for our prototypical users. 96

5.5 Results from resource distance algorithm (distances closer to zero represent more highly related resources). 97

5.7 Results from a performance and stress test. 99

List of Listings

2.1 An example SPARQL SELECT query used by Centruflow to request all information for one specific URI. 18

2.2 An example SPARQL CONSTRUCT query. 18

2.3 An example SPARQL ASK query asking if there is any user with a foaf:name of "Alice". 19

2.4 The simplest SPARQL DESCRIBE query. 19

2.5 Another SPARQL DESCRIBE query asking for all information on a person with the email address "alice@org". 19

2.6 A SPARQL query using an order modifier. 20

2.7 A SPARQL query using an offset modifier a limit modifier and an order modifier. 20

2.8 An example SPARQL/Update query fixing a spelling mistake. 21

2.9 An example SPARQL/Update query deleting all books with a date before the year 2000. 21

2.10 A human readable SWRL rule. 21

2.11 The XML form of the human-readable SWRL rule. 22

4.1 The default SPARQL query used by Centruflow to request all information for one specific URI. 80

4.2 An example of the SPARQL query used to search the Centruflow Server. The <search-token-n> fields should be replaced with a single search paramter. . 84