LINKING INFORMATION RESOURCES
WITH AUTOMATIC SEMANTIC
EXTRACTION

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Humanities

2016

Daniel Joseph
Alliance Manchester Business School
# Contents

Abstract 11

Declaration 12

Copyright Statement 13

Acknowledgements 14

## 1 INTRODUCTION

1.1 Lightweight Ontologies and Concept Selection 16
1.2 Industry Case Study 20
1.3 Aim and Objectives 22
1.4 Research Questions 22
1.5 Statement of Contributions 23
1.6 Thesis Structure 23

## 2 BACKGROUND

2.1 Introduction 25
2.2 Formal Concept Analysis 26
2.3 Formal Concept Analysis As Ontology 30
2.4 Concept Lattice Reduction 34
   2.4.1 Pre-lattice reduction 34
   2.4.2 Post-lattice reduction 37
   2.4.3 Lattice Reduction Validation 43
2.5 Concept Definition by Patterns of UIC 45
2.6 Background of Recommender Systems 47
2.7 Conclusion 53
3 METHODOLOGY

3.1 Introduction .................................................. 54
3.2 Design Research .............................................. 55
3.3 Research Questions .......................................... 56
3.4 Case Study ..................................................... 60
3.5 Conclusion ..................................................... 62

4 EXTRACTING SEMANTICS USING CONCEPT RELEVANCY FOR FCA LATTICES

4.1 Introduction .................................................. 63
4.2 Creating Formal Contexts ................................. 64
   4.2.1 Attribute Selection .................................... 64
   4.2.2 Role of Links in Attribute Selection ............... 65
4.3 Selection of Relevant Concepts .......................... 67
4.4 Generating Concept Hierarchy ............................ 68
4.5 Semantic Extraction With Concept Relevancy .......... 70
4.6 Conclusion ..................................................... 71

5 CONCEPT RELEVANCY AND COLLAPSE INDEX

5.1 Introduction .................................................. 73
5.2 Introduction toCollapse Index .......................... 74
5.3 Derivation of Collapse Index ............................. 75
5.4 Demonstration of Collapse Index ....................... 79
5.5 Adding Uniform Object Instances ....................... 81
5.6 Lower Neighbours Distribution ......................... 89
5.7 Algorithm and Efficiency .................................. 96
5.8 Conclusion ..................................................... 99

6 COLLAPSE INDEX VALIDATION EXPERIMENTS

6.1 Introduction .................................................. 101
6.2 Choice Of Attributes from BT TV Dataset ............ 102
6.3 Pre-Processing of Dataset ................................. 103
6.4 Construction of Formal Context ......................... 104
6.4.1 Genre Formal Context ........................................ 104
6.4.2 Postal Area Formal Context ................................. 105
6.4.3 Postal District Formal Context ............................. 109
6.5 Experiment 1: Comparison Of Measures .......................... 110
   6.5.1 Preliminaries ............................................. 110
   6.5.2 Description of Experiment ................................ 110
   6.5.3 Hypotheses ............................................... 110
6.6 Experiment 1: Results .......................................... 111
   6.6.1 Lattice Summaries ....................................... 111
   6.6.2 Formal Context: Genres ................................ 112
   6.6.3 Formal Context: Postal Areas ........................... 113
   6.6.4 Formal Context: Postal Districts ........................ 115
6.7 Experiment 2: Consistency Across Samples ..................... 117
   6.7.1 Preliminaries ............................................. 117
   6.7.2 Description of Experiment ................................ 118
   6.7.3 Hypotheses ............................................... 119
6.8 Experiment 2: Results .......................................... 120
   6.8.1 Genre Lattice ............................................ 120
   6.8.2 Postal Area Lattice ...................................... 124
6.9 Experiment 3: Noise Recognition ............................... 127
   6.9.1 Preliminaries ............................................. 127
   6.9.2 Description of Experiment ................................ 128
   6.9.3 Hypotheses ............................................... 131
6.10 Experiment 3: Results ......................................... 132
   6.10.1 Results: Type I Noise ................................... 132
   6.10.2 Results: Type II Noise .................................. 134
6.11 Experiment 4: Efficiency of Collapse Index .................... 136
   6.11.1 Preliminaries ............................................. 136
   6.11.2 Description of Runtime Experiment ....................... 137
   6.11.3 Description of Steps Experiment ......................... 137
   6.11.4 Hypotheses .............................................. 138
6.12 Experiment 4: Efficiency: Results ............................ 139
6.12.1 Results: Runtime .............................................. 139
6.12.2 Results: Steps ................................................. 141
6.13 Conclusion ....................................................... 143

7 VALIDATION OF THE SEMANTIC EXTRACTION APPROACH144

7.1 Introduction ....................................................... 144
7.2 Basic Collaborative Recommender System ......................... 146
7.3 Recommender System: FCA Lattice as User Profile ............... 148
  7.3.1 User Profiles .................................................. 148
  7.3.2 Choice of Movie Characteristics ................................ 150
  7.3.3 Similarity Measure ........................................... 152
  7.3.4 Pruning Profiles ............................................ 155
  7.3.5 Predicted Rating ............................................. 156
  7.3.6 Recommender System and Proposed Semantic Extraction Ap-
       proach ............................................................. 157
  7.3.7 Architecture for Inclusion of RS in IP TV Service ............. 159
7.4 Conclusion ....................................................... 161

8 RESULTS OF APPROACH VALIDATION 162

8.1 Introduction ....................................................... 162
8.2 Experiment Design ............................................... 162
8.3 Dataset Separation .............................................. 165
8.4 Movie Descriptions ............................................. 166
8.5 Experiment Results .............................................. 168
  8.5.1 Preliminaries .................................................. 168
  8.5.2 Size of Nearest Neighbours Set ................................ 169
  8.5.3 Number of Movies in User Sample Set ......................... 170
  8.5.4 Pruning User Profiles with Collapse Index .................... 172
  8.5.5 Movie and Attribute Contributions ............................ 174
  8.5.6 Effect of UIC Descriptions ................................... 175
  8.5.7 RS in Implementation Set ..................................... 177
8.6 Conclusion ....................................................... 179
# List of Tables

2.1 Formal Context of Animals ........................................... 28  
2.2 Research Papers Describing Pruning Techniques .................. 43  
3.1 Example of BT TV Records ........................................ 61  
5.1 Formal Context for Collapse Index Demonstration ............... 79  
6.1 Formal Context for Genres ........................................ 105  
6.2 Frequency Matrix: Postal Area .................................... 106  
6.3 Tf-idf Matrix: Postal Area ........................................ 107  
6.4 Cosine Normalised Tf-idf Matrix: Postal Area .................... 108  
6.5 Formal Context After Tf-idf and Cosine Normalisation .......... 108  
6.6 Summaries of FCA Lattices ........................................ 111  
6.7 Summaries of Correlation Coefficients ............................ 116  
6.8 Correlation Across Lattices (Genre) ............................... 121  
6.9 Number Of Formal Concepts In Subcontexts ........................ 126  
6.10 Correlation Across Lattices (Postal Area) ....................... 126  
6.11 Formal Context for Genres ........................................ 128  
6.12 Random Value Assigned to Binary Relations ..................... 129  
6.13 Noisy Formal Context ............................................ 129  
6.14 Recall Using Type I Noise ......................................... 133  
6.15 Recall Using Type II Noise ........................................ 134  
6.16 Runtimes for Genre-based Lattice ................................ 139  
6.17 Runtimes for Postal Area-based Lattice ......................... 140  
6.18 Runtimes for Postal District-based Lattice ...................... 140  
6.19 Average Amount Of Times Each Formal Concept Is Accessed ... 142
List of Figures

1.1 Ontology Spectrum (taken from Giunchiglia and Zaihrayeu (2009)) . . 18
2.1 FCA Lattice of Animals . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
3.1 Design as Mapping for General Design Theory . . . . . . . . . . . . . 55
3.2 Conceptual Framework of Design Research in Information Science (von
Alan et al., 2004) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 57
4.1 Types of Links . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 66
4.2 Links: User Interaction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 67
4.3 Removal Of Formal Concept . . . . . . . . . . . . . . . . . . . . . . . . . 69
4.4 Formal Concept Removed . . . . . . . . . . . . . . . . . . . . . . . . . . . . 70
4.5 Model for Semantic Extraction With Pruned Lattice . . . . . . . . . . 72
5.1 Lattice Showing Concepts Of High Support but Low Relevance . . . . 80
5.2 Adding Object Instances to Largest Lower Neighbour . . . . . . . . . 83
5.3 Growth of $ci(A, B)$ with respect to $|\Gamma|$ . . . . . . . . . . . . . . 87
5.4 Adding Object Instances to Smaller Lower Neighbours . . . . . . . . 87
5.5 Adding Object Instances. Smaller Lower Neighbour(s) Now Larger . . 88
5.6 Lattices Showing $|C_n| = 1$ Maximises Collapse Index Value . . . . . 90
5.7 Lattices Showing Various Distributions of $\Theta$ . . . . . . . . . . . . . 92
5.8 Effect of Shared Subconcepts of Lower Neighbours on Collapse Index . 92
5.9 Lattices With Different $|C_i|$ Distributions Exhibiting Maximum $ci(A, B)$ 94
5.10 Distribution: 14-1-1-1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 95
6.1 Stability Index Correlations for Genre Lattice . . . . . . . . . . . . . . 112
Knowledge is a critical dimension in the problem solving processes of human intelligence. Consequently, enabling intelligent systems to provide advanced services requires that their artificial intelligence routines have access to knowledge of relevant domains. Ontologies are often utilised as the formal conceptualisation of domains, in that they identify and model the concepts and relationships of the targeted domain. However complexities inherent in ontology development and maintenance have limited their availability.

Separate from the conceptualisation component, domain knowledge also encompasses the concept membership of object instances within the domain. The need to capture both the domain model and the current state of instances within the domain has motivated the import of Formal Concept Analysis into intelligent systems research. Formal Concept Analysis, which provides a simplified model of a domain, has the advantage in that not only does it define concepts in terms of their attribute description but object instances are simultaneously ascribed to their appropriate concepts.

Nonetheless, a significant drawback of Formal Concept Analysis is that when applied to a large dataset, the lattice with which it models a domain is often composed of a copious amount of concepts, many of which are arguably unnecessary or invalid. In this research a novel measure is introduced which assigns a relevance value to concepts in the lattice. This measure is termed the Collapse Index and is based on the minimum number of object instances that need be removed from a domain in order for a concept to be expunged from the lattice.

Mathematics that underpin its origin and behaviour are detailed in the thesis showing that if the relevance of a concept is defined by the Collapse Index: a concept will eventually lose relevance if one of its immediate subconcepts increasingly acquires object instance support; and a concept has its highest relevance when its immediate subconcepts have equal or near equal object instance support.

In addition, experimental evaluation is provided where the Collapse Index demonstrated comparable or better performance than the current prominent alternatives in: being consistent across samples; the ability to recall concepts in noisy lattices; and efficiency of calculation. It is also demonstrated that the Collapse Index affords concepts with low object instance support the opportunity to have a higher relevance than those of high support.

The second contribution to knowledge is that of an approach to semantic extraction from a dataset where the Collapse Index is included as a method of selecting concepts for inclusion in a final concept hierarchy. The utility of the approach is demonstrated by reviewing its inclusion in the implementation of a recommender system. This recommender system serves as the final contribution featuring a unique design where lattices represent user profiles and concepts in these profiles are pruned using the Collapse Index. Results showed that pruning of profile lattices enabled by the Collapse Index improved the success levels of movie recommendations if the appropriate thresholds are set.
Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s Policy on Presentation of Theses.
Acknowledgements

Yeh! Shout out to all de Massive and Crew. From St. George’s ”Knowledge is Power” to Manchester’s ”Original Thinking Applied”. The sword man from Paran was sent forth to output a clear prospective view and be seen just by a few...people.

Shout out to my parents. Always supportive. Shout out to Sammy. Always setting the bar high. Shout out to Vivi. Made sure that I was ok. Shout out to Bradley and Dariane and Jordan. Setting the bar high for you guys. Shout out to Monique. Well always have Paris. Shout out to Tiffany. Always makes me smile. Shout out to Deidre. Need some of that homemade bread. Shout out to Russia. Really enjoyed giving you a hard time. Shout out to the K.M.U.L. King of N.Y. And yeah...Shout out to Chivone. You’re amazing. Best part of PhD.

Shout out to the University of Manchester for giving me a chance to study here both for my MSc and PhD. It was a great experience each time.

Shout out to the folks at BT. Really appreciate the opportunity and funding you provided to allow me to pursue a PhD. Shout out to John, Alistair, Roger, Ian Thurlow and Ian Kegel all of whom helped me in different capacities. Shout out to the entire Semantics Research team at BT Research Labs.

And finally...Shout out to my supervisors Babis and Nikolay. Was a great honour and privilege working under you guys. Really cannot thank you gentlemen enough for the opportunity and ever-available support.
Chapter 1

INTRODUCTION

Data is increasingly acknowledged as being a key driver of human economic and social advancement. It adds critical value to decision-making processes of individuals and organisations. Such is its importance as a corporate asset, Mayer-Schönberger and Cukier (2013, p.16) describe data as being “the oil of the information economy”. Consequently, a concerted effort in management and exploitation of data is an investment worthy of consideration for businesses in competitive industries.

The data which a business seeks to exploit may be datasets internal to the business - accumulated output of the business’ processes, or in other cases the data may be that of external content accessible through the Internet. In both cases the business is tasked with making sense of the data. While human analysis of data provides useful insight, the large scale increase in the volume of data within a variety of fields over the last two decades (Chen et al., 2014) has meant that software is increasingly given more responsibility and independence in utilising knowledge found in datasets to make decisions.

In fact this forms the basis of The Semantic Web - an evolution of the Internet conceived by Web visionary Tim Berners Lee, where software agents\(^1\) seek, retrieve, collaborate, and make decisions based on data retrieved from the Internet (Berners-Lee et al., 2001). One can imagine an individual instructing a Semantic Web agent to find an ideal vacation package for a weekend getaway. The agent would then navigate a variety of online airline, hotel, restaurant, car-rental, and weather datasets, then

\(^1\)Agents are software units that act autonomously or semi-autonomously on behalf of another entity (Bădică et al., 2011)
present a set of options to the individual.

To realise this vision of a machine-readable Internet, a necessary precursor is that the knowledge in datasets that the software is expected to utilise be captured, processed, and represented in a manner that would make the knowledge reusable and communicable. The heterogeneity of content in datasets and the need for interoperability between software in intelligent systems has meant that representing knowledge in a way shareable between machines is a priority. The view of Berners-Lee et al. (2001) and Horrocks (2008) was that this representation of knowledge would be implemented through assorted domain ontologies, each of which is created by experts intimately involved in their respective fields.

In the domain of Information Science ‘ontology’ is defined as an explicit specification of a domain conceptualisation (Gruber, 1995) - a conceptualisation being an abstract, simplified view of a domain. Individuals familiar with a domain would identify the domain’s concepts and their relationships and would, for the purposes of standardisation, represent these as a formal ontology using W3C’s Web Ontology Language (OWL) (McGuinness et al., 2004). Software parsing datasets would thus gain a semantic understanding of content by being able to associate entities found in a dataset with concepts in an available ontology.

Ultimately the expected achievements of the Semantic Web have so far proved to be somewhat overly optimistic. Although some industries such as the natural sciences have seen fit to develop detailed ontologies of their respective domains, the availability of ontologies across the Internet remains too limited to achieve the original vision of the Semantic Web (Shadbolt et al., 2006). Shadbolt et al. (2006) argue that the inability to arrive at large scale agent-based mediation is due to multiple factors, one of which is the slow pace of creating formal ontologies.

1.1 Lightweight Ontologies and Concept Selection

The development and maintenance of endorsed ontologies, accomplished manually even with increasingly sophisticated tools, remains a tedious, time-consuming, and costly task (Maedche and Staab, 2001) (Simperl et al., 2006). Shadbolt et al. (2006) acknowledge the difficulty of using ontologies noting that term (concept) definitions
1.1. LIGHTWEIGHT ONTOLOGIES AND CONCEPT SELECTION

are often in flux due to changes in norms brought about by subtle sociological processes of communities and practices.

The formal ontologies on which the Semantic Web was expected to be based are seen as ‘a-priori’ understandings of a domain and its concepts (Aberer et al., 2004). Such a viewpoint is one where the meaning of a concepts is decided in advance of deployment of an ontology. While this may be the most straightforward approach, the meaning of a concept is varied and it is unrealistic to expect that ontology engineers would have awareness of each interpretation beforehand. Moreover, the meanings of concepts are continuously in flux (Shadbolt et al., 2006) limiting the usefulness of a-priori type ontologies and calling for ontologies that reflect this dynamism.

In order to address these demands on resources induced by ontology construction, recommended solutions include the development of semi-automated processes to extract ontologies from datasets or large document corpora (Maedche and Staab, 2001). The creation of an ontology from text is based on the assumption that if a sufficiently large document corpus is available, an ontological representation of the domain has some implicit presence in the corpus (Cimiano et al., 2004); ontology-learning software thus has the responsibility of making the implicit explicit.

Given that any document corpus is limited in size, the conceptualisation of the domain is similarly limited as some concepts and relationships may not be present in the corpus. Secondly, although in some cases documents in the corpus may contain concepts and their explicit definitions, Cimiano et al. (2005) conclude that the more specialised the domain the less likelihood of documents containing definitions of domain concepts. They argue that, in the absence of these definitions, an alternative approach may be to garner knowledge of concepts by analysing their usage throughout the corpus. Embracing an approach where an ontology is learned from a corpus introduces a level of conjecture to ontology construction.

To accommodate uncertainty and reduce the complexity of ontology construction, one strategy is to target the creation of a lightweight ontology rather than aiming for a more elaborate ontological description of the domain. A lightweight ontology is seen as a set of basic relations along with “few unchanging terms that organize very large amounts of data” (Shadbolt et al., 2006, p. 99). These lightweight ontologies would eschew the details present in more complete ontologies but retain enough information
so as to be sufficiently fit for purpose.

Using Giunchiglia and Zaihrayeu (2009)’s ontology spectrum (Fig 1.1), an example of these lightweight ontologies which may serve in several roles as an explicit semantic descriptor of a domain, is that of a taxonomy. A taxonomy is the classification of objects into specially named groups based on class-subclass \((\text{is-a})\) relationship (Hakeem and Shah, 2004). While these less intricate taxonomies are not expected to produce detailed conceptualisations they still facilitate machine understanding of data semantics.

At the same time, although a simpler process than a more complete ontology, creating taxonomies may yet remain a fairly involved, time-consuming prospect. Therefore, as with more detailed ontologies, some automated or semi-automated solutions are of value for creating a taxonomy and using it to annotate instances of data in the domain of interest.

An elegant form for deriving these hierarchical taxonomies, and the focus of this thesis, is Formal Concept Analysis (FCA) (Wille, 2009). Formal Concept Analysis is a mathematical theory of data analysis that retrieves concepts from a set of objects and their individual attribute descriptions. Each concept is defined by a maximal set of attribute common to a maximal set of objects.

In addition to FCA’s ability to retrieve these concepts, concepts are structured in a concept hierarchy exhibiting the class-subclass relationship of hierarchical taxonomies.
This hierarchical representation may be expressed diagrammatically as a lattice or Hasse diagram (Wille, 2009). This concept lattice may also be transformed into a more formal representation of a concept hierarchy (Bendaoud, Napoli and Toussaint, 2008) using Description Logic (Baader, 2003) as necessary.

The duality in FCA concept definition offers the advantage in that not only is a class being defined by the concept’s attribute set, but object instances are simultaneously being assigned to their appropriate class(es). This is especially important as semantic reasoning cannot occur if the class memberships of object instances in the domain are unknown. As an example, the Semantic Web agent booking a vacation on one’s behalf would need to know not only that ‘a hotel is a type of accommodation’ but also that ‘The Waldorf Astoria is a hotel’.

However, despite its usefulness, Formal Concept analysis does have its affiliated caveats. At the forefront is the double-edged feature intrinsic to FCA where an FCA lattice represents every object-attribute relationship retrieved in a dataset (Roth et al., 2008b). By embodying each object-attribute relationship, especially those of a large dataset, an FCA lattice is likely to be comprised or a large amount of concepts. This is problematic as high volumes of concepts directly impact the legibility of the lattice, impeding semantic reasoning. Moreover many of these concepts are due to noise or anomalies in the dataset.

For concepts extracted via FCA, domain experts may intercede making judgements on legitimacy and relevancy of concepts (Bendaoud, Toussaint and Napoli, 2008). However for large datasets and more specifically Big Data (Chen et al., 2014), human analysis of legitimacy and relevancy of each individual concept in FCA lattices is a very precarious proposition in the face of a mass of concepts. Taking into consideration the fact that increasing the size of the dataset may result in an exponential growth of the number of concepts in an FCA lattice (Babin and Kuznetsov, 2012), not only should any automated\(^2\) oriented solutions be able to mitigate the large number of concepts in an FCA-derived taxonomy, but this must be accomplished in an efficient manner.

Existing solutions which target the pruning of FCA lattices such as the Support Value of Stumme et al. (2002) and the Stability Index of Roth et al. (2008b) assign

\(^2\)Although the solutions are described as automatic, the uncertainties inherent in semantics may still require assessment of relevancy by experts before final inclusion of concepts in a formal ontology: e.g. Bendaoud, Toussaint and Napoli (2008).
a numerical relevancy value to FCA-derived concepts. However the Support Value is biased towards concepts with high object instance representations and although the Stability Index provides more nuance by affording concepts with low object instance support to have high relevance values, its calculation is prohibitively inefficient for large datasets (Buzmakov et al., 2014a).

Another very different approach exists where the set of attribute-object relationships in the domain are reduced before the FCA lattice is generated. This typically involves the usage of matrix factorisation solutions which can be fairly complex computations, especially for large datasets (Snasel et al., 2008).

What is needed are alternative solutions to determining concept relevancy in FCA lattices which offer accurate representations of the relevancy of concepts in FCA lattices, maintain the benefits of the Stability Index, but are also efficient enough to be usefully incorporated into analysis of large datasets. This constitutes the main component of the thesis - the development of a concept relevancy measure, the Collapse Index (CI), which satisfies the aforementioned criteria, the validity of which is explored in the context of an industry case study. This case study is in the form of an industry dataset described in Section 1.2.

1.2 Industry Case Study

BT is a leading global telecommunications provider, operating in at least 170 countries worldwide and the leading communications service provider in the UK (Btplc, 2015). BT provides traditional telecommunication services such as fixed-line and mobile. In tandem with their traditional services BT has diversified their offerings to now include many broadband and Information Technology oriented services.

Among these services, and the case study of the thesis, is their TV product $BT TV$ (originally titled ‘BT Vision’). $BT TV$ is a subscription-based Internet Protocol Television (IPTV)$^{3}$ media delivery service. Users$^{4}$ of the service may access a variety of Freeview channels as well as rent or buy movies to view at their convenience (Productsandservices, 2015). Each subscriber to the service is provided with an Internet-connected set-top box which, in addition to delivering content to the user,

---

$^{3}$Television services delivered using the Internet protocol suite

$^{4}$User represents a household which may be comprised of one or more individuals
also retrieves a variety of data on the user’s utilisation of the service, then relays this data back to BT where it is archived.

Within this archived BT TV dataset, BT expects there lies knowledge which may add value to their television service. The premise of this expectation is that the accumulated dataset on usage of the BT TV service would likely have captured patterns of customer interaction; patterns of user characteristics; or patterns of user context; that are previously undocumented. Through the exploration of movies linked by these patterns, concepts defined by these patterns of User Interaction and Context (UIC) would emerge. These UIC-defined concepts likely indicate implicit non-UIC characteristics of the movies that are instantiations of these concepts.

With the additional insight on movies brought about by these concepts, recommendation of relevant movies to users could be more effectively realised, leading to improved user-satisfaction with the service. Everything else remaining constant, higher user satisfaction and greater engagement with the content BT TV offers should lead to greater revenue for BT.

Formal Concept Analysis is an ideal mechanism in this scenario due to its ability to extract concepts defined by shared attribute sets, as well as FCA’s immediate association of object instances with their concept(s). However, given the size of the BT TV dataset, application of FCA inevitably leads to lattices with vast quantities of concepts, many of which are largely inconsequential and only serve to hamper computational processing and semantic reasoning.

Restricting the relevant of these concepts to being only concepts with high object instance support, as per Stumme et al. (2002), does reduce the number of concepts to consider, however this would unfortunately disregard many interesting concepts with low object support. Granting the proposed Collapse Index the responsibility of determining the relevance of concepts in the BT TV dataset would reduce this bias towards concepts with high object instance support while accomplishing the calculations of concept relevance in a necessarily efficient way.
1.3 Aim and Objectives

The primary aim of this research is to obtain an efficient solution for determining the relevancy of concepts obtained through FCA, in order to enable automatic extraction of lightweight ontologies from large datasets. Emphasis on large datasets, in tandem with FCA processes, may lead to a concept hierarchy with a problematically large set of concepts. While many of these concepts represent key classifications in the concept hierarchy, others are superfluous and their existence may be due to noise in datasets. Previous popular techniques of determining the significance of concepts in the lattices have been computationally expensive or are unfairly biased towards classes with high object membership. There is a need for an FCA concept relevancy measure which is simple to calculate while affording concepts of low object instance support the possibility to possess higher relevance values than those of high support.

The second aim is to develop an approach where this concept relevancy measure may be utilised in the extraction of semantics from a dataset. This would be where FCA is the mechanism which identifies concepts and their relations in the dataset.

The final aim is to validate the concept relevancy measure and the approach through their successful application in a recommender system (RS) where movies are recommended to users based on the prior movies they have viewed.

1.4 Research Questions

In light of these aims, three (3) central Research Questions (RQ) are formulated.

\textbf{RQ 1.} Is there an alternative and efficient way of determining the relevance of a formal concept in an FCA lattice that is not overly biased to formal concepts with high object instance support?

\textbf{RQ 2.} For the concept relevancy solution developed, can an approach be developed for its useful inclusion in the extraction of semantics from a dataset?

\textbf{RQ 3.} Can an FCA concept relevancy measure along with an approach for its inclusion in semantic extraction be usefully incorporated in the design of a recommender system?
1.5 Statement of Contributions

Through the various tasks and processes involved in carrying out this research, there emerged several contributions to knowledge.

- This thesis introduces the Collapse Index, a novel solution which determines the relevance of a concept in an FCA lattice with respect to the minimum number of objects that need be removed from a formal context (domain) in order for the concept to collapse. The mathematical basis of its derivation is provided along with the mathematics that govern several aspects of the Collapse Index’s behaviour. Empirical evidence is provided of the Collapse Index’s efficiency, consistency across dataset samples, and ability to retrieve relevant concepts in noisy lattices.

- The thesis provides a description of an overall approach for the inclusion of the Collapse Index in the extraction of relevant concepts from large datasets.

- A user-based recommender system is introduced where the user profile is represented as the set of formal concepts of an FCA lattice, and these profile-lattices are pruned using the Collapse Index. Although based on the use of implicit data, the recommender system emphasises the approach of collaborative recommender systems. A revised taxonomy-similarity measure was used to compare user profiles with the aim being to find users with similar tastes.

1.6 Thesis Structure

To better facilitate the navigation of this thesis an outline is provided of the structure of the remainder of the document, and is as follows:

Chapter 2 has the responsibilities of providing a backdrop of several components of the thesis. This includes the discussion of literature that have informed the research’s motivation, as well as key FCA definitions.

The research methodology selected as an appropriate means of answering the research questions is discussed in Chapter 3. A description of the methodology is provided along with an alignment of the Research Questions of this thesis with the processes of the research methodology. Chapter 3 also provides a detailed description of
the industry dataset on which the proposed solutions of the research questions were applied.

Chapter 4 discusses an approach for extracting semantics from a large dataset using FCA where the concepts in the FCA lattice are pruned based on the use of a concept relevancy measure.

Chapter 5 is dedicated to the main contribution of this research which is the novel measure for assessing the relevancy of a formal concept. The derivation of this measure, termed the Collapse Index, and rules governing its behaviour are discussed in the chapter.

Chapter 6 contains descriptions and results of a variety of experiments conducted to assess various aspects of the Collapse Index measure. These experiments compare the performance of the Collapse Index in a series of tasks to the other popular concept relevancy solutions.

Chapter 7 provides a validation of the approach to semantic extraction discussed in Chapter 4. This takes the form of a recommender system designed to make use of FCA concepts selected by the proposed concept relevancy measure.

Chapter 8 is dedicated to results from the semantic extraction approach experiments conducted via the inclusion of the approach in the proposed recommender system solution.

Discussion and analysis of overall results are found in Chapter 9. The thesis then concludes in Chapter 10 with a summary of the work accomplished and a look towards further work.
Chapter 2

BACKGROUND

2.1 Introduction

This chapter reviews the background which frames the key components of the research. Formal Concept Analysis serves as the central mechanism we utilise for extracting concepts, as well as the platform for the proposed concept relevancy measure. For this reason FCA is the initial point of focus.

In Section 2.2 a formal description of FCA is given, entailing several of its key definitions which will be used at various points in the remainder of this paper. Further clarification is added to these definitions in Section 2.2 by the use of an example formal context and lattice.

With an understanding of what FCA is, the progression of FCA from its origin as a mathematical formalism to its utilisation as a tool for Knowledge Description and Data mining (KDD) is detailed. This is explored in Section 2.3. The theoretical rationale for its success as a means of ontological representation is also included, as are examples of FCA’s actual implementation as a domain descriptor in a diverse set of domains.

The problem of high numbers of resultant concepts is the motivation for our work on FCA. Section 2.4 focuses on prior research conducted in the area of identifying and/or removing low-relevance concepts from an FCA lattice. The literature for this area is grouped into three subsections a.) solutions where the steps to reduce the lattice are effected prior to lattice construction b.) post-lattice pruning and c.) validation of concept relevancy measures.
Section 2.5 meanwhile looks at how patterns of User Interaction and Context (UIC) may contribute to the definition of a concept.

Finally, a background on recommender systems is provided in Section 2.6, touching on the main aspects of recommender systems and research. This is done as a recommender system application was selected to serve as a means of validating a proposed approach to semantic extraction which utilises the Collapse Index concept relevancy measure.

2.2 Formal Concept Analysis

Formal Concept Analysis is a mathematical formalism which produces a lattice of concepts derived from the attribute description of a set of objects. The concepts in these FCA lattices are termed formal concepts, and are themselves individually defined by a set of objects and a set of attributes. A visual depiction of an FCA lattice is in the form of a Hasse diagram where the lines are representative of a partial order relation between a pair of formal concepts. Formal definitions of key aspects of Formal Concept Analysis are provided.

Definition 1. (Formal Context) A formal context is a set of the binary relationships between the set of objects and the attributes they individually posses. It is usually depicted in a table format where rows represent the set of objects and columns the set of attributes. An ‘X’ in a cell is indicative of the object of that row possessing the attribute of the corresponding column. Formally a formal context is defined as a triple $K := (G, M, I)$ where $G$ is the set of objects, $M$ the set of attributes, and $I$ the set of binary relations, $I \subseteq G \times M$.

Definition 2. (Formal Concept) A formal concept is a pair $(A, B)$, $A \subseteq G$ and $B \subseteq M$, such that $A = B^p$ and $B = A^q$, where $A^q = \{m \in M | \forall g \in A : (g, m) \in I\}$ and $B^p = \{g \in G | \forall m \in B : (g, m) \in I\}$. $(g, m) \in I$ reads as object $g$ has attribute $m$. The set of objects $A$ is referred as the extent or the extension of the formal concept whereas the set of attributes $B$ is referred to as the intent or intension of the formal concept. (Note that the operator $\preceq$ is also defined for $x \in G$ such $x^q = \{m \in M : (x, m) \in I\}$)

Definition 3. (Closed Attribute and Object Sets) Both operators $\triangleright$ and $\triangleleft$ individually create a Galois connection between the powerset of $G$, $\mathcal{P}(G)$, and the powerset
2.2. FORMAL CONCEPT ANALYSIS

of \( M, \mathcal{P}(M) \). Both of the composite operators \( \triangleright \triangleleft \) and \( \triangleleft \triangleright \) are closure operators for attribute and object sets respectively. i.e. For any \( A \subseteq G \), \( A \) is a closed object set if \( A^\triangleright = A \), and \( B \subseteq M \) is a closed attribute set if \( B^{\triangleright \triangleleft} = B \). The extent of a formal concept is as closed object set, and the intent a closed attribute set.

**Definition 4. (Subconcept and Superconcept)** A partial order relation, \( \leq \), for context \( K \) may exist between a pair of formal concepts and is described by subset relation between their extents or a superset relation between their intents.

Formally \((A_1, B_1) \leq (A_2, B_2) \iff A_1 \subseteq A_2 (\iff B_1 \supseteq B_2)\). Essentially \((A_1, B_1) \leq (A_2, B_2)\) if all objects of \( A_1 \) are found within the set \( A_2 \) or if all attributes of the set \( B_2 \) are within the set \( B_1 \). \((A_1, B_1)\) would be a subconcept of \((A_2, B_2)\), whereas \((A_2, B_2)\) is described as a superconcept of \((A_1, B_1)\).

**Definition 5. (Upper and Lower Neighbour)** If \((C, D) \leq (A, B)\) and there exist no \((E, F)\) such that \((C, D) < (E, F) < (A, B)\) then \((C, D)\) is a lower neighbour of \((A, B)\) and \((A, B)\) is an upper neighbour of \((C, D)\) - represented as \((C, D) \prec (A, B)\) and \((A, B) \succ (C, D)\) respectively.

**Definition 6. (Infimum and Supremum)** For each FCA lattice there exists a unique greatest common subconcept, \((\text{infimum})\) and a unique least common superconcept \((\text{supremum})\). Given that \( J \) is the set of formal concepts, the infimum and supremum are respectively defined as

\[
\bigwedge_{j \in J} (A_j, B_j) = \left( \bigcap_{j \in J} A_j, (\bigcup_{j \in J} B_j)^{\triangleright \triangleleft} \right) \quad \text{and} \quad \bigvee_{j \in J} (A_j, B_j) = \left( (\bigcup_{j \in J} A_j)^{\triangleright \triangleleft}, \bigcap_{j \in J} B_j \right).
\]

**Definition 7. (Concept Lattice)** Alongside the infimum and supremum, all other formal concepts and their partial order relations derived from the formal context \( K \) can be displayed as a Hasse diagram or lattice. \( L(K) = (K, \leq) \) is the concept lattice of the formal context \((G, M, I)\). The edges of the Hasse diagram are representative of the \( \prec \) relation whereas the meet and join in the lattice are denoted by \( \wedge \) and \( \vee \) respectively.

**Definition 8. (Attribute Implication)** Given \( B \subseteq M \) and \( D \subseteq M \), \( B \rightarrow D \), is an attribute implication if all the objects from \( G \) that have all attributes from \( B \) also have all the attributes from \( D \). This is the case if \( B^{\triangleright \triangleleft} \subseteq D^{\triangleright \triangleleft} \) or equivalently, \( D^{\triangleright \triangleleft} \subseteq B^{\triangleright \triangleleft} \). The set of all implications is summarized by the Duquenne-Guigues basis.
To lend further clarity to these FCA definitions an example of a *formal context* and its resultant lattice are provided. Here the *formal context*, $K$, shown in Table 2.1, is describing a subset of animals. This subset of animals present in the first column are the set of object instances $G$ of the formal context $K := (G, M, I)$, whereas the attribute set in the first row are the set of possible attributes $M$. The binary relations, $I$, between objects and the attribute set are the set of $X$s in the appropriate table cells.

Table 2.1: Formal Context of Animals

<table>
<thead>
<tr>
<th></th>
<th>flies</th>
<th>feeds young milk</th>
<th>feathers</th>
<th>warm-blooded</th>
<th>carnivorous</th>
</tr>
</thead>
<tbody>
<tr>
<td>COW</td>
<td></td>
<td>$X$</td>
<td></td>
<td>$X$</td>
<td></td>
</tr>
<tr>
<td>DOG</td>
<td></td>
<td>$X$</td>
<td></td>
<td>$X$</td>
<td>$X$</td>
</tr>
<tr>
<td>PENGUIN</td>
<td></td>
<td>$X$</td>
<td>$X$</td>
<td>$X$</td>
<td></td>
</tr>
<tr>
<td>RAVEN</td>
<td>$X$</td>
<td>$X$</td>
<td>$X$</td>
<td>$X$</td>
<td></td>
</tr>
<tr>
<td>FRUIT BAT</td>
<td>$X$</td>
<td></td>
<td></td>
<td>$X$</td>
<td></td>
</tr>
<tr>
<td>ALLIGATOR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$X$</td>
</tr>
</tbody>
</table>

Processing these binary relations between objects and attributes will produce the FCA lattice shown in Figure 2.1. In this visualisation, nodes in the lattice represent *formal concepts* while the edges connecting nodes are the partial order sub(super)concept relationship between various *formal concepts*. The uppermost node is the *supremum* whereas the lowest node is the *infimum*.

As an example of a *formal concept* we look at the 5\textsuperscript{th} node in the lattice. For this *formal concept* its *extent*, or set of object instances, is $\{\text{COW, DOG, FRUITBAT}\}$. This is our set $A \subseteq G$ in the representation of a *formal concept* $(A, B)$. Meanwhile the attribute set $B \subseteq M$ in $(A, B)$ is $\{\text{warmblooded, milk}\}$ and is the *intent* of the *formal concept*. The attribute set $B$ is the largest set of attributes in $M$ which are shared by all animals in $A$, and the largest set of animals in $G$ which share the attribute set $B$ is $\{\text{COW, DOG, FRUITBAT}\}$, demonstrating the *closure* property of Definition 3. Given the integral role human interpretation plays in Formal Concept Analysis (as described in Section 2.3) or more specifically plays in *formal concepts* (Wille, 2005), one may make the case that this 5\textsuperscript{th} *formal concept* is a representation of the concept of ‘mammals’.
In investigating the sub(super)concept partial order relationship in the lattice, formal concept 7 is selected. The 7th formal concept is a subconcept of the previous mammal formal concept, in that, as per Definition 4, the extent of the formal concept 7, \{Fruitbat\}, is a subset of the extent of the 5th formal concept, \{Cow, Dog, Fruitbat\}. Alternately the 7th formal concept is described as a subconcept of the 5th formal concept due to the fact that the attribute set of ‘mammals’, \{warmblooded, milk\}, is a subset of the attribute set of the 7th formal concept, \{flies, milk, warmblooded\}. Conversely, the 5th formal concept is a superconcept of the 7th formal concept.

Lower neighbours are illustrated with respect to formal concept 2. This second formal concept has, including itself, 9 subconcepts, (2, 4, 5, 6, 7, 8, 9, 10, 11). Of these subconcepts, the lower neighbours of 2 are 4, 5, and 6, given that there are no formal concepts between these three (3) formal concepts and formal concept 2.

Finally the Duquenne-Guigues attribute implications, as described Definition 8, is demonstrated with respect to the 5th and 2nd formal concepts. We may accept the implication that an animal that feeds its young milk is also warm-blooded, given that all objects(animals) which possess the attribute ‘milk’ (represented as the extent of
the 5th formal concept) also possess the attribute ‘warm blooded’ which is intent of the 2nd formal concept, the sole superconcept of formal concept 5. It is important to note that these implications are ‘correct’ only with respect to the formal context from which the FCA lattice originated.

2.3 Formal Concept Analysis As Ontology

Having described the key definitions of Formal Concept Analysis, attention is now shifted to an examination of the qualities of Formal Concept Analysis in the role of a lightweight ontology. This commences with information on FCA’s origins.

Formal Concept Analysis was invented by Rudolf Wille, first presented in 1982, and republished in Wille (2009). Initially seen as within the purview of mathematics (Priss, 2006), FCA eventually expanded into a varied set of domains Stumme (2002b).

While Stumme (2002b)’s rationale for FCA’s ease of transition across domains was with the intent of accounting for the then novel inclusion of FCA in Computer Science, the framework for the applicability of FCA is arguably of equitable bearing where the conceptualisation of a domain is required. The framework chosen by Stumme (2002b) was that of Davis et al. (1993), a set of five (5) principles of knowledge representation. A knowledge representation should be:

1. a medium of human expression,
2. a set of ontological commitments
3. a surrogate
4. a fragmentary theory of intelligent reasoning
5. a medium for pragmatically efficient computation

The first of two stated motivations underpinning Wille (1982)’s FCA lattice theory contribution, was to bridge communication gaps between lattice theorists and the non-mathematical audience who would be using said theory (Priss, 2006). The visuals of an FCA lattice are an effective and convenient expression of the underlying mathematics of lattice theory to a casual observer. Further, the second motivation of Wille
(1982) was the notion that the formalization of knowledge is subject to argumentation protocol, thus requiring formats that facilitate easily communicable representations of knowledge. A person being able to interpret and analyse the concept hierarchy in an FCA lattice facilitates negotiation on the formalisation of a domain of which the FCA lattice serves as a representation of. Human input could then consolidate a final representation or at least be embodied in changes in the domain’s representation.

In addition, for the second principle (ontological commitments), Stumme (2002b) argues that FCA formal concepts make for an axiomatic approach to the definition of what something is. This FCA knowledge representation is framed in terms of formal concepts and a concept hierarchy. These formal concepts, we recall, are composed of the intent (attribute set) and the extent (object set). Priss (2006) argue that although there have been valid arguments against the idea of a concept being defined in terms of its features, this classical view remains dominant as a knowledge representation in information systems because of its simplicity in management and implementation. Naming or providing a lexical label to these formal concepts is not a primary motivation of FCA, however this could be achieved when FCA is combined heuristically with other considerations.

