Mining Functionally Novel Association Rules: A Domain-Driven Data Mining Approach using Biomedical Literature and Ontology

Yakub Sebastian

School of Engineering, Computing and Science
Swinburne University of Technology Sarawak Campus
Kuching, Sarawak
Malaysia

A thesis submitted for the degree of
Masters of Science by Research

2012
Dedicated to my God, Jesus Christ

and to my parents Faisal Kusnan and Yuliana Ririana, my lovely sisters Hanna Rosella Fitriana and Hanni Stella Angelica, and my grandma for their unceasing prayers and unwavering love all throughout.
Abstract

Introduction. Good medical hypotheses are the key to most medical breakthroughs. Association rules produced by medical data mining systems may lead to the formulation of valid medical hypotheses. An important requirement is that these rules must be novel. However, evaluating rule novelty based on the traditional pairwise approach has always been challenging. This study introduces a new non-pairwise novelty criterion termed the functional novelty.

Objective. To develop a new knowledge discovery framework and techniques for discovering functionally novel association rules from cardiovascular data sets.

Methods. Association rules were mined from two cardiovascular data sets using the FP-Growth algorithm with sufficiently low minimum support and confidence thresholds. The rules were semantically filtered against semantic relations defined in Unified Medical Language System ontology. By applying a modified $\chi^2$-based correlation measure called $\chi^2_{lit}$ the filtered rules were then validated against Pubmed literature to determine their compliance with the existing medical domain knowledge. Only the domain knowledge-compliant rules were selected for the final rule post-processing. In the conducted experiments, a cardiologist prescribed four pairs of medical hypotheses. The functional novelty of each association rule was determined based on its likelihood in mediating these hypotheses using the $Min\chi^2$ score.

Results. KELAM, a novel domain-driven knowledge discovery framework was constructed. Two experiments were conducted with the cardiologist’s evaluation results as the gold standard. In the Experiment I, $\chi^2_{lit}$ exhibited a high recall rate for domain knowledge-compliant rules even though it suffered from a low precision. In the Experiment II, $Min\chi^2$ proved to be useful for ranking the rule functional novelty because all candidate functionally novel rules were found among the top-10 rules. One interesting result showed that KELAM suggested a potential relationship between von willebrand factor and intracardiac thrombus via the association rule diabetes mellitus $\leftrightarrow$ coronary arteriosclerosis.
Conclusion. This thesis successfully produced an effective domain-driven knowledge discovery framework for discovering the functionally novel rules by combining the domain knowledge from the medical literature, the ontology, and user-defined hypotheses. The proposed post-mining evaluation technique and measures proved to be useful in predicting candidate functionally novel rules as shown in experiments validated by a cardiologist. The outcome of this work is expected to become the first step towards medical knowledge discovery systems that can effectively aid medical researchers in rapidly testing and validating the potential medical hypotheses at the initial stage of a medical discovery endeavour.
Acknowledgements

My gratitude first goes to my supervisor, Dr. Patrick Then Hang Hui, for his guidance, inspiration, and friendship. Thanks also to my co-supervisor, Associate Professor Enn Ong for her valuable moral support.

I thank Dr. Alan Fong Yean Yip MRCP of the Cardiology Department, Sarawak General Hospital for the time and energy for supplying this study with the valuable medical expertise. Without it, this work would not have been possible.

Thanks to Mr. Valliappan Raman and Brian Loh Chung Shiong for proofreading the thesis manuscript.

I acknowledge my fellow postgraduate students Lai Chee Ping and Wendy Japutra Jap for their positive criticisms and technical inputs. Thanks to Sam Seo Wei Jye, Vong Wan Tze and Lesley Lu for their friendships.

My special gratitude goes to Evan Khoo Kay H’erh for his enduring friendship throughout many challenging times.
Declaration

I declare that this thesis contains no material that has been accepted for the award of any other degree or diploma and to the best of my knowledge contains no material previously published or written by another person except where due reference is made in the text of this thesis.

Yakub Sebastian
Publications Arising from this Thesis

Parts of this thesis have been published in the following publications:

Journal papers


Conference papers


# Contents

1 Introduction .............................................. 1
   1.1 Background of the Study ................................. 1
   1.2 Research Motives and Contributions ..................... 3
      1.2.1 Research problem .................................. 3
      1.2.2 Research objective ................................ 4
      1.2.3 Significance and contributions .................... 4
   1.3 Thesis Structure ...................................... 5

2 Literature Review ........................................ 7
   2.1 Hypothesis Generation in Medical Knowledge Discovery .... 7
      2.1.1 Observation-based method .......................... 8
      2.1.2 Literature-based method ............................ 9
      2.1.3 KDD-based method .................................. 13
      2.1.4 The integration of the literature-based and KDD-based discovery methods ........................................ 21
   2.2 Chapter Summary ...................................... 22

3 Methodology and Design ................................... 23
   3.1 Functional Novelty .................................... 23
   3.2 The Proposed Hypothesis Discovery Method .............. 24
   3.3 Notations and Definitions ............................... 26
      3.3.1 Association rules ................................. 26
      3.3.2 Estimating the domain knowledge-compliant rules: $\chi^2$-based correlation measures ...................... 27
CONTENTS

3.3.3 Functional novelty score ........................................... 31

3.4 KELAM Framework Design ............................................. 34

3.4.1 Stage 1: Semantic-based attribute discretization and mapping .................. 35

3.4.2 Stage 2: Generating the association rules with FP-Growth algorithm ........... 37

3.4.3 Stage 3: Semantic-based filtering ..................................... 38

3.4.4 Stage 4: Discovering the domain-knowledge compliant rules ..................... 41

3.4.5 Stage 5: Evaluating and ranking functionally novel rules ........................ 41

3.5 Experimental Designs .................................................. 41

3.5.1 The medical expert .................................................... 42

3.5.2 Data sets descriptions ............................................... 42

3.5.3 Design of Experiment I .............................................. 44

3.5.4 Design of Experiment II ............................................. 45

3.6 Chapter Summary ....................................................... 47

4 Implementation and Results ............................................ 48

4.1 Technological Architecture ........................................... 48

4.1.1 Data preparation ..................................................... 50

4.1.2 Data mining ......................................................... 53

4.1.3 User hypothesis ..................................................... 53

4.1.4 Rule scoring ......................................................... 54

4.1.5 Rule ranking and visualization .................................... 54

4.1.6 Literature details visualization module .................................. 54

4.1.7 Database access module .......................................... 55

4.2 Description of KELAM Prototype .................................... 55

4.2.1 KELAM prototype system ....................................... 55

4.2.2 Comparison with other prototypes ................................ 61

4.3 Experiment I ........................................................... 63

4.3.1 CHD_DB .......................................................... 63

4.3.2 ECHO_MSCT .................................................... 69

4.4 Experiment II ........................................................... 71
## Contents

4.5 **Discussions** ................................................................. 80  
4.5.1 Discussions on results obtained in Experiment I .................. 80  
4.5.2 Discussions on results obtained in Experiment II ............... 80  
4.6 **Chapter Summary** ...................................................... 81  

5 **Conclusions** ................................................................. 83  
5.1 Recapitulation ............................................................... 83  
5.2 Significance and Contributions ......................................... 85  
5.3 Limitations ................................................................. 87  
5.4 Future Works ............................................................... 89  
5.5 Conclusions ............................................................... 90  

