Computer Science Department
Faculty of Computer & Information Sciences
Ain Shams University

Computational Intelligence Techniques for Big Data Analytics

A Thesis Submitted to Computer Science Department,
Faculty of Computer & Information Sciences
Ain Shams University, Cairo, Egypt

In Partial Fulfilment of the Requirements for
the Degree of Doctor of Philosophy in Computer Science

By

Mahmoud Ibrahim Elbattah
Doctoral Researcher
Ain Shams University

Under Supervision of

Prof. Dr. Abdel-Badeeh Mohamed Salem
Computer Science Department
Faculty of Computer & Information Sciences
Ain Shams University

Prof. Dr. Mohamed Roushdy
Computer Science Department
Faculty of Computer & Information Sciences
Ain Shams University

Prof. Dr. Mostafa Aref
Computer Science Department
Faculty of Computer & Information Sciences
Ain Shams University

April 2017

Declaration

This is to certify that this work has not been accepted in substance for any academic degree and is not being concurrently submitted in candidature for any other degrees.

Any other portions of this thesis for which the author is indebted to other sources are mentioned and explicit references are given.

Scholar Name: Mahmoud Elbattah

Signature:

Elbattah

Acknowledgements

This work would not have been possible without all who have helped me along the way. First, the author would like to dedicate this work to his parents for their unbreakable faith and moral support over long years. Their constant encouragement helped the author to continue to be persistent and diligent in his work.

Second, the author would like to express his sincere gratitude to the thesis advisors Prof. M.Roushdy, Prof. M.Aref, and Prof A.Salem for their guidance, advice, criticisms, encouragements and insight throughout the research development. In particular, Prof. A.Salem showed immense understanding and patience for the research project during difficult times.

Furthermore, the author would like to thank Prof. Dima Shepelyansky (Université Paul Sabatier Toulouse, France) who organised the Summer School for Advanced Sciences of Luchon in July 2014. The author believes that the knowledge gained from that summer school provided useful insights to address the research problem from a different angle. In particular, the lectures and tutorials provided by Prof. Andreas Kaltenbrunner (Technology Centre of Catalonia, Spain) on network analysis provided a solid foundation to develop an important part of the research project.

Next, the author wishes to thank the providers of the MOOCs (massive open online course) such as Coursera.org and edX.org. The author cannot be more fortunate to have gained a learning experience from such a supporting and knowledgeable community. Specifically, the author is grateful to Prof. Andrew Ng for his comprehensive course on Machine Learning provided on Coursera.org. Special thanks also to Pro.f Yaser S. Abu-Mostafa (California Institute of Technology) who provided the wonderful course "Learning from Data".

Equally important, the author would like to give special thanks to Dr. Steve Elston (Quantia Analytics) who delivered the "DAT203x Data Science and Machine Learning Essentials" module on edX. The author can confidently confirm that Dr Steve provided him with valuable technical knowledge on handling Big Data processing tasks using the Azure Machine Learning Studio.

Abstract

The world of Big Data continues to expand, and forge the landscape of decision making and analytics. Datasets are rapidly growing in size and complexity, and there is a pressing need to develop solutions to harness this deluge of data for producing useful insights. This study addresses the tasks in relation to storing, querying, analysing and visualising Big Data from a graph-based perspective. Through the study, datasets extracted from the immense knowledgebase of Freebase are utilised. Initially, a web-based tool for data visualisation was developed, named as FreebaseViz, for visually exploring the schema of Freebase data. The visualisation design is built upon node-link network layouts, which can facilitate exploring connectivity, visual search and analysis, and visualising patterns underlying the schema graph. FreebaseViz is claimed to enable users to interact with the schema visualisations to filter and drill into lower levels of detail, and highlight subsets of the schema graph. In addition, a graph database-oriented approach is embraced in a further bid to boost the visualisation query-ability using graph-based query operations.

Subsequently, the study conducted a graph-driven methodology for the analysis and visualisation of Freebase complex schema. Specifically, our methodology utilised Freebase schema objects in order to construct a directed weighted graph. The schema graph is employed to perform a modularity-based analysis in order to detect communities underlying Freebase data. In light of that, the detected communities were effectively used for the purpose of revealing unobserved or implicit domain relationships.

In terms of storing and querying large-scale datasets, a graph database-oriented approach is proposed, which considered Freebase data as a large graph. The proposed approach endeavoured to address the limitations encountered within traditional relational models. Furthermore, scalability and query efficiency of the approach are verified based on empirical experiments using a subset of Freebase data that comprised a large-scale graph consisting of more than 500K nodes, and 2M edges

Furthermore, the study addresses the problem of entity clustering within large-scale knowledge graphs with application to the knowledgebase of Freebase. Particularly, the clustering task is approached form a mere graph-driven perspective. Entities were aimed to be clustered based on structural similarity within a knowledge graph. In this manner, entities were clustered in an unsupervised fashion by matching their link-based structure rather than relational attributes.