FCA is applicable to Davis et al. (1993)’s third principle (surrogate) in that a formal concept in FCA acts as an internal substitute, on which surrogate reasoning may be conducted, of an external entity. One can also consider the formal context as not only a declaration of attributes of individual object instances but also as a means for restricting factors used to characterize these objects, giving some flexibility in object characterisation. Noteworthy is allowance also being made for factoring varying elements of uncertainties. Fuzzy FCA is a representative example of this (Shao et al., 2007).

Stumme (2002b, p.7) relates the fourth principle (intelligent reasoning) to Formal Concept Analysis explaining that FCA focuses on “reasoning with concepts” which entails instance implications, clauses, as well as hypothesis generation. The communicative properties of FCA mean that this reasoning is often done on a human level.

Finally, for the fifth principle, although Stumme (2002b) explain that the closure and partial order properties of FCA lattice theory make for efficient computing of tasks such as association rule mining and clustering, the increasing size of datasets
and the often corresponding large size of FCA lattices generated from large contexts may present problems moving forward. This is especially of interest in light of the human engagement in interpretation of FCA lattices stressed by Stumme (2002a) in the previous four principles of knowledge representation.

Despite this, the strengths of Formal Concept Analysis as a knowledge representation still remain, and for this reason Formal Concept Analysis has seen itself successfully utilized in a miscellany of disciplines. Examples include data analysis, information management, and knowledge management (Priss, 2006); linguistics (Priss, 2005); data in digital ecosystems (Fu, 2006); and domestic violence (Poelmans et al., 2009). FCA’s applicability across diverse domains is further illustrated in the following examples where FCA was chosen as the representation of their knowledge base.

Soon and Kuhn (2004) utilise FCA as a means of creating a user-action ontology in the domain of water monitoring. These user actions (e.g. assess, identify, select) are actions afforded by what are termed as affording objects in a text document. A single document was used as the source of actions and their corresponding affording objects facilitating manual extractions from the document. With the actions as attributes an FCA lattice was created from which semantic relationships between actions were sought. These semantic relations between actions, analysed from the ontology, were specific to verbs, including: Troponymy, Proper Inclusion, Backward Presupposition, and Causation. Due to the small size of the formal context, given that it originates from a single document, there was a manageable amount of formal concepts in the FCA lattice and no effort was made in or real need existed for pruning the lattice.

Citing a need for ontologies in the medical domain Jiang et al. (2003) set out to use FCA as ontological support in medicine. They apply Natural Language Processing (NLP) techniques to ‘discharge summaries’ documents in the domain of cardiovascular medicine, in order to retrieve relevant medical terms. With the documents as objects and the medical terms as attributes, the computational analysis of the FCA lattice generated was utilised primarily to find the semantic relationships between concepts (sets of terms) by calculating the association rules (Lakhal and Stumme, 2005) between pairs of concepts. Medical experts were then asked to comment on these associations as well as the validity of the formal concepts in the lattice themselves.

On this occasion the automation through NLP could potentially have led to many
superfluous binary relations and overly complex lattices. However as a way of reducing noise in associations derived from the lattice, the medical terms were corroborated by experts in advance of their inclusion as attributes of the discharge summaries.

Formal Concept Analysis is used by Bendaoud, Toussaint and Napoli (2008) for the purposes of both creating and merging two separate ontologies. From a text corpus related to astronomy, the names of celestial bodies as well as the verbs which coincide with these bodies were retrieved using NLP tools. These bodies (objects) and their related verbs (attributes) formed the formal context from which an FCA lattice was created to serve as an ontology of the domain. In addition to this, a separate FCA lattice was created from a database with these same bodies ascribed properties by experts. The ontology extracted from text and the ontology extracted from the database were both merged by combining their respective formal contexts and then creating a unified FCA lattice. They assumed that there should be overlap between formal concepts in both lattices with respect to object instances, as well as new previously unrecognised classes emergent from the inclusion of the corpus-derived FCA lattice. Several of these new classes were confirmed by experts in the field of astronomy.

As opposed to Jiang et al. (2003) who utilised expert opinion on the attributes before generating the FCA lattice in order to minimise the complexity of the resultant lattice, here Bendaoud, Toussaint and Napoli (2008) utilise expert opinion after the lattice is created to select the concepts that are legitimate. However the desire to automate semantic extraction in the face of large datasets means that direct human attention to obtaining the most relevant of concepts from FCA lattices is an increasingly unrealistic luxury. In the face of this, several automated solutions have been developed to help obtain only the most relevant concepts from an FCA lattice. Similar to the human-oriented solutions of Jiang et al. (2003) and Bendaoud, Toussaint and Napoli (2008) these automated solutions include techniques which simplify the formal context in advance of the creation of the FCA lattice and techniques which identify relevant concepts after the creation of the FCA lattice.
CHAPTER 2. BACKGROUND

2.4 Concept Lattice Reduction

From the previous examples Formal Concept Analysis proved itself to be adept in identifying concepts as well as structuring these concepts in a hierarchy. Despite these successes, usage of FCA is not without its challenges - a recurring example being how best to deal with large lattices.

A formal context can generate an FCA lattice where the number of formal concepts is exponential with respect to the size of the formal context (Babin and Kuznetsov, 2012), (Buzmakov et al., 2014). Therefore when concepts and taxonomies are being generated from large datasets, solutions are required that reduce the FCA lattice to its core formal concepts in order to facilitate more efficient semantic reasoning by humans or otherwise.

There have been so far, two fundamental approaches on the issue of lattice reduction. The first of these is that of reducing the complexity of the formal context itself before the generation of the FCA lattice. The second approach is the generation of the lattice from the original formal context then assigning each formal concept a numerical relevancy value after. Generally for the latter, the relevancy value is a rational number in the range of $[0, 1]$. A threshold value is then assigned for establishing which formal concepts are relevant and those which are not. The formal concepts of low relevance are then either ignored, removed from any final concept hierarchy, or simply recognised as such. Pre-lattice reduction is the first fundamental approach discussed.

2.4.1 Pre-lattice reduction

For pre-lattice reduction, given that a formal context can be recognized as a matrix or ‘a set of vectors’, matrix factorisation techniques commonly utilised in Latent Semantic Analysis (LSA) (Landauer et al., 1998) were employed to simplify the formal context before the application of FCA algorithms.

For Snasel et al. (2007), familiar methods of matrix decompositions were used to reduce the complexity of the formal context in advance of the generation of the FCA lattice. Singular Value Decomposition (SVD) as well as Non-negative Matrix factorization (NMF) were employed in the task of simplifying the formal context. Both techniques were able to achieve a simpler FCA lattice from the reduced formal
2.4. CONCEPT LATTICE REDUCTION

context by grouping similar object instances together. However, SVD grouped similar objects by removing secondary attributes from objects, while NMF grouped similar objects through the addition of secondary attributes to objects. This is as SVD orders the dimensions in a vector space some of the less significant dimensions for vectors are not represented in the final version of the matrix (formal context) (Wiemer-Hastings et al., 2004). Meanwhile for NMF, a vector is an additive combination of the base latent semantics (Xu et al., 2003). Snasel et al. (2008) follow up this research by looking at the loss of attribute implication rules in the now reduced formal contexts and lattices. They assess the loss in the implication base using both SVD and NMF matrix reduction techniques with and without noise added to the dataset.

Kumar and Srinivas (2010) use a similar approach to that of Snasel et al. (2007), Cheung and Vogel (2005) and Dias and Vieira (2010), where the formal context is reduced prior to the creation of the FCA lattice. While Snasel et al. (2007) utilise SVD as a matrix decomposition technique, Kumar and Srinivas (2010) make the argument that the computation of SVD is highly complex and alternative approximating matrix reduction solutions may be better suited. In their case, Kumar and Srinivas (2010) utilise fuzzy K-Means to simplify the matrix of the formal context. This reduced matrix is then used to generate an FCA lattice which Kumar and Srinivas (2010) presents as comparing favourably with SVD techniques. This argument is made primarily with respect to generating a simpler lattice, including the reducing the number of formal concepts, edges, and height of the lattice.

Alongside measures of assessing the effectiveness of lattice reduction solutions, Dias and Vieira (2010) introduce their specific lattice reduction technique which operates by finding equivalences in the original formal context between objects. Weights are assigned to the attributes in the set $M$ by individuals and objects instances are considered similar if the sum of their attribute weights are greater than a threshold value; this is for the set of attributes which the two objects have in common. From the previous processes, objects deemed sufficiently similar are assigned to groups of a pre-determined size. These groups of objects now replace the individual atomic instances in the original formal context, and the FCA lattice is generated from the now reduced formal context. This lattice reduction solution requires expert understanding of the domain in order to facilitate appropriate assignment of weights to attributes.
The final of these pre-lattice reduction techniques discussed is based on the idea that garbage in implies garbage out. To reduce the probability of noise or unnecessary formal concepts in the resultant FCA lattice Cimiano et al. (2003) make great pains to ensure that each binary relationship present in the formal context is legitimate. Given that their research entails the application of Natural Language Processing (NLP) to retrieve object and verb pairs in a document corpus, they use conditional probability measures based on co-occurrence to decide on whether the binary relationships are legitimate and mandate inclusion.

Methods previously discussed achieved a reduced lattice by first reducing the complexity of the formal context itself. From the simplified formal context an FCA lattice would be generated which has a greater likelihood of being simpler than a lattice generated from the original formal context. This approach has several limitations, the first of which is the complexity of the calculations. Matrix reduction techniques such as SVD and NMF utilised in Snasel et al. (2007) and Snasel et al. (2008) can be computationally complex (Snasel et al., 2008).

Secondly, certain approximations made in the matrix reduction techniques and object similarity measures of Dias and Vieira (2010) lead to loss of information in the formal context itself. Objects are assigned attributes or deducted attributes and/or grouped with similar objects. These processes, while of course being a means to an end, have the disadvantage of changing or losing original information.

A third drawback of pre-reduction techniques is that although the reduced lattice may theoretically contain a higher percentage of relevant concepts from the domain, a numerical or categorical representation of the importance of each formal concept is unavailable. It may be argued that sometimes it is not only sufficient to ‘know what you have’, but also to ‘know what you do not have’. Pre-lattice reduction techniques present the interested party only the end product of a set of concepts, all of which are deemed to be of high relevance. However although concepts may be of low significance knowing what concepts are of low significance may be useful information. The relevancy of low relevancy concepts would be of value when assessing variances of concept relevance across datasets or document corpora, or documenting growth or decline of a concept’s relevancy in a domain over extended periods of time.

In many scenarios a numerical representation of the relevance of a formal concept
2.4. CONCEPT LATTICE REDUCTION

may be necessary for analysts. Such a representation is often obtained by first generating the FCA lattice from the original formal context, then utilising a method of assessing the relevance of a formal concept. This gives access to the relevance of each formal concept, allowing for a very customized pruning of relevant or non-relevant concepts as one can adjust threshold values to reflect 'levels of relevance' required or to reflect percentiles when requiring a minimum or maximum number of concepts in final hierarchy.

2.4.2 Post-lattice reduction

The first of these post-lattice reduction methods considered is that of Belohlavek and Macko (2011). Here, all attributes of the formal context are themselves assigned weights prior to generation of the complete FCA lattice. After the lattice is constructed a function of these weights would then be applied to the intent of a formal concept to determine its level of relevance in the lattice. This technique, as is the case of the pre-lattice reduction of Dias and Vieira (2010), is limited by the fact that a detailed knowledge of the domain must be known beforehand in order to assign appropriate weights to the varied set of attributes; this option and knowledge may not be available in a variety of situations. For their solution, the relevance of the formal concept is determined independent of the object instance support of the formal concept. In such a case a formal concept with an extent of size one in a lattice would have the same relevance as a formal concept with an extent of size one hundred in another lattice, given that both formal concepts share the same attribute set (intent).

It is intuitive to think that, as opposed to Belohlavek and Macko (2011), the extent of a formal concept should play a factor in the relevance of the formal concept. This belief underpins the approach of Stumme et al. (2002) and Stumme (2002a) in determining the relevance of a formal concept. For Stumme et al, the relevance of the formal concept is the ratio of objects in the formal context which are elements of the extent of the formal concept being assessed. In tandem with this Support Value (SV), is a minimum support value (minsupp), where minsupp ∈ [0, 1]. This acts as a threshold for ‘relevant’ formal concepts.

Essentially, given a formal context $K := (G, M, I)$, the Support Value of the formal concept $(A, B)$ is formally defined as
\[ \text{supp}(A, B) = \frac{|A|}{|G|}. \quad (2.1) \]

As the object support in a lattice is monotonously decreasing as one descends the lattice (i.e. if \((A, B) \leq (C, D)\) then \(\text{supp}(A, B) \leq \text{supp}(C, D)\)), the formal concepts of high support would be formal concepts in upper positions of the lattice. With a \text{minsupp} value chosen, the iceberg concept lattice will be the uppermost section of the lattice. In some instances these most significant formal concepts will form a complete lattice, however the iceberg lattice is “general only a join-semi-lattice” (Stumme et al., 2002, p.194) and a bottom element (infimum) needs to be added to the frequent formal concepts to achieve a complete lattice.

The Support Value of Stumme et al. (2002), the iceberg lattice approach, is easily implemented and computationally simple, amounting to no more than a frequency count of the extent of a formal concept. This leads to it being one of the most commonly utilised method to extract the relevant formal concepts from an FCA lattice where a relevance value is assigned to each concept.

However despite its simplicity the iceberg lattice approach is limited to extracting only high-level concepts. There exist formal concepts in the lower levels of an FCA lattice which may be of notable relevance despite having a low(er) object instance support. These formal concepts could be emergent concepts of the domain, and a relevancy measure relying only on object instance support would prevent these formal concepts from being considered as relevant. On the opposite side of the spectrum, there may be formal concepts of high object instance support but a case may be made that their relevance should be low. This could be as a result of them being too similar to other formal concepts or for other reasons specific to the domain.

Another consideration for assessing the relevance of a formal concept was described in Klimushkin et al. (2010). In addition to the Stability Index of Roth et al. (2008b), Klimushkin et al. (2010) represents the relevance of formal concept by the ‘concept probability’. This concept probability is the probability of closure of the intent of the formal concept i.e. from a formal context \(K := (G, M, I)\), the formal concept \((A, B)\) is closed if \(B = B^{\uparrow A}\) and the concept probability is \(P(B = B^{\uparrow A})\). Given an extent of size \(|A| = k\), and \(|G| = n\), the concept probability must take into consideration three criteria: there are \(k\) objects in the \(G\) which have all attributes of \(B\); of the remaining
2.4. CONCEPT LATTICE REDUCTION

\( n - k \) objects none possess any attribute in \( B \); and no attribute external to \( B \) is shared by all \( k \) objects. Where \( p_m \) is the probability of an object having the attribute \( m \in M \), and assuming \( P_B \) is the probability of a random object having all attributes from \( B \subseteq M \) (\( p_B = \prod_{m \in B} p_m \)), the concept probability has a complexity of \( O(|G|^2|M|) \) and can be calculated thus:

\[
P(B = B^{\cong}) = \sum_{k=0}^{n} \left[ \binom{n}{k} p_B^k (1 - p_B)^{n-k} \prod_{m \notin B} (1 - p_B^m) \right]. \tag{2.2}
\]

Concept probability was then utilised in the removal of ‘noisy’ formal concepts from a variety of types of FCA lattices. In their assessment of the concept probability in such a role, Klimushkin et al. (2010, p. 265) conclude that “low or high probability of a concept does not say much of its importance”. Concept probability was not self-sufficient in filtering noise however it showed promise as a normalizing measure for the stability index (Roth et al., 2008b) - the combination being the most successfully noise removal approach of methods attempted in Klimushkin et al. (2010).

Addressing a variety of the concerns/limitations of the iceberg lattice approach is the concept of ‘stability’ of a formal concept. The stability of a formal concept was introduced by Kutsenov in Kuznetsov (1990) and Kuznetsov (2007). In Kuznetsov (2007) Kutsnenov et al. formulate the idea of stability built on the notion that a good hypothesis should be independent from the randomness of the dataset. Using the extrapolation of polynomials to fit a given set of points \( X = \{x_1, x_2, ..., x_n\} \) as an example, Kuznetsov (2007) posit that a polynomial extrapolated from a smaller subset \( Y \) of \( X \), where \( |Y| = k \), and the polynomial fits all points of \( Y \) as well as all points in \( X \), would be a more reliable polynomial in explanation of all points in \( X \) than a polynomial derived using all \( n \) points in \( X \).

Kuznetsov (2007) transferred to Formal Concept Analysis this idea of a hypothesis being better substantiated if originating from subsets of a dataset rather than the complete dataset itself. The Levelled Stability Index and an Integral Stability Index for a formal concept \((A, B)\) were introduced. For the Levelled Stability Index the \( j \)-level stability index is the ratio of subsets, \( Y_i \), of \( A \), where \( |Y_i| = j \) and \( Y_i \sqsupseteq B \), with respect to the number of subsets of size \( k \) possible in \( A \). The Integral Stability Index is the ratio of subsets of \( A \), of any size, for which their individual intents are also \( B \). In possession of a definition of both stability indices, examples of their calculations...
were provided on a formal context representing the attributes of snow tires. Analysis
was also provided on the behaviour of both indices as new object instances are added to
the formal context, deriving maximum and minimum values of both indices if \( k \) objects
are added. The paper concludes by showing that the determination of both indices is
\( \#P \)-complete and are “optimal within a factor polynomial in the input (namely,
\( O(|M| \cdot |G|^2) \))” (Kuznetsov, 2007, p. 113). Although providing the mathematical
and conceptual foundation of Stability Index, validation of the stability indices was
not explored in a context, only a basic comparison of a high stability vs low stability
formal concept was done.

The core papers of the Stability Index are that of Roth et al. (2008b) and Roth
et al. (2008a). Here, using a slightly different definition than that of Kuznetsov (2007),
Roth et al. define both an intensional stability index which targets the stability of
the intent of a formal concept and an extensional stability index which targets the
stability of the extent. However their main focus is on the intensional stability,
which henceforth in this thesis is used as the de facto definition of ‘Stability Index’.

As described in Roth et al. (2008b, p.5), the Stability Index “indicates how much a
concept intent depends on the particular objects of the extent” and is defined as follows:
Given a formal concept \((A, B)\) obtained from the context \(K := (G, M, I)\), the Stability
Index, \(\sigma\), is

\[
\sigma(A, B) = \frac{|\{C \subseteq A | C^a = B\}|}{2^{|A|}}.
\] (2.3)

This translates as the Stability Index of a formal concept \((A, B)\) being the ratio
of subsets of the extent \(A\) for which the shared maximal attribute set, \(C^a\), of all
elements of the subset is equal to the intent \(B\). A stable intent would thus be one
where its intent is not overly dependent on a particular set of objects; the removal of
objects from the extent of a stable formal concept would not easily cause the expulsion
of formal concept from the lattice. On the other hand an ‘unstable’ formal concept
would collapse quite easily if a few objects were to be removed from its extent. Roth et
al. were also able to show that the Stability Index of a formal concept \((A, B)\) can also
be described as the ratio of formal contexts (each formal context varies by the chosen
subset of objects) for which there would exist a formal concept in the corresponding
FCA lattices with \(B\) existing as the formal concept’s intent. After these definitions
the Stability Index was then demonstrated in the role of identifying stable epistemic communities (EC) in a.) an area of embryology and b.) Complex Systems. The formal contexts for the lattices were generated using research papers, where the authors were seen as object instances and key terms in the papers used as the attribute set.

Although the Stability Index has found traction as the primary method of assessing relevance of a formal concept in the FCA community (Buzmakov et al., 2014a), concerns about its efficiency in processing large lattices have arisen. A general algorithm was presented in Roth et al. (2008a) that calculates the Stability Index values of all formal concepts in an FCA lattice. However Roth et al.’s algorithm has a time complexity for calculating the Stability Index of a given formal concept that is quadratic, $O(|G|^2|M|L^2)$, with respect to the size of the lattice $L$, and cubic, $O(|G|^2|M|L^3)$, for all formal concepts in the lattice (Zhi, 2014). This algorithm, one of the fastest, often takes more time than the generation of the lattice itself (Buzmakov et al., 2014b). As a consequence, several attempts have been made in simplifying the calculations or approximating the Stability Index values.

Citing the #P-complete complexity of the Stability Index computation, Buzmakov et al. (2014b) derive the maximum and minimum values (stability bounds) of a formal concept’s stability instead. If $(C, D)$ is a lower neighbour of the formal concept $(A, B)$, Buzmakov et al. (2014b) use the premise that ‘if the set $E$ is such that $E \subseteq C$, then $E^c \neq B$’ to arrive at these boundary values of the stability index of $(A, B)$. The exclusion of these subsets of $C$ would be sufficient to calculate the stability value itself, however the fact that there may be duplicates of these subsets across multiple lower neighbours of $(A, B)$, meant that the authors settled with establishing an upper and lower bound for the stability index. Determining these bounds has a time complexity of $O(|G||M|^2)$, which is the equivalent of finding the lower neighbours of a formal concept (Buzmakov et al., 2014b). The stability bounds of a formal concept $c$ can be obtained using

$$1 - \sum_{d \in DD(c)} \frac{1}{2^{\Delta(c,d)}} \leq Stab(c) \leq 1 - \max_{d \in DD(c)} \frac{1}{2^{\Delta(c,d)}}$$

(2.4)

where $DD(c)$ is the set of lower neighbours of $c$ and $\Delta(c,d)$ is ‘the cardinality of the set difference between the extent of $c$ and the extent of $d$', i.e. $\Delta(c,d) = |Ext(c) - Ext(d)|$. It is worth noting that in their analysis of the bounds of the
stability values, it was observed that stability exhibited exponential behaviour and the values generally approximated 1 as the size of the formal context increases. For this reason in their experiments logarithmic stability ($LStab$) is utilised in their paper, rather than the standard Stability Index. ($LStab$) is defined as:

$$LStab(A, B) = -\log_2(1 - \sigma(A, B))$$  \hspace{1cm} (2.5)

A comparison was undertaken by Buzmakov et al. of the efficiency of their stability bounds solution in comparison to that of the Monte Carlo estimate of Stability Index in Babin and Kuznetsov (2012). In the case of Monte Carlo, when estimating the stability of a formal concept $(A, B)$, a random subset of $A$ is chosen and if it ascertained whether its maximal shared attribute set is equal to $B$ ($A^s = B$). If this is the case, this is counted as a success. This process is repeated for another random subset of $A$. After $N$ iterations the ratio of successes is determined and this ratio is used as the estimate of the stability of $(A, B)$. Assuming an error rate $\delta$, to achieve a precision $\varepsilon$, the required number of iterations is

$$N > \frac{1}{2\varepsilon^2} \frac{2}{\delta}. \hspace{1cm} (2.6)$$

Buzmakov et al. (2014b) show that the number of iterations for the Monte Carlo approach can be prohibitively large when trying to achieve high precision values. Therefore in their experiments to analyse computational efficiency, rather than comparing their stability bounding method directly the Monte Carlo approach, they compared their stability bounding method to ‘a combined method of the stability bounding method and Monte Carlo’. In the combined method, if the stability bounds are too wide, the Monte Carlo solution is utilized. The experiments concluded that the combined method was significantly slower than the stability bounds alone, but can give better results under very specific circumstances.

This poor efficiency of the unbiased methods motivates further research into such unbiased methods which would be suitable for large datasets. The desire is that these new methods be more efficient than the Stability Index as well as be able to provide the advantages that Stability Index has offered over other popular measures such as the Support Value.
Having provided some documentation of previous approaches identifying relevant concepts, of which the key approaches are summarised in Table 2.2, it is important to consider also some form of validation for these approaches. No agreed-upon objective validation procedure for determining effectiveness was found in literature, therefore an assortment of documented approaches are considered and described in the following section.

### 2.4.3 Lattice Reduction Validation

What is evident in the literature is that there exists no formal or widely accepted method of validating the various lattice reduction techniques, for, while Klimushkin et al. (2010) test the noise-reduction capabilities of multiple approaches, other authors use different testing strategies. Although the objective of one’s research would inevitable affect the validation processes chosen, a formal framework for analysis of lattice reduction would not be amiss. In the absence of this, a variety of approaches have been utilised.

The most visible attempt to create a domain-independent objective assessment criteria was that of Dias and Vieira (2010). Given the disparate techniques of reducing the complexity of an FCA lattice Dias and Vieira (2010) sought to contribute not only a novel solution but to establish also some formality in determining the success of the reduction strategies employed. Four requirements were proposed for a successful lattice reduction strategy. These include: the avoidance of creating new objects, attributes, or binary relations; preserving the original concept hierarchy; maintaining consistency between knowledge derived from reduced lattice and that of the original lattice (high fidelity); and preserving the discriminatory power of the attributes in original context.
In the main Stability Index papers (Roth et al., 2008a) (Roth et al., 2008b), a qualitative method of appraisal of reduction strategy was used. In both instances Formal Concept Analysis was utilised to create an epistemic community taxonomy derived from the publication of scientific papers in the discipline of embryology. Based on a formal context where authors were object instances and key terms in research papers as attributes, the Stability Index values of the resultant formal concepts were seen as measures of the strength of epistemic or knowledge communities (EC) and the most stable of formal concepts selected. Although implemented in a real world context, there was minimal assessment of the legitimacy of the Stability Index values as a reflection of the strength/weakness of the ECs. The commentary offered on the set of stable formal concepts were qualitative comparisons of the stable concepts to existing expert opinion on the known subfields of the industry.

Another qualitative analysis was undertaken by Jay et al. (2008) for both the Support Value of iceberg lattices and the Stability Index. Both measures were utilised to assess the social communities of hospitals where hospitals are linked by sharing treatment of a set of patients for cancer-related illnesses. Jay et al. (2008) focus on the ability of Stability Index to account for rare stable concepts and frequent unstable concepts by looking at the intersection of (un)stable formal concepts with (in)frequent formal concepts. For their lattice, patients served as object instances and hospitals where patients were treated acted as the attribute set. From the lattice the formal concepts deemed rare stable and frequent unstable (via threshold values) were then analysed qualitatively through comparison to what was previously known about the collaborative practices of the hospitals in question and through the value which the stability-derived taxonomy benefited the hospitals.

An alternate approach was that of Buzmakov et al. (2014a) who sought to determine whether the Stability Index is useful in finding significant patterns within the entire population. Their argument posits that, if effective, the Stability Index should provide consistent results for the stability of the individual formal concepts (the intent) across multiple samples of a population. Essentially there should be a linear correlation, where ideally \( y = x \), if the Stability Index values of formal concepts derived from the formal context of a sample of a formal context, are plotted against the
Stability Index values of equivalent *formal concepts* derived from another sample. In their comparison of a test sample and a reference sample, the fact that the Stability Index values were mostly close to 1 and exhibited exponential behaviour, meant that Buzmakov et al. (2014a) utilised the *Logarithmic Stability* (Equation 2.5), in place of the Stability Index itself. Their results using the $LStab$ values approximated a linear function, providing some validation of the Stability Index (via $LStab$) due to its consistency across samples. They propose that the results using the Stability Index be compared with those of other approaches of determining relevance of *formal concepts*.

For the purposes of this research the decision was made to provide some validation of the Collapse Index measure by incorporating its usage into a recommender system. This would involve the usage of the Collapse Index as a way of pruning user profiles. Given that the user profiles will be based on samples of movies that the user has watched, Buzmakov et al. (2014a)’s test for consistency across samples is also used as a validation solution. Finally due to the fact that the dataset has no explicit indicators of movie preference, the ability to recall relevant concepts in noisy lattices, similar to Klimushkin et al. (2010) is also employed as a way to validate the Collapse Index.

### 2.5 Concept Definition by Patterns of UIC

The case study in which this thesis applies FCA and concept relevancy measures is that of the BT TV dataset containing logs of movie-viewings by a set of users. Specifically, the desire is to investigate the contribution of patterns of User-Interaction and Context (UIC) in defining concepts.

While the structuring of the UIC-described concepts in their hierarchical relationship is obtained via Formal Concept Analysis (Wille, 1982), the idea that these patterns inform semantic definitions is built on an assumption of the nature of language known as the *distributional hypothesis*.

The distributional hypothesis originated from the works of Zelig Harris in the field of Linguistics. In his main works, Harris (1968) and Harris (1954), Harris theorises that in scenarios where terms share similar linguistic contexts, these terms are similar. In (Harris, 1968, p.12) he states specifically that “*The meaning of entities and the*
meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities.”

At its most basic, the argument is being made that a pair of terms are similar if, given a sufficiently large amount of text or document corpus, each term is present in a similar context as the other. Context in this case could refer to, inter alia, a set of neighbouring words, or the set of documents where the terms are present.

One of the main criticisms, or at least observances of Harris’ work, is his avoidance of details on what constitutes ‘meaning’ with respect to the proposed distributional hypothesis (Sahlgren, 2008). Nevin (1993) emphasise that despite the difficulties in understanding what ‘meaning’ is, this should not deter the inclusion of meaning in linguistic research.

From the perspective of Harris, while ‘meaning’ may be beyond the purview of linguistics, that does not preclude meaning from being represented in the distributional hypothesis. Sahlgren (2008)’s interpretation of Harris’ view is that if meaning is linguistic, and the distributional hypothesis successfully captures the linguistic, then distributional hypothesis would capture the meaning despite not knowing what the meaning is, or needing to know what meaning is. Nevin (1993) similarly argue that, as opposed to the belief of Harris’ critics that Harris had avoided meaning in formulation of the distributional hypothesis, Harris had more so avoided judgement on meaning rather than avoiding meaning itself. The semantic analysis done through distributional hypothesis could be accomplished without reference to meaning.

Where Harris does venture to discuss (linguistic) meaning, it is to describe it as being differential rather than referential. Referential is a perspective where a term has a specific meaning which may be formalised (e.g dictionary). The differential perspective is where a term’s meaning is represented in how it differs from other terms. The distributional hypothesis which scrutinizes differences in the usage of terms would thus be in a coveted position in capturing the differential meaning of terms. The ability of the distributional hypothesis to capture differential meaning is encapsulated in the statements of Harris in (Harris, 1954, p.43)

...if we consider words or morphemes A and B to be more different than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference in
meaning correlates with difference in distribution.

In this thesis the differential meaning of the distributional hypothesis is the basis for the concept definitions. By investigating the patterns of behaviour and characteristics of users who interact with movies, links are established between sets of movies representative of some relationship. These relationships are used to create a set of concepts where the concepts are described by the patterns of UIC. The meanings of these concepts are brought about by the patterns of UIC of individual concepts being different, or similar, to the patterns of UIC of other concepts.

Previously mentioned was the decision to validate the proposed approach to semantic extraction where a concept relevancy measure is employed, by incorporating the relevancies assigned to concepts into a recommender system (RS). This was deemed especially relevant in the face of the dataset in which the experiments would be conducted consists of movie-viewing details of customers. For this reason Section 2.6 provides background on recommender systems, touching on fundamental approaches and the role defining movies by context and user characteristics might play.

2.6 Background of Recommender Systems

Notwithstanding increased effectiveness of search engine algorithms in retrieving desired content, in light of increasing volume on the digital data, rather than relying on users to find desired content, it may be prudent to pre-empt user-attempts at content retrieval, and recommend content to users instead. Recommender systems materialised as a way to augment the social process of individuals being recommended content from friends, reviews, and guides (Resnick and Varian, 1997).

In addition to WWW content, in the early 90s, for the purposes of recommending consumer goods such as CDs, books, and movies, both academia and enterprise began investing in research into systems that recommend relevant items to users (Adomavicius and Tuzhilin, 2005). These systems became especially critical in the growing e-commerce market, in order to provide personalized recommendations to consumers (Cho et al., 2002), (Sarwar et al., 2000).

Of the assorted recommender systems that were eventually conceived, most were able to be classified into three main categories. Two of these fundamental categories of
CHAPTER 2. BACKGROUND

recommender systems are collaborative filtering (CF) and content-based filtering (CB) (Park et al., 2012), while the third category is a hybrid of the previous two.

CB creates a user-profile for users based on the characteristics of content that the individual had previously consumed. User-profiles are then compared for similarities with the descriptions of content on offer. For CF, popular or highly-rated content common to a user’s Nearest Neighbours (NN) are recommended to the user (Cho et al., 2002), where NN are the set of users most similar to the user in question.

Highly specialised recommendations common to CB may in fact be a disadvantage for e-commerce applications, as responsibilities of recommender systems in such cases may include expanding user interests. CF techniques, where over-specialisation is less of an issue, have gained popularity for this reason (Park et al., 2012) and is seen as the most successful of the two fundamental approaches (Lee et al., 2010).

CF was introduced to recommender system research via the Tapestry system of Goldberg et al. (1992), and initiated the first substantial move away from content-based recommender systems. In their approach, using the explicit feedback of other users on content was an asset in filtering documents for targeted users. For Tapestry a user would be provided with content based on feedback from either trusted participants or participants whom the user deems consumes relevant content. Although giving the user substantial direct input on the content that arrives in their inbox, this came at the cost of requiring the user to be highly engaged in the network of the domain in order to be able to make useful judgements on the value of specific peers. Anonymity also presented a problem in such an arrangement given that users need to be privy to which participant has provided which feedback for which content. (Sarwar et al., 2000) argue that Tapestry would not be suitable for large communities.

Addressing some of the concerns on anonymity was the GroupLens recommender system of Resnick et al. (1994) as well as Shardanand and Maes (1995). For their approach Resnick et al. (1994) do not permit the user to select their neighbours. Instead, GroupLens employs the Pearson correlation coefficient ($\rho$), a mathematical measure of similarity, to find a user’s closest neighbours based on ratings users have provided of items they have consumed. The correlation coefficients between pairs of users acts as weights in calculations of predicted ratings scores of items being offered but not previously rated by user of interest. These ratings are then used by GroupLens
or other adapted algorithms as determinants of what items to recommend.

A matrix containing ‘ratings users have assigned to items’ remains one of the more prominent and successful mechanisms used in CF. In these rating matrices column vectors and row vectors represent either user or items. Correlations between users are calculated between pairs of vectors within the matrix - each ‘vector of ratings’ serving as that user’s profile.

This correlation between users is used in Resnick et al. (1994) however similarly successful has been establishing correlation coefficients between item-vectors in the ratings matrix rather than user-vectors. In fact more efficient and successful algorithms have been produced that use item-based correlations rather than user-based correlations (Sarwar et al., 2001) as it is common that the vector space for items would be of lower dimensions than the vector space for users. Further arguments for item-based correlation include the fact that the vector space for users was routinely more dynamic than items. This is due to the repeated engagement with system by users requiring more computational effort being made to maintain an accurate portrayal of the domain.

At the same time, although demonstratively successful in making recommendations, both the user-based approaches of GroupLens and the item-based approaches of Sarwar et al. (2001) are hampered by scalability issues that calculations of correlation coefficients for large ratings matrices present (Konstan and Riedl, 2012).

Billsus and Pazzani (1998) address scalability issues through Singular Value Decomposition (SVD) of the ratings matrix. The premise of SVD matrix reduction is that there exists unnecessary user or item dimensions in the ratings matrix, and that these could be reduced to a set of underlying latent dimensions. Utilisation of this reduced ratings matrix would still retain the important characteristics of the original ratings matrix but be much easier to process. Koren et al. (2009) also detail matrix factorization techniques that both scale well, in addition to producing greater prediction accuracy.

However although the simplified ratings matrices make for more efficient calculations of predicted ratings, some benefit is lost as a result of the matrix reduction techniques themselves being computationally costly. This is particularly problematic as the addition of new ratings to the matrix normally requires a re-computation of the
reduced matrix (Konstan and Riedl, 2012). Some attempts have emerged to address this re-computation need, such as folding (Deerwester et al., 1990), but are either similarly computationally intensive or lose their accuracy as more ratings are added.

‘Cold start’, where users new to the system have only rated few products, is another challenge faced by collaborative recommender systems. To reduce the impact of cold start Shardanand and Maes (1995) compel users to provide an initial set of ratings for content before they are allowed to engage with system. This mandatory training period mitigates to some extent the cold start issue, but it has the drawback of burdening the user with extra responsibilities.

The combination of users wanting to avoid such obligations with a large set of users and items, often results in ratings matrices being very sparse, in that many items have not been rated by many users. This leads to a high percentage of empty cells in the ratings matrix. Furthermore, an impact of matrix sparsity is the decreased likelihood of finding many similar users through the usual similarity measures, due to the low probability of a pair of users having rated a shared set of items. Matrix reduction techniques however, do go some way in mitigating sparsity.

In 2006 Netflix the on-demand video streaming service initiated the Netflix Prize, a competition open to individuals or institutions to develop a recommender system better than Netflix’s, which in turn invigorated global investment in RS research. As part of the competition, Netflix released an internal Netflix dataset (Bennett and Lanning, 2007) containing, inter alia, the ratings that users provided for a variety of movies.

Matrix factorisation techniques became a dominant theme among recommender research due to the huge size of the dataset that Netflix made publicly available. 480,000 users and 17,770 movies made more efficient neighbour-generating algorithms a priority. Koren et al. (2009) were able to include implicit feedback, temporal effects, and confidence levels in their matrix factorization techniques and produce better predictions. However to achieve the winning predictive accuracy criteria, ultimately an average of the various individually successful models from a variety of research organisations was utilised.

Throughout all these implementations of CF, a constant is the matrix of explicit ratings. Unfortunately in many situations the user is unwilling to provide any ratings
2.6. BACKGROUND OF RECOMMENDER SYSTEMS

due to the cognitive load effort (Oard et al., 1998) or in some circumstances the interface, such as mobile devices is ill suited (Lee et al., 2010) or does not provide for explicit expression of level of satisfaction (Lee et al., 2008), (Hu et al., 2008). Where avenues exist for providing a rating but user reluctance means few ratings are entered, this contributes to the data sparsity issue which in turn negatively affects the effectiveness of the RS.

A possible solution, especially in the absence of any explicit ratings, is the usage of implicit ratings as a replacement. These ratings are inferences made on user preferences based on observations of user behaviour available to the system (Oard et al., 1998). Utilisation of these implicit ratings would require both solutions which record and extract observations, as well as functions which translate these observations into a rating value.

Some examples of the usage of implicit feedback include Hu et al. (2008), who detail their approach of including implicit feedback into their SVD-matrix factorization models. User data is first mined for concepts then the frequency of these concepts values are translated to a preference value as well as a confidence value. Lee et al. (2008) use temporal information, including time expired between product launch and user purchase, as well as the product release date, in the composition of a ratings function that generates a pseudo-rating for the ratings matrix. Common procedures for CF such as Pearson correlation are then applied to the pseudo-ratings matrix to calculate recommendations.

Choi et al. (2012) also apply typical CF procedures after generation of a ratings matrix from implicit data. In an online shopping mall domain, Choi et al. (2012) generate an implicit user rating for an item \( i \) via a function that takes into consideration the ratio of items which a user purchases that includes item \( i \), to the maximum ratio of said item for all other users. However rather than relying solely on a CF which utilises the implicit ratings to predict the ratings for items the user has not previously purchased, an additional sequential pattern analysis (SPA) solution, based on the order in which a user purchases items, is similarly deployed to create a predicted rating. A final predicted rating is generated by a function which assigns a weight and its complement to both a purchase ratio and SPA predicted ratings.

Lee et al. (2010), in the context of a music purchasing service for mobile devices,
also utilise implicit feedback. In their case the implicit data they utilise is a set of possible actions a user may take during each session using the service. These include: ignore; clicked-through, pre-listened, and purchased. The key difference between their approach however, is that as opposed to the majority of recommender systems who seek a cardinal rating of user preferences, Lee et al. (2010) generate a user profile which is an ordinal ranked representation of user’s music interest. Nearest Neighbours are found based on comparing pairs of user profiles, then predicted ratings for new music are generated.

Some research such as Núñez-Valdéz et al. (2012) do not go as far as making recommendations, but only investigate various parameters of implicit feedback and how well these parameters correspond to explicit ratings. They do this in the context of an eBook reader where implicit feedback on a variety of parameters related to content were recorded, as well as explicit rating provided by the user for the same content. Each parameter was then compared with the explicit rating to determine the level of correlation.

What has been common to all these implicit-feedback approaches, even that of the ordinal-based solution of Lee et al. (2010), is that their analysis is undertaken on the item level, in that a ratings matrix is being generated where the ratings are ratings of items, while the ordinal approach is ranking items. These ratings/rankings of items go on to serve as user profiles.

Where opportunity lies in implicit feedback research is on a more granular level than that of an item, i.e. a user profile may be composed, not of item ratings, but of ‘ratings of characteristics of items’. e.g. rather than utilising implicit feedback to assign a user-rating to a song, an RS can utilise this feedback to assign a rating to a user’s interest in aspects of songs such as drumming, strings, autotune, etc. A user’s set of Nearest Neighbours could then be retrieved based on shared appreciation for specific attributes of a song rather than shared appreciation of the song itself.

For this thesis an implicit-feedback recommender system is developed where Formal Concept Analysis is utilised to identify the concepts\(^1\) of movies which have representation in a user’s profile. As opposed to previous implicit recommender systems which

\(^1\text{concepts in this context are representative of movie characteristics}\)
from implicit data assigned user ratings to the items, here movie concepts are assigned ratings or level of significance. This is done through the incorporation of the FCA concept relevancy measure developed in the course of this research.

2.7 Conclusion

In this chapter extensive description was provided on the thesis’ primary mechanism for semantic extraction, FCA. This includes formal definitions of the key terms/concepts of FCA which are used throughout the thesis as well as a demonstration of these concepts by example. The value of FCA as a means of conceptualising a domain was then shown by positioning FCA with respect to a knowledge representation framework, and by also discussing a diverse set of domains in which FCA was used as such a conceptualisation.