6 **Appendix** ................................................................. 91  
6.1 RapidMiner ............................................................... 91  
6.2 Ontology ................................................................. 91  
6.2.1 Unified Medical Language System (UMLS) ..................... 91  
6.3 Pubmed Literature ....................................................... 95  
6.3.1 Entrez Utilities ....................................................... 97  
6.4 Java ................................................................. 97  

**References** ............................................................... 98
List of Figures

2.1 Observation-based hypothesis generation model ........................................ 8
2.2 Literature-based hypothesis generation model ........................................ 9
2.3 Swanson literature-based discovery model ........................................ 10
2.4 Steps in knowledge discovery in databases (taken from Fayyad et al 1996) ..... 14
2.5 KDD-based hypothesis generation model ........................................ 14

3.1 Functional novelty model ................................................................. 23
3.2 Jensen et al’s ABCD inference model ............................................... 24
3.3 Integration of KDD-based and literature-based methods .......................... 25
3.4 KELAM framework ........................................................................ 34

4.1 KELAM technological architecture ..................................................... 48
4.2 E-Utils data pipeline for real time access to Pubmed ............................ 49
4.3 Semantic map file (.csv): an example ............................................... 51
4.4 Example of .dat file ........................................................................ 52
4.5 Example of .aml file ........................................................................ 52
4.6 Data set and semantic map specification ............................................. 56
4.7 Normal mode setting ........................................................................ 56
4.8 Hypothesis term entry ....................................................................... 57
4.9 The first hypothesis term mapping ..................................................... 57
4.10 The second hypothesis term mapping .............................................. 58
4.11 Displayed rule set ........................................................................... 58
4.12 Rule quality indicators .................................................................... 58
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.13</td>
<td>Pubmed citations retrieved for rule hypercholesterolaemia ↔ hypertension</td>
<td>59</td>
</tr>
<tr>
<td>4.14</td>
<td>Full abstract of a Pubmed literature</td>
<td>59</td>
</tr>
<tr>
<td>4.15</td>
<td>Patient records corresponding to rule hypercholesterolaemia ↔ hypertension</td>
<td>60</td>
</tr>
<tr>
<td>4.16</td>
<td>Drill-down analysis of individual patient data</td>
<td>61</td>
</tr>
<tr>
<td>4.17</td>
<td>CHD_DB: item-to-concept mapping results (.csv)</td>
<td>65</td>
</tr>
<tr>
<td>4.18</td>
<td>Number of rules generated at various minsupp levels</td>
<td>66</td>
</tr>
<tr>
<td>4.19</td>
<td>CHD_DB: evaluation results</td>
<td>67</td>
</tr>
<tr>
<td>4.20</td>
<td>ECHO_MSCT: item-to-concept mapping results (.csv)</td>
<td>70</td>
</tr>
<tr>
<td>4.21</td>
<td>ECHO_MSCT evaluation results</td>
<td>71</td>
</tr>
<tr>
<td>6.1</td>
<td>An example of UMLS Metathesaurus output</td>
<td>92</td>
</tr>
<tr>
<td>6.2</td>
<td>A partial depiction of the hierarchies among the semantic types in the UMLS Semantic Network</td>
<td>92</td>
</tr>
<tr>
<td>6.3</td>
<td>A partial depiction of the hierarchies among the semantic relations in the UMLS Semantic Network</td>
<td>93</td>
</tr>
<tr>
<td>6.4</td>
<td>A partial depiction of the semantic relations among several semantic types</td>
<td>93</td>
</tr>
<tr>
<td>6.5</td>
<td>An example of Pubmed citation record</td>
<td>95</td>
</tr>
<tr>
<td>6.6</td>
<td>An example of Pubmed citation in XML format</td>
<td>96</td>
</tr>
</tbody>
</table>
List of Tables

2.1 A $2 \times 2$ contingency table ................................................. 11

3.1 A $2 \times 2$ contingency table ................................................. 27

3.2 An example of Pubmed query statements for concepts Smoking and Myocardial Infarction ................................................. 27

3.3 Five known associations provided by a medical expert. Total Pubmed count: 19,572,400 (Query date: March 8, 2010) ................................. 28

3.4 Five unknown associations provided by a medical expert. Total Pubmed count: 19,572,400 (Query date: March 8, 2010) ................................. 29

3.5 The comparison between $\chi^2_lit$ and MIM over two independent samples ................................................. 30

3.6 Mann-Whitney u-test results for $\chi^2_lit$ ........................................... 31

3.7 Mann-Whitney u-test results for MIM ........................................... 31

3.8 The comparison between LES and $\chi^2_lit$ on two medical associations. ................................................. 33

3.9 A hypothetical example of the false and true inferences ................................................. 33

3.10 An example of semantic map contents ........................................... 37

3.11 CHD_DB data attributes .................................................. 43

3.12 ECHO_MSCT data attributes .................................................. 43

4.1 Feature comparison of GEMINI and KELAM ........................................... 62

4.2 The domain knowledge usage comparison of ARROWSMITH and KELAM ................................................. 63

4.3 CHD_DB: Discretization of continuous data attributes ................................................. 64

4.4 CHD_DB: Items derived from categorical attributes ................................................. 64

4.5 CHD_DB: Medical expert’s evaluation results ................................................. 68
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6</td>
<td>CHD_DB: performance comparison between $\mu$, $\chi^2$, and $\chi^2_{1d}$</td>
<td>69</td>
</tr>
<tr>
<td>4.7</td>
<td>ECHO_MSCT: Discretization of the continuous data attributes</td>
<td>69</td>
</tr>
<tr>
<td>4.8</td>
<td>MSCT_ECHO: Items derived from categorical attributes</td>
<td>70</td>
</tr>
<tr>
<td>4.9</td>
<td>ECHO_MSCT: performance comparison between $\mu$, $\chi^2$, and $\chi^2_{1d}$</td>
<td>71</td>
</tr>
<tr>
<td>4.10</td>
<td>MSCT_ECHO: Medical expert’s evaluation results</td>
<td>72</td>
</tr>
<tr>
<td>4.11</td>
<td>Functional novelty evaluation obtained by mining CHD_DB data set</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>for smoking $\leftrightarrow$ endothelial progenitor cell levels</td>
<td></td>
</tr>
<tr>
<td>4.12</td>
<td>Functional novelty evaluation obtained by mining CHD_DB data set</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>for smoking $\leftrightarrow$ human anti murine antibody levels</td>
<td></td>
</tr>
<tr>
<td>4.13</td>
<td>Functional novelty evaluation obtained by mining ECHO_MSCT data</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>set for matrix metalloproteinase $\leftrightarrow$ intracardiac thrombus</td>
<td></td>
</tr>
<tr>
<td>4.14</td>
<td>Functional novelty evaluation obtained by mining ECHO_MSCT data</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>set for von willebrand factor $\leftrightarrow$ intracardiac thrombus</td>
<td></td>
</tr>
<tr>
<td>4.15</td>
<td>Medical expert’s final evaluation on the functional novel rules</td>
<td>78</td>
</tr>
<tr>
<td>6.1</td>
<td>A partial depiction of the UMLS Semantic Network content</td>
<td>94</td>
</tr>
</tbody>
</table>
Commonly Used Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D^3M$</td>
<td>Domain-Driven Data Mining</td>
</tr>
<tr>
<td>FP-Growth</td>
<td>Frequent Pattern Growth</td>
</tr>
<tr>
<td>ICD</td>
<td>International Classification of Diseases</td>
</tr>
<tr>
<td>KDD</td>
<td>Knowledge Discovery in Databases</td>
</tr>
<tr>
<td>KELAM</td>
<td>Knowledge Extraction via Logical Association Mining</td>
</tr>
<tr>
<td>MEDLINE</td>
<td>Medical Literature Analysis and Retrieval System Online</td>
</tr>
<tr>
<td>MeSH</td>
<td>Medical Subject Headings</td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>Systematized Nomenclature of Medicine–Clinical Terms</td>
</tr>
<tr>
<td>UMLS</td>
<td>Unified Medical Language System</td>
</tr>
</tbody>
</table>
CHAPTER 1

Introduction

1.1 Background of the Study

The proliferation of digital data in the medical domain demands not only the efficient data storage methodologies and techniques, but also the capabilities in making sense of these data for the greater uses. But the sheer amount of the accumulated data has made the analysis impossible to be performed manually by human analysts. Using the novel data mining algorithms and techniques, large clinical databases can be mined to expand the pool of global clinical knowledge to improve clinical decision making [1].