Eventually, the study aimed to develop an approach for estimating the consistency of knowledgebase triples. The proposed approach was based on utilising machine learning in order to learn the graph-based patterns of the triples. Specifically, the study investigated the feasibility of training a model to learn triples patterns in terms of subject-predicate-object. The validity of the method was experimented using a relatively large-scale subset of Freebase data. The dataset incorporated about 10M triples, which contained 6M true patterns and 4M false patterns randomly generated. The study availed of the cloud platform of Microsoft Azure in order to conduct the large-scale machine learning experiments efficiently. On top of the Azure platform, an Apache Spark cluster was deployed to realise a distributed computing environment. The classifier model evidently demonstrated a relatively high accuracy. Broadly, the study endeavoured to present and emphasise the appropriateness of graph-based methods for dealing with Big Data scenarios in terms of storage, querying, visualisation, and predictive analytics.

Table of Contents

Chapter 1: Introduction **1**

1.1 Motivation 1
1.2 Study Objectives 1
1.3 Research Overview 1
1.4 Contributions of the Study 2
1.5 Thesis Organisation 3

Chapter 2: Background-Big Data and Analytics **4**

2.1 Introduction 4
2.2 Definitions of Big Data 5
2.3 Characteristics of Big Data 6
2.4 Sources of Big Data 8
2.6 Data Analytics 12
2.7 Cloud Computing 14
2.8 Summary 16

Chapter 3: Techniques and Technologies for Big Data Analytics **17**

3.1 Big Data Techniques and Technologies 17
3.2 The MapReduce Paradigm 20
3.3 Massively Parallel Processing 21
3.4. Graph Analytics 21
3.5 NoSQL Databases 22
3.6 In-Memory Analytics 22
3.7 Big Data Case Studies 23
3.8 Summary 23

Chapter 4: Visual Analytics: Designing a Tool for Interactive Exploration of Data Aided by Query-Driven Visualisations **24**

4.1 The Role of Visualisation in Data Analytics 24
4.2 Conceptualisation of Freebase Schema 25
4.3 FreebaseViz Architecture 26
4.4 Visual Design 27
4.5 Interaction Techniques 28
4.6 Visualisation Scenarios 30
4.7 Analysis of Observations 32
4.8 Related Work 34
4.9 Summary 35

Table of Contents (cont'd)

Chapter 5: Relationship Discovery-Finding Implicit Connections Underlying Large-Scale Datasets with Application to Freebase **36**

5.1 Introduction 36

5.2 Data Source: Freebase Knowledgebase 36

5.3 Freebase Data Model 37

5.4 Knowledge Representation in Freebase 39

5.5 Data Preparation 40

5.6 Analysis and Visualisation of the Schema Graph 44

5.7 Modularity-Based Analysis 48

5.8 Observations and Discussion 52

5.9 Methodology Limitations 53

5.10 Summary 54

Chapter 6: Big Data Storage and Query: Large-Scale Ontology Storage Using Graph Database **55**

6.1 Introduction 55

6.2 Ontology Storage Approaches: A Review 56

6.3 The Proposed Approach: Graph Database as Ontology Store 58

6.4 Selection of Graph Database System 59

6.5 Use Case: Ontologising Freebase 60

6.6 Empirical Results 63

6.7 Summary 64

Chapter 7:Large-Scale Entity Clustering Using Graph-Based Structural Similarity within Knowledge Graphs **65**

7.1 Introduction 65

7.2 Background: Knowledge Graphs 66

7.3 Motivation: Entity Clustering in Large-Scale Knowledge Graphs 67

7.4 Data Description 67

7.5 Computing Clusters 69

7.6 Clustering Experiments 73

7.7 Results and Discussion 75

7.8 Related Work 78

7.9 Methodology Limitations 79

7.10 Summary 80

Table of Contents (cont’d)

Chapter 8: Learning Graph-Based Patterns to Predict the Consistency of Knowledgebase Triples **81**

8.1 Introduction 81

8.2 Assessment of Linked Data Quality: A Review 81

8.3 Data Quality Issues of Knowledgebases 82

8.4 Proposed Approach: Learning Patterns from Knowledge Graphs 84

8.5 Data Preparation 84

8.6 Learning Algorithm: Random Forests 85

8.7 Computing Environment 87

8.8 Results 91

8.9 Summary 92

Chapter 9: Conclusions and Future Work **93**

9.1 Conclusions 93

9.2 Future Work 95

References 97

Appendix: Code Snippets 107

List of Figures

Figure 1.1: Research overview. 1

Figure 2.1: Various interpretations of Big Data dimensions. 7

Figure 2.2: Big Data Growth vs. Moore's Law. 9

Figure 2.3: Projected growth in unstructured and structured data. 10

Figure 2.4: Classification of Big Data security challenges. 11

Figure 2.5: A taxonomy of disciplines related to the practice of data analytics. 13