By showing the usage of FCA in modelling domains it was then possible to transition to the ways of addressing complex lattices. The two fundamental approaches to lattice pruning were discussed, - that of simplifying the formal context in advance of generating the lattice and that of assigning a relevance value to each formal concept after the lattice is generated. By showing the efficiency challenges faced by the Stability Index we were able to position the proposed Collapse Index concept relevancy solution. In addition, as we seek to validate the Collapse Index solution, some mention is provided of previous approaches to validate the success of concept relevancy measures.

Along with the previous, a section was dedicated to discussing the role the distributional hypothesis would play in allowing for patterns of UIC to define concepts.

The chapter concluded with a description of the key points of recommender system research. This spoke of the CF and CB approaches to recommender system design as well as the need for recommender systems that are able to operate using implicit data. From this need to operate within implicit data the design of a recommender system based on FCA was shown to be applicable.
Chapter 3

METHODOLOGY

3.1 Introduction

*Design Research* is the chosen methodology for this research as it is felt that questions surrounding the determination of concept relevancy and its usage in extracting semantics from datasets would be best answered through the empirical processes of designing and refining of artefacts. By evaluating the utility of the Collapse Index as an approach grounded in real needs these iterative processes would produce artefacts constituting the three (3) contributions of the thesis. The expected artefacts are a.) concept relevancy measure b.) an approach for the application of concept relevancy measure in extracting semantics from a dataset, and c.) a recommender system which validates both the concept relevancy measure and the approach in a problem of real business significance.

A brief discussion on the motivation of Design Research methodology is provided in Section 3.2. The individual research questions are described with respect to von Alan et al. (2004)’s Conceptual Framework of Design Research in Section 3.3. Section 3.4 then describes the case study (dataset) and the role it plays in resolving the stated research questions.
3.2 Design Research

A variety of research, including Lawson (1979), made a case that the way designers approach a problem is different from scientists (Cross, 1982). Lawson (2006) specifically noted that, as opposed to scientists who set out specifically to study a problem, designers learn about the problem by trying out a variety of solutions.

Design research as a research methodology was birthed as the consequence of the need to understand and formalize the processes of knowledge contribution by the way of the creation of a physical artefact (Bayazit, 2004). ‘Design’ is defined through the theoretical model of the General Design Theory (GDT) descriptive model of Yoshikawa (1981).

General Design Theory describes design as being a mapping from a function space to an attribute space (See Figure 3.1). Essentially, design involves the transition from the design specification which is the set of functionalities the artefact is to accomplish to the design solution which is the artefact which possesses attribute that allow for the desired functionalities to occur.

With ideal knowledge one can move immediately from specification to solution, however limited information, physical or relational constraints between attributes, mean that the solution is not easily attained. Design seeks this solution by operating through a stepwise refinement process where a model of the design object is evolved iteratively, with the expectation that new revisions converge to a theoretical, though
possibly unachievable, complete solution.

For each step of the iterative refinement process the model of the design object is applied to a known context. As much information as is practical is extracted from this process. This information gleaned contains content that is representative of knowledge contributions, as well as being a source of previously unseen challenges some of which are compelling (though of tangential relevance) while other challenges necessitate a resolution to facilitate resolution of the current design objectives. The combined knowledge is incorporated into the current model - the output being the next iteration of the model, one more closely resembling the complete solution. This process is repeated for every new version of the model until a satisfactory proximity to the complete solution is arrived at.

Although initially targeted at physical artefacts, Simon (1996) argue that design is applicable in any course of action where changing an existing situation into a preferred option is the motivation. He adds that the process of creating physical artefacts with desired properties is fundamentally the same intellectual process as developing medicines, devising business strategies, or creating government policy. Design Research has thus been employed in a diverse set of research areas.

### 3.3 Research Questions

In this section the Design Research processes utilised in the resolution of the three (3) research questions of this thesis are positioned with respect to a conceptual framework of Design Research in Information Science as represented in von Alan et al. (2004). This is illustrated in Figure 3.2.

Given the size of the industry dataset containing the information related to the domain, at the forefront of the expected development challenges were the selection of the concepts in FCA lattices which should be considered for inclusion in any final concept hierarchy. From this emerged the main research question (RQ 1) and challenge which was ‘in the face of the inevitable multitude of concepts which would arise if Formal Concept Analysis was applied to the dataset, how would one go about obtaining the relevant concepts in an efficient way’.
Figure 3.2: Conceptual Framework of Design Research in Information Science (von Alan et al., 2004)
The idea proposed was that of a concept relevancy measure which assigns a numerical relevancy value to each formal concept in an FCA lattice while being sufficiently efficient in calculation that it may be usefully applied to complex FCA lattices.

The derivation of the concept relevancy artefact would be achieved through the repeated manipulations of various types of FCA lattices trying to identify the factors which cause a formal concept to collapse. This would draw on the existing knowledge bases related to FCA, concept relevancy measures, and concept relevancy validation techniques. Observances of behaviours while manipulating the lattices would inevitably lead to various hypotheses as to what makes a formal concept collapse. Each hypothesis would then be implemented on synthetic formal contexts i.e. formal contexts specifically designed to exhibit certain traits. Hypotheses which are verified using these synthetic formal contexts would then be applied to various random contexts. If a hypothesis of what makes a formal concept collapse does not operate as expected in all scenarios, the hypothesis is either rejected or subjected to scrutiny and revision. Each refinement would then incorporate useful ideas from the previous proposed solution(s).

This process will continue until a particular solution works consistently for all formal contexts to which it was applied. At this point effort is then made to verify the hypothesis mathematically. If the mathematics does not support the proposed hypothesis then the research process is reverted to trying alternate hypotheses on synthetic datasets while incorporating ideas from previous revisions, including knowledge which would have arisen from attempts at mathematical verification. Eventually a solution would have been achieved that is mathematically verifiable and is consistent for all synthetic formal contexts to which it is applied.

At this point the design model would then be applied in the context of a large industry dataset. Through the application of the model in such a context the size of the datasets would lead to a deeper understanding of the issues related to the efficiency of the proposed relevancy measure that are inherent, specific to the corresponding algorithm, or specific to the characteristics of the formal context. The final outcome from this would be a new concept relevancy measure, an algorithm, and the mathematical support - all to be added to the knowledge bases of FCA and concept relevancy.

RQ 2 represents the original motivation of the research - a desire from industry to
extract useful semantics from a large dataset. Achieving a satisfactory approach to semantic extraction of the dataset not only requires a variety of solutions be considered, developed, and combined to address different challenges, but in many cases the problem definitions and proposed solutions\textsuperscript{1} would have to be revisited. These revisions to problem definition would be made through continued negotiations with industry as they would seek to ensure that their business needs are properly represented in the problem definition and that both pre and post-artefact suggested solutions to semantic extraction continue to address their needs. Knowledge contributions that are emergent through the research but are not necessarily directly addressing the original problem definition would also also be, through continued engagement with industry, positioned as close as possible to the business’ needs.

In the case of RQ 3, given that the case study dataset contains logs of movie viewings by BT TV subscribers, an ideal candidate for the application of an approach where concepts of low relevance are removed from an FCA lattice, was that of an RS. This means either implementing the Collapse Index in a pre-existing RS solution that makes use of the relevancy of formal concepts or implementing the Collapse Index in an RS specifically designed for this thesis to incorporate FCA and concept relevancy. The latter is chosen which would mean the conceiving of such an RS.

Here the processes of Research Design come to the fore as there would now exist a set of functionalities that the RS needs to provide, and an RS design is necessary that would accomplish this. Given that an RS is composed of multiple components, each would need individual attention with respect to their design. This includes the creation of user profiles, the creation of a similarity measure to compare profiles, predicting ratings, and the inclusion of a system to test the success of the system. While each component presents their own unique challenges and would require continuous improvements, the fact that the RS would require an effective combination of the varied set of components, means that great effort would have to be made to ensure that the multiple components form a cohesive whole. This would undoubtedly require a large amount of testing and iterations of the RS.

\textsuperscript{1}This would likely involve decisions related to the use of FCA, attribute selection, and process of validation
3.4 Case Study

Previously mentioned while discussing the fundamentals of Design Research, was that for iterations in the refinement process, the model of the design object is applied to a known context. These contexts could of course be artificially constructed to test very specific things under very specific conditions, however it was thought best to apply any theories developed unto an actual existing industry dataset at various points of analysis. This was also especially pertinent on account of the original research proposal being from industry.

Several datasets were taking into consideration before the final selection. Some were discarded for reasons such as being: incomplete; not sufficiently structured; or lacking a sufficient amount of records. The dataset ultimately selected was that of the logs of the details of movies viewed by subscribers of the BT TV service.

By applying emergent theory to this dataset one is able to test non-functional requirements such as efficiency, be able to see the end-result of the thesis’ contributions, and how best to position the contribution with respect to the needs of the business. Details of the dataset are provided.

Main Dataset (BT TV)

The central dataset for which the analysis was conducted consisted of the logs of movie-viewings by subscribers of an IP-based television service. This dataset was sourced from industry under the restriction of a Non-Disclosure Agreement (NDA). Consumer privacy concerns were mitigated by security restrictions on the locations for storage of dataset; this is in addition to having the user IDs of the dataset anonymised.

The time period from which these logs were collected was the 7-month period from the end of March 2014 to mid-October 2014. Each record in the dataset represents ‘a user watching at least 50% of a movie’ - an amount deemed, by industry, as sufficient to classify a movie as having been watched by a user. Records are described by the following fields:

1. uid: anonymised user id
2. mid: movie id (mapping provided in separate file for corresponding movie titles)
3.4. CASE STUDY

3. playtime: date and time at which user first tuned into program.

4. watched ratio: percentage of movie’s running time watched by the user.

5. postal district: first three characters of the postcode of user.

In addition to the dataset of records a separate file was provided which mapped movie IDs to their formal titles. Mapping information was combined with the BT TV records by the way of a ‘title’ field, producing a final revised set of records resembling that of Table 3.1.

Table 3.1: Example of BT TV Records

<table>
<thead>
<tr>
<th>uid</th>
<th>mid</th>
<th>title</th>
<th>playtime</th>
<th>w. ratio</th>
<th>p. district</th>
</tr>
</thead>
<tbody>
<tr>
<td>253</td>
<td>4</td>
<td>DARK KNIGHT</td>
<td>10/05/2014 21:41</td>
<td>0.56</td>
<td>E14</td>
</tr>
<tr>
<td>67</td>
<td>2314</td>
<td>APOCALYPSE NOW</td>
<td>10/05/2014 12:40</td>
<td>0.82</td>
<td>M13</td>
</tr>
<tr>
<td>34</td>
<td>3125</td>
<td>ZOOLANDER</td>
<td>10/05/2014 21:00</td>
<td>1</td>
<td>B37</td>
</tr>
</tbody>
</table>

In total, the dataset consisted of 644,188 records, from which there existed representation of 4,572 unique users. Although the dataset was described as having 2,534 distinct movies by BT, analysis of the data after reception showed this not to be the case and some revision was done as it relates to distinct movies as well as the number of records in the dataset.

There existed some desire to maintain some homogeneity of content; as a consequence, records related to non-movie content were removed from the dataset including sitcoms and TV shows. Furthermore, in some instances, different movie IDs were found to be referencing the same movie. Resolving this issue along with removal of television programming led to a final BT TV dataset comprised of 624,244 records, 2,122 unique movies, 4,572 unique users, and 1,904 unique postal districts.

In the context of the analysis, the fields of primary importance in this dataset are that of the movie id/title and the postal district. Location-based data is seen as the main avenue of establishing user context and socio-economic characteristics.

Supplementary Dataset (IMDb)

In addition to the BT TV dataset, the online resource Internet Movie Database (IMDb) (IMDb, 2015b) was also utilised as a contributing data source. IMDb is the established online source of information related to movies and television programming. Extensive
information is provided on a variety of facets of movies including: actors, release
dates, directors, writers, quotes, reviews, ratings, etc.; much of which may be obtained
through navigation of imdb.com or through a text-based dataset made available for
free download at IMDb (2015a). Both methods of obtaining dataset were utilised for
obtaining IMDb information. IMDB data obtained was used as supplementary data
source for the BT TV dataset, in that it provided additional descriptive information
on movies found in the BT Vision dataset, mainly genre definitions for movies.

Genre descriptions of movies were chosen as the IMDb-sourced descriptor of movies
as a movie’s genre is a very common and informative way of describing the movie. The
use of genres could then serve as a point of comparison for the UIC described movies.

3.5 Conclusion

In this chapter we have outlined the Design Research methodology utilised to tackle
the research questions of the thesis. Design Research itself was explained in order to
show its applicability to resolve the research questions. The processes of achieving a
resolution of the (3) research questions were then illustrated within (von Alan et al.,

The case study BT TV dataset upon which design models developed are expected
to be applied and refined was then described in detail along with the supplementary
IMDb dataset.

Through the utilisation of the Design Research methodology and applying the
various iterations in both custom and industry datasets, this was deemed suitable and
sufficient to answer all research questions. This thesis represents a documentation of
the answers and the path to arriving at these answers.
Chapter 4

EXTRACTING SEMANTICS USING CONCEPT RELEVANCY FOR FCA LATTICES

4.1 Introduction

A key aspect to the conceptualisation of a domain is the identification of concepts and the creation of the resultant hierarchy. The domain may be implicitly represented in large datasets which contain assorted textual data related to the domain. Where FCA is given the responsibility of extracting and modelling these domain semantics from large datasets, the FCA lattice generated from a formal context may be used as the final taxonomy. However, in the face of large datasets producing a correspondingly large amount of formal concepts in the lattice, concept relevancy measures may be employed to reduce the number of formal concepts present in the final concept hierarchy (See Section 2.4). This chapter discusses an approach for producing a concept hierarchy from such large dataset where low relevance formal concepts are removed from an FCA lattice.

The first component discussed is that of the creation of the formal context; this is included in Section 4.2. Section 4.3 discusses the selection of relevant concepts in the lattice, while Section 4.4 discusses how the final hierarchy of only relevant concepts is produced. A summary of the complete set of steps is presented in Section 4.5.
4.2 Creating Formal Contexts

4.2.1 Attribute Selection

If FCA is to be the mechanism in which concepts are identified and hierarchies created then it is mandatory that a formal context be available from which the lattice may be generated. A formal context \( K := (G, M, I) \), we recall from Section 2.2, is the set of objects \( (G) \), attributes \( (M) \), and the binary relations between objects and attributes \( (I) \).

To arrive at these object-attribute pairs \( (I) \) would first require establishing what would constitute the complete set of objects \( (G) \) in the domain and selecting what would be in the set of potential attribute descriptors \( (M) \) of these objects. Generally domain experts would have come to an agreement on what would be the set of objects being included in the formal context, as well as what attributes would be utilised in their description. The attributes selected by the experts are usually chosen based on a range of criteria including: efficiency, domain being modelled, and the level of detail to which one aims to model the domain (Cigarrán et al., 2004).

In the case of efficiency and level of detail, the more attributes which are taken into consideration in the formal context, the more complex the resultant lattice is likely to be. Complex lattices, or lattices with many formal concepts, model more details of the domain but the runtimes of generating complex lattices may be prohibitively high. Furthermore, the volume of concepts present in complex FCA lattices lead to inefficiencies in semantic reasoning on that many concepts. Ontology engineers are then responsible for finding a balance between efficiency and information present in the formal context when selecting the set of possible attributes.

As it relates to the domain being modelled when selecting the attribute set, it is self evident that the domain would influence what would constitute a viable attribute descriptor. A movie domain may have genre as an attribute since this is a common way to describe movies. Ontology engineers in the automobile domain may select a characteristic of automobile engines as an attribute. Moreover the choice of attributes to utilise in a domain is often dependent on which attributes are deemed more informative or important in describing the objects in the domain (Belohlavek and Vychodil, 2009), (Cigarrán et al., 2004).
4.2. CREATING FORMAL CONTEXTS

On other occasions while an attribute may not be specifically stated in text in the dataset, it may be generated as a function of data present in the dataset. For instance ‘the city in which a movie is most popular’ may not be specifically stated in the dataset but may be retrieved through a combination of text processing and statistical analysis solutions. In the following section we show where the investigation of links between object instances may form the basis of a set of shared attributes.

4.2.2 Role of Links in Attribute Selection

In the formulation of concepts by ontology-creating algorithms, these concepts are usually defined by a shared attribute set between object instances. The examples of FCA-derived taxonomical hierarchies discussed in Section 2.3 describe taxonomies where objects such as celestial bodies (Bendaoud, Toussaint and Napoli, 2008) are described with domain attributes such as ‘flaring’. A flare-orbit (formal) concept would be defined as ‘the set of celestial bodies which exhibit the characteristic flare as well as orbit some other body’, as well as ‘the attribute set {flare, orbit}’ itself.

Each object instance which is a member of the concept is ‘linked’ to other objects by virtue of the fact that they share a common set of attributes. These links between objects are used to group objects into concepts, where distributional hypothesis is then employed to give meaning to each concept in that a concept is defined by the virtue of its attribute description being different from another’s. Links themselves can be of several types, three (3) of which are shown in Figures 4.1a, 4.1b, and 4.1c, of which Figures 4.1a is a representative of links as it relates to the previous celestial body example.

Figure 4.1a shows two objects linked by the fact that they share a common attribute ‘a’. An example of this type of concept is the pair of objects car and bus; both belong to a concept (possibly automobile) as they both share a common attribute ‘wheel’. The second figure (Figure 4.1b) showcases a link between objects where one object explicitly references another. Hyperlinks where a web page contains a reference to another web page are an ideal example. In some instances both web pages may contain different types of content, but the fact that they are linked (via hyperlink) still groups the web pages into a concept as there is some implicit semantic relationship between both web pages. Finally Figure 4.1c represents a link where there is some overt function which
ties the two objects together - an example of such a function being ‘John watched the movie Godfather’.

A link essentially is the existence of a relationship between entities. Semantic information may be obtained from links due to the fact that relationships existing between entities suggest some semantic similarity between the entities. Getoor and Diehl (2005, p.3) state that “attributes of linked objects are often correlated, and links are more likely to exist between objects that have some commonality”.

Given that linked objects tend to be semantically related, if one were to explore ways in which object instances may be linked, this might lead to the identification of novel attribute sets which may provide insight on the semantics of the object instances. Consider a domain where humans are the object instances and FCA is utilised to create a concept hierarchy from humans and their attributes. If one seeks to link humans (objects) by their diets in order to extract concepts, one can imagine that the attribute set chosen to describe humans would be different than the attribute set where the desire is to link humans by their careers. As previously discussed in Section 2.3, FCA grants one this flexibility in selecting the set of attributes used to characterize objects.

One such case is where it is hypothesised that if multiple objects are linked based on shared patterns emergent from users interacting with the objects (Figure 4.2) then these objects share some semantic similarity. These patterns of User Interaction and Context (UIC) are defined both with respect to the actions users take on the objects (Interaction), as well as the characteristics of the users (User) and context of the users who interact with the object (Context). These UIC patterns are patterns emergent.
4.3 Selection of Relevant Concepts

Where the object and attribute set are not pre-determined by experts, Natural Language Processing tools and machine learning algorithms could be employed to process the dataset and return, not only the set of objects and a set of potential attributes, but also the binary relations between objects and attributes. This introduces some level of autonomy in the creation of the formal context.

On acquisition of the set of binary relations between objects and attributes Formal Concept Analysis could then be applied to the formal context to generate the FCA lattice. However, especially for large datasets and for formal contexts generated through software, many of the formal concepts which are present in the FCA lattice are due to noise, anomalies, or are not sufficiently contributory to the domain’s description.

While there will always be an element of subjectivity in decisions on which concepts of the many concepts are relevant, large datasets require some level of automation in

![Figure 4.2: Links: User Interaction](image)
the selection of relevant or important formal concepts. One way of addressing the abundance of formal concepts is to introduce a measure which assigns a numerical value to each formal concept in the FCA lattice which is reflective of the relevance of the formal concept.

Existing solutions such as the Support Value (Stumme et al., 2002) simply select the formal concepts with the highest object instance support\(^1\) as those of high relevance; this in turn leads to many interesting formal concepts with low support being rejected from inclusion in a final hierarchy. Other solutions allow for low support formal concepts to have higher relevancy than high support formal concepts.

With knowledge of the relevance of each formal concept a final concept hierarchy would be built from formal concepts whose relevancy values are above an appropriate threshold. This threshold could be heuristically and/or empirically determined.

### 4.4 Generating Concept Hierarchy

Following the determination of the set of relevant formal concepts in the lattice, one is then tasked with generating the final concept hierarchy. Assuming all formal concepts are determined to be relevant, one may then simply utilise the complete FCA lattice as the taxonomical representation of the domain. In some instances where not all formal concepts are determined to be relevant but the relevant formal concepts are those at the highest levels of the lattice, the lower level formal concepts could be discarded and the remaining formal concepts along with an infimum be used to reconstruct a complete lattice.

In other cases where non-relevant concepts may occur at various levels of lattice, a systematic method would have to be employed to remove the non-relevant formal concepts from the lattice while avoiding compromising the legitimacy of the lattice as a model of the domain. To accomplish this there are three (3) fundamental questions that would need to be addressed when removing a formal concept \((A, B)\) from a lattice:

- **What would be the new lower neighbour(s) of the previous upper neighbour(s) of \((A, B)\)?** As an example, in Figure 4.3, if \((A, B)^2\) was to be removed from

---
\(^1\)support references the cardinality of the extent of the formal concept. See Section 2.2.
\(^2\)Note that we are simply assuming that the concept \((A, B)\) is of low relevance. The actual relevance value was never actually determined and \((A, B)\) is simply being used for illustrative purposes.
the lattice, what would now be the lower neighbours of each of its two upper neighbours.

- What would be the new upper neighbour(s) of the previous lower neighbour(s) of \((A, B)\)? In Figure 4.3, if \((A, B)\) were to be removed from the lattice, what would now be the upper neighbours of each of its two lower neighbours.

- If there existed object instances where their individual attribute sets is the set \(B\), to what formal concept would the objects be added to as an element of their extent? In Figure 4.3, would object \textit{MOVIE}11 be added to a lower neighbour (gain attributes) or to an upper neighbour (lose attributes)? Should the movie be added to multiple formal concepts?

Considering that the example of Figure 4.3 demonstrates the removal of only one formal concept, if multiple formal concepts were to be removed simultaneously one would imagine that reconstructing the hierarchy would require a very systematic solution. There would undoubtedly exist a variety of solutions led by mathematical theory and/or the reason why the taxonomy is required.

The approach proposed for this thesis is as follows:

- Each lower neighbour \((C_i, D_i)\) of the formal concept \((A, B)\) removed is made to be a lower neighbour of all the previous upper neighbours \((X_i, Y_i)\) of \((A, B)\). Conversely each upper neighbour \((X_i, Y_i)\) of the formal concept \((A, B)\) removed
• If \((A, B)\) being removed means that there exists no non-infimum lower neighbour of an upper neighbour \((X_i, Y_i)\) of \((A, B)\) then the new lower neighbour of \((X_i, Y_i)\) is the infimum.

• If \((A, B)\) being removed means that there exists no non-supremum upper neighbour of a lower concept \((C_i, D_i)\) of \((A, B)\) then the new upper neighbour of \((C_i, D_i)\) is the supremum.

• A movie which had as its attribute set the set \(B\) (e.g. MOVIE11), would be assigned to an upper neighbour \((X_i, Y_i)\) which has the highest relevance. The disadvantage of assigning said object to only one upper neighbour is that the object loses one or more of its attributes. As an example, in Figure 4.4 MOVIE11, which was originally described with the attribute set \(\{c, f\}\) is now only described with the attribute set \(\{c\}\).

### 4.5 Semantic Extraction With Concept Relevancy

This section presents a more complete set of processes involved in the creation of a concept hierarchy using an FCA lattice pruned by a concept relevancy measure. This
begins with the availability of a raw dataset and ends with the concept hierarchy. A summary of the approach is provided in Figure 4.5.

In possession of a large dataset, the first responsibility is to perform some pre-processing on the dataset in order to remove errors such as misspellings and corrupted data. This process may also include the usage of language processing tools to: perform Part Of Speech tagging, remove stop words, and lemmatise words where applicable.

Having completed the pre-processing of the dataset the next task is to transform the data into a formal context. In the transformed data stage, based on the goals of the system, decisions are made on the variables which are to be considered for pattern detection. This could be as simple as informing algorithms as to which variables to ignore, or matrix reduction techniques could be used to reduce the dimensionality of large matrices. Along with the ideas expressed in Section 4.2, probabilistic classifiers, weighted measures, and more complex machine learning solutions may be applied to the processed data in order to associate objects with their attributes. The output of this process is the formal context that will be used to create the FCA lattice.

After the creation of the formal context the FCA lattice is generated. A concept relevancy measure is then applied to the lattice to determine the relevance of each *formal concept*. Formal concepts whose relevance values fall below a threshold would be excluded from inclusion in the final concept hierarchy. The final concept hierarchy is then either directly included in an application, or where sharing and reusing of ontologies is a priority; the final concept hierarchy is translated into a formal ontology language.

4.6 Conclusion

In this chapter a description was given of an approach for semantic extraction from a large dataset where the mechanism utilised for the extraction of concepts is FCA and the concepts in the FCA lattice are pruned based on a concept relevancy measure. The steps discussed included the selection of objects and attributes to be used from the dataset, including attributes based on links between objects. Also discussed were the selection of relevant concepts from the FCA lattice; and the formation of the final version of the hierarchy.
Chapter 5

CONCEPT RELEVANCY AND COLAPSE INDEX

5.1 Introduction

A case may be made that the relevance of each formal concept present in an FCA lattice is not the same. Several concepts in the lattice would arguably be the inevitable random artefacts which arise in a system influenced by multiple variables over an extended period of time. In contrast, other concepts could be core defining aspects of a domain and their appearance in the lattice integral. Chapter 2 provided examples of this - where experts were given the responsibility of identifying significant concepts in an FCA lattice generated from a dataset. These ‘approved’ concepts were then included in an ontology representing the domain. This human intervention is feasible when dealing with small datasets, however when FCA is purposed for extracting semantics from large datasets, too often far too many concepts are limiting the ability of humans to make a judgement on the significance of each concept.

Chapter 2 outlines software solutions historically and commonly utilised to assess the relevancy of formal concepts in an FCA lattice. Of the main solutions described, included are the efficiency concerns of the Stability Index (Roth et al., 2008a) (Roth et al., 2008b) and the bias of the Support Value (Stumme et al., 2002) to concepts with high object instance support. In this chapter the Collapse Index is presented as a novel solution for assessing the relevancy of formal concepts. It is designed to address both of the previous concerns.
5.2 Introduction to Collapse Index

The Collapse Index is a measure designed to determine the relevance of individual *formal concepts* in an FCA lattice. Similar to the Stability Index and the Support Value, the Collapse Index sees the *formal concept* as being defined by its attribute set (intent) more so than the set of object instances of its extent. From this vantage point, despite using the term ‘relevancy of the *formal concept*,’ all the three (3) measures are in fact calculations of the ‘relevancy of the *intent* of the *formal concept*’. However the calculation of this relevancy entails the usage of information obtained from the extent.

The Support Value of Equation 2.1 uses the cardinality of the extent of the *formal concept*, expressed as a ratio of all objects in the formal context, as the support of the extent for the intent. For the Stability Index (Equation 2.3) the support for which the extent lends the intent of a *formal concept* is obtained by determining the proportion of sets in the ‘powerset of the extent’ for which the maximal shared set of attributes, of each set, would be the intent of the *formal concept* in question. The higher the proportion of sets the more *stable* the *formal concept* is deemed.

As previously mentioned the calculation of the Stability Index can be a fairly inefficient process in lattices containing many *formal concepts*. The oft-utilised Support Value produces a concept relevancy value with a very simple calculation, however *formal concepts* with low object instance support are too easily discounted using this approach. The Collapse Index is introduced as a compromise between the Support Value and the Stability Index, in that it offers the advantages of the Stability Index as it relates to exhibiting less bias to *formal concepts* with high object instance support, while maintaining the simplicity of calculation as that of the Support Value.

Collapse Index as a *formal concept* relevancy measure, borrows the notion of *stability* from the Stability Index in that it also investigates the notion of a stable or unstable concept. In its approach the Collapse Index investigates the collapsibility of a *formal concept* by determining the ease at which a *formal concept* may be removed from an FCA lattice. A *formal concept* \((A, B)\) is considered removed (collapsed) from the lattice if after some change(s) in the formal context there exists no *formal concept* in the updated lattice where the intent of the *formal concept* is \(B\).

The Collapse Index evokes the expression “*A chain is strong as its weakest link*” as
it entails the determination of the minimum number of object instances that need be removed from a formal context in order for the formal concept in question to collapse. If a formal concept may collapse with the removal of very few object instances from the formal context, then the formal concept is seen as being very unstable. Scenarios where it requires the removal of a large number of object instances for a formal concept to collapse are suggestive of a very stable formal concept.

One may view these concept relevancy measures through the analogy of a house being supported by a set of ten (10) columns. The Support Value defines the stability of the house as the house having 10 columns. The Stability Index defines the stability of the house in terms of the ratio of possible combinations of the 10 columns for which the house would remain upright. The Collapse Index defines the stability of the house by the minimum number of columns that need be removed before the house collapses. E.g. If the house has ten (10) columns but there exists a column that if removed the house will collapse, then the Collapse Index does not consider the house as being stable.

In light of the previous the next step would be to provide a more formal description of the Collapse Index. This commences with the mathematical derivation of the Collapse Index function. Note that a table is provided in Appendix B for quick reference on key symbols utilised in the propositions and proofs present in the thesis.

### 5.3 Derivation of Collapse Index

The first step in deriving the Collapse Index function is to establish that the addition of object instances to a formal context would not affect the presence of a formal concept in the lattice.

**Proposition 1.** Given $K := (G, M, I)$ and $(A, B)$ is a formal concept of $L(K)$, if the object $x$ is added to the formal context, where $x \notin A$, the attribute set $B$ will remain as the intent of a formal concept within $L(K_{x+})$, where $K_{x+}$ is the formal context $K$ with the binary relations of object $x$ included.

**Proof.** If the attribute set of object instance $x (x^a)$ is such that $B \subseteq x^a$ then $(A \cup x)^a = B$ and $B^p = (A \cup x), \Rightarrow (A \cup x, B) \in L(K_{x+})$. Where $x^a \subset B$ then $A^a = B$ and $B^p = A, \Rightarrow (A, B) \in L(K_{x+})$
CHAPTER 5. CONCEPT RELEVANCY AND COLLAPSE INDEX

In light of the previous proof showing that the addition of objects to the formal context will not result in a formal concept’s collapse, the formalisation of the Collapse Index is done solely through investigation of object instances being removed from the formal context. In regards to the removal of objects from the formal context the first proposition is that for an FCA lattice, the removal of an object from the formal context where the object is not an element of the formal concept’s extent, will not result in the collapse of that formal concept.

**Proposition 2.** Given $K := (G, M, I)$ and $(A, B)$ is a formal concept of $L(K)$, if the object $x$ is removed from the formal context, where $x \notin A$, the attribute set $B$ will remain as the attribute set of a formal concept within $L(K_x-)$, where $K_x-$ is the formal context $K$ with the binary relations of object $x$ removed.

**Proof.** If the attribute set of object $x$ is $x^a$ such that $x^a \subset B$ or $x^a \cap B = \emptyset$, this would not affect the presence of $(A, B)$ as $A$ remains a subset of $G \setminus x$. As such, $A^q = B, B^p = A, \Rightarrow (A, B) \in L(K_x-)$. If $x^a = B$ or $B \subset x^a$, then $x$ would have been an element of $A$, as per Definition 3 in Section 2. \hfill \Box

The next step is to show that if there exists at least two objects, each with differing attribute sets, for which this set of objects share a common set of attributes, then the intersection would be the intent of a formal concept in the lattice.

**Proposition 3.** Given formal context $K := (G, M, I)$, if there exists multiple objects $x_i$ in $G$ which share a common set of attributes $B$, where $B$ is the maximal shared subset of their differing overall attribute sets, then a formal concept will exist in $L(K)$ where $B$ is its intent.

**Proof.** Given $C_i = \{ x \in G : x^a = B \cup H_i \}$ then $C_i^a = B \cup H_i$, where $H_i \subseteq M, H_i \notin B$. If $C_1, C_2, \ldots C_n \subseteq G$, where $H_i \neq H_j$, then $(C_1 \cup C_2 \cup \ldots \cup C_n)^a = B, B^p = (C_1 \cup C_2 \cup \ldots \cup C_n) \Rightarrow ((C_1 \cup C_2 \cup \ldots \cup C_n), B) \in L(K)$. \hfill \Box

Another important proposition is that objects for which their individual attribute sets is $B$, must all be removed from the formal context in order for the formal concept whose intent is $B$ to collapse.

**Proposition 4.** Given $K := (G, M, I), (A, B) \in L(K)$, and $O_B = \{ x \in G : x^a = B \}$, if a set of objects is removed from the formal context $K$, such that a subset $O \subseteq O_B$
remains in the formal context, then \( B \) will exist as the intent of a formal concept in \( L(K_{x^*}) \).

**Proof.** If there exists \( P \subset A \) such that \( B \subset P \), then \( (O \cup P)^\circ = B \) and \( B^\circ = (O \cup P) \Rightarrow ((O \cup P), B) \in L(K) \). If there does not exist a \( P \subset A \) such that \( B \subset P \), then \( O^\circ = B \) and \( B^\circ = O \Rightarrow (O, B) \in L(K_{x^*}). \)

Combining Proposition 2 and 3 it is evident that to achieve the collapse of a formal concept \((A, B)\) would thus require at least the removal from the formal context the set of objects \( O_B \), as well as the removal of sufficient \( x \in A \setminus O_B \), such that there exists no pair of \( x_i, x_j \) where \( x_i^\circ \neq x_j^\circ \) and \( x_i^\circ \cap x_j^\circ = B \).

This is demonstrated by showing that, given the removal of \( O_B \) from the formal context, it is not possible for \( B \) to exist as the extent of a formal concept if the number of lower neighbours of \((A, B)\) is 1.

**Proposition 5.** There does not exist \((A, B) \in L(K) \) where \(|O_B| = 0 \) and \(|(C_i, D_i)| = 1 \), where \((C_i, D_i) \prec (A, B)\).

**Proof.** If \(|O_B| = 0 \) then \( A \setminus (C_1 \cup C_2 \cup ... \cup C_n) = 0 \). If \(|(C_i, D_i)| = 1 \) then \( C_1 \cup C_2 \cup ... \cup C_n = C_k \) and \( A \setminus C_k = 0 \Rightarrow A = C_k \Rightarrow (C_k, D_k) = (A, B) \) (contradiction).

The previous represents the conditions which must be met in order for a formal concept \((A, B)\) to collapse. All elements in \( O_B \) must be removed, as well as sufficient elements of the extents of the lower neighbours of \((A, B)\), such that a theoretical point is approached where only one lower neighbour of \((A, B)\) remains. The minimum number of objects from the lower neighbours that need be removed is now explained.

**Proposition 6.** Assuming the prior removal of \( O_B \) from the formal context, and given the set of formal concepts \((C_i, D_i)\) such that \((C_i, D_i) \prec (A, B)\) where \(|C_1| \leq |C_2| \leq ... \leq |C_n|\), then the minimum number of objects that need to be further removed from lattice such that \( B \) is no longer the intent of any formal concept in lattice, is \(|(C_1 \cup C_2 \cup ... \cup C_n)| - |C_n|\).

**Proof.** If \( R_m \) is the set of elements of \((C_1 \cup C_2 \cup ... \cup C_n)\) which remain in the formal context after \( O_B \) is removed, and \( R_m \) is such that \(|R_m| > |C_n|\) then there exists at least one pair of objects, \( x_i, x_j \), where \( x_i \in C_i, x_j \in C_j, x_i \notin C_j, \) and \( x_j \notin C_i \). This
is as the \( \max\{|C_1|, |C_2|, \ldots, |C_n|\} = |C_n| \). For this pair of values \( x_i^a \cap x_j^a = B \), and as per proof of Proposition 2, this implies the presence of a formal concept with \( B \) as its intent. If \( |R_m| = |C_n| \) then it is possible that \( R_m = C_n \), one lower neighbour. From proof of Proposition 4, the theoretical presence of one lower neighbour, and \( O_B \) being absent, is contradictory to the presence in the lattice of a formal concept with attribute set \( B \); the implication being a collapse of such a formal concept.

From the above one concludes that given \((A,B)\) is a formal concept within the lattice of \( K := (G,M,I) \), where \((C_i,D_i)\) is such that \((C_i,D_i) \prec (A,B)\) and \(|C_1| \leq |C_2| \leq \ldots \leq |C_n|\), the minimum number of objects that need be removed from formal context \( K \) in order for there to be no formal concept with \( B \) as its intent, is the number of objects in the extent \( A \), less the number of objects in the largest extent of the lower neighbours, \( C_n \). This difference is the equivalent of ‘\( |O_B| \) plus all other objects which need be removed’.

\[
ci(A, B) = \frac{|A| - |C_n|}{|G|}
\]  

(5.1)

Similar to the Support Value, the Collapse Index shown in Equation 5.1 is expressed as a ratio between the ‘minimum number of objects that need be removed’ and the overall number of objects in the original formal context. While there may be value in expressing the Collapse Index ratio with respect to the size of the formal concept’s extent rather than the overall size of the formal context, in the case of the Collapse Index the overall size \( |G| \) is used.

One of the things the usage of \( |G| \) does is to distinguish between a pair of formal concepts which require the same minimum amount of object be removed from their formal context in order to collapse, but each formal concept is derived from formal contexts of different sizes. As an example, if in a domain of ten (10) objects a concept is dependent on two (2) of these objects to exist, is that concept of similar relevance as ‘a concept in a domain of 100 objects and is similarly dependent on two (2)’? The Collapse Index takes the view that both concepts are not of equal relevance and that the concept in the domain of smaller size is of greater relevance.

Stemming from the Collapse Index ratio being defined in terms of the overall size of the formal context, is that by doing so, a maximum value is set on the relevance
of each formal concept in the lattice. If the formal context contains 50 objects and a formal concept taken from said formal context has an extent of size 5, the maximum value which the relevance of the formal concept may attain is 5/50. More or less the Collapse Index is saying that even if it requires all of the five (5) objects of the formal concept’s extent be removed in order for the concept to collapse, since the formal concept remains represented by only five (5) object instances from the set of 50, therefore the formal concept can be at most 10% relevant. A formal concept, as per the Collapse Index, can only be 100% relevant if all objects have to be removed from the formal context if one wants the formal concept to collapse.

5.4 Demonstration of Collapse Index

To depict implementations of the Collapse index, a lattice based on the formal context of Table 5.1 is provided to show these calculations as well as several key attributes of the Collapse Index. The formal context in this case are: a set of generic movies as the set of object instances $G$; a set of genre descriptions typically associated with movies, $M$; and the set of binary relations between these movies and the genre attributes, $I$. The lattice derived from this formal context is shown in Figure 5.1 and is depicted using the reduced labelling formatting approach described in Appendix C.

Table 5.1: Formal Context for Collapse Index Demonstration

<table>
<thead>
<tr>
<th>Movie</th>
<th>act</th>
<th>adv</th>
<th>dra</th>
<th>ani</th>
<th>cri</th>
<th>rom</th>
<th>wes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOVIE A</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOVIE B</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOVIE C</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOVIE D</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOVIE E</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOVIE F</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOVIE G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOVIE H</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOVIE I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOVIE J</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOVIE K</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Attention is first drawn to the formal concept in the lattice labelled as $(\alpha, \beta)$ which possesses as its intent the attribute set \{crime, adventure\} and an extent of cardinality 7. Of its lower neighbours the largest extent is of cardinality 3. The minimum number
CHAPTER 5. CONCEPT RELEVANCY AND COLLAPSE INDEX

Figure 5.1: Lattice Showing Concepts Of High Support but Low Relevance

of object instances which require removal in order for the concept \((\alpha, \beta)\) to collapse is \(7 - 3 = 4\). With the size of the formal context, \(|G|\), being 11, the collapse index value of \((\alpha, \beta)\) amounts to \(ci(\alpha, \beta) = \frac{7 - 3}{11} = 0.36\).

Central to the motivations of the Collapse Index is a desire that it be able to recognize concepts of ‘low support but high relevance’, as well as objects of ‘high support but low relevance’. Collapse Index’s compliance with such criteria is shown via the example of the formal concept \((\gamma, \delta)\) whose sole attribute description is crime. Given its position near the top of the lattice hierarchy, as expected, \((\gamma, \delta)\) has a very high support, that being of extent cardinality 8. Of its lower neighbours, the largest extent is of size 7. \(ci(\gamma, \delta) = \frac{8 - 7}{11} = 0.09\).

This is a very low relevance value in comparison to the relevance if calculated using the Support Value of Stumme et al. (2002), where the relevance is determined by the ratio of objects in the extent of \((\gamma, \delta)\) relative to all object instances in the formal context \(K\). \((\gamma, \delta)\) would thus have a Support Value of \(8/11 = 0.73\). From the viewpoint of the Support Value \((\gamma, \delta)\) is a very stable concept as a result of 8 of the 11 objects of the formal context having \{crime\} as an element of their intent. However from the perspective of the Collapse index the formal concept \((\gamma, \delta)\) is highly unstable

\[\text{This is the numerator of the Collapse Index function of Equation 5.1}\]
as, although it has high support, the removal of only one object (MOV, I in this case) would lead to the collapse of (γ, δ).