Knowledge discovery in databases (KDD) is the process of identifying valid, novel, useful, and understandable patterns in large databases [2]. Data mining is the central process of KDD that focuses on searching for patterns of interest in various representational forms. One of the most popular representational forms is the association rules [3].

Many major medical breakthroughs were preceded by the formulation of novel and potent hypotheses. The final goal of KDD is to produce testable hypotheses from data that will lead to the formulation of new knowledge in a domain. To the medical experts, the ability to extract the interesting patterns from large medical databases in the form of association rules may hold the key to uncovering new medical hypotheses.

An association rule must exceed certain interestingness thresholds to be considered as knowledge. Interestingness is the "overall measure of pattern value, combining..."
validity, novelty, usefulness, and simplicity” [2]. However, measuring the rule interestingness is difficult as it is influenced by various factors including the task, the individual, and the context [4]. Despite having been extensively studied in data mining field [5, 6, 7, 8], this problem remains highly challenging [9, 10, 4].

There are nine rule interestingness criteria [7]. Novelty is the most important interestingness criterion [11, 12, 2]. Traditionally, a rule is novel if the user “did not know it before and is not able to infer it from other known patterns” [7]. A survey conducted by Geng and Hamilton [7] reported three categories of techniques for discovering novel rules: (a) the post-mining filtering technique [13, 12], (b) the interactive incremental filtering technique [14]; and (c) the in-mining constraint-based filtering technique [15]. The careful examinations of these existing techniques reveal that they normally measure the rule novelty using the pairwise approach. To be interesting, a novel rule ought to consist of the previously unknown and unexpected combinations of the rule antecedent and consequent with respect to the existing knowledge. In this work, this paradigm is referred to as the *pairwise novelty* paradigm. However, the paradigm encounters two major problems:

1. **The rare item problem.**
   Within the support-confidence framework, it is generally assumed that the novel association rules will normally reside with a very low support [13]. But to mine association rules using a low minimum support (minsupp) threshold often generates a large number of other uninteresting rules. On the other hand, the optimal support and confidence threshold values remain unknown [16].

2. **User acceptance problem.**
   Due to the lack of supporting background knowledge, it is difficult to validate rules that possess uncommon item combinations. Moreover, performing validations via experiments is expensive. However, unless validated, the rules are hardly applicable for any clinical usage. In medical informatics, the user acceptance has always been a major issue and evidence has shown that the compliance with the existing domain knowledge helped to increase the acceptance of the learned models by the medical experts [17].

There has been no compelling evidence which suggests the pairwise novelty as the only paradigm to define the rule novelty. Instead, Han et al. [18] recommended
Chapter 1: Introduction

that the future pattern analysis should include the contextual analysis of the patterns and not be limited to the analysis of pattern composition only. Jaroszewicz et al. [19] also suggested that rule evaluation should consider the full joint probability on data. Consequently, one may ask, ‘Is it possible for a common rule (i.e. a rule having a well-known item composition) to be novel to the user at the same time? If yes, how can we determine it? Can the rule novelty be eventually evaluated without relying on the pairwise novelty paradigm?’ These questions are important not only because they demand a rethinking of the existing notion of the rule novelty but also because their answers may provide the means to help overcome the current limitations of the traditional pairwise novelty approach.

1.2 Research Motives and Contributions

This study hypothesizes that the rule novelty can be evaluated as the functional novelty. A rule is functionally novel if the rule mediates a previously unknown relationship between two user hypotheses by the way of chaining several medical associations.

The new approach is considered non-pairwise because the association rule’s novelty no longer depends on its combinations of the rule’s items. The non-pairwise approach has the advantages over the traditional pairwise novelty approach. Firstly, a data mining algorithm does not have to focus on mining the association rules with the uncommon item combinations. Rules that have the common item combinations can be considered functionally novel. As a result, the algorithm does not have to operate at a very low minimum support level. This, in turn, will reduce the severity of the rare item problem. Secondly, functionally novel rules are more likely to be acceptable by the medical experts because the rules may include the well-known item combinations.

1.2.1 Research problem

The main research problem is how to come up with a feasible KDD methodology and measures for discovering the functionally novel rules.
CHAPTER 1: INTRODUCTION

1.2.2 Research objective

The objective of this work is to produce an effective knowledge discovery methodology and application that is capable of mining a small set of the functionally novel rules from a medical data set based on a particular medical hypothesis provided a medical expert. The application shall be capable of automatically incorporating the medical domain knowledge from the literature, ontology, and medical expert’s hypotheses at the rule evaluation stage. The findings obtained using this application shall serve as a proof of concept for the usefulness of functional novelty in the rule interestingness evaluation process.

1.2.3 Significance and contributions

Several key research trends underline the significance of this work. Biology is moving "from an era of data-collection to one of hypothesis-driven research" [20]. Conceptual biology advocates the review of the existing pieces of knowledge in a concept-driven manner and linking them into a testable chain of hypothesis [21]. Significant scientific discoveries often come through forming connections between knowledge in various fields. For instance, Shahaf and Guestrin [22] recently introduced a data mining technique for analyzing the connections between two distinct events reported in news articles.

In KDD, the proponents of domain-driven data mining (D³M) have advocated a metasynthesis of the ubiquitous elements of the domain intelligence to mine actionable knowledge [23, 24]. To discover the actionable knowledge, future KDD systems should integrate information retrieval (IR), data mining methods and knowledge inference methods [10, 25]. One of the primary challenges of the rule interestingness research is to blend various aspects of interestingness into a unified interestingness measure [6, 26].

This work is significant as it aims at discovering a unified functional novelty evaluation framework that is capable of combining the elements of the subjective interestingness and the objective interestingness in the form of a simple user hypothesis (subjective) and the utilisation of correlation measures (objective). The new framework metasynthesizes the unstructured and structured domain knowledge from the literature and the ontology as the basis for rule evaluation. The implemented
CHAPTER 1: INTRODUCTION

application also effectively integrates the IR and data mining methods by remotely querying information from the relevant literature at the post-mining stage. Finally, the work facilitates the practice of conceptual biology by allowing the user to investigate, test and validate medical hypotheses.

There are two main contributions of this work:

1. *Functional novelty*, a new non-pairwise criterion for mining novel rules. No prior research has proposed a similar rule interesting criterion.