Figure 2.6: The spectrum of data analytics. 14

Figure 3.1: The CRISP-DM stages. 17

Figure 3.2: MapReduce architecture. 20

Figure 4.1: Example of the Freebase schema structure. 25

Figure 4.2: Overview of the FreebaseViz architecture. 26

Figure 4.3: The layout view of the FreebaseViz tool. 27

Figure 4.4: Visualisation Scenario: Finding dominant schema Types. 30

Figure 4.5: Visualisation Scenario: Category-filtered schema graph. 31

Figure 4.6: Visualisation Scenario: Type-filtered schema graph. 32

Figure 4.7: In-degree distributions in the schema graph. 33

Figure 4.8: The percentage of isolated Types within the 9 categories of Freebase, which have no connections with other Types. 33

Figure 5.1: An example of how Freebase deals with ambiguity and multiplicity of entities. 37

Figure 5.2: Example of the multi-faceted nature of entities, and how Freebase handles that in flexible manner. 38

Figure 5.3: Hierarchy of Freebase schema. 39

Figure 5.4: An example of knowledge graphs within Freebase. 40

Figure 5.5: Main components of the schema extraction process. 41

Figure 5.6: Sequence of schema extraction procedures. 41

Figure 5.7: Relational schema used for modeling the extracted Freebase schema. 42

Figure 5.8: An example of nodes linking within the schema graph. 43

Figure 5.9: Top 10 ranked Freebase types by in-degree. 44

Figure 5.10: Freebase schema graph with emphasis on high ranked in-degree nodes.. 45

Figure 5.11: Top 10 ranked types by Eigenvector centrality. 45

Figure 5.12: Visualisation of Freebase schema graph with emphasis on nodes of significantly high centrality. 46

Figure 5.13: Domain-based visualisation of the schema graph. 47

List of Figures (cont'd)

Figure 5.14: Top 10 ranked Freebase domains with respect to node in-degree. 48

Figure 5.15: Visualisation of the detected communities. 49

Figure 5.16: MQL code example. 50

Figure 5.17: Plotting Jaccard similarity indices with regard to the detected communities. 52

Figure 5.18: Overlaps between the detected communities. 53

Figure 6.1: Stages of the review process. 57

Figure 6.2: An example of how data is structured within the ontology graph. 62

Figure 6.3: The property graph data model of Neo4j. 62

Figure 7.1: A simple knowledge graph. 66

Figure 7.2: An illustration of Knowledge graphs in Freebase. 68

Figure 7.3: An example of graph-based clustering in a simple knowledge graph. 71

Figure 7.4: Example of R script generated. 74

Figure 7.5: Histogram of the density of clusters computed based on sub-graph density for each cluster. 76

Figure 7.6: Density of entities inside clusters with respect to the original categories. 77

Figure 8.1: Examples of knowledgebase triples. 84

Figure 8.2: Random forest: Combining predictions using the mean. 86

Figure 8.3: Overview of the cloud-based computing environment. 87

Figure 8.4: Overview of Spark Distributed Architecture. 88

Figure 8.5: Overview of the cloud computing environment. 88

Figure 8.6: Components of Apache Spark. 89

Figure 8.7: Using Spark in the cloud through the Azure platform. 90

Figure 8.8: The AUC of the classifier. 91

Figure 9.1: Utilising clusters for topic exploration. 95

List of Tables

Table 1.1: Organisation of the thesis. 3

Table 2.1: Big Data definitions. 5

Table 2.2: Proposed definitions of analytics. 13

Table 2.3: Cloud service models. 16

Table 3.1: Big Data techniques. 18

Table 3.2: Big Data technologies. 19

Table 3.3: MapReduce components. 20

Table 3.4: Graph analytics approaches. 21

Table 3.5: The main categories of NoSQL databases. 22

Table 3.6: In-memory analytics technologies 22

Table 3.7: Examples of Big Data use cases. 23

Table 4.1 The components of the schema model of Freebase. 25

Table 4.2: Related visualisation tools. 34

Table 5.1: Summary of the extracted objects of Freebase schema. 42

Table 5.2: List of MQL queries associated with extraction procedures. 42

Table 5.3: Summary of detected communities. 50

Table 5.4: Freebase Categories and Included Domains. 51

Table 5.5: Jaccard similarity coefficients of detected communities. 51

Table 6.1: Key characteristics of ontology storage models. 56

Table 6.2: Investigative questions and motivations. 56

Table 6.3: Review of graph databases capabilities. 59

Table 6.4: Freebase Categories. 60

Table 6.5: Freebase categories and underlying domains. 61

Table 6.6: Description of query scenarios and response time. 63

Table 7.1: Questions of interest. 67

Table 7.2: Domains per category. The categories selected to partition the Freebase knowledge graph. 69

Table 7.3: Statistics of the dataset used by the study. 69

Table 7.4: Main categories of graph clustering algorithms. 72

Table 7.5: Computational complexity of graph clustering algorithms. 72

Table 7.6: Summary of computed clusters. 75

Table 7.7: Structure of computed clusters with respect to Freebase original categories of entities. 78

Table 8.1: Potential quality issues of Knowledgebases. 83

Table 8.2: Parameters of Random Forests. 86

Table 8.3: Head node computing capabilities. 89

Table 8.4: Worker node computing capabilities. 89

Table 8.5: Accuracy measures of the classifier. 91