Moreover, one may now compare the Collapse Index relevancy values of the formal concept (α, β) to that of (γ, δ). The concept (α, β) is a subconcept of (γ, δ) yet has a higher relevancy value than (γ, δ). As objects accumulate upwards in the lattice\(^2\), the subconcept of a formal concept would have a smaller extent than its parent concept. In this case though, the subconcept (α, β) despite having lower object support has a relevancy value of 0.36 while its superconcept has a relevancy of 0.09. This is an example of the Collapse Index being able to grant formal concepts of lower object support a relevance greater than formal concepts with high object support. Of course this could be interpreted or read as the opposite ‘able to grant formal concepts of higher object support a relevance lower than formal concepts with low object support’.

Having said that, despite the more balanced viewpoint which the Collapse Index has with respect to high or low support object of a formal concept, in most cases formal concepts of high instance support would be recognised as being of high relevance. This is due to the fact that formal concepts with high support are usually at high levels in the lattice. Although the maximum size of the extent of their lower neighbours, |C\(_n\)|, would be high, so too would the size of the extent of the formal concept itself, |A|, ultimately leading to a high value of |A| − |C\(_n\)|. This is in contrast to formal concepts lower in the lattice with low values of |C\(_n\)| but also low values of |A|. However it is important that when situations arise where some nuance is required to recognise formal concepts that go against these norms, that the Collapse Index is able to provide this clarity.

### 5.5 Adding Uniform Object Instances

In possession of a formal definition of the Collapse Index measure, further theoretical insight is provided on the Collapse Index. This commences by investigating the changes in the Collapse Index value of a formal concept (A, B) as object instances are added to the formal context. As the addition of a diverse set of object instances has the potential to affect many different formal concepts, analysis is limited to the addition

\(^2\)i.e. that the cardinality of the extent of any formal concept is greater than or equal to that of any of its subconcepts
CHAPTER 5. CONCEPT RELEVANCY AND COLLAPSE INDEX

of uniform\(^3\) object instances to the formal context.

Firstly the argument is made that, given the presence of the formal concept \((A, B)\) in the lattice, the addition of a set of uniform object instances to the formal context where these object instances do not share the attribute set \(B\) with the extent of the formal concept \((A, B)\), would decrease the Collapse Index value of \((A, B)\).

Proposition 7. Given \(K := (G, M, I)\) and \((A, B)\) is a formal concept in \(L(K)\), if a set of uniform objects \(\Gamma\) are added to the context \(K\) where \(G \cap \Gamma = \emptyset\), \(B \not\subset \Gamma\), and if \((C_1, D_1) \prec (A, B)\) then \(|C_1| \leq |C_2| \leq ... \leq |C_n|\), then \(ci(\Lambda, B) < ci(A, B)\) where \(\Lambda = B^\circ\) in \(L(K_\Gamma^+)\).

Proof. If \(B \not\subset \Gamma\) then \(ci(\Lambda, B) = \frac{|A| - |C_n|}{|G| + |\Gamma|} > \frac{|A| - |C_n|}{|G|} = ci(A, B)\) as denominator has increased after the addition of \(\Gamma\) while \(|A| - |C_n|\) remains constant. \(\square\)

Secondly we show that if the set of object instances being added to the formal context have the same attribute description as the formal concept \((A, B)\), the Collapse Index value of \((A, B)\) will increase.

Proposition 8. Given \(K := (G, M, I)\) and \((A, B)\) is a formal concept in \(L(K)\), if a set of uniform objects \(\Gamma\) are added to the context \(K\) where \(G \cap \Gamma = \emptyset\), \(B = \Gamma^\circ\), and if \((C_1, D_1) \prec (A, B)\) then \(|C_1| \leq |C_2| \leq ... \leq |C_n|\), then \(ci(\Lambda, B) > ci(A, B)\).

Proof. If \(B = \Gamma^\circ\) then \(ci(\Lambda, B) = \frac{(|A| + |\Gamma|) - |C_n|}{|G| + |\Gamma|} = \frac{|A| - |C_n| + |\Gamma|}{|G| + |\Gamma|}\). In a more abstract manner we seek to show that \(\frac{x + my}{y + m} \geq \frac{x}{y}\), where \(|A| - |C_n| = x\), \(|G| = y\), and \(|\Gamma| = m\).

This is true as assuming the opposite leads to the implication that \(y < x\) if \(m > 0\)\(^4\), which is a contradiction in our case as \(|G| \geq |A| > |A| - |C_n|\). \(\square\)

Next the case is made that, all other things being equal, given a formal concept with a known Collapse Index value, the addition of object instances to the formal context, where these newly introduced object instances share the attribute set of the largest lower neighbour of the formal concept, will decrease the Collapse Index value of the formal concept.

Proposition 9. Given \(K := (G, M, I)\) and \((A, B)\) is a formal concept in \(L(K)\), if a set of objects \(\Gamma\) are added to the context \(K\) where \(D_n \subseteq \Gamma^\circ\), \(G \cap \Gamma = \emptyset\), and if

---

\(^3\)uniform object instances are objects which have the same attribute descriptions.

\(^4\)The proof is relatively straightforward so not shown here
5.5. ADDING UNIFORM OBJECT INSTANCES

\((C_i, D_i) \prec (A, B)\) then \(|C_1| \leq |C_2| \leq \ldots \leq |C_n|\), then \(ci(\Lambda, B) < ci(A, B)\) where \((\Lambda, B) \in L(K_{T+})\).

Proof. If \(ci(A, B) = \frac{|A| - |C_n|}{|G|}\) then \(ci(\Lambda, B) = \frac{(|A| + |\Gamma|) - (|C_n| + |\Gamma|)}{(|G| + |\Gamma|)} = \frac{|A| - |C_n|}{|G| + |\Gamma|}\). This is less than \(ci(A, B) = \frac{|A| - |C_n|}{|G|}\) since \(|\Gamma| > 0\). Also see proof of Proposition 7.

Proposition 9 is illustrated using the examples of the addition of object instances to a formal context, where the object instances share the attribute set of the largest lower neighbour of a specific formal concept of focus. Figure 5.2a shows a default subsection of a lattice before the addition of a new set of object instances \(\Gamma\). The uppermost formal concept is the formal concept of interest \((A, B)\). Figure 5.2b represents the addition of object instances \(\Gamma\) where no \(x \in \Gamma\) shares any attribute with any other lower neighbour of \((A, B)\) other than the largest lower neighbour \((C_n, D_n)\). Meanwhile for Figure 5.2c object instances added share attributes with both the largest lower neighbour as well as other lower neighbours of \((A, B)\).

For the purposes of this explanation the lattices in Figures 5.2a to 5.2c are labelled atypically. Labelling of the lattices omits attribute related information and is limited to quantitative information on the object instances. A value inside a node represents the number of object instances whose attribute set is equal to the intent of the formal concept the node represents, while the value outside of the node represents the cardinality of the extent of the corresponding formal concept. This within the thesis is referred to as ‘extent labelling’.

\[\]
For the default lattice shown in Figure 5.2a the Collapse Index value of \((A, B)\) the topmost concept is \(\frac{9-5}{|G|} = \frac{4}{|G|}\) as a result of the largest extent of its lower neighbours being \(|C_n| = 5\). Adding an object instance (Figure 5.2b) that has the same attribute set as the largest lower neighbour increases the cardinality of that largest lower neighbour to 6 as well as increases the cardinality of the extent of \((A, B)\) to 10. The Collapse Index value of the topmost concept, now \((\Lambda, B)\), is now \(\frac{10-6}{|\Gamma|+1} = \frac{4}{|\Gamma|+1}\), a lesser value than \(\frac{4}{|G|}\). If the newly introduced object instance share attributes with other lower neighbours of \((A, B)\), as per Figure 5.2c the same Collapse Index value for \((\Lambda, B)\) is obtained, \(\frac{10-6}{|\Gamma|+1}\).

Essentially the Collapse Index measure is making the argument that if there is a (increasingly) dominant lower neighbour of a particular formal concept, this diminishes the value of the formal concept itself. Translated to a domain this implies that a subclass in a taxonomy with an increasingly high number of associated instances lessens the need for its parent class. If in a movie dataset there are thousands of examples of romantic-comedies, this means that a parent class romance is less significant in the description of the domain. This would be especially true if there is a low representation of other romance-related movies.

A new case is now made that if, for example, the representation of movies in other immediate sub-classes of romance were to be increased to approach the high level of the representation of romantic-comedy, this would strengthen the relevancy of the romance class in the movie taxonomy, as per the Collapse Index measure, if these new movies were not romantic-comedies themselves. In essence, all things being equal, the addition of uniform object instances to the formal context, where \(B\) is a subset of the attribute set of these new object instances, while the intent of the largest lower neighbour of \((A, B)\) is not a subset of the attribute set of the new object instances, will increase the relevancy value of \((A, B)\) as long as no lower neighbour now has an extent of cardinality larger than the previous maximum.

**Proposition 10.** Given \(K := (G, M, I)\) and \((A, B)\) is a formal concept in \(L(K)\) where if \((C_i, D_i) \prec (A, B)\) then \(|C_1| \leq |C_2| \leq \ldots \leq |C_n|\), if a set of objects \(\Gamma\) are added to the context \(K\) where \(G \cap \Gamma = \emptyset\) and \(\exists C_i\) such that \(C_i^a \subseteq \Gamma^a, C_i \neq C_n,\) and \(|C_i \cup \Gamma| \leq |C_n|\), then \(ci(\Lambda, B) > ci(A, B)\) where \((\Lambda, B) \in L(K^+).\)

**Proof.** If \(\not\exists C_i\) where \(C_i^a \subseteq \Gamma^a\) and \(|C_i \cup \Gamma| < |C_n|\) then \(ci(\Lambda, B) = \left(\frac{|A|+|\Gamma|-|C_n|}{|G|+|\Gamma|}\right).\) This
may be rewritten as $\frac{|A| - |C_n| + |\Gamma|}{|G| + |\Gamma|}$ which is greater than $ci(A, B) = \frac{|A| - |C_n|}{|G|}$ as earlier shown in proof of Proposition 8.

It may be the case that the addition of uniform object instances to the formal context results in a new lattice where a lower neighbour of $(\Lambda, B)$ ends up having a greater cardinality than that of the previous largest lower neighbour $C_n$. In such an instance the Collapse Index of $(\Lambda, B)$ will become less than that of $(A, B)$ when the number of objects being added goes past a certain threshold.

**Proposition 11.** Given $K := (G, M, I)$ and $(A, B)$ is a formal concept in $L(K)$ where if $(C_1, D_1) \prec (A, B)$ then $|C_1| \leq |C_2| \leq ... \leq |C_n|$, if a set of uniform objects $\Gamma$ are added to the context $K$ where $G \cap \Gamma = \emptyset$ and $\exists C_i$ such that $C_i^o \subseteq \Gamma^o$, $C_i \neq C_n$, and $|C_i \cup \Gamma| > |C_n|$, then there exists $r \in \mathbb{R}$ such that if $|\Gamma| > r$ then $ci(\Lambda, B) \leq ci(A, B)$

**Proof.** If $C_k^o \subseteq \Gamma^o$, $|C_k \cup \Gamma| > |C_n|$ and $\exists C_j : |C_j \cup \Gamma| > |C_k \cup \Gamma|$ then

$$ci(\Lambda, B) = \frac{(|A| + |\Gamma|) - (|C_k| + |\Gamma|)}{|G| + |\Gamma|} = \frac{|A| - |C_k|}{|G| + |\Gamma|}.$$  
If we let $|A| - |C_n| = x$, $|G| = y$, $|\Gamma| = m$, and $(|C_n| - |C_k|) = n$ then

$$ci(A, B) = \frac{x}{y},$$
$$ci(\Lambda, B) = \frac{x + n}{y + m}$$

If $ci(\Lambda, B) > ci(A, B) \implies \frac{x + n}{y + m} > \frac{x}{y} \implies \frac{n}{m} > \frac{x}{y}$

$$\implies \frac{|C_n| - |C_k|}{|\Gamma|} > ci(A, B) \implies |\Gamma| < \frac{|C_n| - |C_k|}{ci(A, B)}$$

In light of Proposition 11 an additional proposition is made that the continued addition of uniform object instances to another lower neighbour of a formal concept $(A, B)$ will eventually achieve a maximum Collapse Index for $(\Lambda, B)$ before it commences declining and eventually becomes less than the original $ci(A, B)$.
Chapter 5. Concept Relevancy and Collapse Index

Proposition 12. Given $K := (G, M, I)$ and $(A, B)$ is a formal concept in $L(K)$ where if $(C_i, D_i) \prec (A, B)$ then $|C_1| \leq |C_2| \leq \ldots \leq |C_n|$, if a set of uniform objects $\Gamma$ are added to the context $K$ where $G \cap \Gamma = \emptyset$, $C_k \neq C_n$, $C_k^q \subseteq \Gamma^q$ and $\exists C_j : |C_j \cup \Gamma| > |C_k \cup \Gamma|$, then $\max \limits_{|\Gamma| \in \mathbb{Z}} ci(\Lambda, B)$ where $ci(\Lambda, B) = f(|\Gamma|)$ exists for $|\Gamma|$ in the domain $[0, \frac{|C_n| - |C_k|}{ci(A, B)}]$.

Proof. From Proposition 11 it is shown that the addition of $|\Gamma|$ object instances eventually leads to a point where $ci(\Lambda, B) \leq ci(A, B)$ if $\exists C_i : |C_i \cup \Gamma| > |C_n|$ and $C_i^q \subseteq \Gamma^q$. Meanwhile Proposition 10 shows the addition of object instances $\Gamma$ to the formal context increases the value of $ci(\Lambda, B)$ assuming that $\forall C_i : C_i^q \subseteq \Gamma^q$, $C_i \neq C_n$, and $|C_i \cup \Gamma| \leq |C_n|$. Therefore, in light of Proposition 11 where $|\Gamma|$ is sufficiently large to reduce the $ci(\Lambda, B)$ to be no more than $ci(A, B)$ then there must exist a turning point or a $|\Gamma|$ which maximises the value of $ci(\Lambda, B) = f(|\Gamma|)$ in the domain $[0, \frac{|C_n| - |C_k|}{ci(A, B)}]$.

This is said in light of the fact that the Collapse Index function for the scenario described in Proposition 10, $ci(\Lambda, B) = \frac{|A| - |C_n| + |\Gamma|}{|G| + |\Gamma|}$, is a continuously increasing function as $|\Gamma|$ increases, given the fact that both $(|A| - |C_n|)$ and $|G|$ are constants. At the same time the Collapse Index function for Proposition 11’s scenario, $ci(\Lambda, B) = \frac{|A| - |C_k|}{|G| + |\Gamma|}$, is a continuously decreasing function, as $|\Gamma|$ increases while $(|A| - |C_k|)$ remains constant. Consequently it suggests that the maximum value the $ci(\Lambda, B)$ may attain occurs when $|\Gamma| + |C_k| = |C_n|$ or at the theoretical meeting point of the left and right hand limit of $f(|\Gamma|)$.

$$\max \limits_{|\Gamma| \rightarrow |(C_n| - |C_k|)^-} ci(\Lambda, B) = \lim \limits_{|\Gamma| \rightarrow |(C_n| - |C_k|)^+} ci(\Lambda, B)$$

This information is represented in Figure 5.3 showing key aspects of the function $ci(\Lambda, B)$ discussed in the previous propositions including the value of $|\Gamma|$ at which $ci(\Lambda, B)$ achieves it maximum value and the value of $|\Gamma|$ at which $ci(\Lambda, B)$ becomes less than $ci(A, B)$.

Further, this behaviour of the Collapse Index growth is represented, by example, in the Figures 5.4a to 5.5c. Once again referencing the default lattice in Figure 5.2a where the formal concept $(A, B)$ is the uppermost formal concept, in each of these figures a set of uniform object instances, $\Gamma$, are added to a lower neighbour of $(A, B)$ where the attribute set of the current largest formal concept (rightmost formal concept), $C_n$,
5.5. **ADDING UNIFORM OBJECT INSTANCES**

Figure 5.3: Growth of $ci(A, B)$ with respect to $|\Gamma|$

is not a subset of the attribute set of the object instances being introduced to the lattice. More or less the size of another lower neighbour of $(A, B)$ other than $C_n$ is being increased. In the figures this ‘other lower neighbour’ is the left-most node.

---

![Diagram](image)

**Figure 5.4: Adding Object Instances to Smaller Lower Neighbours**

In Figure 5.4a one object instance ($|\Gamma| = 1$) is added to the other lower neighbour of $(A, B)$. As per Proposition 10 this should increase the value of the top most *formal concept* $(\Lambda, B)$. If we assume that the FCA lattice in Figure 5.2a, the default lattice, is a complete lattice, then the Collapse Index of the uppermost *formal concept* is $\frac{9-5}{9} = 0.44$. The addition of the 1 object instance in Figure 5.4a results in $ci(A, B) = \frac{10-5}{19} = 0.5 > 0.44$, as it should. A similar process for Figure 5.4b where 2 object
instances are added to the left-most node also increases $ci(\Lambda, B)$ to $0.54 > 0.5 > 0.44$.

If enough object instances are added to the left-most *formal concept* such that the cardinality of its extent is now equal to the previous largest node, then $ci(\Lambda, B)$ should now be the maximum it could be (assuming all other things being equal) as per Proposition 12. Figure 5.4c represents such a scenario in that the left-most node now has a cardinality of 5 equal to the previous highest. $ci(\Lambda, B)$ is now $\frac{12 - 5}{12} = 0.583 > 0.54 > 0.5 > 0.44$. If this constitutes a maximum value of $ci(\Lambda, B)$ then the addition of further object instances to the left-most node should result in a decline in $ci(\Lambda, B)$.

![Diagram](image)

Figure 5.5: Adding Object Instances. Smaller Lower Neighbour(s) Now Larger

Figures 5.5a, 5.5b, and 5.5c are instances where the left-most node is now of a greater size than the originally highest lower neighbour, producing $ci(\Lambda, B)$ values of 0.538, 0.5, and 0.438 respectively, each showing a successive decline as object instances are added to the left-most node. Particularly noteworthy is the value of 0.438 for Figure 5.5c which is less than $ci(\Lambda, B) = 0.44$ of the default lattice of Figure 5.2a. This corresponds to Proposition 11 where it is posited that $ci(\Lambda, B)$ cannot be greater than $ci(A, B)$ if $|\Gamma| > \frac{|C_n| - |C_k|}{ci(A, B)}$. In the case of Figure 5.5c, $\frac{|C_n| - |C_k|}{ci(A, B)} = \frac{5 - 2}{0.44} = 0.6814$ therefore values of $|\Gamma| > 0.6814$ should result in $ci(\Lambda, B) < 0.44$. $|\Gamma| = 7 > 0.6814$ and $ci(\Lambda, B) = 0.438 < 0.44$ exhibiting the predicted behaviour.
5.6 Lower Neighbours Distribution

Further, as opposed to the previous propositions where it was assumed that the set of object instances added to the formal context are uniform, we now consider a variety of rules governing the behaviour of the Collapse Index value of a formal concept \((A, B)\) with respect to the distribution of the set of object instances \((\Theta)\) across the lower neighbours of \((A, B)\), where these objects are diverse\(^5\) and are instantiations of subconcepts of \((A, B)\).

The first proposition in this regard is that for this set \(\Theta\) of object instances, the Collapse Index of \((A, B)\) achieves a maximum value if the objects are distributed across the lower neighbours of \((A, B)\) in such a manner that size of the extent of its largest lower neighbour is 1.

**Proposition 13.** Given \(K := (G, M, I)\) and \((A, B)\) is a formal concept in \(L(K)\) where \(A = O_B \cup \Theta\) and if \((C_i, D_i) \prec (A, B)\) then \(|C_1| \leq |C_2| \leq \ldots \leq |C_n|\), then the max \(ci(A, B)\) is achieved if \(|C_n| = 1\)

**Proof.** Given that \(|A|\) and \(|G|\) are constant in this scenario, \(ci = \frac{|A| - |C_n|}{|G|}\) maximised if \(|C_n|\) is minimised. Minimum value of \(|C_n|\) is 1 as \(|C_n| \in \mathbb{Z}, C_n > 0\).

\[\Box\]

If the largest lower neighbour of \((A, B)\) is of size 1 then the implication is that all other lower neighbours must also have an extent of cardinality 1 as \(0 \leq |C_i| \leq 1\). If this is to be the case this translates as a concept in a taxonomy being especially important if it has many subclasses, each subclass having only one representative object. As an example in Figure 5.6a the formal concept \((A, B)\) (top-most concept) has a greater Collapse Index value than the same concept in Figure 5.6b, \(\frac{8-1}{|C|}\) vs. \(\frac{8-3}{|C|}\). Note also that if a lower neighbour has an extent of cardinality of 1, it has no subconcepts.

Noticeable in Figure 5.6a was the uniform distribution\(^6\) of the lower neighbours of \((A, B)\), this was explained by the fact that \(|C_n|\) being 1 value meant that all other \(|C_i|\) values had to be 1 as well. However in instances where \(|C_n| \neq 1\) does the distribution of the lower neighbours have to be uniform? The next proposition makes the case

\(^5\) *diverse* in this context means that the objects being considered may have different attributes

\(^6\) *Uniform distribution* refers to each of the lower neighbours having an extent of the same cardinality
that the distribution has to be uniform across the lower neighbours to maximise the Collapse Index value. This assertion is made under the assumption that: \( \Theta \) is the set of object instances represented in the lower neighbours of \((A, B)\), no pair of the lower neighbours share a common subconcept, the number of lower neighbours \((n)\) is a constant, and \( n \) is a factor of \( |\Theta| \).

**Proposition 14.** Given \( K := (G, M, I) \) and \((A, B)\) is a formal concept in \( L(K) \) where \( A = O_B \cup \Theta, |\Theta| = np, p \in \mathbb{Z}, \frac{1}{n}(E_i, F_i), (C_j, D_j), (C_k, D_k) : (E_i, F_i) < (C_j, D_j) \) and \((E_i, F_i) < (C_k, D_k)\), and if \((C_i, D_i) < (A, B)\) then \(|C_1| \leq |C_2| \leq \ldots \leq |C_n|\), then the max \( ci(A, B) \) is achieved if \(|C_1| = |C_2| = \ldots = |C_n|\).

**Proof.** If proposition is true then \( \frac{|\Theta|}{n} \leq max\{|C_i|\} \) where \( \sum_{i=1}^{n} |C_i| = |\Theta|, |C_i| \in \mathbb{Z}, \) and \(|C_i| > 0\). We let \( |C_k| = max\{|C_i|\} \) and assume \( \frac{|\Theta|}{n} > |C_k| \), the opposite of the proposition.

\[
\frac{|\Theta|}{n} > |C_k| \implies |C_k| < \frac{|\Theta|}{n}
\]

\[
\sum_{i=1}^{n} \frac{|\Theta|}{n} = \frac{n|\Theta|}{n} = |\Theta|
\]

\[
|C_k| < \frac{|\Theta|}{n} \implies \sum_{i=1}^{n} |C_i| < |\Theta|
\]

which is a contradiction as \( \sum_{i=1}^{n} |C_i| = |\Theta|\).

Therefore the proposition \( \frac{|\Theta|}{n} \leq max\{|C_i|\} \) is true. \( \square \)

The proof of Proposition 14 has shown that \( ci(A, B) \) is maximised if the extent for its lower neighbours are all equal in size. However this was established under the
assumption that no two lower neighbours share a common subconcept. Proposition 15 posits that if there exists these shared subconcepts between lower neighbours, then the maximum achievable Collapse Index value of \((A, B)\) is less than it would have been in a uniform distribution without.

**Proposition 15.** Given \(K := (G, M, I)\) and \((A, B)\) is a formal concept in \(L(K)\) where \(A = O_B \cup \Theta, |\Theta| = np, p \in \mathbb{Z}, \exists (E_i, F_i), (C_j, D_j), (C_k, D_k) : (E_i, F_i) < (C_j, D_j)\) and \((E_i, F_i) < (C_k, D_k)\), and if \((C_i, D_i) < (A, B)\) then \(|C_1| \leq |C_2| \leq \ldots \leq |C_n|\), then \(ci(A, B)\) is less than \(\max ci(A, B)\) if no such \((E_i, F_i)\) existed.

**Proof.** If there exists in the lattice such an \((E_i, F_i)\) then \(|C_j \cap C_k| > 0\) meaning that there exists objects (duplicates) which appear in both \(C_j\) and \(C_k\). In instances where there are multiple \((E_i, F_i)\)s, at multiple levels of the hierarchy, across multiple \((C_i, D_i)\)s, the calculation of \(|duplicates|\) is fairly complex and outside of the scope of this proof. However we know that \(|duplicates| > 0\); with this in mind and the knowledge that the shared number of object instances between the lower neighbours remains \(|\Theta|\) then

\[
|\Theta| = \sum_{i=1}^{n}|C_i| - |duplicates| \implies \sum_{i=1}^{n}|C_i| = |\Theta| + |duplicates|.
\]

If \(\forall (C_i, D_i) < (A, B), |C_i| \leq \frac{|\Theta|}{n}\) then \(\sum_{i=1}^{n}|C_i| \leq |\Theta| < |\Theta| + |duplicates|\) (contradiction given that \(|duplicates| > 0\)). Therefore \(\exists C_i : |C_i| > \frac{|\Theta|}{n}\).

If that be the case then \(ci(A, B)\) is less in this scenario than it was for a uniform distribution of \(\Theta\) where \(|C_n| = \frac{|\Theta|}{n}\).

Elaborating on Proposition 14 and 15, several examples of lattices (using extent labelling) with varied distributions of \(\Theta\) are provided.

In Figure 5.7a \((A, B)\) is once again the top-most formal concept and has \(n = 2\) lower neighbours. \(|\Theta| = 100, |O_B| = 1\) and there exists no subconcepts common to both lower neighbours. As per Proposition 14 the maximum value of \(ci(A, B)\) would be achieved in such a lattice as the distribution of object instances among the lower neighbours is uniform. i.e. \(|C_i| = \frac{|\Theta|}{n} = \frac{100}{2} = 50\). Assuming the lattice in Figure 5.7a is a complete lattice, \(ci(A, B) = \frac{101-50}{101} = 0.505\).
In contrast the lower neighbours of Figure 5.7b, where $|\Theta|$ remains 100 and $n = 2$, have an uneven distribution with respect to the size of their extents. With both extents being 75 and 25, $|C_n| = 75$ and $ci(A, B)$ now becomes $\frac{101 - 75}{101} = 0.257$, a value less than the 0.505 of uniform distribution.

Figure 5.7: Lattices Showing Various Distributions of $\Theta$

If subconcepts are introduced to the lattice such that the subconcept is common to at least two lower neighbours, then Proposition 15 shows that $ci(A, B)$ should be less than the value of $ci(A, B)$ when no shared subconcepts existed and the distribution was uniform. Figure 5.8a represents such a scenario where such a subconcept exists. In its lattice $ci(A, B) = \frac{101 - 60}{101} = 0.406$ which is similarly less than the 0.505 of uniform distribution with no shared subconcepts.

Figure 5.8: Effect of Shared Subconcepts of Lower Neighbours on Collapse Index

Note that the subconcepts must be shared across at least two (2) lower neighbours.
5.6. LOWER NEIGHBOURS DISTRIBUTION

to lessen the Collapse Index value. In the lattice shown in Figure 5.8b many subconcepts exist for the two lower neighbours of \((A, B)\) however none of these subconcepts are shared across the two lower neighbours. \(ci(A, B) = \frac{101 - 50}{101} = 0.505\), the same as that of uniform distribution and no shared subconcepts. In fact for Figure 5.8b there is uniform distribution and no shared subconcepts, hence the achievement.

Particular attention must be drawn to the fact that the claims in Proposition 14 and 15 are made with the expectation that the number of lower neighbours, \(n\), is a factor of \(|\Theta|\). This will of course not necessarily always be the case. In such a situation where \(n\) is not a factor, the maximum value of \(ci(A, B)\) may be ascertained and is arrived at if the maximum cardinality of the extents of the lower neighbours is the smallest integer greater than \(\frac{|\Theta|}{n}\). The first proposition in this regard is made assuming no shared subconcepts between lower neighbours of \((A, B)\).

**Proposition 16.** Given \(K := (G, M, I)\) and \((A, B)\) is a formal concept in \(L(K)\) where \(A = O_B \cup \Theta, \not\exists p \in \mathbb{Z} : |\Theta| = np, \not\exists (E_i, F_i), (C_j, D_j), (C_k, D_k) : (E_i, F_i) < (C_j, D_j)\) and \((E_i, F_i) < (C_k, D_k)\), and if \((C_i, D_i) < (A, B)\) then \(|C_1| \leq |C_2| \leq ... \leq |C_n|\), then the \(ci(A, B)\) is maximum when \(|C_n| = \lceil\frac{|\Theta|}{n}\rceil\)

**Proof.** If \(\not\exists p \in \mathbb{Z} : |\Theta| = np\) then \(|\Theta| = nq + R\), where \(R = \text{remainder}\), \(n\) is the number of lower neighbours, and \(q \in \mathbb{Z} : q = \lfloor\frac{|\Theta|}{n}\rfloor\). If \(\max\{|C_i|\} = q \implies \sum_{i=1}^{n} |C_i| \leq \sum_{i=1}^{n} q = nq < |\Theta|\) (contradiction as \(|\Theta| = nq + R\))

Since \(\max\{|C_i|\}\) cannot be less than \(q\), we seek to show that \(\max\{|C_i|\}\) can be the next higher integer \(q + 1\). We do this by first letting \(|\{C_i : |C_i| = q + 1\}| = s\). If we subtract all their associated object instances from \(|\Theta|\) we get

\[
|\Theta| - s(q + 1)
= nq + R - s(q + 1)
= nq - sq + R - s
\]

Combining all object instances would result in \(\Theta\), therefore

\[
|\Theta| = s(q + 1) + nq - sq + R - s
= s(q + 1) + (n - s)q + R - s
\]
The previous represents a possible distribution of the extents of the lower neighbours of \((A, B)\). \(s(q + 1)\) represents the object instances of \(\{C_i\}\) where \(|C_i| = (q + 1)\) and \((n - s)q\) represents the object instances of \(\{C_i\}\) where \(|C_i| = q\). Both \(s(q + 1)\) and \((n - s)q\) account for all lower neighbours of \((A, B)\) as \(s + (n - s) = n\). Therefore \((R - s) = 0 \implies R = s\).

Given this, it can be said that there exists at least one scenario where a lower neighbour has an extent of cardinality \(q + 1\). If \(|\Theta| = nq + R\) then there may exist \(R\) lower neighbours of \((A, B)\) where the cardinality of the lower neighbour is equal to \(q + 1\). Since we have shown that \(q\) cannot exist as the minimum value of \(|C_i|\) where \(|\Theta| = nq + R\), and we have shown that \(q + 1\) can exist as the minimum, then the maximum of \(ci(A, B)\) for the scenario described in the proposition occurs when the \(|C_n| = q + 1\).

Examples of lattices which abide by Proposition 16 are shown in Figures 5.9a, 5.9b, 5.9c, and 5.10. In each case \(|\Theta| = 17\) which needs to be distributed amongst 4 lower neighbours of \((A, B)\) (the top-most formal concept in the set of figures). In each case if \(|\Theta| = 17\) and \(n = 4\) then \(q + 1 = \lceil \frac{17}{4} \rceil = 5\). Maintaining the value 5 as the maximum cardinality of an extent leads to Figures 5.9a, 5.9b, and 5.9c achieving the maximum value of \(ci(A, B) = \frac{18 - 5}{18} = 0.722\), despite each having various distributions of \(|C_i|\).

Meanwhile Figure 5.10 where \(|C_n| = 14 > 5\) has a lower value of \(ci(A, B)\) as \(|C_n|\) is much larger than \((q + 1) = 5\). Attention is paid to the fact the variance for the distribution of \(\Theta\) represented in this lattice is 42.25, which is comparatively high.
with respect to those of Figures 5.9a, 5.9b, 5.9c whose variances are 2.25, 0.92, 0.25 respectively. This large difference is due to the fact that the variance function ($var = \sum_{i=1}^{n}(x-\mu)^2$), by squaring the value of $(x - \mu)$, penalises large differences between $x$ values and the mean $\mu$.

The value of this observation is to point out that where the distribution of $\Theta$ cannot be uniform, the Collapse Index value of $(A, B)$ would tend to be high if the variance between the lower neighbours is low relative to variance of other possible distributions. Low-variance distributions would tend to have values of $|C_i|$ being close to the mean, and subsequently to each other. As the mean is such that $q \leq \mu = \frac{|\Theta|}{n} \leq (q + 1)$ then for low variances, the highest values of $|C_i|$ have a high probability of being close to $q + 1$ (the value at which $ci(A, B)$ is maximum). High variance distributions possess in contrast, on average, values which are further away from the mean, increasing the likelihood of a high value of $|C_n|$ in tandem with a low value of $ci(A, B)$.

The proof of Proposition 16 showed that, where $\Theta$ cannot be evenly distributed among the $n$ lower neighbours of a concept, then the maximum value that the Collapse Index of the concept could achieve is attained when the largest lower neighbour is of extent cardinality $q + 1 (q = \lceil \frac{|\Theta|}{n} \rceil)$. This was done under the auspices however that no object instance was a member of multiple lower neighbours of the concept. Proposition 17 shows that the maximum attainable value of the Collapse Index remains $q + 1$ if an object instance is shared across multiple lower neighbours.

**Proposition 17.** Given $K := (G, M, I)$ and $(A, B)$ is a formal concept in $L(K)$ where $A = O_B \cup \Theta$, $\exists p \in \mathbb{Z} : |\Theta| = np$, $\exists (E_i, F_i), (C_j, D_j), (C_k, D_k) : (E_i, F_i) < (C_j, D_j)$ and $(E_i, F_i) < (C_k, D_k)$, and if $(C_i, D_i) < (A, B)$ then $|C_1| \leq |C_2| \leq ... \leq |C_n|$, then the $ci(A, B)$ is maximum when $|C_n| = \lceil \frac{|\Theta|}{n} \rceil$.
Proof. As previously discussed in Proposition 15 if there exists in the lattice such an \((E_i, F_i)\) subconcept then there exists duplicates which appear in both \(C_j\) and \(C_k\). Given that \(q \in \mathbb{Z}: q = \lfloor |\Theta| \rfloor, R = \text{remainder}\) then under the circumstances described \(|\Theta| = qn + R\) and

\[
|\Theta| = qn + R = \sum_{i=1}^{n} |C_i| - |\text{duplicates}|.
\]

If \(\forall (C_i, D_i) < (A, B), |C_i| \leq q\) then \(\sum_{i=1}^{n} |C_i| \leq qn \leq qn + R\) (contradiction). Therefore \(\exists C_i : |C_i| > q\). If having the maximum value of \(|C_i|\) being \(q\) is not possible, and the proof of Proposition 16 shows that having a maximum \(|C_i|\) value being \(q + 1\) is possible\(^7\), then the maximum value of \(ci(A, B)\) when subconcepts are shared among lower neighbours and a uniform distribution amongst lower neighbours is unachievable, is arrived at when \(|C_n| = \lceil |\Theta|/n \rceil = q + 1\).

The final proof, that of Proposition 17, shows that the maximum value of the Collapse Index, for uneven distributions of \(\Theta\), remains \(q + 1\) even if there exists shared subconcepts among the lower neighbours of \((A, B)\). The existence of values of \(|C_i| > (q + 1)\) will result in an ever-decreasing value of \(ci(A, B)\). Similar to examples where there were no shared subconcepts, low variances between the size of lower neighbours of \((A, B)\) increase the probability of having a high \(ci(A, B)\).

5.7 Algorithm and Efficiency

Apart from the theories which frame the construction of the Collapse Index measure, a generic algorithm is presented whose input is an FCA lattice \(L(K)\) where \(K := (G, M, I)\) and output is the Collapse Index value of all the lattice’s formal concepts. Algorithm 1 is not proffered as the most optimised method of obtaining these Collapse Index values, however it serves as a straightforward outline on how these values may be calculated.

In effect the algorithm iterates through every formal concept in the lattice \(L(K)\). For each of these formal concepts, \((A, B)\), the algorithm determines the maximum

\(^7\)The section in Proposition 16’s proof that shows that \(q + 1\) is possible, is not affected by the absence or presence of shared subconcepts by lower neighbours of \((A, B)\)
5.7. ALGORITHM AND EFFICIENCY

Algorithm 1 Collapse Indices Algorithm

1: procedure CALCULATE COLLAPSE INDICES
2: \( \text{size} \leftarrow |G| \)
3: for each concept \((A, B)\) in lattice do
4: \( \text{extent} \leftarrow |A| \)
5: \( \text{max} \leftarrow 0 \)
6: for each lower neighbour \((C, D)\) of \((A, B)\) do
7: \( \text{if } |C| > \text{max} \text{ then} \)
8: \( \text{max} \leftarrow |C| \)
9: end if
10: end for
11: \( \text{ci}(A, B) = (\text{extent} - \text{max})/\text{size} \)
12: end for
13: end procedure

cardinality value of the extent of all of \((A, B)\)'s lower neighbours. This maximum value alongside the cardinality of \(A\) and the cardinality of \(G\) are used to calculate the Collapse Index value of \((A, B)\) using Equation 5.1.

Construction of an FCA lattice has a time complexity of \(O(|G|^2|M|L)\) (Zhi, 2014), (Kuznetsov and Obiedkov, 2002), where \(L\) is the number of formal concepts in the lattice. Usage of Algorithm 1 to determine the Collapse Index of a specific formal concept means accessing the formal concept and all its lower neighbours, which can at most be \(L\) formal concepts being accessed. Thus the time complexity of calculating the Collapse Index of a formal concept is \(O(|G|^2|M|L^2)\). More importantly the calculation of the Collapse Indices of all concepts would, as per Algorithm 1, require the repetition of the previous for all \(L\) formal concepts in the lattice. This translates to a time complexity of \(O(|G|^2|M|L^3)\) to calculate all Collapse Index values.

Note that the time complexity being described is a worst case scenario where the entire lattice would have to be regenerated each time a formal concept is accessed. This is unlikely to be the case for algorithm implementations; in such cases the original lattice is expected to be maintained but traversed, as necessary, in an appropriate manner to access the applicable formal concepts.

While the Collapse Index is expected to assign a relevancy value to a formal concept that reflects or shows high correlation with the Stability Index value of that formal concept, the actual numerical value they both output are not expected to be the same. Basically, the Collapse Index is an alternative method of determining the relevance of
a *formal concept*, not an alternative method of calculating the Stability Index.

Despite this, it is still worth comparing the efficiency of both approaches as they target similar ends - though not necessarily value. The main algorithm used in the calculation of the Stability Index (Equation 2.3) is presented in (Roth et al., 2008b), and is reproduced here as Algorithm 2.

**Algorithm 2 Stability Indices Algorithm**

1. **procedure** Compute Stability
2.  
3.  
4.  
5.  
6.  
7.  
8.  
9.  
10.  
11.  
12.  
13.  
14.  
15.  
16.  
17.  
18.  **end procedure**

The premise of Algorithm 2’s approach in the calculation of the Stability Index of a generic *formal concept* \((A, B)\)\(^8\) written as \(\sigma(A, B)\), is that by starting at the base of the lattice and calculating the Stability Index of each subconcept of \((A, B)\), when one eventually arrives in their traversal of the lattice to the concept \((A, B)\), one would then have accumulated sufficient information to determine \(\sigma(A, B)\) itself. By virtue of this, to determine \(\sigma(A, B)\) requires accessing all subconcepts of \((A, B)\), which means accessing at most all \(L\) *formal concepts* in the lattice. Therefore as with the algorithm of \(ci(A, B)\), determining \(\sigma(A, B)\) using Algorithm 2 has a time complexity\(^9\) of \(O(|G|^2|M|L^2)\), the same as \(ci(A, B)\) via Algorithm 1.

That being said, despite this similar time complexity, generally the calculation

---

\(^8\)Note that the *formal concept* \((A, B)\) being referred to here, and for the remainder of this chapter, is representative of a generic *formal concept* and not the \((A, B)\) in Algorithm 2.

\(^9\)Once again this refers to a worst case scenario where the lattice is regenerated each time a *formal concept* needs to be accessed
of $ci(A, B)$ via Algorithm 1 requires accessing less formal concepts than $\sigma(A, B)$ via Algorithm 2, as the former only requires accessing the lower neighbours of $(A, B)$ which is an amount most likely to be less than or equal to the number of subconcepts of $(A, B)$. This applies in worst case scenarios where the lattice is regenerated each time a formal concept is accessed as well as where the original lattice is traversed as required.

Furthermore, while the bulk of Stability Index calculations, $\sigma(A, B)$, in Algorithm 2 take place in the while-loop (steps 7-17), the instantiations of several variables are done in the initial for-loop (steps 3-6) which iterates through every formal concept in the lattice. Within this for-loop the number of lower neighbours of each formal concept is acquired. Now this could be done under the assumption that each data-type representation of a formal concept also contains information on its lower and/or upper neighbours (such as the amount of lower neighbours). If that is to be the case then the lower neighbours themselves would not have to be accessed.

Assuming otherwise it means a separate iteration needs to be done in Algorithm 2 to go through each of $(A, B)$’s lower neighbours to count the amount that exists. In such a case Algorithm 2 in (steps 3-6) already has the same time complexity as the Collapse Index algorithm (Algorithm 1) as those steps iterate through all formal concepts and also iterates through all of the lower neighbours of these formal concepts - the same process as Algorithm 1. Note that this is before even commencing on the main calculations of steps 7-17 in Algorithm 2.

Ultimately, due to these uncertainties, it may be better to compare the efficiency of the Collapse Index and Stability Index algorithms empirically. Comparison on efficiency is one of a variety of Collapse Index experiments conducted in Chapter 6.