2. *KELAM* (Knowledge Extraction via Logical Association Mining), data mining methodology and framework that synchronize the three major sources of domain knowledge: literature, ontology, and medical expert user. This allows the user of the system to interpret the significance and validity of each association rule within a highly context-sensitive environment.

The scope of this work is limited to the interestingness evaluation of 2-item association rules. The possibility of applying the available IR techniques for expanding literature queries using biomedical thesaurus is considered and excluded from this study for the lack of substantial evidence in the IR field that such method would improve the information retrieval performance.

1.3 Thesis Structure

This thesis is structured as follows:

Chapter 2 Literature Review reviews the literature. The review is presented and discussed from the perspective of hypothesis generation, which is the final goal of the exploratory knowledge discovery systems.

Chapter 3 Methodology and Design presents the proposed KDD methodology, measures, and experimental designs.

Chapter 4 Implementation and Results describes the implemented KDD system. It subsequently describes and discusses the experimental results.

Chapter 5 Conclusion concludes this thesis. Limitations of the work and recommended areas for future research are highlighted.
Chapter 1: Introduction

Chapter 6 Appendix describes in detail the software components used by the proposed solution.
Chapter 2

Literature Review

2.1 Hypothesis Generation in Medical Knowledge Discovery

Hypothesis generation plays the central role in the overall process of biomedical knowledge discoveries [20]. The formulation of high quality hypotheses led to many important medical discoveries [27]. Louis Pasteur’s germ theory, for example, was the result of the initial hypothesis that fermentation and putrefaction might have been caused by microorganisms instead of by spontaneous generation [28]. Marshall and Warren, winners of Nobel Prize in Medicine 2005, based their medical discovery on the hypothesis that bacteria *helicobacter pylori* might be the cause of peptic ulcers [29].

Formulating medical hypotheses involves the "novel juxtaposition of concepts not previously connected" [27] by logically assembling the independent pieces of known relationships in order to make an inference about potentially new knowledge. When proven, the hypotheses become the new medical discoveries. Marshall and Warren’s hypothesis, for example, was the result of logically assembling three pieces of known but previously-disconnected information: (1) Pasteur’s theory that bacteria could cause disease; (2) frequent observations that *helicobacter pylori* co-occurred with gastritis; and (3) gastritis was often symptomatic of peptic ulcers.

There are three major methodologies for generating medical hypotheses: the observation-based method, the literature-based method and the KDD-based method. The first relies on observing frequent occurrences in nature, the second focuses on literature analysis, whereas the last method focuses on database analysis. Discussions on the first two methods shall provide the context for more thorough discussions on the third method which is the main research focus of this study.
2.1.1 Observation-based method

In the traditional observation-driven model, medical hypotheses are formulated by observing the interesting or novel occurrences in nature or in patients being treated. Robert Koch postulated that in order to identify the cause of infectious diseases, parasitic organisms must be observed constantly in characteristic form and arrangement in the diseased tissue \[30\]. For example, Pasteur’s germ hypothesis emerged from the frequent observations in fermenting solution which contained optically active compounds which was known to be the product of living organisms \[31\]. Similarly, Marshall and Warren’s discovery originated from Warren’s frequent observation that spiral bacteria were constantly present in the stomachs of patients suffering from gastritis.

Figure 2.1: Observation-based hypothesis generation model

Figure 2.1 shows the general model of observation-based method. Novel medical observations lead to the formulation of the potential medical hypotheses. When sufficiently validated through experiments, the hypotheses become new medical discoveries.

In many cases, however, the observation-based method takes considerably long time to concur on a worthwhile hypothesis because it relies on the accumulation of sufficient amount of the right data from the right subjects. Medical discovery process can be accelerated if the hypotheses are generated, tested, and validated by automatically connecting previously disconnected pieces of the existing medical knowledge in the literature.
2.1.2 Literature-based method

The proliferations and the increasing accessibility to the electronic medical literature make it possible to automatically synthesize new medical knowledge by reviewing biological facts and information in a concept-driven manner and linking them into testable chains of hypotheses [21, 32]. Figure 2.2 shows the generic model of literature-based method.

Swanson [33] analyzed biomedical literature to uncover the hidden relationships between two or more previously-disconnected medical concepts. He discovered from a set of MEDLINE literatures that fish oil (A) could lower patient’s high blood viscosity (B). In a separate set of literature, he also discovered that high blood viscosity (B) was often associated with Raynaud’s Syndrome (C). As a result, he syllogistically hypothesized that the regular intake of fish oil (A) might potentially alleviate Raynaud’s Syndrome (C), i.e. $A \rightarrow C$. However, relation $A \rightarrow C$ was unknown in the literature and no literature had been found to co-cite both fish oil and Raynaud’s Syndrome literature. Swanson’s hypothesis was novel and was later confirmed through a randomized placebo-controlled clinical trial by Digiacomo et al. [34].

Swanson’s work (Figure 2.3) demonstrates how new medical knowledge can be discovered by systematically analyzing the “complementary but disjoint” sets of literature [35, 36]. It also suggests that the hypothesis generation process can eventually be performed computationally on the literature. A number of subsequent studies further advanced Swanson’s discovery model [37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48].
Literature-based method uses two techniques to search for the potential medical hypotheses: one-node search and two-node search [36]. In the one-node search, the user starts with a medical concept $A$. Then a list of $B$ concepts is generated that co-occur most frequently with $A$. A list of frequently co-occurring $C$ concepts is generated for each top-$n$ $B$ concepts, provided that they do not co-occur with $A$. The aim is to discover the strongest $ABC$ inferential connection that allows the user to infer a hidden relationship between $A$ and $C$.

In contrast, the two-node search starts with a pair of hypothesis concepts $A$ and $C$ to be tested, $A \cap C = \emptyset$. Then a list of $B$ concepts most frequently co-occurring with both concepts is generated to allow the user to infer hidden relations between $A$ and $C$.

In general, ranking the inferred relations is done by filtering and ranking the concept co-occurrences in the literature according to certain information correlation measures. Two widely used measures are chi-square ($\chi^2$) measure and mutual information measure ($MIM$).

**Chi-Square**

Chi-square ($\chi^2$) is a well-established statistical measure for measuring the correlation strength between two or more categorical variables [49]. $\chi^2$ is used to find statistically significant correlations between two concepts in literature. Statistical significance is derived from the deviation of the observed frequency from the expected frequency of concept co-occurrence. The more the observed frequency deviates from the expected frequency, the more statistically significant is the co-occurrence between two
CHAPTER 2: LITERATURE REVIEW

concepts.

First, a $2 \times 2$ co-occurrence matrix for two concepts, $A$ and $B$, is generated (Table 2.1).

Table 2.1: A $2 \times 2$ contingency table

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>~B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>~A</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

$\chi^2$ is calculated as follows (Eq. 2.1.1):

$$\chi^2 = \sum \frac{(f_e - f_o)^2}{f_e},$$

(2.1.1)

where,

$f_e$ = the expected frequency. For cell $a$, it is calculated as $f_e = \frac{(a+c)(a+b)}{a+b+c+d}$

$f_o$ = the observed frequency

Given $p = 0.05$ and 1-degree of freedom, the critical value is $\alpha = 3.84$. The correlation between $x$ and $y$ is statistically significant if $\chi^2 > \alpha$. Chi-square measure is reliable when the relationship is strong, the sample size is large, and the number of co-occurrence of two concepts is large [50], which is a typical situation when analyzing concept correlation in a large literature database.