5.8 Conclusion

This chapter had the responsibility of providing the details of the Collapse Index relevancy measure. It began with descriptions of the premise of the Collapse Index, which is that ‘the minimum number of object instances whose removal from the formal context would cause a formal concept to collapse is reflective of the relevance of the formal concept’. This idea was then translated to FCA through the mathematical
derivation of the Collapse Index function. A demonstration by example of a key feature of the Collapse Index was then presented - which was the ability of the Collapse Index to afford *formal concepts* of low instance support to have a higher level of relevance than *formal concepts* of high support.

Following this, additional mathematical backing was provided which showed that the relevancy value of a *formal concept* will eventually perpetually decline if the addition of uniform objects to the formal context increases the support level of one of the *formal concept*’s lower neighbours. In addition, this chapter showed that a *formal concept* has its highest levels of relevance when the distribution of object instances among its lower neighbours is even or near even.

Finally the chapter ended with a time complexity comparison between a proposed algorithm of the Collapse Index and the main algorithm of the Stability Index. Both algorithms were of the same time complexity and it was thought best to compare the efficiencies empirically.
Chapter 6

COLLAPSE INDEX

VALIDATION EXPERIMENTS

6.1 Introduction

Within the previous chapter the underlying mathematics which govern the calculation and behaviour of the Collapse Index function were presented. This chapter takes the responsibility of examining empirically the validity of Collapse Index as a relevancy measure. The evaluation takes the form of a set of experiments, each targeting different aspects of concept relevancy.

The first of these experiments (Experiment 1) is found in Section 6.5 and is tasked with determining how well do the relevancy values produced by the Collapse Index measure correlate with the values of other popular concept relevancy measures. Section 6.7 (Experiment 2) investigates whether the values the Collapse Index assigns to formal concepts are consistent across lattices generated from different samples of a dataset. Section 6.9 (Experiment 3) details experiments related to the ability of the Collapse Index to recall concepts in noisy lattices, while Section 6.11 (Experiment 4) contains experiments designed to assess the computational efficiency of the Collapse Index measure.

Firstly however, it is noted that the experiments are conducted on a select set of formal contexts, each of which was derived or partly derived from the BT TV dataset previously described in Section 3.4. This dataset is that of the accumulated set of records on movie-viewing by users and is expected to contain implicit information on
movie concepts as well as their relations. In Section 6.2 the choice of objects and attributes from the BT TV dataset that would be utilised in the formal contexts of the experiments are discussed.

### 6.2 Choice Of Attributes from BT TV Dataset

For the purpose of the formal contexts the desire is to link movies by patterns emergent from users interacting with the movies. These patterns of User Interaction and Context (UIC) are defined with respect to: the actions users take on the movie; the characteristics of the users who show high affinity with the movie; and the contexts of the users who show high affinity towards the movie. By grouping or linking movies who share similar patterns of UIC, unique insight on movies may be obtained through the emergence of movie concepts defined by these UIC patterns.

Of the three (3) components of UIC (*interaction, user characteristics, context*) *interaction* speaks to the set of actions users take on a particular object/entity. If there exists a pair of movies and users tend to perform a similar set of actions on each, then there might be reason to believe that these movies are similar.

The second component of UIC which may serve as a semantic link between object instances is that of the *characteristics of the users* who interact with the object. This simply refers to the traits of the users who show a high affinity for or have high levels of engagement with the movie. High levels of engagement with a movie by a specific demographic may suggest a certain characteristic of the movie. If this same demographic also has high engagement with another movie, both of these movies are believed to share some semantic similarity and are instances of a UIC-defined movie concept.

The final component of UIC used in the establishment of links between movies is that of *context*. For *context* we utilise the definition of (Musto et al., 2013, p.130) who point out that the concept of context has been historically difficult to define but settled on context being “a set of (external) factors able to influence user perception of the utility of a certain item”. If the users who have high levels of engagement with a movie do so from within a certain context, then this arguably says something about the movie.
6.3. **PRE-PROCESSING OF DATASET**

Of the three (3) UIC components emphasis is placed on *contexts* and *characteristics of users* more so than *interaction* with the movies. This decision was made based on the data which was available, the nature of the entities being considered, and beliefs on what UIC factors may or not be reflective of movie characteristics. The belief is that socio-economic realities of individuals reflected in their postal address and house value would influence their tastes in movies. Therefore if one were to link movies by socio-economic patterns in movie-viewing then useful movie concepts would emerge.

From the case study dataset the field *postal district* is the main point of focus and source of UIC data. From the postal district the postal area of the user is easily extracted and knowing the postal district also allows the determination of additional context-related fields by mapping the BT TV dataset to external data sources through the postal district field. *House prices* was selected as one of these additional external-to-BT-TV-dataset fields which would serve as a useful description of users and their context.

For the experiments conducted in this chapter which directly assess the functionality of the Collapse Index the UIC-factor *house price* was not used\(^1\). Instead genre descriptions for the BT TV movies, sourced from IMDb, were utilised as a point of comparison to the UIC-defined movies. The formal contexts utilised are those where objects are described by the *postal area*, *postal district*, or *genre* attribute sets. Section 6.4 outlines the transformation of the dataset into the three (3) formal contexts utilised in the experiments following the process described in Chapter 4.

### 6.3 Pre-Processing of Dataset

In advance of creating the formal contexts through transformation of the dataset, effort was made in cleaning the data to remove errors and inconsistencies. The structured nature of the dataset meant that there was not a large enough need to apply advanced NLP routines to the dataset, however there were several minor anomalies that needed to be addressed.

This included the removal of content that were classified as movies but were not. There were also cases where the same movie was been referenced by multiple movie

\(^1\)Note that *house price* is utilised in Chapter 8 where the thesis’ proposed approach to semantic extraction is applied in a recommender system.
IDs. While spelling errors were minimum, there were instances where the misuse of upper case and lower case letters led to distinctions being made for the same movie or movie attribute.

6.4 Construction of Formal Context

The formal context from which an FCA lattice is eventually built requires the determination of the complete set of binary relations between objects and attributes. Each binary relation can be seen as a pair \((m, n)\) where \(m\) is an object and \(n\) is an attribute that the object \(m\) possesses. The set of pairs can be represented visually as a table where an ‘X’ in a table cell substitutes for the pair \((m, n)\), \(m\) is the corresponding object row, and \(n\) the corresponding attribute column.

For all formal contexts used in Experiments 1 to 4, movies are always seen as the set of objects under consideration, whereas the attribute categories used to describe the movies are either genres, postal districts, or postal areas. Further details on each formal context are presented in the following subsections.

6.4.1 Genre Formal Context

The first of these formal contexts utilised is that of the genre formal context. All 2,122 unique movie represented in the BT TV dataset are described by a set of genres (e.g comedy, romance, western, etc.). Owing to the fact that the BT TV dataset did not contain genre descriptions, movie genre descriptions were obtained from the website IMDb.com (Section 3.4). For their genre descriptions IMDb generally describes each movie by a set of three (3) or fewer genre tags. As an example IMDb describes the movies PREDATOR and POSEIDON as such:

- PREDATOR: \{Action, Horror, Sci-Fi\}
- POSEIDON: \{Action, Adventure, Drama\}

From the movie PREDATOR the three pairs (PREDATOR, Action), (PREDATOR, Horror), and (PREDATOR, Sci-Fi) are obtained. Each represents a binary relation in the formal context \(K := (G, M, I)\) where \(I\) is the set of all binary relations.
6.4. CONSTRUCTION OF FORMAL CONTEXT

These binary relation pairs correspond to an ‘X’ in the tabular representation of a formal context as pictured in Figure 6.1. The complete set of 2,122 unique movies, each described by a set of movie genres, serves as the genre formal context utilised in the experiments.

Table 6.1: Formal Context for Genres

<table>
<thead>
<tr>
<th></th>
<th>action</th>
<th>adven</th>
<th>drama</th>
<th>horror</th>
<th>sci-fi</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDATOR</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>POSEIDON</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6.4.2 Postal Area Formal Context

The postal area formal context is created with motivation, as described in Chapter 4, to describe object instances with patterns of User Interaction and Context (UIC). The UIC factor being considered here as a descriptor of movies is the geographical locations(s) in which the movie is popular.

While the genre descriptions of movies were assigned to movies through expert opinion, the creation of the formal context for the UIC factor postal area is a much more involved process. As expected, such location descriptors of movies would not be readily available as genre information was from IMDb. For purposes of this research, the BT TV dataset containing the records of movies viewed by users would be utilised in determining the location-description of movies in empirical fashion. As an example of the output, similar to the genre-description of the movie PREDATOR, (Action, Horror, Sci-Fi), a location-description is sought e.g. PREDATOR (London, Stockport, Suffolk) where London, Stockport, and Suffolk are the top three (3) postal areas where Predator shows especially high popularity.

Essentially the premise of location descriptors of movies is that the popularity of specific movies in certain geographical regions, reflects some implicit characteristics of the movies. A function of the viewing frequency for which a movie is viewed in a specific location, would then determine how relevant a location is in describing said movie. The most relevant locations are then used in the movie’s location-description.

Of the five (5) fields present in the BT TV dataset, postal district is the field which represents locational data. Postal codes are alphanumerical representations, adopted
by the United Kingdom between 1959 and 1974 (Postcodearea, 2015), of a location comprised of a set of addresses or a major delivery point. This alphanumeric code (e.g. M13 9PL) is composed of two sections; the outward and inward code. The outward code is the set of characters before the space in the postal code (e.g. M13), while the inward code is the characters after the space (9PL). The postcode area is the first one or two characters of the outward code. Examples include L (Liverpool), AB (Aberdeen), E (Edinburgh), and M (Manchester).

In order to determine the postal areas in which the movies were most popular, the entirety of the BT TV dataset was parsed to acquire the frequency count of movie being watched, per postal areas, for each unique movie. Parsing of said dataset identified 121 unique postal areas and produced a frequency matrix of the format shown in Table 6.2 where the value in the table cells represent the number of times the movie (row) has been viewed by a user/household in a postal area (column)\(^2\).

<table>
<thead>
<tr>
<th>Movie</th>
<th>AB</th>
<th>B</th>
<th>M</th>
<th>CB</th>
</tr>
</thead>
<tbody>
<tr>
<td>GODFATHER</td>
<td>11</td>
<td>3</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>MAN OF STEEL</td>
<td>2</td>
<td>1</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>BATMAN</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>SHAWSHANK</td>
<td>3</td>
<td>0</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>PALE RIDER</td>
<td>4</td>
<td>1</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>101 DALMATIONS</td>
<td>7</td>
<td>0</td>
<td>12</td>
<td>1</td>
</tr>
</tbody>
</table>

From the BT TV-derived frequency matrix, the desire is to identify the postal areas in which the movies are most popular. If one considers the movie GODFATHER in Table 6.2 one might say that it is most popular in Manchester (M) and Aberdeen (AB) therefore these two postal areas can be used to characterise the movie GODFATHER. However given that Manchester (M) seems to be a dominant postal area for all movies, it is arguable that describing GODFATHER as a Manchester movie is not providing informational value about the movie GODFATHER. To address this, a term frequency-inverse document frequency (tf-idf) numerical statistic was utilised to adjust the weight of a postal area as a movie descriptor in light of how common the postal area is across all movies. The tf-df statistic employed as a replacement for the raw frequency was obtained from Salton and Buckley (1988) and is written as

\(^2\text{Data in table are not actual values obtained from the BT TV dataset but were chosen for demonstration purposes}\)
6.4. CONSTRUCTION OF FORMAL CONTEXT

\[ w = tf \cdot \log \frac{N}{n} \]  

(6.1)

where \( tf \) = location frequency, \( N \) = number of movies in dataset, \( n \) is number of movies where at least one user has watched the movie from the location in question, and \( w \) is the newly revised weight of the location with respect to the movie. Application of this tf-idf statistic transforms the frequency matrix into the new table of weights presented in Table 6.3.

Table 6.3: Tf-idf Matrix: Postal Area

<table>
<thead>
<tr>
<th></th>
<th>AB</th>
<th>B</th>
<th>M</th>
<th>CB</th>
</tr>
</thead>
<tbody>
<tr>
<td>GODFATHER</td>
<td>0.871</td>
<td>0.903</td>
<td>0</td>
<td>0.880</td>
</tr>
<tr>
<td>MAN OF STEEL</td>
<td>0.158</td>
<td>0.301</td>
<td>0</td>
<td>0.176</td>
</tr>
<tr>
<td>BATMAN</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SHAWSHANK</td>
<td>0.238</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PALE RIDER</td>
<td>0.317</td>
<td>0.301</td>
<td>0</td>
<td>0.704</td>
</tr>
<tr>
<td>101 DALMATANS</td>
<td>0.554</td>
<td>0</td>
<td>0</td>
<td>0.176</td>
</tr>
</tbody>
</table>

One must also bear in mind that the postal areas represent different geographical locations of different sizes, different populations, and/or different number of subscribers to the BT TV service. A movie exhibiting a high number of views from a specific postal area does not necessarily imply that that postal area has great affinity towards the movie but may simply be the inevitable consequence of the postal area having a large population or a large number of BT TV subscribers.

For this reason not only was the tf-idf measure employed but the tf-idf value was normalised to account for the variations in populations/subscribers. Cosine normalisation (Salton and Buckley, 1988, p.518) was the normalisation method of choice, leading to a final weight \( w_f \) for a location for a movie being

\[ w_f = \frac{tf \cdot \log \frac{N}{n}}{\sqrt{\sum_{\text{vector}} \left( tf_i \cdot \log \frac{N}{n_i} \right)^2}} \]

(6.2)

Application of the cosine normalisation to the tf-idf matrix of Table 6.3 leads to a final table of weights pictured in Table 6.4. If the desire was to describe each movie with the two (2) locations in which the movie was most popular, the raw frequency matrix of Table 6.2 would describe GODFATHER as a \{Aberdeen(AB), Manchester(M)\} movie. If the tf-idf matrix of Table 6.3 is used, GODFATHER would be described as
a \{Birmingham(B), Cambridge(CB)\} movie; whereas if the final normalised matrix is used, GODFATHER would be described as an \{Aberdeen(AB), Birmingham(B)\} movie.

Table 6.4: Cosine Normalised Tf-idf Matrix: Postal Area

<table>
<thead>
<tr>
<th></th>
<th>AB</th>
<th>B</th>
<th>M</th>
<th>CB</th>
</tr>
</thead>
<tbody>
<tr>
<td>GODFATHER</td>
<td>0.778</td>
<td>0.905</td>
<td>0</td>
<td>0.762</td>
</tr>
<tr>
<td>MAN OF STEEL</td>
<td>0.142</td>
<td>0.302</td>
<td>0</td>
<td>0.152</td>
</tr>
<tr>
<td>BATMAN</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SHAWSHANK</td>
<td>0.213</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PALE RIDER</td>
<td>0.284</td>
<td>0.302</td>
<td>0</td>
<td>0.61</td>
</tr>
<tr>
<td>101 DALMATIONS</td>
<td>0.496</td>
<td>0</td>
<td>0</td>
<td>0.152</td>
</tr>
</tbody>
</table>

Throughout the process it is observed that although Manchester (M) has the highest frequency representation for all movies, the fact that it was common to all movies meant that it was removed from consideration as a descriptor for any in the tf-idf process. Moreover there were eleven (11) viewings of GODFATHER in Aberdeen (AB) which is much larger than the amount of viewings of GODFATHER in other cities. The viewings of other movies in Aberdeen also being relatively high is probably suggestive of a large population (or number of BT TV subscribers) in Aberdeen. The normalisation process reduces the weight of the Aberdeen viewings of GODFATHER such that the final weight (0.778) is lower than the weight of the viewings of GODFATHER in Birmingham (0.905), despite the fact that there were only three (3) viewings of GODFATHER in Birmingham. However the fact that the viewings of all movies from Birmingham were generally low (0s or 1s), adds weight to there being three (3) viewings of GODFATHER in Birmingham. The formal context obtained in the end is shown in Table 6.5.

Table 6.5: Formal Context After Tf-idf and Cosine Normalisation

<table>
<thead>
<tr>
<th></th>
<th>AB</th>
<th>B</th>
<th>M</th>
<th>CB</th>
</tr>
</thead>
<tbody>
<tr>
<td>GODFATHER</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAN OF STEEL</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>BATMAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHAWSHANK</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PALE RIDER</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>101 DALMATIONS</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

For the postal area frequency matrix derived from the complete BT TV dataset, the
previous tf-idf and normalisation processes were applied producing a formal context where the 2,122 unique movies were each described by (at most) 10 postal areas where the movie was most popular. However lattices generated from this formal context for the experiments were limited to using only the top four (4) postal areas. Four (4) was the chosen maximum in order to bring the postal area descriptions more in-line with the genre descriptions of the genre formal context where each movie was described with at most three (3) genre attributes. Moreover, there were efficiency concerns in using large amounts of attributes per movie, as the Colibri Java FCA libraries utilised for processing lattices exhibited prohibitively large runtimes as the amount of attributes per object instance increased.

The usage of 4 of a possible 121 attributes per movie, does have an impact on lattices generated from such a formal context, given the decreased likelihood of objects/movies sharing the same attribute set. However this discussion is left until arriving at points in the experiments where this knowledge comes into play.

6.4.3 Postal District Formal Context

While households/users from a certain postal area may share different characteristics the size of the postal areas may mean that uniqueness of households is ‘diluted’ amongst a large and likely demographically diverse populace. For this reason postal districts, which are subdivisions of postal areas, are also utilised as descriptors of movie instances given that demographical consistency may be more likely be present at a more granular level of the postal code.

The postal district itself is represented as the outward code of the postal code. It is a combination of one (1) or more digits appended unto the postal area, and represents a subdivision of the geographical area of the postal area. Examples of postal districts are AB13 (Aberdeenshire), CM0 (Southminster), DY1 (Dudley) or PE27 (St. Ives). In total there were 1,904 postal districts represented in the BT TV dataset.

Similar to the postal area, the frequency matrix of postal districts was obtained from the BT TV dataset, which was then subjected to the tf-idf and cosine normalisation processes. The four (4) districts which had the highest weights for a movie were used to describe each of the 2,122 unique movies in the BT TV dataset, producing in the end the postal district formal context.
6.5 Experiment 1: Comparison Of Measures

6.5.1 Preliminaries

Given the successful usage of the Stability Index and the Support Value as concept relevancy measures in prior research, the Collapse Index adds to its legitimacy if there should exist a high correlation between the relevancy values derived from the Collapse Index and those of these other measures. As the Stability Index is seen as the ‘gold standard’ in the FCA community (Buzmakov et al., 2014), Experiment 1 compares the ‘Collapse Index against the Stability Index’ and the ‘Support Value against the Stability Index’. However, since the Stability Index values exhibits exponential behaviour (Jay et al., 2008), it was deemed appropriate to compare the other measures (Collapse Index and Support Value) against the Logarithmic Stability (LStab) of Equation 2.5 rather than the pure Stability Index.

6.5.2 Description of Experiment

An FCA lattice is first constructed from a formal context. For each formal concept in the lattice the Support Value, Collapse Index, and Stability Index are all calculated. The LStab value is calculated from the Stability Index. With these values in hand, a check is made for the Pearson product-moment correlation coefficient \( r \) for ‘Stability Index v. Support Value’ and for the ‘Stability Index v. Collapse Index’ plot. More importantly, \( r \) for ‘LStab v. Support Value’ and for ‘LStab v. Collapse Index’ are also calculated.

To add further credence to the outcome of such an experiment the same process is repeated over multiple formal contexts. Formal contexts utilised are the movies described by genres; movies described by postal areas; and movies described by postal districts.

6.5.3 Hypotheses

Consistent results across the various formal contexts constitute evidence to support various hypotheses - the first of which is the following:
**Hypothesis 1.** There is a statistically significant correlation between the values of theCollapse Index and the Logarithmic Stability for formal concepts in the same lattice

Collapse Index shares the ability of Stability Index to identify formal concepts of ‘low support but high relevance’ as well as formal concepts of ‘high support but low relevance’; this is not a property of the Support Value. With that in mind the second hypothesis for Experiment 1 is

**Hypothesis 2.** Pearson’s coefficient (r) is greater for ‘LStab v. Collapse Index’ than ‘LStab v. Support Value’

### 6.6 Experiment 1: Results

Experiment 1 focuses on establishing whether the Collapse Index has a strong correlation with the ‘gold standard’ Stability Index (via Logarithmic Stability). To address this, the values of both the Collapse Index and the Support Value are compared to the Stability Index.

#### 6.6.1 Lattice Summaries

FCA lattices were generated for the genre, postal area, and postal district formal contexts using the Colibri-Java libraries. The size of the dataset meant that these resultant lattices are too complex to be usefully visualised in ConExp\(^3\) or presented here. However summaries of the various lattices are presented in Table 6.6.

<table>
<thead>
<tr>
<th></th>
<th>No. of Concepts</th>
<th>Max extent</th>
<th>Avg. extent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genre</td>
<td>358</td>
<td>918</td>
<td>30</td>
</tr>
<tr>
<td>P. Area</td>
<td>5439</td>
<td>195</td>
<td>4.1</td>
</tr>
<tr>
<td>P District</td>
<td>3911</td>
<td>53</td>
<td>2.92</td>
</tr>
</tbody>
</table>

Note that *No. of Concepts* is a count of the number of formal concepts in the lattice. *Max extent* is the size of the largest extent from the set of formal concepts, while *Avg. extent* is the average size of extents in the lattice\(^4\).

\(^3\)See Appendix C.1
\(^4\)The values obtained for these fields are obtained after excluding the infimum and supremum in the respective lattices.
6.6.2 Formal Context: Genres

The lattice derived from the genre-based formal context is first up for scrutiny. The resultant FCA lattice of this context consisted of 360 formal concepts including the infimum and supremum. Given that supremum is too generic to contribute to being a meaningful class and the infimum is unlikely to exist in the domain of study\textsuperscript{5}, the infimum and supremum were ignored when assessing correlation between relevancy measures.

Due to the exponential behaviour of the Stability Index as previously mentioned, Stability Index values tend to be over the value of 0.5 and the majority were close to the value of 1 (e.g. 0.999999999999, etc.). In fact, some Stability Values were in such close proximity to 1 that the Java calculations using Java’s primitive data type double, a double-precision 64-bit IEEE 754 floating point (Java, 2015), returned the value 1.0 as the Stability Index value for several formal concepts. Problems arose for such formal concepts when calculating their Logarithmic Stability as per Equation 2.5. A Stability Index of 1 would lead to the calculation of \( L_{Stab} = \log_2(0) \) which is undefined. The decision was made to ignore such formal concepts with undefined LStab values, alongside the infimum and supremum, as the quantity of formal concepts for which this was true was relatively low in comparison to the total amount of formal concepts in the lattice. This reduced the number of formal concepts used in assessing correlation from the original 360 formal concepts to 324.

![Figure 6.1: Stability Index Correlations for Genre Lattice](image)

(a) Stability Index v. Support Value: Genre  
(b) Stability Index v. Collapse Index: Genre

Immediately evident from the results is the fact that the relationship between the

\textsuperscript{5}No movie is likely to be characterised by all genres or postal districts/areas represented in the formal context
6.6. EXPERIMENT 1: RESULTS

Collapse Index and the Stability Index as well as the Support Value and the Stability Index is not a linear relationship. Figures 6.1a and 6.1b show that this more closely resembles a logarithmic relationship than a linear. Therefore as previously decided, the Logarithmic Stability values were plotted against both the Support Value and the Collapse Index values - the results being better approximations of a linear function.

![Graphs](a) LStab v. Support Value: Genre  (b) LStab v. Collapse Index: Genre

Figure 6.2: Logarithmic Stability Correlations for Genre Lattice

Although plotting ‘LStab v. Support Value’ (Fig. 6.2a) produces a higher linear correlation and a better linear relationship \( r^2 = 0.6486 \) than ‘Stability Index v. Support Value’ \( r^2 = 0.2503 \) (Fig. 6.1a), the scatterplot is fairly inconsistent. This is in contrast to the results of plotting ‘LStab v Collapse Index’ (Fig. 6.2b). Here a much more ‘tight’ linear relationship was observed, and an \( r^2 \) value of 0.999 demonstrates that for the genre formal context the Collapse Index is better able to account for the values of the Logarithmic Stability than the Support Value would\(^6\).

6.6.3 Formal Context: Postal Areas

The processes of Experiment 1 were also conducted on a lattice constructed from a formal context where the movies were described by the postal areas in which the individual movies showed highest levels of popularity. Table 6.6 shows several differences between the postal area lattice and the genre lattice. Postal area produced 5,441 formal concepts in contrast to genre’s 360, and the formal concepts of the postal area lattice generally had extents with a lower cardinality.

Similar to the genre lattice, calculating the Stability Index for several formal concepts returned the value 1.0, resulting in the inability to determine their LStab values.

\(^6\)Bear in mind that the Logarithmic Stability is derived from the gold-standard Stability Index
These formal concepts were once again disregarded when determining correlation values, reducing the number of formal concepts to 5,381, a loss of 1% of formal concepts compared to the genres lattice’s 9.5% loss.

Of particular interest in regards to the postal area lattice, is that when LStab Values are plotted against these Support Values (Fig. 6.3a) there is a stronger linear correlation \( (r = 0.974) \) than for the same in the genres lattice \( (r = 0.805) \)(Fig. 6.2a). A possible factor in this improvement is the size of the extents of formal concepts. Table 6.6 shows that the extents of the postal area lattice tend to be smaller than those of the genres lattice.

![Figure 6.3: Logarithmic Stability Correlations for Postal Area Lattice](image)

Indication of this effect was found in a sub-experiment conducted for the genres lattice where the formal concepts in the lattice were separated into two disjoint subsets. One contained the 40 formal concepts with extents of highest cardinalities, the other subset being a random sample \( (n_s = 40) \) of the remaining formal concepts. For the subset with high extent cardinalities the correlations were determined for ‘LStab v. Support Value’ and for ‘LStab v. Collapse Index’. This process was repeated for the other subset containing lower extent cardinalities, the results of which are shown in Figures 6.4a, 6.4b, 6.5a, and 6.5b.

Under consideration first are the plots of ‘LStab v. Support Value’ in Figures 6.4a and in 6.5a. As previously assumed, the correlation between these two variables becomes noticeably worse when the cardinalities of the extents are high. The correlation coefficient \( (r = 0.8) \) for ‘LStab v. Support Value’ in Figure 6.4a is noticeably better than the coefficient \( (r = 0.1) \) when formal concepts have extents of high cardinalities (Fig. 6.5a). One may go as far as to say that this correlation for high cardinalities is...
6.6. EXPERIMENT 1: RESULTS

(a) LStab v. Support Value: Low  
(b) LStab v. Collapse Index: Low

Figure 6.4: Logarithmic Stability Correlations for Low Cardinalities

(a) LStab v. Support Value: High  
(b) LStab v. Collapse Index: High

Figure 6.5: Logarithmic Stability Correlations for High Cardinalities

very poor as it relates to Support Value and the Logarithmic Stability.

In contrast, when undertaking the same procedures for ‘LStab v. Collapse Index’
the Collapse Index performs slightly worse for low cardinality formal concepts (Fig.
6.4b) than for high cardinality formal concepts (Fig. 6.5b), obtaining correlation
coefficients of $r = 0.996$ and $r = 0.9996$ respectively. However these are still very high
correlation coefficients, indicating a more consistent relationship between the Collapse
Index and LStab than between the Support Value and the LStab, across variances in
extent cardinality.

6.6.4 Formal Context: Postal Districts

The final formal context used in Experiment 1 was that of movies being described by
the postal districts. Table 6.6 shows that the resultant FCA lattice is comprised of
less formal concepts than that of postal areas despite there being a greater number of
possible attributes (postal districts) of which may be used to describe movies.
A possible explanation of this low amount of *formal concepts* in the postal district lattice is that because of the large number of possible postal districts\(^7\) in contrast to the only four (4) postal districts maximum being assigned to each movie, meant that few movies shared a common set of postal districts. As a consequence an FCA lattice is produced for postal districts which is very ‘wide’ but very shallow, and with a relatively low number of *formal concepts*. The wide-flat lattice hypothesis is somewhat corroborated in Table 6.6 where the sizes of the extent of the postal district FCA lattice tend to be lower than that of the postal area lattice. If the cardinality of the extents are low this could signify *formal concepts* not having many subconcepts given that objects accumulate upwards in an FCA lattice.

Table 6.7: Summaries of Correlation Coefficients

<table>
<thead>
<tr>
<th>No. of Concepts</th>
<th>Avg. extent</th>
<th>(r): supp v lstab</th>
<th>(r): ci v lstab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genre</td>
<td>358</td>
<td>30</td>
<td>0.805</td>
</tr>
<tr>
<td>P. Area</td>
<td>5439</td>
<td>4.1</td>
<td>0.974</td>
</tr>
<tr>
<td>P. District</td>
<td>3911</td>
<td>2.92</td>
<td>0.971</td>
</tr>
</tbody>
</table>

Plotting ‘LStab against Support Value’ for postal district produced an \(r\) value of \(r = 0.971\) whereas the correlation coefficient for the equivalent plot for postal area was \(0.974\). These results as well as those of the previous genre and postal area are summarised in Table 6.7, from which conclusions for the experiment are drawn.

Hypothesis 1 asked that the significance of the correlation between Collapse Index and the LStab values be corroborated. This produces the null and alternate hypotheses:

\(H_0 : \rho = 0\) No correlation between variables

\(H_1 : \rho \neq 0\) Significant correlation between variables.

The test statistic utilised is the *t-test for correlation coefficient* given as:

\[
t = \frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}}.\tag{6.3}
\]

where \(n\) = number of pairs of values used, \(r\) = *product-moment correlation coefficient*, and the *degrees of freedom* (df) = \(n - 2\).

In each case (genre, postal area, and postal district formal contexts), the test statistic produced p-values less than \(\alpha = 0.05\). The null hypothesis that there is no

\(^7\)1,577 postal districts had representation in the formal context utilised for postal district lattice
6.7. EXPERIMENT 2: CONSISTENCY ACROSS SAMPLES

The previous section made a case for the legitimacy of the Collapse Index as a measure of the relevancy of a formal concept, via the comparison of the relevancy values obtained by the Collapse Index to those values of a known successful approach (Stability Index). Here a second case is made for the value and validity of the Collapse Index.

In Kuznetsov (1990) an argument is made that a pattern is more legitimate if the pattern is present in multiple subsets of a dataset. Building on this notion, a similar argument may be made that not only should a legitimate pattern be present in multiple subsets, but assuming a pattern is present in multiple subsets of the dataset, the relevance value of the pattern in one subset should correlate proportionally with the relevance of the pattern in another subset.

In the case of Formal Concept Analysis, as advanced in Buzmakov et al. (2014a), a method of determining the relevance of a formal concept (e.g. Support Value, Stability Index, etc.) is granted some validity if the values it produces for the relevance of formal concepts are consistent across lattices generated from multiple samples of the dataset.

More specifically, if two FCA lattices are created, each from a different subset of the complete formal context, the relevancy value assigned to a formal concept in one lattice by a relevancy measure should approximate the relevancy value of its

---

Given a formal context $K := (G, M, I)$, a subset of a formal context (subcontext) can be more formally defined as $K_i := (G_i, M, I_i)$ where $G_i \subseteq G$, and $I_i$ is the set of binary relations, $I_i \subseteq G_i \times M$ such that $I_i \subseteq I$.
equivalent formal concept in the other lattice. Theoretically there should exist a linear correlation if the relevancy values of formal concepts in one lattice are plotted against the corresponding relevance of their equivalent formal concepts in the second lattice.

Although the formal concepts extracted by FCA from a formal context are defined by the dual properties of a set of objects and a set of attributes, the set of attributes (intent) is the emphasised pattern of interest. ‘Equivalent’ formal concepts across two lattices would then be formal concepts whose intents are equal. Formal concept relevancy measures such as the Support Value, Stability Index\(^9\), and the Collapse Index which ascertain the relevancy of the intent of the formal concept more so than the complete formal concept, are thus worthy candidates for having their consistency in ‘assigning relevance values to the same pattern (intent) across subsets’ assessed.

Ideally for these equivalent formal concepts, the correlation across lattices should produce a linear function \(y = x\). This function is unlikely to be achieved for a variety of reasons in real world application, however the gradient of the best-fit linear function should approach 1 and the \(y\)-intercept approach 0. The coefficient of determination \((r^2)\) should also approximate 1.

At least two (2) scenarios related to the pair of subcontexts used are employed in this experiment to investigate the correlation between relevancy values across pairs of lattices produced by their respective subcontexts.

- **disjoint**: the pair of subcontexts of the formal context from which the two lattices are created are disjoint. Subcontexts being disjoint translates to no shared object instances between the pair of subcontexts. More formally given that \(K_i := (G_i, M, I_i)\) and \(K_j := (G_j, M, I_j)\) are both subcontexts of \(K := (G, M, I)\), then \(K_i\) and \(K_j\) are disjoint if \(G_i \cap G_j = \emptyset\)

- **sample-population**: one of the subsets of the formal context is a subset of the second subset. More formally given that \(K_i := (G_i, M, I_i)\) and \(K_j := (G_j, M, I_j)\) are both subcontexts of \(K := (G, M, I)\), then \(K_i\) is a subset of \(K_j\) if \(G_i \subseteq G_j\)

### 6.7.2 Description of Experiment

The following outlines the steps taken when carrying out the experiment.

\(^9\)Note that although the Stability Index is mentioned here, it is its derivative, LStab, that is once again utilised when calculating the correlation across lattices.
6.7. EXPERIMENT 2: CONSISTENCY ACROSS SAMPLES

1. From a central formal context, two (2) subcontexts are selected. If disjoint subcontexts are being investigated, $|G_i|$ of each subcontext is 1,000. If consistency across sample-population is being investigated, the sample subcontext has a $|G_i|$ value of 500, each object instance being chosen by random sampling. The population formal context consists of all 2,122 object instances available to us in the BT TV dataset.

2. An FCA lattice is generated from each of the pair of subcontexts.

3. The Collapse Index, Stability Index, LStab, and Support Value are calculated for all formal concepts in both lattices.

4. For all formal concepts for which their intent is common to both lattices, the relevancy of that formal concept (as per a specific relevancy measure e.g. Support Value) is plotted against the relevancy of its equivalent formal concept in the other lattice.

5. The coefficient of determination ($r^2$) is calculated for each plot as well as the best-fit line obtained via linear regression. The results for the various relevancy measures are then analysed and compared.

6. The previous steps are carried out for both the genre formal context (Section 6.4.1) as well as the postal area formal context described in Section 6.4.2.

6.7.3 Hypotheses

As the Collapse Index is presumed to be less susceptible to noise than the Support Value there is an expectation that the Collapse Index would be more consistent across subcontexts allowing for better prediction of concept relevancy across similar datasets.

Hypothesis 3. If the relevancy values of formal concepts generated from a subcontext are plotted against the relevancy values of the equivalent formal concepts in a lattice from another subcontext, the coefficient of determination ($r^2$) would be greater if the relevancy measure used was the Collapse Index rather than the Support Value.

Meanwhile Collapse Index values are expected to correlate or reflect the Stability Index values, more so than reflect the Support Values. Given this the consistency
of the Collapse Index measure is expected to match closely the consistency of the Stability Index and hence the Logarithmic Stability.

**Hypothesis 4.** If the relevancy values of formal concepts generated from a subcontext are plotted against the relevancy values of the equivalent formal concepts in a lattice from another subcontext, the coefficient of determination \( r^2 \) if the Collapse Index was the relevance measure used would be comparable to if the Logarithmic Stability was used.

### 6.8 Experiment 2: Results

Tests for consistency of relevancy values across lattices were conducted for both the genre formal context and the postal area formal contexts. Results are first presented for the genre formal context.

#### 6.8.1 Genre Lattice

In the case of genre formal context, results when investigating disjoint subcontexts are the first points of discussion. For this scenario the genre formal context was divided into two disjoint subcontexts, \( S_A \) and \( S_B \), each having a representation of 1000 movie instances (\(|G_A| = |G_B| = 1000\)). The first of these subcontexts, \( S_A \), produced an FCA lattice comprised of 269 formal concepts (inclusive of the infimum and supremum) whereas the other subcontext \( S_B \) produced 299 formal concepts. Of these sets of formal concepts, 217 were common to both lattices, i.e. the intents of 217 formal concepts were represented in both lattices. The relevancy values of these 217 formal concepts in the \( S_A \) lattice were calculated using the Support Value, Stability Index, and Collapse Index; this was repeated for the very same 217 formal concepts in the \( S_B \) lattice.

A comparison is first made for the consistency across genre lattices for the Support Value measure. When the Support Values for the shared 217 formal concepts in \( L(S_B) \) are plotted against the values of the same formal concepts in \( L(S_A) \) this produced a best-fit linear equation of \( y = 0.9906x - 0.0003 \). This linear relationship, shown in Figure 6.6a, also exhibited a very high coefficient of determination \( r^2 \) value.
6.8. EXPERIMENT 2: RESULTS

of 0.9828. When the equivalent processes were carried out utilising the Collapse Index measure this also produced a linear relationship, \( y = 1.0057x - 0.0004 \) (Figure 6.6b) and an \( r^2 \) value of 0.9814. For both the Support Value and Collapse Index, their correlation functions closely approximated the ideal \( y = x \) and showed no real substantial differences between their linear function or \( r^2 \) values. A summary is provided in Table 6.8 of these and all other best-fit lines and \( r^2 \) values obtained for the genre formal context scenarios.

In the case of the Stability Index, the values for both subcontexts were plotted against each other creating the relationship displayed in Figure 6.7a. It is immediately noticeable that the relationship across lattices for the Stability Index is not as identifiably linear as that of the Support Value and Collapse Index plots. Also notable is that the bulk of the Stability Index values tend to be very close to 1. A low \( r^2 \) value of 0.4386 suggests a poor ability of the Stability Index value in one lattice to account for the Stability Index value assigned to the equivalent formal concept in another lattice. However, as previously noted in Section 6.7.1, it may be more useful to utilise the Logarithmic Stability.

<table>
<thead>
<tr>
<th></th>
<th>Best-fit Equation</th>
<th>( r^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Support Value</strong></td>
<td>( y = 0.9906x - 0.0003 )</td>
<td>0.9828</td>
</tr>
<tr>
<td><strong>Collapse Index</strong></td>
<td>( y = 1.0057x - 0.0004 )</td>
<td>0.9814</td>
</tr>
<tr>
<td><strong>Stability Index</strong></td>
<td>( y = 0.6514x + 0.2865 )</td>
<td>0.4386</td>
</tr>
<tr>
<td><strong>Logarithmic Stability</strong></td>
<td>( y = 0.8668x + 0.721 )</td>
<td>0.8799</td>
</tr>
<tr>
<td><strong>Sample-Population</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Support Value</strong></td>
<td>( y = 0.9887x - 0.0002 )</td>
<td>0.9916</td>
</tr>
<tr>
<td><strong>Collapse Index</strong></td>
<td>( y = 0.9977x + 0.00004 )</td>
<td>0.988</td>
</tr>
<tr>
<td><strong>Stability Index</strong></td>
<td>( y = 0.3779x + 0.6403 )</td>
<td>0.2992</td>
</tr>
<tr>
<td><strong>Logarithmic Stability</strong></td>
<td>( y = 3.7209x + 1.1808 )</td>
<td>0.7647</td>
</tr>
</tbody>
</table>

Figure 6.6: SV and CI Correlations for Disjoint Subcontexts: Genre
For several *formal concepts*, the Stability Index returned a value of 1, leading to the inability to calculate the Logarithmic Stability values\(^\text{11}\). Consequently this reduced the amount of *formal concepts* utilised for the correlation calculations across lattices from 217 to 204. This however is not expected to have any real impact on the \(r^2\) or linear regression calculations. Results for the LStab plots are shown in Figure 6.7b.

![Figure 6.7: SI and LStab Correlations for Disjoint Subcontexts: Genre](image)

Utilisation of the LStab produced a relationship across lattices more closely resembling a linear function. The outcome was a best-fit linear function \(y = 0.8668x + 0.721\) and an \(r^2\) value of 0.8799. This relationship improves upon the one achieved from the Stability Index as the best-fit line is now closer to the ideal \(y = x\). However LStab correlation performs worse than the Collapse Index and Support Value with respect to maintaining consistency across subcontexts. Both Collapse Index and Support Value resulted in linear functions which better approximated \(y = x\), and both had higher values of \(r^2\).

For the same genre formal context, a second investigation was conducted for the *sample-population* subcontext relationship. In this case the Support Value, LStab, and Collapse Index values for *formal concepts* in a lattice based on the complete formal context \(S_P\) are plotted against the values of equivalent *formal concepts* in a lattice from a sample subcontext \((S_S)\). The sample subcontext which was comprised of 500 movie instances produced 209 *formal concepts*, whereas the entire population of 2,122 movies produced a lattice of 360 *formal concepts*. Of both lattices 207 *formal concepts* were held in common.

The scatterplot of the Support Value measure is shown first in Figure 6.8a. Once

\(^{11}\)The reason for this was previously explained in Section 6.6.2
again the relationship is linear producing a best-fit line $y = 0.9887x - 0.0002$ and an $r^2$ value of $= 0.9916$. Exhibiting similar performance the Collapse Index produced a best-fit linear equation $y = 0.9977x + 0.000045$ and had an $r^2$ value of 0.988 (Figure 6.8b)

![Graphs showing correlations between Support Value and Collapse Index](image)

(a) Support Value $S_P$ v. $S_S$  
(b) Collapse Index $S_P$ v. $S_S$

Figure 6.8: SV and CI Correlations for Sample-Population Subcontexts: Genre

Proceeding to the usage of the Logarithmic Stability, using 174 of the 207 formal concepts common to both lattices\(^{12}\), LStab produces a best-fit line $y = 3.7209x + 1.1808$ and a coefficient of determination $r^2 = 0.7647$ (Figure 6.9). These are both worse than results of LStab experiments conducted using disjoint subcontexts. Since the range of LStab values is large given that $LStab(A, B) \in [0, \infty]$, it is understandably more difficult to achieve a $y$-intercept of 0 in a best-fit approximation of a scatterplot. However the gradient of the linear function if there is consistency for concept relevancy across FCA lattices should at least be close to 1.

From Figure 6.9 it is observed that the LStab values are as high as 50, however these were achieved for the complete $S_P$ formal context ($|G| = 2, 122$) represented on the $y$-axis. For the $x$-axis, the sample formal context $S_S$ ($|G_S| = 500$) the highest LStab value achieved is less than 14. This suggests that the Stability Index value of a formal concept is not independent of the size of the formal context.

While this should decrease expectations for a formal concept having the same\(^{13}\) LStab value if the formal concept was present in two lattices each built from formal contexts of different sizes, the LStab values from one subcontext should at least account for the variances in the other. This should translate to the LStab scatterplot diagram with a best-fit linear regression equation, which although not approaching $y = x$, |

\(^{12}\)Some LStab values were undefined
\(^{13}\)same as in 'very close' but not necessarily equal
still possessing a high $r^2$ value. As previously stated, the sample-population scenario produced a best-fit line for the LStab scatterplot of $y = 3.7209x + 1.1808$ and a coefficient of determination $r^2 = 0.7647$.

While the $r^2$ value is fairly decent it is noticeably worse than the $r^2$ value when the subcontexts were of equal size ($r^2 = 0.8799$), lending credence to the suggestion that LStab is inconsistent across subcontexts of different size.

### 6.8.2 Postal Area Lattice

To corroborate results or knowledge emergent from Section 6.8.1, the equivalent processes were carried out on the postal area formal context. Disjoint subcontexts $S_A$ where $|G_A| = 1,000$ and $S_B$ where $|G_B| = 1,000$ were extracted from the postal area formal context and lattices generated from each subcontext. For formal concepts found common to both lattices the Support Value, LStab, as well as Collapse Index values were calculated for each formal concept. Scatterplot diagrams were created where the relevancy values (obtained via a specific relevancy measure) of all common formal concepts in one lattice $L(S_B)$ were plotted against the relevancy values of the equivalent formal concepts in the other lattice $L(S_A)$. The best-fit linear functions were obtained for the Support Value, LStab, as well as Collapse Index scatterplot diagrams. In addition the coefficient of determination ($r^2$) of each diagram was calculated. This was repeated for the sample-population subcontexts where $S_S$ is such that $|G_S| = 500$, and
$S_P$ is the complete formal context such that $|G_P| = 2,122$. Results for these tests are summarised in Table 6.10.

Initially however, a few basic statistics specific to the subcontexts of the postal area formal context (See Table 6.9) are outlined. For the disjoint subcontexts $S_A$ and $S_B$, each produced lattices containing 2,582 and 2,565 formal concepts respectively. Of the union of these formal concepts 708 were common to both lattices, exhibiting a Jaccard Coefficient of 0.16. This is quite low considering that when the equivalent was done for the disjoint genre subcontexts the Jaccard Coefficient for common formal concepts across lattices was 0.62. Such a low amount of shared formal concepts may be better understood by considering the characteristics of the postal area lattice described in Table 6.6.

Information in Table 6.6 related to the lattice generated by the complete postal area formal context, show that not only is the postal area lattice composed of the highest number\(^\text{14}\) of formal concepts (in comparison to the genre and postal district formal contexts) but the average size of the extents is low (4.1). While 4.1 is greater than the 2.1 of the postal district lattice, it is much less than the 30 of the genre lattice. The low extent cardinalities means that many of these formal concepts owe their existence to very few object instances, and that many of the formal concepts in the lattice are likely at the lowest levels of the lattice. If ‘the few object instances they owe their existence to’ are in one particular subcontext, then a formal concept representing their attribute set would not be present in a lattice generated from a disjoint subcontext. However this does not necessarily hinder the experiment and results.

In the case of sample-population scenario every formal concept present in the lattice of $S_S$ is also present in the lattice of $S_P$. The lattice from the subcontext $S_S$ produced 1,206 formal concepts and the lattice of the entire formal context is composed of 5,441 formal concepts, however the infimum and supremum are ignored in the analysis resulting in only 1,204 formal concepts being utilised in the scatterplot diagrams.

From Table 6.10 it is observed that in the case of disjoint subsets both the Support Value and the Collapse Index produced best-fit lines that are close to the ideal $y = x$. The Stability Index is as expected a poor fit for a linear correlation. Meanwhile LStab produces a best-fit linear equation of $y = 0.9688x + 0.101$, which, although not as close

\(^{14}\)5,439 formal concepts excluding the infimum and supremum
CHAPTER 6. COLLAPSE INDEX VALIDATION EXPERIMENTS

Table 6.9: Number Of Formal Concepts In Subcontexts

<table>
<thead>
<tr>
<th>Condition</th>
<th>No. Of Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>disjoint ((S_B v. S_A))</td>
<td></td>
</tr>
<tr>
<td>(S_A,</td>
<td>G_A</td>
</tr>
<tr>
<td>(S_B,</td>
<td>G_B</td>
</tr>
<tr>
<td>in common</td>
<td>708</td>
</tr>
<tr>
<td>defined for LStab</td>
<td>692</td>
</tr>
<tr>
<td>sample-population ((S_P v. S_S))</td>
<td></td>
</tr>
<tr>
<td>(S_S,</td>
<td>G_S</td>
</tr>
<tr>
<td>(S_P,</td>
<td>G_P</td>
</tr>
<tr>
<td>in common</td>
<td>1204</td>
</tr>
<tr>
<td>defined for LStab</td>
<td>1146</td>
</tr>
</tbody>
</table>

to \(y = x\) as the previous relevancy measures, is still fairly successful.

Table 6.10: Correlation Across Lattices (Postal Area)

<table>
<thead>
<tr>
<th>Plot</th>
<th>Best-fit Equation</th>
<th>(r^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>disjoint ((S_B v. S_A))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Support Value</td>
<td>(y = 0.9836x + 0.0001)</td>
<td>0.9439</td>
</tr>
<tr>
<td>Collapse Index</td>
<td>(y = 0.9889x + 0.0005)</td>
<td>0.9409</td>
</tr>
<tr>
<td>Stability Index</td>
<td>(y = 0.6141x + 0.2238)</td>
<td>0.3753</td>
</tr>
<tr>
<td>Logarithmic Stability</td>
<td>(y = 0.9688x + 0.101)</td>
<td>0.9106</td>
</tr>
<tr>
<td>sample-population ((S_P v. S_S))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Support Value</td>
<td>(y = 0.9993x - 0.0017)</td>
<td>0.9574</td>
</tr>
<tr>
<td>Collapse Index</td>
<td>(y = 1.0449x - 0.001)</td>
<td>0.9488</td>
</tr>
<tr>
<td>Stability Index</td>
<td>(y = 0.42x + 0.4631)</td>
<td>0.1463</td>
</tr>
<tr>
<td>Logarithmic Stability</td>
<td>(y = 4.8542x - 1.5906)</td>
<td>0.7935</td>
</tr>
</tbody>
</table>

Sample-population subcontexts also result in best-fit linear equations for Support Value and Collapse Index relevancy measures that are close to \(y = x\) while having values of \(r^2\) close to 1. The LStab scatterplot however is further away from this ideal linear equation and an \(r^2\) value of 0.7935. Analysis of the genre scenarios in Section 6.8.1 had led to the observation that LStab performs poorly when the subcontexts are of different sizes, however for that genre formal context, the LStab scatterplot in the sample-population scenario was still able to maintain a relatively high \(r^2\) value (0.7647). Here, for the postal area formal context, the LStab scatterplot results in a similar \(r^2\) value of 0.7935.

In summary, from the experiments conducted for consistency across subcontexts, with respect to Hypothesis 3, it is not immediately clear which of the Support Value or the Collapse Index is more consistent across datasets or produce higher values of \(r^2\). Both concept relevancy measures are consistent across disjoint subcontexts as well as for sample-population subcontexts, each concept relevance measure producing very high coefficients of determination.
Meanwhile, while producing high values of $r^2$ (0.8799, 0.7647, 0.9106, 0.7935) for the four instances in which the $r^2$ were calculated for Logarithmic Stability scatterplot diagrams, these $r^2$ values were consistently less than the corresponding $r^2$ values for the Collapse Index scatterplots (0.9814, 0.988, 0.9409, 0.9488). With respect to Hypothesis 4 the results of the experiments lead to a conclusion that Collapse Index would have higher $r^2$ values than the Logarithmic Stability across subcontexts. If this is the case then, if given a subcontext, using the Collapse Index to assess the relevancy of a formal concept in a lattice based on the subcontext would serve as a better predictor of the relevance of the formal concept in another subcontext, than utilising the Logarithmic Stability of the formal concept in one subcontext to predict the Logarithmic Stability of the same formal concept in another subcontext.

When the subcontexts are of different sizes, predicting the Logarithmic Stability across subcontexts is more difficult. This could be mitigated to some extent by incorporating the size of the subcontexts in prediction models, however at its most basic the Logarithmic Stability measure is less successful at predicting values across subcontexts when the subcontexts are of different sizes.

6.9 Experiment 3: Noise Recognition

6.9.1 Preliminaries

Experiment 3 involves a set of experiments comparing the performances of Support Value, Logarithmic Stability, and the Collapse Index in recalling formal concepts from noisy FCA lattices. Fundamentally these experiments compare the ability of the three concept relevancy measures to recall/identify the intensional descriptions of formal concepts from a noisy formal context where the intent was originally present in a lattice derived from an original non-noisy formal context. In order to limit somewhat the possible combinations of sub experiments, focus is kept only on the genre formal context, utilising it as the ‘pure’ or non-noisy formal context.

The FCA lattice derived from this genre formal context is a taxonomical representation of the domain composed of BT TV movies described by IMDb-sourced genre

\footnote{Although the term ‘formal concept’ is being used throughout the discussion of this experiment, the intent of the formal concept is once again what is being referred to}
description. Although a representation of a real domain, albeit imperfectly so, and labelled pure for the purposes of the experiments, the dataset/context may already contain some noise due to collection, transcription, or other human/software related factors.

In light of this, when testing for recall between the pure and noisy lattices a subset of the formal concepts derived from the pure formal context are disregarded. Specifically, the formal concepts ignored are those formal concepts with only one associated object (movie) instance. The set of all other formal concepts is titled ‘Valid’ and the presence of all formal concepts in the set Valid is sought in the noisy lattice.

In the case of the noisy formal context, two types of noise are taken into consideration:

- Type I: Binary object-attribute relationships in formal context are randomly changed
- Type II: Random objects are added to the formal context.

### 6.9.2 Description of Experiment

Investigating the usefulness of each measure in recalling formal concepts from noisy lattices, the following steps are undertaken.

1. A lattice is generated from the original formal context. This lattice represents the ‘pure’ or correct lattice. A subset of such a formal context would resemble Table 6.11.

<table>
<thead>
<tr>
<th></th>
<th>action</th>
<th>adven</th>
<th>drama</th>
<th>horror</th>
<th>sci-fi</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDATOR</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>POSEIDON</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALIEN</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

2. If Type I noise is to be added to the formal context, for each possible object-attribute pair in the formal context, a random value in the range $[0, 1]$ is assigned to each cell (Table 6.12).
6.9. EXPERIMENT 3: NOISE RECOGNITION

Table 6.12: Random Value Assigned to Binary Relations

<table>
<thead>
<tr>
<th></th>
<th>action</th>
<th>adven</th>
<th>drama</th>
<th>horror</th>
<th>sci-fi</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDATOR</td>
<td>0.8404</td>
<td>0.3117</td>
<td>0.6262</td>
<td>0.0510</td>
<td>0.0223</td>
</tr>
<tr>
<td>POSEIDON</td>
<td>0.8101</td>
<td>0.6805</td>
<td>0.8629</td>
<td>0.0196</td>
<td>0.9197</td>
</tr>
<tr>
<td>ALIEN</td>
<td>0.00975</td>
<td>0.7200</td>
<td>0.7311</td>
<td>0.4922</td>
<td>0.2027</td>
</tr>
</tbody>
</table>

If a cell’s random value is less than a chosen noise level, the binary relation in the cell is switched from original value in the pure formal context to its alternate. e.g. if a noise level of is 10% (0.1) is utilised, then the formal context with noise added would now be as shown in Table 6.13. The cell \((ALIEN, action)\) now displays \(X\) where it was empty in the pure formal context. The cell \((PREDATOR, sci-fi)\) is now blank in the noisy formal context, a change from the value ‘\(X\)’ in the pure formal context. These cells were changed as the random value assigned to these cells were 0.00975 and 0.0223 respectively - both less than the 10% noise level. All other cells remain as they were in the pure formal context. The noise levels chosen for the Type I experiments are 1%, 5%, 10%, and 20%.

Table 6.13: Noisy Formal Context

<table>
<thead>
<tr>
<th></th>
<th>action</th>
<th>adven</th>
<th>drama</th>
<th>horror</th>
<th>sci-fi</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDATOR</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>POSEIDON</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALIEN</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. If Type II noise is to be added, several new object instances are added to the pure formal context. Each of these newly introduced object instance is described by a random set of genres from the set of all possible genres. To ascribe a genre to a movie (a) each genre is assigned a probability of occurring as a descriptor of a movie and (b) the probability, \(p\), assigned to each genre should be chosen to best increase the likelihood of each movie being assigned three (3) genres.

In reference to (a), although \(p\) of each specific genre may be obtained experimentally from the IMDb dataset, we assume the \(p\) is equal for all possible genres for the purpose of simplicity. More importantly this is done in order to allow for more unorthodox combinations of movie genre descriptions - a better representation of noise.

The latter, (b), is implemented as movies in IMDb tend to be represented by no
more than three (3) unique genres, something which we largely seek to replicate. Assuming independence\textsuperscript{16} between genre attributes, \( p \) may be obtained via binomial distribution. If \( n = 23 \) represents the number of possible IMDb genres, and \( k \) is the number of genres assigned to a movie, then the \( p \) value which gives us the best opportunity of three movies being assigned to a movie, is the maximum value of the function

\[
f(p) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{23}{3} p^3 (1-p)^{23-3}.
\]

This occurs at \( p = \frac{3}{23} = 0.13 \)

As it relates to the experiments conducted, the number of movie instances added to the pure formal context as Type II noise, were 1\%, 5\%, 10\%, 20\%, and 40\% of movies in the pure formal context.

4. After creation of a noisy formal context an FCA lattice is created from this noisy formal context from which the Support Value, Logarithmic Stability, and Collapse Index are calculated for each formal concept in the lattice.

5. For each relevancy measure the formal concepts derived from the noisy formal context are arranged in numerical order according to their relevance. Given that the number of formal concepts in the set Valid\textsuperscript{17} is \( n_v \), the top \( n_v \) formal concepts from the noisy formal context are classified as the set ‘Relevant’. The ratio of Valid formal concepts recalled serves as the measure of success of the concept relevancy measures and may be defined as

\[
\text{recall} = \frac{|\text{Valid} \cap \text{Relevant}|}{|\text{Valid}|}
\]

Figure 6.10 shows these processes. The block on the left represents the FCA lattice generated from the pure formal context. The upper section of the same block represents the set of Valid formal concepts while the lower section represents the formal concepts whose extents have a cardinality of 1. The block on the

\textsuperscript{16}Assuming independence between genres also makes the random genre combinations assigned to movies more weird or unusual.

\textsuperscript{17}Valid, as earlier described, is the set of formal concepts derived from the pure formal context, that have an extent cardinality greater than 1.
Figure 6.10: Assessing Recall from Noisy Formal Contexts

right represents the FCA lattice produced from the noisy formal context and the formal concepts produced are ranked with respect to their relevancy. The top $n_v$ formal concepts in this lattice are the set of Relevant formal concepts. Within this Relevant set the Valid formal concepts are sought. In this instance, of the six (6) formal concepts in the Valid set, four (4) formal concepts, $\{B, Y, H, D\}$, have been retrieved in the noisy lattice, giving a recall value of $4/6 = 0.67$

Note that once again the Logarithmic Stability is being used instead of the Stability Index. As the LStab function of Equation 2.5 is continuously increasing with respect to the domain $SI \in [0, 1]$ (Figure 6.11), ordering the noisy formal concepts by LStab will produce the same order as ordering the same formal concepts by the Stability Index.

6.9.3 Hypotheses

Building on the expectation of Experiment 1 where the Collapse Index values for formal concepts in an FCA lattice are expected to exhibit high correlation with the
Stability Index values, it is also the expectation here that because of this similarity in values, the Collapse Index should be similarly successful as the Stability Index in recognizing noise in a lattice. We derive the hypothesis for the noise experiment from this basis.

**Hypothesis 5.** *Collapse Index is able to recall a similar amount of formal concepts in a noisy lattice as the Stability Index.*

### 6.10 Experiment 3: Results

Included in this section are the results of the experiments related to assessing the capabilities of the Collapse Index, Support Value, and the Logarithmic Stability in identifying noise in a noisy FCA lattice. Experiments using Type I noise (randomly changing binary relations in formal context) are first up for discussion.

#### 6.10.1 Results: Type I Noise

Firstly we are reminded that as a way of reducing what may be noise in the pure lattice, *formal concepts* for which their extents had a cardinality of only 1 were excluded from being taken into consideration when seeking to recall *formal concepts* in the noisy lattices. For the genre FCA lattice which served as the *pure* lattice, this reduced the number of *formal concepts* taken under consideration from 358 to 270. Hence the *Valid set of formal concepts* we seek to retrieve in the noisy lattices is a set of 270 *formal concepts.*
Varying levels of Type I noise were added to the pure formal context—the first of which was 1% noise. Essentially each binary relation in the pure formal context had a 1% chance of switching to its alternate state. After infusion of 1% noise on the pure formal context an FCA lattice was generated which was composed of 782 formal concepts (excluding the infimum and the supremum). The relevancy values of these 782 formal concepts were determined using Support Value, Logarithmic Stability, as well as the Collapse Index.

When arranging the 782 formal concepts in order based on their Support Value, the percentage of the 270 Valid formal concepts from the pure formal context found in the top-270 of the ordered noisy formal concepts was 83.7% (recall 0.837). Ordering the noisy formal concepts by LStab and the Collapse Index produced recall values of 0.870 and 0.855 respectively. Of these the LStab performed the best with a recall value of 0.870 but was not markedly better than the Collapse Index at 0.855 - meanwhile the Support Value performed the worst.

With 5% noise added to the pure formal context, the result was a lattice with 2,546 formal concepts. Following similar processes as above in retrieving the 270 Valid formal concepts, using the Support Value had a recall value of 0.726; using LStab had a recall value of 0.755; and using the Collapse Index gave an equal recall value of 0.755. This shows a gradual and expected decline in the efficacy of all concept relevant measures as the level of noise incorporated into the pure formal context is increased. However it is noted that both LStab and Collapse Index exhibited better performance than the Support Value.

Using 10% noise further reduced the recall values of all three relevancy measures, details of which may be found in Table 6.14. Noteworthy is that the Collapse Index (0.696) once again exhibits comparable performance to the the Logarithmic Stability (0.689) in recall, and both measures performing better than the Support Value. 20% noise, seen in Table 6.14, is also reflective of this trend.

<table>
<thead>
<tr>
<th>Noise</th>
<th>1%</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td># Concepts in Noisy Lattice</td>
<td>782</td>
<td>2446</td>
<td>6730</td>
<td>28869</td>
</tr>
<tr>
<td>Support Value</td>
<td>0.837</td>
<td>0.726</td>
<td>0.652</td>
<td>0.533</td>
</tr>
<tr>
<td>LStab</td>
<td>0.870</td>
<td>0.755</td>
<td>0.689</td>
<td>0.574</td>
</tr>
<tr>
<td>Collapse Index</td>
<td>0.855</td>
<td>0.755</td>
<td>0.696</td>
<td>0.574</td>
</tr>
</tbody>
</table>

Table 6.14: Recall Using Type I Noise
CHAPTER 6. COLLAPSE INDEX VALIDATION EXPERIMENTS

6.10.2 Results: Type II Noise

Noise in Type II experiments took the form of adding object instances to the pure formal context. The amount of objects added was expressed as a percentage of the number of object instances in the pure formal context. The set of percentages corresponding to the noise level were 1, 5, 10, 20, and 40%.

Similar to the Type I experiments, the desire was to retrieve the 270 valid formal concepts originating from the pure lattice in the noisy lattice. This was done using the Support Value, Logarithmic Stability, and the Collapse Index as varying ways of ordering the formal concepts in the noisy lattice in terms of their relevance. Results are summarised in Table 6.15.

Table 6.15: Recall Using Type II Noise

<table>
<thead>
<tr>
<th>Noise (% objects added)</th>
<th>1%</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td># Concepts in Noisy Lattice</td>
<td>389</td>
<td>546</td>
<td>703</td>
<td>1109</td>
<td>1892</td>
</tr>
<tr>
<td>Support Value</td>
<td>0.959</td>
<td>0.863</td>
<td>0.778</td>
<td>0.704</td>
<td>0.607</td>
</tr>
<tr>
<td>LStab</td>
<td>0.956</td>
<td>0.956</td>
<td>0.881</td>
<td>0.774</td>
<td>0.656</td>
</tr>
<tr>
<td>Collapse Index</td>
<td>0.963</td>
<td>0.930</td>
<td>0.878</td>
<td>0.752</td>
<td>0.637</td>
</tr>
</tbody>
</table>

Our first point of interest is that adding Type I noise (randomly changing binary relations in the formal context) adds a greater amount of formal concepts to the lattice derived from the noisy formal context than adding Type II noise (adding a percentage of random object instances). Although these two types of noise are not necessarily the converse of each other, there is still value in comparing their effects on the lattice.

From Figure 6.12 it is noted that the growth of the number of formal concepts in the noisy lattice due to Type I noise is much faster than that of Type II. While the growth exhibited using Type II noise appears linear, that of Type I appears at least polynomial if not exponential. This suggests that noise which affects attribute descriptions of object instances is a greater threat to an accurate taxonomical representation of a domain using FCA, than the addition of illegitimate object instances to the domain’s context.

It may be necessary to take the previous into consideration when examining the results of recall experiments for Type II noise shown in Table 6.15. Although this table shows that the recall values for all relevancy measures are generally higher for Type II noise than those of Type I noise in Table 6.14, it may not necessarily be due to a better ability of these relevancy measures in differentiating between noise and
legitimate concepts for Type II noise. Fewer concepts to consider may be the main contributing factor to having these high recall figures.

At the same time there are several other notable aspects of the Type II results in Table 6.15. Representation of the information contained in Table 6.15 as a set of line graphs (Fig. 6.13) shows that the Support Value generally under-performs the Logarithmic Stability and the Collapse Index for the various noise levels. Comparing the performance of the Collapse Index to that of the Logarithmic Stability reveals that the Logarithmic Stability performs better than the Collapse Index, although not significantly so.

As it relates to Hypothesis 5, Figure 6.13 shows that for Type II noise both the Collapse Index and LStab perform similarly, with neither showing a clear advantage. Both however do produce better recall results than the Support Value. Similarly for Type I noise, while neither the Collapse Index nor LStab showed any real improvement over the other (Table 6.14), both were consistently better than the Support Value in recalling formal concepts in the noisy lattice.
6.11 Experiment 4: Efficiency of Collapse Index

6.11.1 Preliminaries

Some theoretical investigation into the efficiency of the Collapse Index was undertaken in Section 5.7, and how this efficiency compares with that of the Stability Index. However due to the fact that the Collapse Index is not a calculation of the Stability Index value, and the uncertainties of elements contained in the data type used to represent formal concepts in FCA libraries, it was deemed more useful to undertake empirical assessment of the Collapse Index’s efficiency.

These empirical assessments take the form of two types of tests, the first of which being simply recording the time taken for the Collapse Index and the Stability Index to calculate individually the relevancy values for all formal concepts in the lattice. While great care was made to replicate the algorithm provided for the Stability Index in Algorithm 2 it was also necessary to ensure that any inconsistencies had between said algorithm and the one implemented in Java for the purposes of this experiment were mitigated. As such a second experiment was designed where a count was recorded for each time a formal concept was accessed when processing the relevancy values of the entire lattice. These counts, and averages of these counts, were done for both the Collapse Index and the Stability Index algorithms.
6.11.2 Description of Runtime Experiment

The runtime experiment is fairly straightforward and is conducted by simply measuring the time taken for each algorithm to calculate a relevancy value for all formal concepts in a lattice, assuming the lattice is generated beforehand. The runtime experiment is conducted for the genre, postal area, and postal districts formal contexts described in Section 6.4. The steps in experiment undertaken are now described.

1. Implement Stability Index algorithm (Algorithm 2) and Collapse Index algorithm (Algorithm 1) in Java.
2. Generate an FCA lattice for the genre, postal area, and postal district formal contexts.
3. For each lattice apply the Stability Index algorithm and record the time taken to calculate relevancy values for all formal concepts. Repeat this process five (5) times for each lattice. Calculate average time-taken for each lattice.
4. For each lattice apply the Collapse Index algorithm and record the time taken to calculate relevancy values for all formal concepts. Repeat this process five (5) times for each lattice. Calculate average time-taken for each lattice.

Of relevance is a description of the computing environment in which the runtime experiments were conducted. Details are as follows:

- **computer model**: Samsung LAP-HPZS91NC100
- **memory**: 8GB Memory
- **cpu**: Intel Core i7-2675QM CPU @2.20GHz
- **OS**: Windows 7 Enterprise; Service Pack 1. 64-bit

6.11.3 Description of Steps Experiment

‘Steps’ in this regard is intended to represent the number of times a formal concept is accessed by the algorithm. However since ‘access’ may be inconsistently defined
between algorithms, for the purpose of the experiments, specific points of access to formal concepts in the algorithms are being referenced.

In the case of the Collapse Index of Algorithm 1 the only information required from the formal concepts are the cardinality of their extents. Each time this cardinality of a formal concept is read by the algorithm this is counted as ‘the formal concept being accessed’.

Algorithm 2 which determines the Stability Indices accesses the formal concepts in a variety of ways. Examples include: within the for-loop of lines 3 to 6, the formal concepts are counted in the process of knowing the number of lower neighbours each formal concept has; Line 9 calculates the stability by reading the size of the formal concept’s extent; while line 12 performs a calculation on a newly introduced aspect (Subsets) of a formal concept.

For the purposes of the experiment only the 14th line of the algorithm is counted as a step. Basically the while-loop (lines 7 - 17) loops until the Count value of each formal concept is 0. While a formal concept may be ‘accessed’ in various ways at different points in the algorithm, we know it is no longer accessed after its Count is equal to 0. Given that line 14 decrements the Count value, the number of times the Count is decremented for Algorithm 2 should correspond to the minimum amount of times the formal concept was accessed. Therefore, each time the Count value of a formal concept is decremented in line 14 this step is counted as ‘the formal concept being accessed’.

### 6.11.4 Hypotheses

Formal representations of the expected outcomes of the experiments are represented in the following hypotheses.

**Hypothesis 6.** Calculation of the Collapse Index values of all formal concepts in an FCA lattice using Algorithm 1 takes less time than that of the Stability Index via Algorithm 2.

---

18 This is done mainly for the purpose of determining whether cardinality of the extent of a lower neighbour is the maximum of all the lower neighbour

19 Note that the word Count here represents a variable introduced in Algorithm 2 and is different from other usages of the word ‘count’
Hypothesis 7. Calculation of the Collapse Index values of all formal concepts in an FCA lattice using Algorithm 1 accesses each formal concept in an FCA lattice, on average, less than that of the Stability Index via Algorithm 2.

6.12 Experiment 4: Efficiency: Results

This section includes the results for the runtime and step-counting efficiency-oriented experiments described in Section 6.11.2 and 6.11.3 respectively.

6.12.1 Results: Runtime

Runtime experiments were carried out for lattices created from three separate formal contexts; genres, postal areas, and postal districts. The first results on display are those of the runtimes for calculation of the Stability and Collapse Index values for the FCA lattice derived from the genre-based formal context. Table 6.16 shows these results for the five (5) occasions on which runtime experiments were conducted for both Collapse and Stability Index implementations.

<table>
<thead>
<tr>
<th></th>
<th>runtime(CI)</th>
<th>runtime(SI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>115ms</td>
<td>4mins 82ms</td>
</tr>
<tr>
<td>2</td>
<td>56ms</td>
<td>3mins 429ms</td>
</tr>
<tr>
<td>3</td>
<td>41ms</td>
<td>3mins 369ms</td>
</tr>
<tr>
<td>4</td>
<td>22ms</td>
<td>3mins 334ms</td>
</tr>
<tr>
<td>5</td>
<td>18ms</td>
<td>3mins 357ms</td>
</tr>
<tr>
<td>average</td>
<td>50.4ms</td>
<td>3mins 514 ms</td>
</tr>
</tbody>
</table>

It is evident from these results that the acquisition of the relevance value of all formal concepts in an FCA lattice via the current implementations of the Collapse Index and Stability Index algorithms is accomplished in a lower time span for the Collapse Index than that of Stability Index implementations.

In the case of the postal area lattice whose results are shown in Table 6.17, as with the genre-based lattice, the Collapse Index implementation showed much lower runtimes than that of the Stability Index implementation. An average runtime of 583 milliseconds was obtained compared to the approximately 30 minute average of the Stability Index implementation.
Table 6.17: Runtimes for Postal Area-based Lattice

<table>
<thead>
<tr>
<th></th>
<th>runtime(CI)</th>
<th>runtime(SI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>766 ms</td>
<td>30 mins 22 s</td>
</tr>
<tr>
<td>2</td>
<td>545 ms</td>
<td>30 mins 11 s</td>
</tr>
<tr>
<td>3</td>
<td>569 ms</td>
<td>30 mins 04 s</td>
</tr>
<tr>
<td>4</td>
<td>490 ms</td>
<td>30 mins 08 s</td>
</tr>
<tr>
<td>5</td>
<td>546 ms</td>
<td>30 mins 01 s</td>
</tr>
<tr>
<td>average</td>
<td>583 ms</td>
<td>30 mins 09 s</td>
</tr>
</tbody>
</table>

Table 6.18: Runtimes for Postal District-based Lattice

<table>
<thead>
<tr>
<th></th>
<th>runtime(CI)</th>
<th>runtime(SI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3s 963 ms</td>
<td>1hr 54 mins 06 s</td>
</tr>
<tr>
<td>2</td>
<td>3s 612 ms</td>
<td>1hr 53 mins 18 s</td>
</tr>
<tr>
<td>3</td>
<td>3s 589 ms</td>
<td>1hr 52 mins 34 s</td>
</tr>
<tr>
<td>4</td>
<td>3s 616 ms</td>
<td>1hr 51 mins 01 s</td>
</tr>
<tr>
<td>5</td>
<td>3s 559 ms</td>
<td>1hr 37 mins 07 s</td>
</tr>
<tr>
<td>average</td>
<td>3s 668 ms</td>
<td>1hr 49 mins 37 s</td>
</tr>
</tbody>
</table>

Finally the results of the runtime calculations for the postal district-based FCA lattice are shown in Table 6.18. Continuing with the previous behaviours is the poor performance of the Stability Index implementation. While the implementation of the Collapse Index accomplishes its goals in approximately 3 and a half minutes, the implementation of Stability Index algorithm approaches two hours in execution time.

If one were to contrast these runtimes with the number of formal concepts in each lattice (Table 6.6), it is noticed that for both the Stability and Collapse Index implementations, the runtime does not correlate directly with the number of formal concepts in the lattice. This is evident in the fact that although the postal district FCA lattice has only 3,913 formal concepts, the runtimes in calculating the set of Collapse and Stability Index values of the entire postal district-based lattice is greater than when undertaking the same for the postal area FCA lattice which is composed of a greater amount (5,441) of formal concepts.

It must be noted that time complexities are theoretical maximums and ‘real world’ restrictions and factors mean that time complexity does not necessarily translate directly to runtime. As an example, consider that the time complexity for determining the concept relevancy of all formal concepts in an FCA lattice using the Stability Index algorithm employed in Algorithm 2 is $O(|G|^2|M|L^3)$ (Zhi, 2014). Of the three

---

20In Table 6.6 the infimum and supremum are excluded
variables, increasing the number of number formal concepts in the lattice $L$ has the greatest proportional effect on the theoretical maximum time. This however does not necessarily imply the runtime for a lattice with a high $L$ value would be greater than the runtime for a lattice with a lower value of $L$; it comments only on the theoretical maximum.

For both the postal district and postal area lattices the number of movies in their respective formal contexts was $|G| = 2,122$. While $M$, the number of possible attributes, is larger for the postal district formal context than the postal area, the restriction of the number of attributes to four (4) meant a very shallow postal district lattice as discussed in Section 6.6.4. This shallow-and-wide lattice may have implications for the other aspects for the implementation of Algorithm 2, such as functions of FCA Java libraries which enable traversal of the lattice in implementation. This is neither a claim nor hypothesis on the operations of these libraries, as the determination of these may be outside of the scope of this research, but as an illustration that values for variables in the time complexity function may affect variables not represented in the time complexity function but which affect the runtime during implementation.

Ultimately however, to produce the relevancy values of all formal concepts in an FCA lattice, the runtimes using the implementation of the proposed Collapse Index algorithm (Algorithm 1) are much less than the runtimes using the implementation of the main Stability Index algorithm (Algorithm 2).

6.12.2 Results: Steps

As mentioned in the preliminaries, while best attempts were made to replicate Algorithm 2 in code, there exist possibilities for some slight differences. For this reason a separate test was constructed where a count was done for every time a formal concept was accessed in the implementations of the Collapse Index and the Stability Index algorithms.

Fundamentally what this ‘steps’ experiment does is to determine the minimum number of times each formal concept is accessed in the implementation of Stability Index algorithm of Algorithm 2, as well as the maximum number of times each formal concept is accessed in the implementation of the Collapse Index of Algorithm 1. Both algorithms were applied to the genre, postal area, and postal district lattices. For
each lattice, the average amount of times each formal concept was accessed using the various algorithms are shown in Table 6.19

Table 6.19: Average Amount Of Times Each Formal Concept Is Accessed

<table>
<thead>
<tr>
<th></th>
<th>Collapse Index</th>
<th>Stability Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>genre</td>
<td>4.05</td>
<td>6.17</td>
</tr>
<tr>
<td>postal area</td>
<td>4.41</td>
<td>7.11</td>
</tr>
<tr>
<td>postal district</td>
<td>4.03</td>
<td>4.64</td>
</tr>
</tbody>
</table>

From the table it is observed that for all three types of formal contexts, in the calculation of the relevancy values for all formal concepts in their respective FCA lattices, the unoptimised Collapse Index algorithm accesses each formal concept a lesser amount of times than the Stability Index algorithm of Algorithm 2. In the case of the postal district lattice, the maximum amount of times each formal concept is accessed is on average 4.03 times; if the Stability Index implementation was applied, the minimum number of times a formal concept was accessed is 4.64 times. Of the three lattices the performance of both algorithms, with respect to the average amount of times formal concepts are accessed, is closest for the postal district. This may potentially be related to the fact of the postal district lattice being wide and shallow, since the Stability Index algorithm visits repeatedly visits the superconcepts of formal concepts as the algorithm iterates upwards in the lattice. A formal concept in a shallow lattice will have very few superconcepts.

While the improvement of the access-count efficiency for the Collapse Index over the Stability Index is not as dramatic as the runtimes experiment, attention is paid to the fact that for the Collapse Index the maximum amount of times a formal concept is accessed is what is being recorded, whereas for the Stability Index implementation the minimum. There inevitably exists multiple other points in the Stability Index algorithm at which formal concepts are accessed in a variety of ways and for different purposes. This will undoubtedly increase the range between the frequency at which each formal concepts is accessed by the Collapse Index algorithm and the frequency accessed by the Stability Index algorithm.

To summarise, the application of both the Collapse Index and Stability Index algorithms to three (3) different formal contexts, each of different size and characteristics, has shown that the proposed Collapse Index algorithm exhibits much lower runtimes
than the central Stability Index algorithm used in FCA research. In addition it was also shown that the Collapse Index algorithm accesses each formal concept in the lattice a lesser amount of times than the Stability Index algorithm. These statement act as responses to the hypotheses of the efficiency experiment, Hypothesis 6 and Hypothesis 7.

6.13 Conclusion

Included in this chapter were four experiments designed to contrast empirically the performance of the Collapse Index to the other prominent measures of determining the relevancy of formal concepts, the Support Value and the Stability Index. Due to the fact that the Stability Index behaves exponentially and the Collapse Index was not expected to, the Logarithmic Stability was used in the place of the Stability Index in most of the experiments. The experiments were carried out on the genre, postal area, or postal district formal contexts.

Experiment 1 showed that the relevancy values produced by the Collapse Index show higher correlation with those of the Logarithmic Stability than those of the Support Value do with the Logarithmic Stability. Experiment 2 showed that the Collapse Index values generated from a subcontext can better account for variances in the relevancy values of formal concepts taken from another subcontext of the same domain; this is in contrast to the Logarithmic Stability. Experiment 3 showed that the Collapse Index is able to retrieve concepts in a noisy lattice at a comparable level of success as the Logarithmic Stability. And finally Experiment 4 showed that the Collapse Index algorithm achieves much lower runtimes than the main Stability Index algorithm when calculating the relevancy values of each formal concept in the an FCA lattice, and on average accesses formal concepts in the FCA lattice less often.
Chapter 7

VALIDATION OF THE
SEMANTIC EXTRACTION
APPROACH

7.1 Introduction

This chapter explores the approach towards semantic extraction described in Chapter 4 where a concept relevancy measure is applied to an FCA lattice to select the relevant concepts. More specifically the notion of a lattice pruned of low-relevance concepts is imported into a recommender system (RS) designed to recommend movies to users based on their past movie-viewing habits. While Chapter 8 speaks to the implementation and results of the RS, this chapter provides a description of the incorporation of the proposed semantic extraction approach by describing an RS where pruned lattices are used.

Recommender systems as described in Section 2.6 are implemented in two different ways: collaborative filtering (CF) and content-based filtering (CB). Because of its success CF has been subject to the majority of recommender system research. However CF has historically been associated with making recommendations based on the user viewing habits which have been formalised in the form of explicit ratings of content the user has previously viewed. In the case of the available BT TV dataset explicit ratings are not available.
At the same time the over-specification issue common to content-based recommender systems (Véras et al., 2015) compels one to embrace the $k$-NN approach utilised by collaborative recommender systems. For $k$-NN solutions, recommendations are made taking into consideration movies viewed by a user’s most similar neighbours, and similar neighbours are determined by similar patterns of movie ratings by multiple users. However, granted that for the BT TV dataset users have not provided explicit ratings of movies, a hybrid RS\footnote{A combination of aspects of content-based and collaborative recommender systems} is chosen as the RS approach.

In addition to the need to incorporate the Collapse Index into the RS design, there was also a desire to include concepts defined by patterns of UIC. A recommender solution was thus required that made use of Formal Concept Analysis. Although there exists some use of FCA in collaborative RS solutions such as du Boucher-Ryan and Bridge (2006) where FCA was used to group users who rated similar items, in the case of this thesis a novel application of FCA is introduced where FCA lattices are utilised as user-profiles. This would allow for the Collapse Index to be used as a pruning solution in user profiles, as well as have UIC-defined concepts be used to represent aspects of user interest.

It is important to note that for the implementation of the RS in the course of this thesis, emphasis is not being placed on the RS being more efficient or more effective than the current state of art; the main priorities are to demonstrate the successful application of the approach described in Chapter 4 (lattices constructed from datasets are pruned from low-relevance concepts) in the domain of a recommender systems as well as the successful utilisation of concepts retrieved by taking into consideration UIC links.

As the proposed RS largely channels collaborative recommender systems, Section 7.2 provides a brief overview of CF. Section 7.3 then provides the details of the expected operation of the proposed RS concluding with an overview of its design with respect to the semantic extraction approach of Chapter 4, and a proposed architecture for the inclusion of the RS in a business case.
CHAPTER 7. VALIDATION OF APPROACH

7.2 Basic Collaborative Recommender System

Although technically a hybrid RS, granted that CF recommender systems are generally characterised by the presence of a ratings matrix and such a matrix is unavailable, the recommender system designed for the purpose of this thesis takes most of its cues from CF. For this reason, this section provides a more detailed understanding of CF in order to show how the thesis’ proposed RS parallels CF.

CF is built on the basis that there exists a set of \( m \) users, \( U = \{u_1, u_2, ..., U_m\} \), a set of \( n \) items, \( I = \{i_1, i_2, ..., i_n\} \), and a set or user-rating, \( R = \{r_{u,i}\} \), where as an example \( r_{2,1} \) is the rating that User 2 (\( u_2 \)) has assigned to Item 1 (\( i_1 \)). The complete scenario may be represented as matrix where each column corresponds to an item and each row a user; an example is provided in Table 7.1. For illustrative purposes we may state that with respect Table 7.1 that \( r_{2,1} = 5 \), i.e. User 2 rated Item 1 a 5.