Mutual information measure

Mutual information measure (MIM) is a popular probabilistic measure for measuring the significance of associations based on information theory [51, 52].

Given:

$T_A$: literature count of $A$ concept,
$T_B$: literature count of $B$ concept,
$T_{AB}$: literature count of $A \cap B$,
$T$: total number of literature,
$P_A$: probability of observing the occurrence of $A$, $P_A = \frac{T_A}{T}$,
$P_B$: probability of observing the occurrence of $B$, $P_B = \frac{T_B}{T}$,
CHAPTER 2: LITERATURE REVIEW

\[ P_{AB}: \text{probability of observing the occurrence of } A \cap B \text{ in literature}, \ P_{AB} = \frac{T_{AB}}{T}. \]

Let \( P_A \) and \( P_B \) be two independent events such that \( P_{AB} = P_A \cdot P_B \). If \( A \) and \( B \) are dependent events (i.e. genuinely correlated) then \( P_{AB} \gg P_A \cdot P_B \), such that \( \frac{P_{AB}}{P_A \cdot P_B} > 1 \)\[53\]. MIM is calculated as follows:

\[
MIM(A, B) = \log_2 \frac{P_{AB}}{P_A \cdot P_B}
\] (2.1.2)

\( MIM(A, B) \gg 0 \) implies the association between \( A \) and \( B \), otherwise \( MIM(A, B) \approx 0 \). In literature-based hypothesis discovery, MIM was used to measure the correlation strength of \( AB \) as well as \( BC \)\[43, 50\]. Wren\[43\] removed the log function to avoid a negative weighting (Eq. 2.1.3). The higher the MIM score, the stronger the correlation between two concepts in literature.

\[
MIM(A, B) = \frac{P_{AB}}{P_A \cdot P_B}
\] (2.1.3)

Comparing \( \chi^2 \) and MIM

Hu et al\[50\] found that \( \chi^2 \) was suitable for validating the associations found in a very large information collection such as MEDLINE where the co-occurrences of medical concepts tend to be large. \( \chi^2 \) could also filter out frequent but non-statistically significant concept co-occurrences, such as the co-occurrence between \textit{humans} and \textit{Raynaud’s Syndrome}. It could identify the most semantically related concepts although favoring the more general medical concepts. When used in conjunction with the association rules, \( \chi^2 \) pruned away many non-meaningful rules. Its critical value serves as a useful threshold for determining statistical significance.

\( MIM \) calculated concept dependency differently from \( \chi^2 \) as it does not take into consideration the absence of terms\[54\]. It favored more specific medical concepts. Unlike \( \chi^2 \), \( MIM \) has no precise significance threshold, varying from one application to another\[55, 56, 57, 58\]. Overall,\[54\] found that \( \chi^2 \) performed better than \( MIM \) in predicting the truly significant term associations.

Besides the correlation measures, the semantic information from the ontology could
be used to determine the relevance of term associations in literature analysis\textsuperscript{59, 50} and to reduce the literature search space. Ontology is the "explicit specification of a conceptualization" in a domain\textsuperscript{60}. For the medical literature-based discovery, the user’s query statements could be mapped to the standard medical concepts and semantic types. Then, the semantic relations between the semantic types were determined and used to filter out the semantically unrelated medical concepts. For discovering the unknown drug mechanism associations in literatures, Ahlers et al\textsuperscript{47} demonstrated that utilizing the semantic predications that exist between concepts in the Unified Medical Language System (UMLS) was more effective than simply detecting the concept co-occurrences.

The literature-based method differs from the observation-based method in that the hypothesis discovery process is initiated by the user’s prescribing some novel hypotheses rather than by being initiated by the natural observations\textsuperscript{21} (see Figure 2.1 and Figure 2.2). The candidate hypotheses are supported or dismissed by rapidly evaluating the disjoint but complementary pieces of medical knowledge from the literature. Subsequently, the user can judge whether the suggested hypothesis deserves to be further validated through medical experiments. As a result, the literature-based method is able to accelerate the initial process of generating the candidate hypotheses beyond what can be achieved by the traditional observation-based method.

2.1.3 KDD-based method

Similar to the literature, healthcare data is also proliferating at an increasingly rapid pace. The computerization of patient health records and the advancement of medical diagnostic systems result in a huge accumulation of the structured and semi-structured medical data that bear on diagnosis, prognosis, and treatment of patients\textsuperscript{61}. Many potential medical hypotheses may exist in the form of hidden data patterns in the data and remain unnoticed by the medical practitioners.

Knowledge discovery in databases (KDD) is the "nontrivial extraction of implicit, previously unknown, and potentially useful information from data"\textsuperscript{11}. The useful information is extracted from a data set in the form of patterns or rules which can become the potential hypotheses. In general, KDD systems ought to be capable
of automatically selecting meaningful medical hypotheses to be examined and experimentally tested by the medical experts \[2, 62\].

The KDD-based method differs from the literature-based method in a few respects. Firstly, KDD targets the hypothesis generation from structured databases instead of literature. Secondly, the conventional KDD-based method is not particularly guided by a specific hypothesis as the input.

The KDD-based method is promising because the collections of healthcare data are increasingly viewed as the central asset for improving the future healthcare services \[63\]. Figure 2.5 shows the KDD-based methodology for generating hypotheses from databases.

**Figure 2.4:** Steps in knowledge discovery in databases (taken from Fayyad et al 1996)

**Figure 2.5:** KDD-based hypothesis generation model
CHAPTER 2: LITERATURE REVIEW

Association rules

The association rule data mining algorithm aims at finding the meaningful co-occurrence of two or more attributes in a data set [3]. Although initially used primarily for solving the market basket analysis problems, the association rule mining technique has also been applied to discover meaningful associations from medical data sets [64, 65, 66, 67, 68] as well as for identifying the previously unknown relationships among entities in the medical literature [69]. When there is a sufficient ground to support their validity, these associations may eventually become important medical hypotheses which can be tested through medical experiments and clinical trials.

For an association rule to qualify as a potential hypothesis, it must be interesting [11]. Interestingness is the “overall measure of pattern value, combining validity, novelty, usefulness, and simplicity” [2]. Because large data sets can easily produce tens or hundreds of thousands of association rules, it is impossible for the user to manually examine each rule. This problem can be addressed by automatically calculating the interestingness score for each rule based on various rule interestingness criteria.

But to automatically and objectively estimate the rule interestingness without the user’s involvement is very challenging as interestingness eventually depends on the user’s tasks, intentions, and domain knowledge [9, 4]. It is a product of the user’s cognitive factor, bias, and background knowledge [70, 71, 72, 17], and hence may vary from one user to another depending on their subjective point of views.

Rule novelty

Novelty is one of the most important rule interestingness criteria [11, 12, 2]. A novel rule must not be previously known to the user and must not be immediately inferable from the other known rules [7]. Novelty is an important factor in the hypothesis discovery mechanism because a hypothesis assumes a previously unknown relationship between two or more entities. But novelty has received the least attention from researchers and remains the most challenging among all interestingness criteria [7].

Most rule interestingness evaluation methods followed the pairwise novelty paradigm which emphasizes on the structure and syntactic properties of the rule items [73, 18].
To be novel, it is required that the composition of a rule item should syntactically or logically deviate from the existing knowledge. However, this paradigm produces two major problems.