<table>
<thead>
<tr>
<th></th>
<th>( i_1 )</th>
<th>( i_2 )</th>
<th>( \cdots )</th>
<th>( i_0 )</th>
<th>( \cdots )</th>
<th>( i_{n-1} )</th>
<th>( i_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u_1 )</td>
<td>3</td>
<td>-</td>
<td>( \cdots )</td>
<td>3</td>
<td>-</td>
<td>( \cdots )</td>
<td>-</td>
</tr>
<tr>
<td>( u_2 )</td>
<td>5</td>
<td>1</td>
<td>-</td>
<td></td>
<td>-</td>
<td>( \cdots )</td>
<td>-</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>-</td>
<td>-</td>
<td>( \cdots )</td>
<td>-</td>
<td>-</td>
<td>( \cdots )</td>
<td>-</td>
</tr>
<tr>
<td>( u_j )</td>
<td>5</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>( \cdots )</td>
<td>2</td>
</tr>
<tr>
<td>( u_k )</td>
<td>2</td>
<td>-</td>
<td>3</td>
<td>2</td>
<td></td>
<td>( \cdots )</td>
<td></td>
</tr>
<tr>
<td>( \vdots )</td>
<td>-</td>
<td>-</td>
<td>( \cdots )</td>
<td>-</td>
<td>-</td>
<td>( \cdots )</td>
<td>-</td>
</tr>
<tr>
<td>( u_{m-1} )</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>3</td>
<td>-</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>( u_m )</td>
<td>1</td>
<td>-</td>
<td>5</td>
<td>4</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This ratings matrix could be interpreted as either ‘a set of user vectors’ (set of rows) or ‘a set of item vectors’ (set of columns). In the case of the former, the user vectors are the formal representation of a user-profile for the purposes of CF. In possession of this formalisation of a user-profile, a similarity measure could then be employed to find the set of nearest neighbours (NN) of each user by calculating the similarity between user vectors. Conversely, as per Section 2.6, if undertaking an item-based collaborative solution, one may use the similarity measures to determine the similarity between item vectors rather than users. However the implementation is limited to only that of the user-based approach.

While there exists a variety of similarity measures, the two most commonly utilised to determine similarity between user vectors are that of Pearson’s correlation and the
cosine similarity. Given \( u, v \in U \), the Pearson correlation similarity measure, \( \rho(u, v) \), can be formally expressed as

\[
\rho(u, v) = \frac{\sum_{i \in I} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I} (r_{v,i} - \bar{r}_v)^2}}. \tag{7.1}
\]

Alternatively the cosine similarity measure, \( \cos(u, v) \), is

\[
\cos(u, v) = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}| \times |\vec{v}|} \tag{7.2}
\]

where \( \vec{u} \cdot \vec{v} \) is the dot-product of the vectors of user \( u \) and user \( v \). Given \( \vec{v} = (r_{v,1}, r_{v,1}, ..., r_{v,n-1}, r_{v,n}) \) then \( |\vec{v}| \) is the length of the vector \( \vec{v} \) and may be calculated as

\[
|\vec{v}| = \sqrt{\sum_{i=1}^{n} r_{v,i}^2} \tag{7.3}
\]

Utilising either of these similarity measures one may obtain the set \( NN \) of the most similar users to each specific user \( u \). A predicted rating, \( p_{u,i} \), is then calculated for items which user \( u \) has not previously rated (i.e. blank cells in Table 7.1). A simple weighted average such as Equation 7.4 may be used to calculate this predicted rating where \( w(j, u) \) is representative of the similarity between users \( j \) and \( u \).

\[
p_{u,i} = \frac{\sum_{j \in N} r_{j,i}w(j, u)}{\sum_{j \in N} |w(j, u)|} \tag{7.4}
\]

For each user, the items the user has not previously consumed are ranked according to the predicted ratings. Of these ranked items, the highest rated are recommended to the user.

While research is ongoing on improving several aspects and limitations of CF, several of which were previously described in Section 2.6 including ratings matrix factorisation and proposed solutions for addressing sparsity and cold start, there remains three (3) core components of user-based CF. These are summarized below.

1. A ratings matrix or a triple \( M = (U, I, R) \) where \( U \) is a set of users, \( I \) a set of items, and \( R \) the set of ratings users have provided for items.
2. A similarity measure (e.g. Pearson’s, cosine) which provides a value that represents how similar a user is to another user.

3. A means of making a ratings prediction based on user-similarity (e.g. simple weighted average).

With these (3) core components of CF in mind, the design of the proposed RS is explored.

### 7.3 Recommender System: FCA Lattice as User Profile

#### 7.3.1 User Profiles

With the lack of explicit ratings and prioritising the usage of concepts in mind, the first novel aspect of the proposed RS is that each user-profile not be a vector of ratings but instead be composed of a set of formal concepts. The premise of this ‘lattice representation of a user’ approach is that characteristics of the user would be reflected in the concepts and hierarchy of the user’s FCA lattice.

For each user, the formal concepts used to represent their profile are those present in an FCA lattice created from a formal context built on a user’s past movie-viewing records. The object instances for these formal contexts are the movies the user has previously viewed, while the attribute set are characteristics of the movies the user has previously viewed (e.g. genre description).

Note that movies users have watched and their attribute descriptions are sourced from one or more datasets. This would require, in advance of the creation of user formal contexts, undergoing the ‘pre-processing of the dataset’ phase of semantic extraction approach outlined in Section 4.5. Specific to the BT TV case study the steps mentioned in Section 6.3 would be done to clear the dataset from inconsistencies and errors.

The generation of a user-profile lattice is demonstrated by the examples of User A and User B. The movies which User A and User B have previously viewed are shown, formatted as formal contexts, in Table 7.2 and Table 7.3 respectively.

Formal Concept Analysis is then applied to both formal contexts generating lattice
Table 7.2: Formal Context Showing Genre Attributes of Movies Viewed by User A

<table>
<thead>
<tr>
<th>Movie</th>
<th>action</th>
<th>fantasy</th>
<th>adventure</th>
<th>crime</th>
<th>drama</th>
<th>mystery</th>
</tr>
</thead>
<tbody>
<tr>
<td>BATMAN</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAN OF STEEL</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GODFATHER</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCEPTION</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>SHAWSHANK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 7.3: Formal Context Showing Genre Attributes of Movies Viewed by User B

<table>
<thead>
<tr>
<th>Movie</th>
<th>action</th>
<th>fantasy</th>
<th>adventure</th>
<th>crime</th>
<th>drama</th>
<th>mystery</th>
</tr>
</thead>
<tbody>
<tr>
<td>BATMAN</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DARK KNIGHT</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHERLOCK HOLMES</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>INCEPTION</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>SHAWSHANK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Each profile is then viewed as being ‘a set of formal concepts’. The infimum and supremum of the lattices are ignored as it is unlikely for a movie to not have any attribute descriptions or for a movie to have all possible attributes as its description. Consequently the final representations of the profiles of User A and User B are available in Table 7.4 where the 2nd and 4th columns are the formal concepts in the profiles in User A and User B respectively, and columns 1 and 3 are IDs of the formal concepts.

---

2 Reduced labelling is described in Appendix C
7.3.2 Choice of Movie Characteristics

The previous section described the process of composing a user-profile from a sample of movies where each movie has been previously viewed by the user. Construction of such an FCA profile is dependent on the formal context representation of the sample utilised.

In the current movie recommendation scenario, for a formal context $K := (G, M, I)$, $G$ is the set of movies in the sample set of movies viewed by the user whereas $M$ represents the set of possible attributes which a movie may be described as having. Although the examples provided in Section 7.3.1 were limited to genre descriptions of movies, there exists a much wider set of possible attribute values which may characterize movie instances.

The Internet Movie Database (IMDb), as an example, provides a vast amount of descriptive information about individual movies. These include the aforementioned genre attribute category\(^3\) as well as other attribute categories such as: actors, directors, writers, average ratings, quotes, trivia, user reviews, release dates, among other things. For a variety of reasons it may not be useful or practical to utilise all or most of these attributes categories. Categories such as user-reviews are unstructured and may require great effort in extracting key concepts or themes within the text. In the case of the attribute category actors, there are and have been tens of thousands of movie actors throughout history; a specific movie may have a cast of hundreds.

Formal contexts where there exist a large set of possible attribute values and/or object instances are described with a large set of attribute values lead to computational efficiency concerns in the implementation of proposed RS. Where $|M|$ is large this has

\(^3\)attribute category refers to a set of values used to describe an aspect of a movie. e.g genre. An attribute value is an element within the attribute category set e.g. action, comedy, etc.
a direct impact on the time complexity of the construction of the FCA lattice given that the time complexity for constructing an FCA lattice is $O(|G|^2|M|L)$ (Zhi, 2014). Object instances being described by a large number of attribute values increases the possibility of a large variety of shared attribute-values across object instances, i.e. a large variety of intents and subsequently formal concepts $L$ in the lattice. The computational complexity of these large lattices also carries forward to lattice pruning techniques and lattice similarity measures employed in the RS implementation.

Due to the aforementioned efficiency concerns some restrictions are necessary on what attributes are used to describe the movies. One solution may be the usage of matrix factorisation techniques on the formal context where $|M|$ is large. This would reduce the size of $|M|$ in the final formal context as well as addressing the sparsity issue whose presence would coincide with a large value of $|M|$.

An alternate approach is simply to restrict $M$ to a limited set of attribute categories which are felt to contain either implicit or explicit information on important characteristics of the movie. If, for example, the formal context of user profiles is built solely on the genre attribute category, a user’s profile containing the concepts romantic-comedy and western-adventure may be enough to represent the tastes of the user and find similar users.

However since we are also interested in concepts obtained via UIC links rather than relying solely on typical movie attribute categories such as genres as the description of the movies, UIC-defined concepts are used to describe user profiles. More specifically, patterns of UIC are to be applied in the capacity of the attribute set $M$ in the formal context $K := (G, M, I)$ which forms the basis of a user-profile. The UIC attribute category could be the sole attribute category used in $M$ or could be combined with other common attribute categories such as genres.

Implementing such an RS also requires a decision on what factors of UIC one believes would effectively serve in characterising aspects of movies. Some UIC factors which are candidates include:

- *interaction*: percentage of movie watched, frequency of channel change, etc.
- *user characteristics*: e.g age, gender, profession, geographical location, etc.
- *context*: time of day movie is watched, day of week movie is mostly watched,
geographical location.

Certainly the contribution of concepts based on the previous factors would vary in their ability to characterise aspects of movies and create useful concepts. Movies which share an age demographic pattern of the 10-15 age group may represent the movie being an animated movie, whereas channel-switching patterns may not be really saying much about the movie apart from the movie’s quality\textsuperscript{4}.

Furthermore, one must bear in mind that data related to a variety of these factors may not be readily available. Some customer data such as age and location may not be collected or available due to privacy-related regulations. Attempts to include these factors as a source of patterns of UIC may be hampered by the absence of the necessary data.

Due to the need for a sufficient amount of accumulated data before a movie can be ascribed these patterns, newly introduced movies to service provider would be largely devoid of these types of descriptors. However the descriptors of movies are not restricted to only patterns of UIC but also to the earlier IMDb-type attribute categories. When UIC data are available they could then be combined with IMDb type attribute categories.

Assuming that lattice characterisation of a user-profile is indeed representative, one is then tasked in addressing the second core component of user-based collaborative recommender systems, which is the development of a function to assess the similarity between user profiles.

### 7.3.3 Similarity Measure

For CF a distance metric must be available to assign a similarity value between each possible pair of users. Furthermore, usage of FCA lattices as user-profiles requires a similarity measure oriented-towards lattice similarity. For such a purpose ideas are appropriated from both Maedche and Staab (2002) and Formica (2006) - papers offering solutions for measuring similarities between ontologies and/or concepts.

Maedche and Staab (2002) obtain the similarity between ontologies $O_1$ and $O_2$ by comparing each concept present in $O_1$ to its equivalent concept (lexical representation)

\textsuperscript{4}These are neither claims nor hypotheses but simply mentioned for explanatory reasons
in $O_2$. This concept comparison, referred to as the Taxonomic Overlap ($TO$) is done using the Jaccard coefficient of the Semantic Cotopy ($SC$) of the both concepts; the Semantic Cotopy being the set of hypernyms and hyponyms of the concept in question.

If the lexical representation of a concept $C_1$ in $O_1$ is not found in $O_2$ then the highest $TO$ value obtained when $C_1$ is compared with each concept in $O_2$ is selected as the $TO$ value of $C_1$. The overall similarity between $O_1$ and $O_2$ is then obtained by comparing the average Taxonomic Overlap value ($\overline{TO}$) for all concepts in $O_1$.

For our Formal Context Analysis purposes there are several key differences, the first of which is that formal concepts in an FCA lattice do not have lexical representations, therefore a direct comparison cannot be knowingly carried out between a formal concept in lattice $A$ and its equivalent formal concept in lattice $B$. It is therefore required, in the process of comparing the similarity between lattice $A$ and $B$ ($sim(K_A, K_B)$), to compare each formal concept in $K_A$ to all formal concepts in $K_B$. If the formal concept in $K_A$ of current interest is $(C_A, D_A)$ then the formal concept $(E_B, F_B)$ in lattice $K_B$ which is most similar to $(C_A, D_A)$ is retrieved. The concept $(E_B, F_B)$ which is deemed most similar is described as the ‘best match’.

The second key difference for the proposed lattice-oriented similarity measure is that in the process of comparing concepts across FCA lattices, as opposed to the Jaccard coefficient of the Semantic Cotopies, the Jaccard coefficient of the intent as well as the Jaccard coefficient of the extent are now calculated. If one seeks to find the similarity between the formal concept $(C_A, D_A)$ in $K_A$ to the formal concept $(E_B, F_B)$ in $K_B$ then the similarity value is a function of the $J(C_A, E_B)$ and $J(D_A, F_B)$. Essentially

$$J(\{(C_A, D_A), (E_B, F_B)\}) = wJ(C_A, E_B) + (1 - w)J(D_A, F_B)$$  \hspace{1cm} (7.5)

where $w$ is a weight whose value may be adjusted through heuristic processes or based on the level of priority one seeks to assign to object instances or attributes.

Given a formal concept $(C_A, D_A)$ in lattice $K_A$, its best match in lattice $K_B$ is

$$J'(\{(C_A, D_A), K_A, K_B\}) = \max_{(E_B, F_B) \in K_B} J(\{(C_A, D_A), (E_B, F_B)\})$$  \hspace{1cm} (7.6)

The overall similarity between lattices $K_A$ and $K_B$ is then obtained by averaging the similarity of all the best matches. Formally
CHAPTER 7. VALIDATION OF APPROACH

\[ \text{sim}(K_A, K_B) = \frac{\sum_{(C_A, D_A) \in K_A} J'( (C_A, D_A), K_A, K_B) }{n} \]  

(7.7)

where \( n \) is the number of number of concepts \((C_A, D_A) \in K_A\). Note that for our purposes the number of \((C_A, D_A) \in K_A\) and \((E_B, F_B) \in K_B\) are reduced due to the removal of the infimum and supremum from consideration. In some instances non-infimum or non-supremum formal concepts are also removed from \(K_A\) and \(K_B\) for pruning purposes, however the set of remaining formal concepts from each lattice are still referred to as \(K_A\) and \(K_B\) respectively.

Further clarity is provided to the lattice similarity function by the way of an example. A structured way of assessing the similarity between the User profiles A and B (earlier described in Table 7.4) is presented in Table 7.5. All related calculations are not shown, however, included is the calculation of Jaccard Coefficient values between formal concept \(A_1\) in User A’s profile to the formal concepts in User B’s profile.

Table 7.5: Calculating Similarity between User A and User B

<table>
<thead>
<tr>
<th></th>
<th>(B_1) (ext)</th>
<th>(B_1) (int)</th>
<th>(B_2) (ext)</th>
<th>(B_2) (int)</th>
<th>...</th>
<th>...</th>
<th>(B_7) (ext)</th>
<th>(B_7) (int)</th>
<th>(J')</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A_1)</td>
<td>1/3</td>
<td>1/2</td>
<td>0/6</td>
<td>0/3</td>
<td></td>
<td></td>
<td>1/2</td>
<td>1/2</td>
<td>0.5</td>
</tr>
<tr>
<td>(A_6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(A_6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To better explain, focus is placed on the intersection of the first three columns and first three rows in Table 7.5; this subsection is duplicated in Table 7.6.

Table 7.6: Calculating Similarity between formal concepts \(A_1\) and \(B_1\)

<table>
<thead>
<tr>
<th></th>
<th>(B_1) (ext)</th>
<th>(B_1) (int)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A_1)</td>
<td>1/3</td>
<td>1/2</td>
</tr>
<tr>
<td>(A_1)</td>
<td></td>
<td>0.42</td>
</tr>
</tbody>
</table>

Here the formal concept \(A_1 = \{GFATHER, SHSHNK\}, \{cri, dra\}\) in Profile A \((K_A)\) is being compared with formal concept \(B_1 = \{SHSHNK, DK\}, \{cri\}\) of Profile B \((K_B)\). The Jaccard Coefficient of their extents is 1/3. likewise the Jaccard Coefficient of their intents is 1/2. If equitable weights are assigned to the extent and intent of formal concepts then \(w = \frac{1}{2}\) leading to an overall similarity between
7.3. RECOMMENDER SYSTEM: FCA LATTICE AS USER PROFILE

A₁ and B₁ of \(\frac{11}{23} + \frac{11}{22} = 0.42\). Similarly A₁ is compared to B₁, B₂, ..., B₆, B₇. The *formal concept* which A₁ is most similar to is that of B₇ having a similarity of 0.5. The similarity value of this bestmatch is placed in the final column of Table 7.5 titled \(J'\). A₂ is then compared with all formal concepts in Profile B and the best match obtained and placed in the \(J'\) column. This is repeated for all Aᵢ in Profile A. The overall similarity between User A and User B \((\text{sim}(K_A, K_B))\) would then be the sum of the values in column \(J'\) divided by 6 (the number of formal concepts in \(K_A\)).

Make note that the similarity function is not symmetric i.e. that \(\text{sim}(K_A, K_B)\) is not necessarily equal to \(\text{sim}(K_B, K_A)\). Case in point, if \(K_A \subset K_B\) then Jaccard Coefficient of the bestmatches for all formal concepts in \(K_A\) will be 1, leading to \(\text{sim}(K_A, K_B) = 1\). However, if assessing \(\text{sim}(K_B, K_A)\), while the formal concepts \(K_B\) that are also present in \(K_A\) will have a bestmatch whose Jaccard Coefficient is 1, the other formal concepts in \(K_B\) which are not \(K_A\) will have bestmatch values less than 1.⁵

While there may be potential benefits for asymmetrical similarity measures (Bao et al., 2003), in order to facilitate more fair comparisons to previous collaborative approaches which utilised symmetric similarity functions such as Pearson and Cosine similarity, the previous similarity function is updated to be symmetric. This symmetric similarity function is simply an average of \(\text{sim}(K_A, K_B)\) and \(\text{sim}(K_B, K_A)\)

\[
\overline{\text{sim}}(K_A, K_B) = \overline{\text{sim}}(K_B, K_A) = \frac{\text{sim}(K_A, K_B) + \text{sim}(K_B, K_A)}{2} \quad (7.8)
\]

The symmetric similarity measure is then used to find the set of Nearest Neighbours (NN) of each user.

Knowing the set NN, one may then explore the third core-component of a user-based collaborative recommender system - predicting item/movie ratings. In advance of this the role and process of pruning user profile lattices is discussed.

### 7.3.4 Pruning Profiles

The fundamental premise of the current RS approach is the representation of a user-profile by the formal concepts present in the corresponding FCA lattice e.g. \(K_u\). While for the current implementation the *infimum* and *supremum* of the FCA lattice

⁵Note that the *infimum* and *supremum* are removed from both \(K_A\) and \(K_B\)
are always excluded from consideration, in some instances other formal concepts may also be excluded from a user’s profile. Considering the fact that explicit ratings are not available to us, and a user having watched a movie may not necessarily be indicative of their liking said movie, effort may be required to mitigate ‘noise’ in a user’s profile.

The formal concepts being used as a representation of a user profile are emergent from the set of movies the user has watched and the characteristics of said movies. Given that the addition of objects to formal concepts may lead to exponential growth of the corresponding lattices (Babin and Kuznetsov, 2012), many additional formal concepts may arise in a user profile from the user having watched a movie. Many of these additional formal concepts may be seen as ‘noise’, especially those formal concepts which arise from the user watching a different type of movie the user does not necessarily like. The implementation of the RS would obtain value from removing these noisy formal concepts from the set of formal concepts $K_u$ which constitute the user profile.

Available for the reduction of the set of of formal concepts in a user profile to its core formal concepts is the newly introduced Collapse Index which determines a numerical value for the relevance of each formal concept in an FCA lattice. Formal concept being assigned relevancy values give control over the number of formal concepts which are included in the final user-profile $K_u$.

### 7.3.5 Predicted Rating

Section 7.2 described the simple weighted average of Equation 7.4 as a candidate for calculating a predicted rating for an item not previously consumed by a user. In light of the characteristics of the dataset including the absence of the set of explicit ratings, \{r_{u,i}\}, for movies, this simple weighted average is customised to suit the characteristics of the case study dataset. Given that $a_{v,i}$ represents whether User $v$ has watched movie $i$, the adjusted weighted average becomes

$$p_{u,i} = \frac{\sum_{K_v \in NN} \text{sim}(K_v, K_u)a_{v,i}}{|NN|}$$

(7.9)

where $p_{u,i}$ is the predicted rating for movie $i$ by User $u$ and $a_{v,i} \in \mathbb{Z}_2$. $a_{v,i} = 1$ if User $v$ has watched movie $i$. $a_{v,i} = 0$ otherwise.
For each user the set of movies available for viewing are assigned a predicted rating. These movies are ranked with respect to their predicted ratings and the highest ranked movies are recommended to the user in question.

### 7.3.6 Recommender System and Proposed Semantic Extraction Approach

Having described the positioning of the proposed RS with respect to the core components of a user-based collaborative recommender system in the previous section, an overall summary of these processes of the RS are described in the flowchart of Figure 7.2. This is done with respect to a user *User A*.

Further, the steps in the flowchart are positioned with respect to the approach to semantic extraction proposed in Chapter 4 and depicted in Figure 4.5.

- Step 1 in the RS processes outlined in Figure 7.2 would correspond to the ‘pre-processing of the dataset’ as well as include some elements of ‘transforming the dataset’. To retrieve the correct movies and attributes all issues related to typographical errors and incorrectly labelled content must be resolved. The ‘transforming the dataset’ element being included here is related to choosing the candidates for the attribute set.

- Step 2 is focused on the ‘transformation of the pre-processed data’ into the formal context of the users. Here more sophisticated analysis is performed in determining the final attributes which are utilised to describe the individual movies. This information is translated into a formal context.

- Step 3 is the ‘construction of the FCA lattice’ from the transformed data. As stated, here the formal context is translated into an FCA lattice featuring the *formal concepts* which would go on to serve as the user profile.

- Step 4 represents the ‘usage of a concept relevancy measure’ in order to identify concepts of low relevance in an FCA lattice. Step 4 can also be said to, at least partly, represent the ‘construction of the final hierarchy’. It is described as ‘partly’ as the hierarchy is not reconstructed as a lattice or non-lattice hierarchy,
1: Retrieve a sample set of movies viewed by User A and their attributes

2: Create formal context based on User A’s sample

3: Create FCA lattice $K_a$ from the formal context

4: Remove undesired concepts from User A’s lattice

5: Compare User A’s lattice to the profile-lattice of all other users

6: Find the set of $|NN|$ users most similar to User A

7: Calculate predicted ratings of unwatched movies using similarity to users in the set $NN$ as weights

8: Recommend to User A the unwatched movies with the highest predicted ratings

Figure 7.2: Flowchart Showing Processes of Recommender Engine
however the less relevant *formal concepts* have been removed from the original set of *formal concepts*.

- For the purposes of the RS there was no real need to provide formal representation of the concepts and their hierarchy, therefore the optional formalisation component of the semantic extraction approach is excluded from the RS.

- At this point the revised user profile, composed of only the relevant *formal concepts*, is then utilised in an application. This ‘application’ is composed of the steps 5-8 and involve the making of recommendations now that semantics of the user profile is known.

### 7.3.7 Architecture for Inclusion of RS in IP TV Service

In addition to the flowchart showing how recommendations are made to a user, an architecture is proposed for incorporating such an RS in the context of an IP-TV type service where a set-top box in households relays information back to a central server. Figure 7.3 shows this proposed architecture. As the use of UIC-defined concepts is not a necessity, the inclusion of the UIC related components of the architecture is represented with dotted lines to signify that these components may be excluded.

In the proposed architecture, information on the movies users have viewed is collected from set-top boxes and is relayed to a central location. This data is stored in two locations: Db1 which focuses on ‘the set of movies users have viewed over a recent period of time’, and Db3 which stores ‘long term accumulated data on user viewing records with emphasis on UIC fields’. Data from Db3 is used in the UIC Analytics Engine to retrieve patterns of UIC that are highly associated with different movies. This information is combined with data from another database, Db2, to provide individual movies with a set of attributes. Db2 contains information about movies such as IMDb type data (genres, actors, etc.). This information is then fed into the Recommender Engine (described in Figure 7.2) which builds user lattice-profiles and makes recommendations.
Figure 7.3: Architecture for Implementation of Proposed Recommender System.
7.4 Conclusion

This chapter describes an RS designed to incorporate the approach of semantic extraction using a concept relevancy measure on an FCA lattice. Key to the RS design is the representation of a user profile by a set of formal concepts that originate from the movies the user has previously viewed and their attributes. It is made clear the role the Collapse Index would play in pruning the user lattices of noisy concepts.

As the RS is designed to be user-based CF, a measure is provided that would enable one to calculate the similarity between user profiles; a function to calculate the predicted rating for an item is also presented. Included also is a proposed architecture for the implementation of such an RS in an IP TV service.
Chapter 8

RESULTS OF APPROACH VALIDATION

8.1 Introduction

This chapter details the implementation of the RS described in Chapter 7 and the results obtained when implemented. Section 8.2 gives an overview of the experiment design; Section 8.3 describes the separation of BT TV dataset into subsets for the purposes of training; and Section 8.4 speaks on the various attribute categories utilised to describe movies in the experiments. Finally Section 8.5 contains the results of said experiments including results specifically related to the usage of the Collapse Index.

8.2 Experiment Design

Before describing the experiment a decision is made on the way in which the success of the RS is assessed since this would drive the experiment’s design. Typically user ratings are available for CF; these ratings are used not only for representing the user profile but also in assessing the success of the recommendations made. The predicted ratings of a user generated by the CF are compared against actual ratings the user had assigned to movies. Evaluation metrics which provide a statistic based on the difference between observed and expected values, such as the Mean Absolute Error (MAE) or the Root Mean Square Error (RMSE), are embraced.

However the BT TV case study dataset does not include user ratings. The only
knowledge available is what the user watched and that the user has watched at least 50\% of a movie. No explicit information is present which indicates user interest in a movie. In light of this, rather than an utilising MAE or RMSE, recall is used in assessing the success level of the recommendations made.

In essence, based on a set of movies that a user has watched, a prediction is made on what other movies the user would have also watched. To accomplish this the complete set of movies a user has watched is split into two (2) disjoint sets - one which is labelled the Sample set and another which is labelled the Absent set. The Sample set is the set of movies used to construct the user’s profile. The Absent set is the set of movies which, while technically it is known that the user has watched these movies, it is assumed that this knowledge is not available. The RS seeks to predict the movies which would be in this Absent set.

To better reflect the real world context of BT, the separation of a user’s dataset is done on a temporal basis. The Sample dataset are the logs of movie viewings by a user before a certain date-time threshold, and the Absent dataset are the logs of movie viewings after said date. The earliest date in the dataset was 27\textsuperscript{th} March 2014 and the most recent date 19\textsuperscript{th} October 2014, amounting to approximately seven (7) months of data. The Sample set was chosen as ‘movies viewed for the four (4) month period from the earliest date in the dataset till July 31\textsuperscript{st} 2014’. The Absent consisted of ‘movies viewed after the 31\textsuperscript{st} July till the final date 19\textsuperscript{th} October 2014’.

The original intention was to perform this Sample-Absent split for each user and make the predictions for their respective Absent sets, however there arose several reasons for not following through with this ambition. Firstly, while some effort was undertaken to make certain processes more efficient, the overall size of the dataset and inherent limitations in the RS design, meant that processing a large number of users had problematically large runtimes. Secondly, it was not always the case that users had watched movies both before and after the date-time threshold; this led to the inability to create a Sample and Absent set for several users. Ultimately the experiments were conducted for a limited set of users \(
\{\text{\textit{u}}_1, \text{\textit{u}}_2, ... \text{\textit{u}}_{n-1}, \text{\textit{u}}_n\}
\) with a criterion being each user must have ‘watched at least one movie before and after the date-time threshold’.

Other than selecting the users in the experiments, an additional decision had to
be made on what movies should be candidates for (predicted) membership in a user’s Absent set. The overall BT TV dataset contained 2,122 unique movies but does not state specifically what movies were available for viewing at any point in time. As the goal is to predict the set of movies in the Absent set of users (which are movies viewed after the date-time threshold), the set of movies which served as candidates for prediction \( I_A = \{i_1, i_2, \ldots, i_{m-1}, i_m\} \) were any movie viewed by any user after the date-time threshold. Movies having being viewed after this date meant that the movie would have been available for viewing.

From the set \( U \) of users ultimately utilised, for each user \( u_i \) their Sample set is converted to a formal context from which an FCA lattice is created representing the user’s profile. This lattice-profile is then pruned with the Collapse Index concept relevancy measure. The revised user profile is then compared to all other user profiles from the set \( U \), a process which determines \( u_i \)’s set of Nearest Neighbours (NN). User \( u_i \)’s similarities with the members of NN are utilised in calculating a predicted rating \( (p_{u,i}) \) for each movie in set \( I_A \). The movies in \( I_A \) are then ranked with respect to their predicted ratings.

The top 50 \( (I_{u_i,50}) \) highest ranked movies in \( I_A \) are then compared to \( u_i \)’s Absent set for overlap\(^1\). Note that the size of a user’s Absent set is restricted to being 50 or less movies in order to correspond with the 50 highest ranked movies being recommended. The cardinality of the set intersection of the top 50 ranked movies and \( u_i \)’s Absent set is obtained. The cardinality is the number of movies correctly predicted that the user \( u_i \) would have viewed. This cardinality value is obtained for all users in the set \( U \). The average cardinality value, \( \overline{rec} \), for all users in the set \( U \) is the measure of the overall success of the system, and is shown in Equation 8.1. This value is labelled the average recall; however, note that the word ‘recall’ here represents an absolute value and as such differs from the relative value that ‘recall’ is typically defined as (e.g Equation 6.5).

\[
\overline{rec} = \frac{\sum_{u_i \in U} |I_{u_i,50} \cap Absent_{u_i}|}{|U|} \tag{8.1}
\]

\(^1\)50 is chosen as a convenient number greater than the average size of users Absent set. The average size of the Absent sets in the BT TV dataset is approximately 19 movies.
8.3 Dataset Separation

A key fact in the implementation of the RS and measuring its success is that the experiments were not conducted on the entire BT TV dataset. In advance of calculations of the *average recall* value, the dataset was separated into three disjoint subsets. This was done for two reasons, the first being a desire for a training dataset to assess various aspects of the RS. Knowledge developed here would be applied to a second subset of the dataset to confirm theories.

The second reason for splitting the dataset is related to the UIC-defined concepts we seek to acquire. Dedicating a subset of the dataset to the extraction of patterns of UIC prevents circular reasoning scenarios where information learned about a movie by a user’s viewing habits is used to make predictions for a user based on that same data. At the heart of it is the preference that the patterns of UIC used to characterise movies be obtained from a separate dataset than the dataset being used to make user recommendations.

In order to accomplish this, the following steps were taken:

1. The order of the records in the BT TV logs is randomised.

2. The records are then separated into three (3) equal\(^2\) sections (subsets). Section 1 represents the *Training* dataset. Section 2 represents the *Implementation* dataset, and Section 3 represents the dataset utilised for extracting patterns of UIC.

3. Section 1 and Section 2 are both ordered with respect to date-time of movie viewing.

4. Section 1 and Section 2 are then each separated into two subsections based on the date-time threshold 31st July.

5. Section 3 is mined to extract the patterns of UIC associated with individual movies.

\(^2\)equal means that the amount of records in each subset of records is equal. *Equal* sized subsets was chosen to maximise the amount of data available in each section. This was necessary in order for each user to have sufficiently large Sample and Absent sets in Section 1 and Section 2, as well as to have sufficient data to mine UIC patterns in Section 3.
CHAPTER 8. RESULTS OF APPROACH VALIDATION

The separation of the BT TV dataset into the three sections is shown in Figure 8.1.

8.4 Movie Descriptions

UIC dataset was mined for patterns related to two aspects of UIC; these were the postal areas in which a movie was very popular and the average house price in the postal area in which the user is located. Although the smaller geographical area covered by a postal district is thought likely to be a better representation of unique socio-economic characteristics, the volume of possible postal districts in the dataset - combined with the average number of movies in a user’s profile - combined with each movie only being described by four (4) postal districts, would result in a low probability of a
postal district-defined concept being common to multiple user profiles. This affects
the ability to find similar neighbours. Postal area, although representative of a wider
geographical area with more varied socio-economic characteristics, was used since there
are fewer unique postal areas.

In the case of the postal area, the processes of associating postal areas to specific
movies were the same as the frequency→tf-idf→normalisation steps outlined in Section
6.4.2. However in this case rather than using the entire BT TV dataset to extract the
frequency matrix, only the data found in the UIC section of the BT TV dataset was
utilised. In the experiments conducted each movie was assigned 4 postal areas.

One of the issues which arises from utilising only a subsection of the BT TV dataset
for the purposes of extracting patterns of UIC, is that not all movies in the BT TV
dataset had representation in the UIC subsection of the dataset. Of the 2,122 unique
movies in the BT TV dataset only 2,112 were present in the UIC subsection. To
address this the movies absent from the UIC subsection were ignored in the user profiles
constructed in the Training or Implementation set. As an example, if a user watched
25 movies before the date-time threshold, one of which had no UIC definitions, this
movie was ignored and the user is instead considered to have only watched 24 movies.

The second UIC factor utilised in the creation of a formal context was the house
prices of users of the BT TV service. The house price of a BT TV user is arguably
reflective of their socio-economic characteristics and is expected to have a correlation
with the types of movies preferred by the BT TV user. While the BT TV dataset
does not contain the house prices, the postal district is available in the dataset. The
web resource home.co.uk (Home, 2015) was able to provide the average house price for
all postal districts in the BT TV dataset. It is this average house price for the postal
district of user and not the actual house price of the BT TV user’s house that is used
to represent the user’s house price.

The house prices provided for the BT TV dataset postal districts ranged from a
minimum of £54,565 to a maximum of £3,185,870. Using the deciles of the set of
house prices, the set of house prices for all postal districts was segmented into ten
(10) categories so that each category represents 1/10 of the number of postal districts
present. Each postal district in the BT TV dataset is assigned to one of the ten (10)
categories.
In possession of these newly created categories for house prices, and the mean house
prices for each postal district, a frequency matrix similar to that of the postal area
was created from the BT TV dataset where the amount of times a movie is watched
from ‘a geographical location where the average house price falls within a certain
category’ is included as a value in a cell in the frequency matrix. The tf-idf and cosine
normalisation processes are then applied with the final outcome being a formal context
where movies are described by the house price categories for which the movie shows
great popularity. Lattices generated from this formal context utilise movies described
by four (4) house price categories.

As a point of comparison of the contributions of postal area and house price as
movie descriptors, genre descriptions were also utilised to describe movies in the RS
experiments. Results obtained from UIC factors postal area and house price would
be compared to the non-UIC factor genre. Each movie is assigned three (3) genre
attributes. Although genre descriptions were available for all movies in the BT TV
dataset, to maintain consistency in the RS implementation, the movies for which there
were no UIC descriptions available were treated as if there were no genre descriptions
and ignored when building user profiles.

Finally, for a limited set of experiments, UIC attribute categories were combined
with the genre attribute category. Genres were combined with postal areas leading to
each movie being described by seven (7) attributes. Genres were also combined with
house price categories where similarly each movie ends up being described by seven
(7) attributes.

8.5 Experiment Results

8.5.1 Preliminaries

Part of the process of comparing user lattice profiles is the comparison of a pair of
formal concepts for similarity. The function with the responsibility for this (Equation
7.5, simplified in Equation 8.2) is composed of an object component which compares
the extent of one formal concept to the extent of the other formal concept, and an
attribute component which compares the intent of one formal concept to the intent
of the other formal concept. The extent and intent component were each assigned
weights that are the complement of the other, \( w \) and \( 1 - w \) respectively.

\[
similarity = w_{\text{Ext}} + (1 - w)_{\text{Int}}
\] (8.2)

In the experiments conducted in this chapter the weight assigned to the extent component is set to \( w = 0 \) unless otherwise stated.

### 8.5.2 Size of Nearest Neighbours Set

The first RS experiment was designed to determine empirically the number of Nearest Neighbours\(^3\) which provides the best recommendations - bearing in mind that the measure of success of the proposed RS is the average recall value of all users \( U \) being assessed.

The maximum number of movies used in a user’s Sample set was set as 50. This experiment was conducted using the genre descriptions of movies, implemented within the Training section of the BT TV dataset, and was done for \( |U| = 100 \). The average recall value for the set \( U \) was recorded as the size of the set \( NN \) was incremented. In order to ascertain whether optimum \( |NN| \) is dependent on the number of users the same experiment was carried out for \( |U| = 200 \). Results are displayed in Table 8.1.

| \( |NN| \) | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 |
|-----|----|----|----|----|----|----|----|----|----|----|
| \( U = 100 \) | 1.11 | 1.21 | 1.22 | 1.3 | 1.33 | 1.34 | 1.35 | 1.31 | 1.34 | 1.32 |
| \( U = 200 \) | 1.115 | 1.075 | 1.21 | 1.275 | 1.26 | 1.24 | 1.245 | 1.25 | 1.26 | 1.255 |

If the results are plotted when 100 users are used in the experiment (Figure 8.2a) it appears that the average recall peaks and flatlines when the number of Nearest Neighbours is in the mid-30s. A similar view of the relationship between \( |NN| \) and average recall when the \( |U| = 200 \) (Figure 8.2b) shows that, likewise, at around 30 Nearest Neighbours, the performance of the recommendations shows no more signs of improvement.

While the experiment was conducted using genre descriptions of movies there was little expectation that optimum \( |NN| \) would be very different from the results using

\(^3\)As a reminder the set of Nearest Neighbours (NN) are the users with similar movie-viewing habits as the user in question
non-genre descriptions. In light of this, for the ongoing implementations of the RS the number of Nearest Neighbours is kept at 30.

8.5.3 Number of Movies in User Sample Set

A second parameter in the RS implementation whose value would require some optimisation is the number of movies in a user’s Sample set. This is the set of movies used to create a user’s profile and is comprised of the set of movies watched before the date-time threshold in the BT TV dataset. Not unexpectedly, the number of movies in the set Sample varies depending on the user. Although for testing purposes it may be useful to keep the amount of movies in a user’s Sample set constant for all users, the desire to replicate real world scenarios where users would watch varying amounts of movies in a certain time span resulted in the decision to allow the size of user Sample sets to vary. However a maximum amount was set for the cardinality of the Sample sets.

The experiment was conducted for $|U| = 100$; the descriptor of movies was limited to being genres only; and the experiment was conducted in the Training section of the BT TV dataset. In this Training set the smallest Sample size for a user was 1 and the largest 109. The average Sample size was 27. Figure 8.3 shows the level of success of the RS when the maximum Sample size is changed.

The figure shows growth in the success level of the RS as the maximum allowable Sample size approaches 40, then a marked decline after. Eventually the success level of the RS becomes fairly constant past a maximum Sample size of 50. What is believed
however, is that performance of the RS will perform increasingly poorly as the maximum Sample size increases past a specific number (40 in this case). This is because as the number of movies in a user’s Sample set increases, the lattice built on this Sample set which represents a user’s profile, will eventually acquire too many concepts.

As an example, if a user watches an action-comedy movie, an action-comedy concept will now be in the user’s profile. If the user watches an adventure-fantasy movie, an adventure-fantasy concept will be in a user’s profile. Even if a user has very specific interests the user will eventually at some point watch other types of movie out of curiosity or coincidence. As long as that movie is watched once, considering that from the BT TV dataset there is no way of knowing whether the user liked the movie or not, a concept will be present in the user’s lattice profile based on the movie’s description. Therefore, assuming that $w = 0$ for the similarity measure\(^4\), every user’s profile, as their respective Sample size increases, converges to a common uniform profile where all possible concepts are present in the user profile.

As user’s profiles converge to a common point, the inability to distinguish the specific tastes of a user $u$ will produce Nearest Neighbours that offer no unique insight on the interests of User $u$, leading to poor recommendations. The reason why performance flatlines in Figure 8.3 rather than show consistent decline as maximum Sample increases, is that many users simply have not watched that many movies. The average size of the Sample set for users in the Training dataset was 27; increasing the maximum Sample size past 50 will not result in any real change in the success level.

\(^4\)i.e. ignoring the extents when comparing formal concept similarity across user profiles
as very few users have watched that many movies.

8.5.4 Pruning User Profiles with Collapse Index

One of the takeaways from the previous section is that all user profiles converge to a common profile as the set of movies used to create their individual profiles (Sample) increases. One way of maintaining some uniqueness to user profiles is to prune noisy concepts from a user’s profile. For this we utilise the Collapse Index measure which determines the relevance of each concept in an FCA lattice.

Each formal concept in a user’s profile lattice is assigned a relevancy value determined by the Collapse Index. A threshold value is set for the relevancy value and all formal concepts which fall below that threshold are ignored in a user’s profile. The threshold value for relevancy can be varied to empirically assess its effect on the RS success. In the first pruning experiment conducted the relevancy threshold, $CI_{\theta}$, was set based on the size of the user’s Sample set, specifically

$$CI_{\theta} = \frac{x_{\theta}}{|\text{Sample}|}$$  \hfill (8.3)

In effect, based on the premise of the Collapse Index, this translates to ‘if a concept may formal collapse due to the removal of only $x_{\theta}$ movie(s) in the user’s Sample set, this formal concept is removed from a user’s profile’.