1. **Rare item problem**
   Within the support-confidence framework, it is generally assumed that the novel association rules will normally reside with a very low support \[^{13,74}\]. Mining the association rules with a low minimum support (\textit{minsupp}) threshold often generates a large number of other uninteresting rules. On the other hand, the optimal support and confidence threshold values remain unknown \[^{16}\].

2. **User acceptance problem**
   It is difficult to validate the novelty of rules due to the lack of supporting background knowledge. In medical informatics, the user’s acceptance of rules is crucial. The evidence showed that compliance with the existing domain knowledge helped increase the acceptance of the learned models by the medical experts \[^{17}\]. On the other hand, the learned models that do not comply with the existing user beliefs are often less acceptable by the experts. Unfortunately, the pairwise novelty inherently requires a trade-off between the rule acceptance and the rule novelty because, by definition of this approach, the rule should violate the pre-existing user beliefs in order to be novel and interesting.

There is no compelling evidence to suggest that the pairwise novelty is the only paradigm for evaluating rule novelty. Pattern analysis should include the contextual analysis of the patterns and not be limited to the mere analysis of the pattern composition \[^{18}\]. Jaroszewicz et al \[^{19}\] suggested that rule evaluation should consider the full joint probability on the data. Introducing a new non-pairwise paradigm to evaluate the rule novelty may overcome the current limitations of the pairwise paradigm.

**Current approaches to evaluating rule novelty**

There are three existing main approaches to evaluating and determining the rule novelty \[^{4}\]:

- **Objective approach.** The objective rule interestingness approach emphasizes on rule statistical properties as the main interestingness indicators and relies on the
probability theory, statistics, and information theory \[7\]. There are approximately forty different objective measures \[75, 76, 8\], including support, confidence, lift, chi-square, accuracy, etc. Objective measures alone cannot discover the truly novel rules because they do not capture the entire domain contexts and the user’s background knowledge \[77\]. Statistically strong rules are usually well-known to user, hence not novel \[78, 6\].

**Subjective approach.** The subjective approach takes into consideration the user’s background knowledge to determine the rule novelty \[79, 12, 80, 81, 14, 73\]. The background knowledge provides the point of reference and the context against which the rule novelty is evaluated. Klemettinen et al. \[79\] asked the user to construct a set of inclusive templates to express his or her interests. Any rule whose items match items in the inclusive templates is viewed as interesting. Liu et al. \[82\] compared the syntax of the discovered classification rules with the syntax of the general impressions, i.e. a formal representation of the user’s background knowledge. The rules are considered to be unexpected if its antecedent and/or consequent do not match the antecedent and/or consequent of the general impression. The unexpected rules are interesting because they challenge the pre-existing user beliefs. In \[14\], a set of ancestor rules were mined and manually classified by the user as True-Not-Interesting, Not-True-Interesting, Not-True-Not-Interesting, or True-Interesting. The remaining rules which belonged to the not-interesting categories were automatically eliminated so that only the potentially interesting rules were left for the final user inspection.

In general, the subjective rule interestingness evaluation approach suffers from the serious bottlenecks because the user must manually construct their background knowledge in the machine-understandable formats \[77\]. It requires extensive human efforts to interpret the final data mining results \[83\] and its discovery capability is restricted by the extent to which the user can sufficiently define the required background knowledge \[6\].

**Domain-driven approach.** The notion of rule interestingness ties closely with the notion of domain knowledge \[9\]. Novelty only makes sense if it is interpreted within the context of what is already known in the domain. Hence incorporating the domain knowledge is essential for discovering the novel rules especially in a complex domain \[84\]. The domain-driven approach explicitly incorporates domain intelligence as the basis for a rule novelty evaluation. The domain intelligence is a collection of
"domain resources that not only wraps a problem and its target data but also assists in the understanding and problem-solving of the problem" [23]. Domain intelligence includes the user-specific knowledge source, e.g. user hypothesis, and the other domain knowledge sources, e.g. ontology and literature. In contrast to the objective approach, by acquiring domain intelligence into the data mining process, the domain driven approach is capable of evaluating the rule novelty within an entire context of the domain knowledge. Unlike the subjective approach, the user does not need to manually specify his background knowledge because the domain knowledge can be automatically extracted from the available knowledge sources.

Most domain-driven techniques focused on using semantic information from the ontology to validate the association rules [85, 86, 67, 87]. Ping et al [67] matched the association rules with the association edges of the UMLS semantic network to determine whether a rule is correct yet trivial, or whether it is correct yet unknown. The latter was considered interesting to the user. From among 26,141 of the correct yet unknown rules mined from the Heartfelt Study data set, the technique successfully recovered 8 out of 10 association rules that agreed with the actual original findings of the Heartfelt Study. Unfortunately, the immense number of the other uninteresting rules being generated suggests that the precision of the proposed technique is poor. Kuo et al [86] used the UMLS ontology to categorize data attributes according to their semantic groups. The authors found that the ontology was helpful in solving the duplicate meaning among attributes. It was also useful for discriminating the irrelevant attributes in the data preprocessing step. These may eventually contribute to generating the association rules that are more meaningful to the user. Marinica and Guillet [87] concluded that, as a formalism, the domain ontologies can improve the integration of user domain knowledge into the data mining postprocessing.

Ontology alone may not sufficiently provide the domain knowledge needed to fully determine the interestingness of rules. Ontology was described as just one of the components in the entire domain knowledge sphere [86]. Marinica and Guillet [87] introduced a new formalism called Rule Schema to allow the user to express their expectations and goals which cannot be represented by the contents in the ontologies.

The published literature is a valuable source of domain knowledge for data mining [77, 88]. In addition, automatically extracting the relevant pieces of the domain knowledge from the published literature can replace the manual acquisition of
Chapter 2: Literature Review

the user’s background knowledge and will effectively reduce the bottleneck of the
subjective approach [77]. Compared to the use of the ontology, works that utilized
the literature remain relatively few. Fodeh and Tan [88] automatically acquired
the medical knowledge from the online MEDLINE database to evaluate the quality of the
mined association rules. For items in the target rule \( x \rightarrow y \), the corresponding medical
concepts \( X \) and \( Y \) were acquired. The literature-based support score of the rule was
subsequently calculated based on the number of MEDLINE abstracts in which the
concept \( X \) and \( Y \) co-occurred. By adjusting the threshold value to an appropriate
level, this method could predict the obvious rules with a 74.3% accuracy against a
domain expert’s actual classification. One main challenge of automatically extracting
the domain knowledge from the published literature is the accuracy with which the
relevant literature can be identified.

Besides relying on the intricate formalisms to represent the user’s subjective interest
(for example [79, 82]), a simple representation of the user hypothesis may provide
a useful frame of reference to determine the user’s interests (in this context, a
hypothesis refers the unknown yet suspected relationship between two or more
entities). The theory of inference control [89] which is based on the notion
of interestingness supports this. Ram argued that a reasoner’s interest closely
corresponds to his or her knowledge goals. Subsequently in [90], the author
formulated the theory of question and question asking in which he demonstrated that
the knowledge goals are often expressed in the form of questions. Since it is plausible
to consider hypothesis as one type of questions, a user hypothesis may become the
simple and elegant way for incorporating the user’s subjective interest into the data
mining that neither the ontology or the literature can provide. Some papers have
demonstrated how the intuition, creativity, and experience of the domain expert can
be incorporated into the rule interestingness evaluation in the form of high quality
hypotheses [91, 92, 93].

Assessing the effectiveness of the rule interestingness evaluation techniques

How does one know that a rule interestingness evaluation technique is effective? Methods
for assessing the effectiveness of rule interestingness evaluation techniques can be
categorized into three [94]:

Assessing the effectiveness of the rule interestingness evaluation techniques

How does one know that a rule interestingness evaluation technique is effective? Methods
for assessing the effectiveness of rule interestingness evaluation techniques can be
categorized into three [94]:

19
Chapter 2: Literature Review

1. **Confirmation with the prior knowledge, without the additional information.**
   This method assesses the ability to recover the known associations by mining the same data sets as the ones used in the original medical experiments. Ping et al. [67] used the medical findings previously published in peer-reviewed journals as the evaluation gold standard. Using the original study data set, the association rules were mined and automatically categorized. Rules that were categorized as *Unknown Correct* were considered interesting. Eight out of ten of the original findings were successfully mined out of a total output of 26,141 rules. Another work [95] evaluated the discovered candidate disease genes against a list of disease genes provided by the medical experts. Kuo et al. [86] identified two known associations from a real medical case study as the basis for comparing two association rule mining algorithms. The main drawback of this evaluation technique is that it is difficult to use. This is because the medical study data sets containing the original medical findings are not easily accessible.

2. **Exception to or conflict with the prior knowledge.**
   Hussain et al. [74] evaluated the exception rules against common sense and the reference rules mined from the data. Even though the exception rules are inherently novel, they are difficult to validate because they cannot be supported by the existing knowledge. Pazzani et al. [17] experimentally demonstrated that the user preferred rules that were consistent with the existing medical knowledge.

3. **Discovery of the new knowledge compatible with the prior knowledge.**
   This requires the domain expert to classify or label the discovered association rules based on his or her interest. In turn, the result of the classification is used as the evaluation gold standard.

   Fodeh and Tan [88] asked the user to classify the rules as either *obvious* or *unexpected*. The evaluation was based on the degree of match between the machine classification and the user classification. The other works compared the association rules against the domain expert’s actual interests [76, 85, 26]. Even though this evaluation technique is rather subjective, it can be applied to more varieties of the available medical data sets. Most importantly, it can be used to evaluate the capability of a KDD system to discover the potentially new knowledge, instead of just re-discovering the old findings.
2.1.4 The integration of the literature-based and KDD-based discovery methods

The effectiveness of the KDD systems can be enhanced by incorporating domain knowledge from the literature. The effective clinical decision support systems (CDSS) should capture both the literature-based and practice-based information to provide the more convincing evidence [96]. The current research trend points towards integrating the database methods and the information retrieval methods [25] in which the future biomedical hypothesis generation systems can integrate the data mining and literature-mining techniques. For example, the highly-structured genomic sequences and experimental databases could be mined to improve the accuracy of the literature-based discovery results and to generate other complementary knowledge that is not included in the literature [97, 98]. Tiffin et al [95] integrated a text mining technique applied on the Pubmed literature with a data mining technique applied to the human gene expression data set to successfully select disease gene candidates. Pospisil et al [99] applied a data mining technique on the textual and structured data to identify cancer-related enzymes. Alves et al [100] showed how the hypotheses regarding the behavior of a molecular biological system can be automatically generated and tested by integrating all the relevant information sources including the sequence data, structural data, literature, gene expression data, proteomics data, and metabolomics data. Prabhu et al [101] incorporated additional concepts from a structured knowledge base as the metadata to enhance the Pubmed literature search.

Nevertheless, the current integrative models are limited in at least two respects:

1. The hypotheses were discovered by integrating the literature with the human-curated structured knowledge bases, such as the databases of annotated gene and protein sequences, gene expression data, and proteomics data [100]; gene ontology [97], PharmGKB [101]; or Ensembl database [95]. In constrast, it can be argued that it is more challenging to discover good hypotheses from the large practice-based medical databases because their contents are not specially curated by the human experts. In this case, the rule interestingness plays a very important role to filter out spurious and meaningless rules.

2. Most of the systems being reviewed were usually constructed with the purpose
of solving a very specific biomedical problem, e.g. in [95]. They did not specifically address the rule interestingness problem.

2.2 Chapter Summary

Hypothesis formulation has always been the important catalyst for major medical breakthroughs. Methodologies for generating the medical hypotheses have evolved from the observation-based methods to the literature-based methods and on to the KDD-based methods.

The main interest of this work is on the KDD-based methodology. KDD aims at providing the methodologies capable of automatically mining the novel rules from the medical data sets which may constitute the new potential hypotheses. However, it is always difficult to automatically measure the rule novelty. The literature review suggests that the traditional pairwise novelty paradigm is problematic as it leads to the rare item problem and the lower user acceptance of the rules. A new non-pairwise definition of the rule novelty may offer solutions to the rule novelty evaluation problem.

The existing rule interestingness evaluation approaches have their own strengths and weaknesses. The domain driven approach that uses an integrative knowledge discovery model to synthesize various domain intelligence (data sets, ontologies, literature, and the user’s subjective inputs) is the most promising.
CHAPTER 3

Methodology and Design

3.1 Functional Novelty

This work proposes functional novelty, the new paradigm for evaluating and determining the rule novelty [102]. An association rule is functionally novel if it establishes a previously unknown relationship between a pair of user hypotheses. Given a pair of user hypotheses $A$ and $D$, an association rule $x \Rightarrow y$ is functionally novel if it fulfills the criteria below (see also Figure 3.1):

1. $x \Rightarrow y$ complies with the existing domain knowledge;

2. the associations between $A$ and $x$, and between $y$ and $D$ are supported by the domain knowledge; and

3. the association between $A$ and $D$ is assumed to be previously unknown to the user.

To understand the motivation behind the rule functional novelty as the basis for determining the rule interestingness, consider again Swanson’s $ABC$ inference model [33] previously discussed in Chapter 2. Note that at the time of Swanson’s discovery, the blood viscosity did not represent a novel piece of knowledge by its own. This human blood property was already known to the medical experts. Nevertheless, one

$$A \Leftrightarrow x \Leftrightarrow y \Leftrightarrow D$$

Figure 3.1: Functional novelty model
could argue that blood viscosity was functionally novel as it was previously not known that it could be the mediating agent between the fish oil and the Raynaud’s Syndrome. The medical experts can view the blood viscosity as the novel and interesting piece of knowledge; not on the account of its existence (i.e. by being blood viscosity), but on the account of its functionality (i.e. by being the mediating agent of a previously unknown relation).

Jensen et al. [98] extended the ABC model to the more complex ABCD model. A previously unknown relation between the Rim11 and Erg9 proteins (AD) was established by assembling several known protein-to-protein relations, i.e. between the Rim11 and Ume6 (AB), Ume6 and Ino2 (BC), and between Ino2 and Erg9 (CD) (Figure 3.2) proteins. Although the relation AD was not directly known, however provided that the AB, BC, and CD relations were known, one could eventually infer and hypothesize the existence of the AD relation. Unlike Swanson’s model, the AD was established not by an intermediate concept (B) but by the mediating relationship between two concepts, i.e. BC.

How is this applicable to the rule novelty evaluation problem? Referring to the Figure 3.1 above, it suggests that the novelty of an association rule is not primarily dependent on having the unknown rule item composition (i.e. x and y). Instead, the novelty of a rule can be evaluated based on its likelihood in mediating the hypothesis AD. In this case, the rule with a well-known item composition can even be considered as novel following this framework, provided that there is a sufficient evidence to show that it is capable of establishing a previously unknown relationship between the two hypothesis terms prescribed by the user. Consequently, the rule novelty evaluation algorithm can be set to discover novel rules from among the high-support and high-confidence rule sets.