One of the expectations of pruning the profile lattices is that pruning will become increasingly relevant as the size of user Sample sets increase. For this reason the experiments as conducted in Section 8.5.3 where the maximum size of the Sample sets of users was gradually increased was repeated. However, now the $x_{\theta}$ value is set in the relevancy threshold function to be $x_{\theta} = 1$ and $x_{\theta} = 2$. Results are plotted in graph shown in Figure 8.4.

The No Pruning plot shows the average recall value as the max Sample size increases when no pruning was done for the user lattices. These are the same values obtained in Section 8.5.3 and are shown in the current graph for purposes of comparison. Pruning 1 shows the average recall when user lattices are pruned for a Collapse Index threshold value where $x_{\theta} = 1$. Using said threshold, there was no real improvement when the maximum Sample size of users was in the range of 10 and 20. However there is a relatively substantial improvement (in comparison to No Pruning) in the
recommendations when the range was 20 - 30. However this improvement is lost as the max Sample size increases past 30. Eventually the performance becomes worse than No Pruning.

A possible explanation for the decline in performance as the maximum Sample size increases is that the threshold is too low at $x_\theta = 1$. As the set of movies used to create a user’s profile (Sample) increases, the Collapse Index threshold value in their lattice would have to be increased. When $x_\theta = 2$ (Pruning 2 in Figure 8.4) there is not only an improvement in the average recall values when ‘maximum Sample size is greater than 40’ over Pruning 1, but the average recall values are also better than when there is no pruning.

To confirm, for the fourteen (14) maximum Sample sizes used in the pruning tests, a 1-tailed t-test was done between the average recall values of the non-pruned profiles and the pruned profiles where $x_\theta = 2$. This produced a p-value less than 0.05. This indicates the usage of the Collapse Index to prune profile lattices can indeed improve recommendations. Note that the threshold for concepts pruned should be proportional to the number of movies from which the user’s lattice is built.

Nevertheless, from Figure 8.4 it appears that regardless of pruning with $x_\theta = 1$ or $x_\theta = 2$ the best results are obtained when the maximum size of user Sample sets is 40. As a final component of this experiment, the maximum Sample size was set at 40 and all variables were held constant while $x_\theta$ was set at 3 and then 4. This resulted in an average recall value of 1.16 and 1.05 respectively, exhibiting a decline in performance of the RS. What this shows is that the concept relevancy threshold has to achieve a
CHAPTER 8. RESULTS OF APPROACH VALIDATION

balance between too much or insufficient pruning. In this case \( x_\theta = 2 \) seems to be the ideal threshold.

Most importantly, the adjustment of the Collapse Index threshold to allow or remove concepts based on their relevancy is a gauge of the effectiveness of the Collapse Index as a way of rating user interest in item attributes.

8.5.5 Movie and Attribute Contributions

In this section an investigation is undertaken on the success of the RS if the weights \( w \) and its complement \( 1 - w \) are adjusted to include some contribution of the extent component of Equation 7.5. Using 100 users, genre descriptions, \( |NN| = 30 \), Collapse Index threshold \( x_\theta = 2 \), and maximum Sample size of 40, changes in average recall are observed as the weights \( w \) and \( 1 - w \) for the extent and intent components of the formal concept similarity measure (simplified in Equation 8.2) are adjusted.

As \( w \) is the coefficient of the extent similarity, \( w \) being 0 means that the extent (set of movies) is completely ignored when calculating the similarity between formal concepts, and thus the concept similarity is exclusively based on the Intent (genre descriptions). Incrementing \( w \) leads to a greater emphasis being placed on extent similarity and less on attribute similarity. The culmination of the incrementing of \( w \) is the point when similarity depends solely on the extent and not the movie attributes.

In Figure 8.5 where results are shown, as greater weight is assigned to the extent the performance of the RS declines slightly. A best fit linear equation produces a
negative gradient. This is not sufficient evidence to say that including the extent when assessing formal concept similarity across user lattice profiles (i.e. $0 < w \leq 1$) will not contribute to and/or benefit making recommendations. It is believed that as Sample sizes increase the use of the extent will become increasingly important. However the desire to limit several parameters involved in making recommendations led to $w$ being 0 for the RS experiments.

### 8.5.6 Effect of UIC Descriptions

One of the goals related to the usage of the BT TV dataset was to determine whether patterns of UIC may produce useful concepts. Having determined values for parameters that are thought to optimise the performance of the RS, attention is now turned to the usage of the lattices of UIC-defined concepts in user profiles rather than the genre-based concepts used in the previous RS experiments. Here movie descriptions are based on either the postal area or house price category of Section 8.4. Also included are the attribute category combinations genres-combined-with-postal-area, or genre-combined-with-houseprice-category. The experiment was conducted in the Training section of the BT TV dataset for 100 users, $\max(|\text{Sample}|) = 40$, $|\text{NN}| = 30$, $x_\theta = 2$, and the weight ($w$) for formal concept similarity being 0. Results are shown in Table 8.2.

**Table 8.2: Average Recall for UIC-derived Lattice Profiles**

<table>
<thead>
<tr>
<th>Description</th>
<th>Rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>genre</td>
<td>1.41</td>
</tr>
<tr>
<td>postal area</td>
<td>1.11</td>
</tr>
<tr>
<td>house price</td>
<td>1.21</td>
</tr>
<tr>
<td>genre-area</td>
<td>1.31</td>
</tr>
<tr>
<td>genre-houseprice</td>
<td>1.09</td>
</tr>
</tbody>
</table>

One notices that where movies are defined solely by genre the RS achieves the best results. However in previous experiments the values for the parameters thought to obtain best recommendations were determined based on the use of genre descriptions. It may be possible that these optimised parameter values differ depending on the attribute category being used to describe movies. With this in mind earlier experiments were repeated for the postal area and the house price descriptors to obtain optimum
values for $x_\theta$ and maximum Sample size (Table 8.3). It was not felt that the optimum $|NN|$ would vary with respect to movie descriptor, therefore $|NN|$ was held constant at 30. The combinations genre-area and genre-houseprice were also not optimised.

Table 8.3: Optimum Values of $x_\theta$ and Maximum Sample Size

<table>
<thead>
<tr>
<th>Genre</th>
<th>$x_\theta$</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>postal area</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>house price</td>
<td>1</td>
<td>55</td>
</tr>
</tbody>
</table>

The re-calculation of RS average recall values using these optimised parameters leads to the results shown in Table 8.4

Table 8.4: Average Recall for UIC-derived Lattice Profiles: Optimised

<table>
<thead>
<tr>
<th>Genre</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>genre</td>
<td>1.41</td>
</tr>
<tr>
<td>postal area</td>
<td>1.22</td>
</tr>
<tr>
<td>house price</td>
<td>1.2</td>
</tr>
</tbody>
</table>

In Table 8.4 improvement is shown for postal area after having optimised the parameters. However house price shows a decline, albeit a slight decline. The decline (or lack of improvement) could arguably be due to expected variances in results with so many factors in play. One must bear in mind that the values for parameters in the previous experiments that were found to achieve best results are not ‘final’ but are part of the heuristic processes of improving the RS. While one may optimise a parameter as other parameters are held constant (as was mostly done here), when multiple parameters are allowed to change the values which provide for best performance for the parameters may likely be different\footnote{This is referring to being different from when the optimising parameter values were obtained by keeping all other parameters constant.} as each parameter may affect others. More sophisticated optimisations routines could be done to find the best combination of $\max(|Sample|)$, $|NN|$, $x_\theta$, and $w$, however this is outside of the current scope of the current work.

Ultimately however, for values of the parameters that were so far found to give best results, the success levels for both postal area and house price are a bit below genres. This however does not signify that the UIC-defined concepts are not useful in defining movie concepts (or that non-UIC concepts are useful). In order to show
that UIC links do indeed create useful concepts, a comparison of the results against some baseline success level is required. As the motivation of the RS implementation is primarily to show the usefulness of the Collapse Index and UIC concepts more so than being an improvement over current state of the art RS solutions, the results of UIC-based recommendations are compared against a scenario where 50 random movies are recommended to users rather than the 50 highest ranked movies ($I_{u,50}$) of the proposed RS.

The average recall for 100 users if random movies are recommended to users is 0.6. If for recall values used to obtain the average recall values in Table 8.4 a 1-tailed t-test is done for genre recall values versus the random recommendation recall values, the p-value is less than 0.05. Likewise if t-tests are done for postal area v. random, and house prices v. random, in each case the p-values are less than 0.05. This indicates that using the genres as a descriptor of movies in the proposed RS produces recommendations better than if the users were recommended random movies. Moreover it also indicates that describing movies with the UIC contextual data such as geographical location or house price where movies are popular also contributes to making recommendations.

### 8.5.7 RS in Implementation Set

While the previous experiments have shown that the RS was able to produce useful recommendations using genre, postal area, and house price, what is evident is that the average recall is low. The average recall is below 1.5 for all movie attribute categories, bearing in mind that recall represents the number of movies successfully predicted from a set of at most 50 movies a user has watched. However while the value appears low at first glance, there are several explanations for this.

A first contribution to average recall being low is that the previous RS experiments have used very small user sets. Although the Training dataset contains a maximum of 4,459 users who have watched movies both before and after the date-time threshold, set $U$ has been restricted to 100 users in the majority of the experiments for practical reasons. However, as the approach utilises a user $u$’s Nearest Neighbours, restricting the set of users to 100 lessens the probability of finding another user with high similarity to user $u$. If the set $U$ is expanded, retrieval of other users more similar to user $u$ should be more likely and bring about better recommendations.
In examining the previous claim, experiments are now executed within the Implementation dataset as opposed to the Training dataset as most of the necessary hypotheses have been developed in the Training set but need confirmation in an alternate dataset. The optimising parameter values previously determined are used to calculate the average recall when the RS is executed with $|U| = 100$ and $|U| = 1,000$. This is done for genre, postal area, and house price descriptors. Results are provided in Table 8.5

<table>
<thead>
<tr>
<th>parameter values</th>
<th>$\overline{REC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$</td>
</tr>
<tr>
<td>genre</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>40</td>
</tr>
<tr>
<td>postal area</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>50</td>
</tr>
<tr>
<td>house price</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>55</td>
</tr>
</tbody>
</table>

As predicted there is an increase in the average recall ($\overline{REC}$) value if the size of $U$ is increased. Comparing the results of the RS when the number of users is 100 to the results when the number of users is 1,000 - for genre, the average recall moved from 1.33 to 1.402. Similarly, for postal area, the average recall moved from 1.29 to 1.523 and for house price the average recall moved from 1.25 to 1.321.

However the increase in the average recall in each instance is not statistically significant. This was determined by comparing the average recall for $|U| = 100$ and $|U| = 1,000$ using a 2-tailed t-tests. In each case (genre, postal area, and house price) the p-value was greater than a 0.025 ($\alpha = 0.5$). Therefore, for the 1,000 users, while there was an increase in performance for the RS, the increase is not that substantial.

However, although the increases were not that significant, due to the fact that there was in improvement in recommendations in each case, one may consider further increasing the size of $U$ with the desire to achieve a statistically significant improvement. Efficiency concerns would eventually come into play with such large sets of users however.

---

$^6$As this test requires that the sets be of the same size, a random set of 100 users from the set $|U| = 1000$ were used in the t-test to compare with the users from the set where $|U|$ was originally 100.
8.6 Conclusion

Whereas Chapter 7 showed the design of the RS which incorporated the proposed semantic extraction approach which utilises the Collapse Index, this chapter showcases the results of the implementation of the proposed RS. The chapter was able to show the success of the RS by comparison of the recommendations made via the proposed RS to a user being recommended random movies. We were able to show that by defining user profiles by *formal concepts* deemed relevant by the Collapse Index one was able to successfully make recommendations of movies to users.
Chapter 9

DISCUSSION

9.1 Collapse Index

This thesis has proposed a novel measure of relevance of formal concepts and has provided details on its theoretical and empirical behaviour and usefulness in an approach to extract semantics in large datasets. From the empirical assessments of the Collapse Index we were able to determine the following: the relevancy values determined by the Collapse Index show high correlation with those of the Logarithmic Stability; the Collapse Index values for concepts present in lattices generated from different samples of a dataset are more consistent than those of the Logarithmic Stability; the Collapse Index shows as much success in retrieving relevant concepts in a noisy lattice as the Logarithmic Stability; and that the runtimes of the proposed Collapse Index algorithm are less than those of the Stability Index. These indicate that the Collapse Index is indeed a useful assessment of the relevancy of a formal concept addressing the issues of noisy and large datasets. However there are some limitations to the Collapse Index.

Firstly the Collapse Index, like the Support Value and the Stability Index is more a measure of the relevance of the intent of the formal concept rather than the complete formal concept. For the purpose of the research conducted in this thesis, this was sufficient as the emphasis was on the attribute descriptors of objects (movies) but deeper insight on the relevance of a formal concept may be ascertained if the relevance of the extent of the formal concept is also taken into consideration.

As the duality of formal concept lattices (Ganter and Wille, 2012, p.4) has facilitated the creation of an Extensional Stability Index (Roth et al., 2008b) which
determines the significance of the extent of a formal concept, so too does the duality enable the creation of a Collapse Index measure which emphasises the extent. Given a formal concept \((A, B)\) in \(L(K)\) where \(K := (G, M, I)\), and \(\{(X_i, Y_i)\}\) are upper neighbours of \((A, B)\) the Extensional Collapse Index of \((A, B)\) can be defined as

\[
\text{ci}_e(A, B) = \frac{|B| - |Y_n|}{|M|}
\] (9.1)

The (Extensional) Collapse Index is reflective of the minimum number of attribute that may be removed from formal context, which would result in the removal of any formal concept with the extent \(A\). This Extensional Collapse Index could be combined, through weights or otherwise, with the Intensional Collapse Index and its utility examined in alternative scenarios, however this is best left to be explored in further research.

A second limitation of the (Intensional) Collapse Index is that for a formal concept \((A, B)\) the number of possible relevancy values which may be assigned to the formal concept is \(|A|\). As an example, the formal concept \((\{X, Y, Z\}, \{a, b\})\) can have three (3) possible Collapse Index values due to the Collapse Index measure being based on the minimum number of objects which need to be removed in order for the formal concept to collapse. There are a total of three (3) objects in the extent, therefore there can be at most three (3) possible Collapse Index values for the formal concept.

However, in contrast, the number of possible Stability Index values of a formal concept is exponential with respect to the cardinality of the concept’s extent. If a formal context is comprised of twenty (20) objects, the maximum cardinality possible for the extent of any formal concept in its lattice is 20. This would mean that there exists \(2^{20} = 1,048,576\) possible Stability Index values for formal concepts in the lattice; as a comparison there can be at most twenty (20) possible Collapse Index values for formal concepts derived from the same formal context.

This diversity of values gives the Stability Index measure an advantage in some scenarios and types of analysis. In cases where ranking of concepts with respect to their relevance is more important than knowing the actual relevance values, the greater volume of possible relevancy values is useful in distinguishing rank. Where there are many formal concepts and few possible relevancy values, many of the formal concepts end up being tied in terms of their relevancy value, leading to a decreased ability to
have the concepts distinguished by rank. However such a high level of nuance may often be unnecessary. Where this level of detail is unnecessary the Collapse Index is well suited due to its simplicity and effectiveness.

A third limitation of the Collapse Index is that as opposed to the Support Value, if the formal concepts\footnote{Reference is being made here to the intent of the formal concept} whose relevancy values fall below a relevance threshold are removed from the FCA lattice, the lattice that remains is not necessarily ‘intact’. Intact in this case means that a lattice comprised of only the formal concepts deemed relevant may not be a complete FCA lattice with respect to the underpinnings of FCA theory.

This is not only a challenge for the Collapse Index measure but also the Stability Index which may also identify concepts of high and low relevance at various hierarchical levels of the FCA lattice. While a possible solution was mentioned in Chapter 4, there exist opportunities for further research in investigating the re-creation of an FCA lattice based on the highly relevant formal concepts - while minimising the loss of these relevant formal concepts - while minimising the creation of new formal concepts - and while maintaining a legitimate taxonomical model of the domain. This however is likely to be a non-trivial task. At the same time there is not always the need for the final hierarchy to be an FCA lattice. It may be that only the concepts themselves are needed or it may be a situation where the relevant concepts could be utilised in the construction of a non-FCA-lattice concept hierarchy.

An aspect of the Concept Index that has not been previously mentioned is that the Collapse Index, like the Support Value and the Stability Index, is an objective measure of a subjective idea. In essence the Collapse Index is assigning a numerical value to the abstract concept ‘relevance’. How accurate these relevancy measures are in capturing the relevance (of the intent) may be secondary to the question of “what exactly is relevance?”.

As a case in point, each of the main concept relevance measures has a different premise as to what constitutes ‘relevance’ of a concept. The Support Value deems relevance as being determined by ‘the number of object instances which possess the attribute set’. The Stability Index deems relevance as being determined by ‘the ratio of combinations of object instances which possess the attribute set as their maximum
9.1. COLLAPSE INDEX

shared attribute set’. Meanwhile the Collapse Index deems relevance as being determined by ‘minimum number of object instances which need to be removed from the attribute set to no longer be a concept’. Each is different, and each author would have arguments for the legitimacy or supremacy of their respective premise.

The experiments having showed that the values the different relevancy measures produce for formal concepts in a lattice exhibited high correlation, suggests that there is some shared element in what the relevancy measures are capturing, and that this shared element may be ‘what relevance is’. However there are conflicts where, for example, one measure assigns a high relevance value to a formal concept and another measure a low. Which is correct? Are they both correct?

One way of assessing the correctness of the relevancy value is the implementation of the concept relevancy measure into an application where concept relevancy matters. This was the case for the thesis where the RS was designed to incorporate the usage of concept relevancy with respect removal of noise in user profiles. However, although the Collapse Index was utilised in the recommender system, due to project constraints, the other concept relevancy measures were unfortunately not included so as to compare the various understandings of relevance inherent in each measure.

Sometimes the relevancy of a concept may be best interpreted in the eyes of the beholder. Although automation is the ultimate target when seeking the conceptualisation of a domain, human input still has its place. A useful undertaking would be a comparison of the various relevancy measures in a study, where domain experts are asked to give their opinions of the relevancy of formal concepts in an FCA lattice. This feedback could then be compared with the relevancy values determined by the three (3) relevancy measures to determine which best matches human opinion. Further work is planned which will do exactly this. Moreover, similar to the work of Wille (2005) where formal concepts were positioned with respect to the idea of ‘concepts’ in a human mind, it would be useful to undertake a similar philosophical study on the meaning of ‘concept relevance’ in the human mind.

In the end, as it relates to RQ 1, the Collapse Index has shown itself to be a successful measure of the relevance of a formal concept in an FCA lattice, satisfying the criteria related to bias and efficiency.
9.2 Semantic Extraction Approach

In addition to the development of a concept relevancy measure, an approach was developed for the measure’s inclusion in the semantic extraction process of a dataset.

Of the various steps within the approach, several were not explored to a large extent within this thesis. The RS in which the approach was implemented did not require the reconstruction of the hierarchy after low-relevance formal concepts had been removed from the lattice, therefore this step was not researched extensively. A basic solution to the reconstruction was described in Chapter 4 however the merits of the proposed solution to reconstructing the lattice was not explored in the context of the RS. A more substantial analysis of lattice reconstruction is required with respect to lattice theory and with respect to what characteristics of proposed lattice reconstruction solutions are suitable for which domains.

One of the key motivations of ontologies is the ability to share the conceptualisations of a domain; this typically requires the formalisation of the conceptualisation. While this is included as a step in the semantic extraction approach outlined in Chapter 4, no formalisation was undertaken, or necessary, when incorporating the semantic extraction approach in the RS. While the novel use of patterns of UIC as attributes (postal district, house price) in the thesis may make it difficult to map concepts developed in this research to available (movie) ontologies, the proposed approach to semantic extraction can easily be applied to research in domains with ‘typical’ attributes. In such cases formalisation may be highly useful if not necessary. The resultant concept hierarchy at the end of the process could then be used as a standalone ontology or its concepts and relations merged with other existing ontologies.

We of course note that the semantic extraction approach is not limited to the use of the Collapse Index; other concept relevancy measures may replace the Collapse Index in the semantic extraction approach. The Stability Index is an ideal candidate given its ability to identify relevant concepts at multiple levels of a lattice.

Ultimately, although not all steps in the proposed approach were explored in depth in the thesis, as it relates to RQ 2, the proposed approach to semantic extraction has shown itself to be useful by providing a clear logical path to a concept hierarchy from a dataset, as well as being successfully implemented in a recommender system.
9.3 Recommender System

As a way of validating the proposed approach to semantic extraction of large datasets, a recommender system was designed which incorporates concept relevancy by representing user profiles by the set of *formal concepts* in an FCA lattice. The formal contexts from which the profile lattices are created are derived from a large industry dataset, and these profile lattices are pruned for noise using the Collapse Index relevancy measure.

Implementation of the RS showed that the number of correct recommendations\(^2\) was low. While an earlier suggested explanation for this was the usage of a smaller pool of users in making recommendations, there are several other ways of rationalizing the low figure.

Firstly the pool of movies from which we are choosing the fifty (50) movies to recommend to the user is fairly large. In the *Implementation* dataset the number of movies which are candidates for recommendation is 1,162. Having a large pool of movies from which to choose decreases the probability of a user having watched any of the 50 movies recommended. The user may certainly prefer the 50 recommended over a set of 50 non-recommended movies if given the chance to view the combined 100, but the user having a wide pool of movies to choose from decreases the probability of the 50 predicted being the 50 viewed. Despite this however, the performance of the RS is still arguably low, and only nearly doubles the success level if random movie were recommended to the user.

A second consideration is that the RS recommends movies to the user based on what the RS predicts the user would prefer. A user not watching a movie does not necessarily indicate that the user does not like the movie. A user may not have watched a movie for a variety of reasons that do not indicate preference. Some of these include lack of awareness of the movie’s existence or, given that the movies were watched on a television service, the movie was scheduled when the user was asleep or at work, or busy. Therefore a movie being recommended by the RS but not viewed by the user does not indicate that the user would not like the movie.

Likewise, for the dataset, the user having watched a movie does not necessarily

\(^2\)i.e., of the fifty (50) movies predicted that the user would have viewed - the amount that the user did in fact view
imply that the user liked the movie. This is an especially important point as it is central to the RS extraction of implicit information on user preference. The generation of the user profile is based on the idea that the user having watched the movie indicates some level of preference for the movie - in absence of an explicit indicator. However, once again, given that the dataset represents television viewing, a user having watched a movie may have been the result of the movie being the only movie available when the user had free time, or it could be that the television is kept on in the background throughout the day despite the user sleeping or performing miscellaneous activities in or outside of the same room as the television. This builds up a volume of movies unrepresentative of a user’s interest, serving to dilute the user profile and not present a clear picture of user preference. Certainly this makes the usage of pruning techniques especially applicable, and the successful usage of the Collapse index to improve recommendations in Section 8.5.4 demonstrated this, as well as serves as a very direct response to RQ 3.

If there existed some presence of a ‘like’ or ‘dislike’ field in the BT TV dataset, enabled through the BT TV set-top box interface, this may have very well made a big difference in the success of the recommendations. If the movie viewing logs were from a pay-per-view type service or movie theatre viewings, the fact that the users would be more pro-active and invested in selecting movies based on their preference would better indicate preference and lead to more successful recommendations.

Furthermore, the word ‘user’ has been utilised to represent the viewer of the movies, but ‘user’ in the case of the BT TV television service is more representative of a ‘household’ than a single individual. The subscriber may be one individual but there are likely to be, in many households, multiple individuals who watch television. Multiple individuals utilising the same BT TV subscription would mean that varying tastes and preferences are being embedded in the logs related to that account, once again diluting the ‘user’ profile. This dilution of the profile creates many generic profiles, where although the set of Nearest Neighbours will be found, there is not anything sufficiently unique to the profiles to recommend targeted content.

Another potential avenue for improving the recommendations is the similarity measure developed for the purposes of comparing user lattice-profiles in the RS. To reduce the number of combinations of parameters in the experiments only one similarity
measure was tested. The similarity measure outlined in Section 7.3.3 is one option for comparing lattices; there may exist other similarity solutions or variances of the current similarity measure which are more effective in capturing similarities of lattice profiles.

With respect to the currently implemented similarity solution, with a view to improve recommendations, one may ask several questions including: “Should variances in the number of movies watched per user be factored into the comparison of profiles?”; “Granted that user lattice profiles contain different number of concepts, should only the top-n concepts in each lattice be compared for best-match?”, “Should the similarity measure be symmetric?”, and “Are there other ways of incorporating the concept relevancy measure rather than for pruning low-relevance concepts?” While incorporating into the RS solutions that address the previous questions may not guarantee improvement in the recommendations to a large degree, they would inevitably have some impact and may be worth investigating moving forward.

In addition, although a non-functional aspect, the similarity measure utilised in the implementation of the RS is fairly inefficient in its calculations of similarity. This has hampered implementation of experiments for large sets of users and/or users with large Sample$^3$ sets. Creation of a more efficient version of the current similarity solution will be a necessity moving forward if the similarity measure in this thesis is to be utilised in large datasets.

Finally the implementations of the RS mostly considered one attribute category of a movie at a time. e.g. genre, house price etc. This is not likely to be enough to understand the appeal of a movie to an individual. The winners of the Netflix prize in (Bell and Koren, 2007, p.75) state that

...peoples’ taste for movies is a very complicated process. A particular user’s rating of a movie might be affected by any countless factors ranging from the general - e.g., genre - to quite specific attributes such as the actors, the setting, and the style of background music.

To produce better results would require incorporation of multiple movie attribute categories into the description of movies used in the construction of user profiles.

ootnote{Sample we recall is the set of movies used to construct a user’s profile lattice}
While this was not extensively explored in the research of this thesis, the experiment in Section 8.5.6 featured scenarios where multiple aspects of the movie were combined (genre-house price, genre-postal area). However the performance declined when attributes were combined. The reason for this may be specific to the attribute categories being combined however it more suggests that improvement of the RS success levels may not be achieved simply by the addition of other attribute categories to a movie’s description, but may require additional research in how to best include compounded attribute categories of a movie into the proposed RS.

9.4 Linking Objects By UIC

In Chapter 4 where the approach to semantic extraction was outlined, one of the steps discussed was the selection of attributes that serve as candidates for describing object instances in the domain. There it was mentioned that by considering specific ways of linking objects this may inform the selection of the appropriate attribute sets to describe the objects. This idea was incorporated and tested in the recommender system which utilises the proposed semantic extraction approach. The results for these experiments are described in Section 8.

Specific to the BT TV dataset the desire was to investigate the objects (movies) being linked by patterns of UIC. This was investigated in the context of the RS where user profiles were composed of UIC-defined concepts. Results showed that the UIC factors postal area and house price do create useful concepts given that the representation of user profiles by concepts defined by UIC factors, does in fact lead to useful movie recommendations.

Despite the low average recall values of the RS in general, the exclusive use of UIC factors in defining movies still produced statistically significant recall values over random recommendation of movies. Therefore the patterns revealed over time on the characteristics of users who interact with content, including the user’s geographical location, and the average house price in their postal district, do provide some semantic insight on the content the user is consuming.

House price was chosen as it was felt that this would reflect the socio-economic realities of the viewer and that these characteristics may correlate with the types of
movies users prefer. The successful recommendation of movies using movies exclusively described by the house price attributes makes a strong case that these socio-economic factors do correspond with characteristics of movies.

Something of note is that, for the house price related experiments conducted, there were ten (10) house price categories. The price range for each house price category was determined based on having an equal amount of postal districts in each house price category. However it may be argued that deviations in socio economic conditions does not progress that uniformly. Studies such as (Savage et al., 2013, p.230) identify the set of latent social classes that exist in the UK and the average house price for these classes. The percentage of the UK population that each class constitutes is not even. It is expected that using these more formal boundaries for social classes would serve better in finding unique associations between social class and movie traits.

Postal area was utilised as a reluctant replacement to postal district. Postal district was originally envisioned as an aspect of context which would be reflected in a user’s choice of programming. Not only would income levels be reflected in a user’s postal district but the high granularity of postal districts would better capture nuances in social behaviour and demographics. However despite this, postal area did show ability to define movies as the RS results for the exclusive usage of postal areas to define movies did perform better than if users were recommended random movies.

Of the three (3) components of UIC (interaction, user characteristics, and context) user characteristics (postal area, house price) was the main component investigated. A debate may be had on whether the user location is a ‘description of the user’ or a ‘description of the user context’ given that the user for the television service is always in a fixed geographical location when watching movies but at the same time the user’s location may influence the utility he/she derives from the movie. In the grand scheme of things however this question is not that significant; patterns emergent from both factors are due to users interacting with the movies - which is the key aspect of UIC.

9.5 General Limitations

Despite the successes achieved in the thesis, there existed several limitations to the study. The first of these is that the experiments were conducted on the dataset of one
domain. Effort was made to diversify the formal contexts by using varied attribute categories such as postal area, genre, house price, and postal district to describe movies, since each would produce FCA lattices of various shapes and complexity. However the domain is still one of movies. Being able to apply the Collapse Index as well as the RS in another domain and produce similar results would lend more credence to several theories developed in the thesis.

Another limitation is that by separating the BT TV dataset into three subsections this meant that the number of movies watched per user in the Training or Implementation sets declined. Smaller user Sample sets used to build profiles lessen the ability to test the effectiveness of pruning via Collapse Index as pruning would become increasingly relevant as more concepts are added to the user lattice. It would have been useful to have a larger dataset where users on average had watched more movies.

One of the arguments made when discussing the success of the UIC factors in the RS is that the average recall is low, at least as an absolute value. In response it was mentioned that the user not having watched the movies does not indicate that the user would not have liked the movies. The user having not watched the movie could be due to a variety of reasons not related to their like or dislike of said movie. Ideally to better confirm the success of the RS a dataset could be employed where both UIC data and explicit user ratings are available. The predicted ratings based on these UIC concepts could then be compared to the actual rating the users have provided for movies.

The final limitation is that due to inefficiencies inherent in the RS design and those brought about in the implementation, processing large sets of users had extremely lengthy runtimes. This limited experimentation on these large user sets. While access was granted to the university’s more powerful computing service in the latter moments of the research, there were no immediate gains in performance, likely suggestive of a need to customise code to take advantage of multiple processors. However at that point there would not have been enough time to accomplish this.
Chapter 10

CONCLUSION AND FUTURE WORK

10.1 Summary

The main contribution of this thesis is that of a measure which assigns relevancy values to the formal concepts in an FCA lattice. This relevancy measure, labelled the Collapse Index, defines the relevance of a formal concept as being proportional to the minimum number of object instances which need to be removed from the formal context in order for the formal concept in question to collapse. In the thesis a formal definition of the Collapse Index measure was provided along with its mathematical derivation.

Mathematical support was also included that outlines how several aspects of the presence of object instances in the formal context affect the relevancy, as per the Collapse Index, of a formal concept. The most notable of these being a.) a formal concept will eventually lose relevance if there are an increasing amount of object instances in the extent of one of the formal concept’s lower neighbours, and b.) a formal concept achieves its highest levels of relevance if the distribution of object instances across the lower neighbours is equitable or near equitable.

In addition, empirical assessment was also done of the Collapse Index measure through a series of experiments which compared the performance of the Collapse Index to the dominant existing concept relevancy measures: the Support Value and the Stability Index. However the Logarithmic Stability, which is derived from the Stability
Index, served as a substitute to the Stability Index within most of the experiments.

The experiments on formal contexts derived from the case study dataset showed that a.) if the Collapse Index is used to calculate the relevancy values of all concepts in an FCA lattice, the values it produces show very high correlation with the values if gold-standard-derived Logarithmic Stability was the measure used b.) the Collapse Index is comparable in performance to the Logarithmic Stability in retrieving relevant concepts in a noisy lattice and better than the Support Value c.) the Collapse Index values of a formal concept taken from a lattice generated from a subset of a dataset can better explain the variances of relevancy values of their equivalent formal concepts in a lattice taken from another subset and d.) the Collapse Index produces lower runtimes and accesses, on average, each formal concept in the lattice a fewer number of times than the Stability Index when calculating the relevancy values for all formal concepts in an FCA lattice.

The second contribution of the thesis is that of an approach to semantic extraction from a large dataset where FCA is the mechanism used to identify concepts within the dataset and the Collapse Index is utilised in selecting the relevant concepts. The proposed approach begins with the pre-processing of the raw dataset and ends with the creation of a final concept hierarchy which contains only the most relevant formal concepts of the domain.

In light of the available case study dataset a recommender system solution was developed as a means of demonstrating the validity of the proposed approach to semantic extraction. Unique to the recommender system is the representation of individual user profiles being that of the set of formal concepts from an FCA lattice. Each lattice was generated from a formal context comprised of the movies which a user viewed and the attribute descriptions of the respective movies.

The experiments conducted demonstrated that in using the Collapse Index to prune undesirable concepts from a user’s profile, improvements in recommendations can be achieved when thresholds for concept relevance are set relative to the number of movies in a user’s profile. Although reconstructing the concept hierarchy from the final set of relevant concepts was unnecessary in the context of the recommender system, the recommender system’s utilisation of the other steps in the semantic extraction approach serve as an effective validation of the approach.
The final contribution of the thesis is the design of the recommender system itself - which is centred around user profiles being represented as a set of formal concepts. The thesis has shown the success of this design and has importantly shown opportunities for its improvement in future research.

10.2 Further Work

In conducting the research a variety of unanswered questions arose creating opportunities for additional research. The main options considered are:

- **Recommender System Research**: The Collapse Index in the recommender system was utilised as a pruning mechanism of the user lattice-profiles. However much of the value of the Collapse Index was lost as it simply separated the concepts of the profile lattice into ‘relevant’ and ‘not relevant’. A more nuanced view of a user profile is likely to be achieved if the relevancy value assigned to each formal concept is used in the calculation of similarity between profiles. This would constitute the continuation of recommender system research.

  Note however that one of the main drawbacks of the present research was that the dataset did not provide any explicit information on user preference. Further research would utilise a dataset where there exists some indicator of user preference; this is in order to test the levels of success achieved in a more objective way.

- **Concept Relevancy Research**: In the thesis it was mentioned that, as opposed to the Support Value, the formal concepts which remain after the removal of formal concepts identified by the Collapse Index as being of low relevance, do not necessarily constitute a complete FCA lattice. Although a basic solution was offered in the thesis this was not extensively explored. Further research will be undertaken to determine more effective ways to regenerate an FCA lattice or at least a concept hierarchy from the remaining relevant formal concepts.

  Additional research will also be done on the validation of the Collapse Index as a concept relevancy measure by applying the Collapse Index measure to similar domains and datasets that have previously been subject to analysis by the other
concept relevancy measures. This includes noise recognition and efficiency tests. Part of this additional validation of the Collapse Index would be having experts decide on the validity of concepts identified through the Collapse Index.

A final research option relates to the optimisation of the Collapse Index algorithm. The current algorithm is a straightforward way of performing the calculation, however there exists opportunities for improving said algorithm. This is partly driven by the Collapse Index of a formal concept being dependent on information obtained about lower neighbours of the formal concept, therefore if one were to iterate upwards through the lattice from the infimum, one may reduce the number of times each formal concept is accessed. Moreover one may pursue embedding the calculation of the Collapse Index in the lattice-generating algorithm itself, allowing the direct transition from a formal context to a lattice where each formal concept is already assigned a relevancy value.

10.3 Conclusion

This thesis has produced three important contributions to knowledge; these being a.) the Collapse Index - a measure of the relevance of a formal concept in an FCA lattice b.) an approach for extracting semantics from a dataset where the Collapse Index is employed to determine the relevant concepts of a final concept hierarchy, and c.) a recommender system where user profiles are composed of a set of formal concepts deemed relevant by a concept relevancy measure such as the Collapse Index.

These contributions serve as direct responses to the three research questions. RQ 1 asks for an efficient measure of the relevance of a formal concept which is not overly biased to formal concepts with high object instance support; these criteria are met by the Collapse Index. The second research question (RQ 2) asks for an approach for inclusion of this concept relevancy measure in extracting semantics from a dataset. This approach was outlined in the thesis and validation of the proposed approach was provided by its successful inclusion in a recommender system. The final research question (RQ 3) asks for a recommender system designed to incorporate concept relevancy of FCA concepts. This question was answered through a recommender system introduced in the thesis where a key aspect of its design is the representation of a user
profile as a set of formal concepts. This allows for the usage of the Collapse Index concept relevancy measure as well as the application of the thesis’ proposed approach to semantic extraction within the implementation of the recommender system.
Bibliography


Appendix A

Acronyms

<table>
<thead>
<tr>
<th>CB</th>
<th>Content Based Filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF</td>
<td>Collaborative Filtering</td>
</tr>
<tr>
<td>CI</td>
<td>Collapse Index</td>
</tr>
<tr>
<td>FCA</td>
<td>Formal Concept Analysis</td>
</tr>
<tr>
<td>LStab</td>
<td>Logarithmic Stability</td>
</tr>
<tr>
<td>NN</td>
<td>Nearest Neighbours</td>
</tr>
<tr>
<td>RS</td>
<td>Recommender System</td>
</tr>
<tr>
<td>RQ</td>
<td>Research Question</td>
</tr>
<tr>
<td>SI</td>
<td>Stability Index</td>
</tr>
<tr>
<td>UIC</td>
<td>User Interaction and Context</td>
</tr>
</tbody>
</table>
Appendix B

Key Symbols Used In Proofs

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A, B)</td>
<td>Used as representation of the formal concept of interest.</td>
</tr>
<tr>
<td>(C_i, D_i)</td>
<td>Used as the representation of a lower neighbour of the formal concept (A, B).</td>
</tr>
<tr>
<td>(A, B)</td>
<td>Formal concept with B as its intent after changes have been made to the ‘original formal context which had produced the formal concept (A, B)’.</td>
</tr>
<tr>
<td>O_B</td>
<td>The set of objects where the complete attribute set of each object is the intent of the formal concept (A, B).</td>
</tr>
<tr>
<td>x^a</td>
<td>The attribute set of object x.</td>
</tr>
<tr>
<td>K := (G, M, I)</td>
<td>Formal context representing the object set G, the attributes set M, and the set of binary relations I.</td>
</tr>
<tr>
<td>L(K)</td>
<td>FCA lattice generated from formal context K.</td>
</tr>
<tr>
<td>K_{x^+}</td>
<td>Formal context where the object x and its object-attribute relations of object have been added to the formal context K.</td>
</tr>
<tr>
<td>K_{x^+}</td>
<td>Formal context where the object x and its object-attribute relations of object have been removed from the formal context K.</td>
</tr>
</tbody>
</table>
Appendix C

Formal Concept Analysis

Technology Choices

Given the importance of Formal Concept Analysis in this thesis, so too were several FCA-related software. These were required to accomplish the critical tasks of a.) creating FCA lattices from formal contexts b.) facilitating traversal of the lattices and c.) providing a visual of lattices generated. The first FCA tool utilised was that of Concept Explorer.

C.1 Concept Explorer

Concept Explorer (Yevtushenko, 2000) or ConExp is an open source FCA software application available at SourceForge.net. The software is under the administration of Sergey Yevtushenko and the most current version of the software is 1.3. ConExp provides interfaces for users to input formal contexts then outputs the resultant lattice. Throughout the thesis, unless stated otherwise, lattice diagrams are presented in the format of ConExp lattices.

Central to ConExp’s depictions of FCA lattices are:

- reduced labelling. As is evident in Figure 2.1, the labelling of each formal concept’s intent and extent may produce a fairly cluttered lattice. Reduced labelling is a less busy way of labelling lattices without loss of information on the extents and intents. Attribute labels when utilising reduced labelling are only assigned to the most general formal concept which has in its intent that specific attribute;
object label are only assigned to the most specific formal concept which has in its extent that specific object. The intents and extents are interpreted using the notion that in FCA lattices attributes accumulate downwards while objects accumulate upwards.

- where nodes are large, this indicates there exists at least one object instance whose complete set of attributes is equal to the intent of the formal concept the node represents.

- where the upper semicircle of a node is gray/blue, this indicates that an attribute label is assigned to the formal concept.

- where the lower semicircle of a node is black, this indicates that an object label is assigned to the formal concept.

- Object labels are expressed in uppercase

- Attribute labels are expressed in lowercase

The lattice of Figure 2.1 is re-imagined in the format of ConExp in Figure C.1. The formal concept with the labels ‘COW’ and ‘milk’, as per reduced labeling, would have as its intent \{milk, warmblooded\} as it inherits the attributes of all of its super-concepts. Likewise it has as its extent \{COW, DOG, FRUITBAT\} as it inherits the extents of its subconcepts.

C.2 Java Libraries (Colibri)

Beyond visualisation of lattices Colibri-Java (Götzmann, 2015) libraries were incorporated into Eclipse IDE to provide more computational access to components of an FCA lattice as well as customisability. Colibri-Java is an FCA JAVA library developed by Daniel Götzmann in conjunction with Christian Lindig. It exists as an open source project on Google Code under GNU GPLicense v2. Important functionalities enabled through Colibri-Java libraries are the ability to:

- input a formal context (in the form of binary relation pairs)

---

1Not specific to ConExp however this standard is utilised for purposes of clarity

2Not specific to ConExp however this standard is utilised for purposes of clarity
• create a lattice based on the relations

• traverse the lattice in varying manners based on a set of iterators provided (top-down, bottom-up, etc.).

• find specific concepts (infimum, supremum, lower neighbours of specified concepts)