### 3.2 The Proposed Hypothesis Discovery Method

The proposed hypothesis discovery method is a hybrid between the literature-based method and the KDD-based method (Figure 3.3). At the initial stage, the user first
prescribes a previously unknown relationship between a pair of medical hypotheses. Next, the association rules are mined from the target data set. The association rules $x \leftrightarrow y$ are generated from the target data set using the minimum support ($\text{minsupp}$) and minimum confidence ($\text{minconf}$) thresholds. The maximum itemset size is limited to 2-itemset in order to simplify the rule analysis. The functional novelty of each rule is subsequently evaluated with reference to the prescribed hypotheses by incorporating the domain knowledge from the literature and ontology.

There are two major tasks to be achieved by the proposed framework. The first task is to mine the domain knowledge-compliant rules. The goal is to produce rules that are acceptable by the medical experts. A semantic-based rule filtering technique that utilises the semantic relations in ontology is applied to prune away the non-compliant rules. A new measure, $\chi^2_{\text{lit}}$, is calculated for each rule. $\chi^2_{\text{lit}}$ is a literature-calibrated $\chi^2$ score calculated based on the co-occurrence of the biomedical terms in Pubmed. These terms correspond to the rule items $x$ and $y$.

The second task is to measure the functional novelty of the rules. A pair of hypotheses $AD$ will be acquired from the medical expert. For each rule, the correlation strengths of $A \leftrightarrow x$ and $y \leftrightarrow D$ will be calculated using $\chi^2_{\text{lit}}$ scores. The weakest of these $\chi^2_{\text{lit}}$ scores will be selected to produce a $\text{Min}\chi^2$ score. $\text{Min}\chi^2$ is the final indicator of a rule’s functional novelty and will be used as the basis for ranking the rule set.

Note that from this section onwards the notation $\Rightarrow$ will be replaced with the notation $\leftrightarrow$ to avoid misinterpreting the meaning of a rule’s association. The association rules are computed solely based on the co-occurrences of items and do not necessarily imply a particular type of relation between items (e.g. causal relation).
CHAPTER 3: METHODOLOGY AND DESIGN

An interactive Web-based KDD application will be built as the proof-of-concept. This application will be capable of automatically retrieving the relevant medical literature and patient records to provide the context-rich rule evaluation environment for the user.

3.3 Notations and Definitions

3.3.1 Association rules

Let \( I = \{i_1, i_2, \ldots, i_n\} \) be a set of \( n \) binary items in a data set and \( D = \{t_1, t_2, \ldots, t_m\} \) be a set of all transactions over \( I \). \( X = \{i_1, i_2, \ldots, i_k\} \) is a non-empty \( k \)-itemset where \( X \subseteq I \) and \( 0 < k \leq n \). An association rule is an implication \( x \Rightarrow y \), where \( x \) and \( y \) are items from a similar \( x' \) itemset, \( x \cap y = \emptyset \). Item \( x \) is referred to as the antecedent and \( y \) as the consequent.

**Definition 1.** Support of rule \( x \Rightarrow y \), denoted as \( \text{supp}(x \cup y) \), is the number of transactions in \( D \) containing \( x \cup y \), divided by the total number of transactions \( m \), \( 0 \leq \text{supp} \leq 1 \).

**Definition 2.** Confidence of a rule, \( \text{conf}(x \Rightarrow y) = \frac{\text{supp}(x \cup y)}{\text{supp}(x)} \), is the likelihood that transactions containing \( x \) also contains \( y \), \( 0 \leq \text{conf} \leq 1 \).

**Definition 3.** \( \mu \) is the harmonic mean of \( \text{supp} \) and \( \text{conf} \) shown in Eq. 3.3.1 \([50]\). In the proposed experiment, \( \mu \) will serve as the most basic rule strength measure.

\[
\mu = \frac{2(\text{supp})(\text{conf})}{\text{supp} + \text{conf}} \tag{3.3.1}
\]

**Definition 4.** \( \text{maxitem} \) is the maximum number of items allowed in the candidate itemsets, \( 0 < k \leq \text{maxitem} \).

**Definition 5.** Ontology is “an explicit specification of a conceptualization” \([60]\). Let \( C = \{c_1, c_2, \ldots, c_n\} \) be a set of all concepts in ontology, \( S = \{s_1, s_2, \ldots, s_m\} \) be a set of all semantic types defined over \( C \), and \( R = \{r_1, r_2, \ldots, r_p\} \) be a set of all semantic relations defined over \( S \). The correlation between two concepts, \( c_1 \) and \( c_2 \), is semantically valid if a relation between the respective semantic types, \( s_1 \) and \( s_2 \), is found in the ontology, \( r_{s_1 \rightarrow s_2} \subset R \). Note that the relation between two semantic types is asymmetric, \( r_{s_1 \rightarrow s_2} \neq r_{s_2 \rightarrow s_1} \). Let \( x \Rightarrow y \) be an association rule, and \( c_x \) and \( c_y \) be concepts derived for item \( x \) and \( y \), respectively, the rule is semantically valid if \( r_{s_x \rightarrow s_y} \subset R \).
3.3.2 Estimating the domain knowledge-compliant rules: $\chi^2$-based correlation measures

**Table 3.1:** A $2 \times 2$ contingency table

<table>
<thead>
<tr>
<th></th>
<th>$y$</th>
<th>$\neg y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>$\neg x$</td>
<td>$c$</td>
<td>$d$</td>
</tr>
</tbody>
</table>

**Definition 6.** Given the antecedent $x$ and consequent $y$, and a $2 \times 2$ contingency table, cell $a$ is the number of transaction in which item $x$ and $y$ co-occurs (see Table 3.1). Based on this contingency table, $\chi^2$ is calculated for rule $x \leftrightarrow y$ (Eq. 3.3.2).

$$\chi^2 = \sum \frac{(f_e - f_o)^2}{f_e},$$  \hspace{1cm} (3.3.2)

$f_e$ is the expected frequency of a cell and $f_o$ is its observed (actual) frequency. For cell $a$, $f_e$ is calculated as: $f_e = \frac{(a+c)(a+b)}{a+b+c+d}$. Given $p = 0.05$ and 1-degree of freedom, the critical value is set at 3.84 ($\alpha = 3.84$). The correlation between item $x$ and $y$ is statistically significant in the data set if $\chi^2 > \alpha$.

**Definition 7.** Given the concept $c_x$ and $c_y$ derived for the item $x$ and $y$, respectively, a $2 \times 2$ contingency table is constructed based on the co-occurrence of the concepts in Pubmed. Table 3.2 gives an example of a Pubmed query that searches for the co-occurrence of Smoking and Myocardial Infarction. Based on this, $\chi^2_{lit}$ (Eq. 3.3.3) is calculated. Given $\alpha = 3.84$, the correlation between $x$ and $y$ is statistically significant in the literature if $\chi^2_{lit} > \alpha$.

$$\chi^2_{lit} = \sum \frac{(f_e - f_o)^2}{f_e},$$  \hspace{1cm} (3.3.3)

**Table 3.2:** An example of Pubmed query statements for concepts Smoking and Myocardial Infarction

```
("smoking"[MeSH Terms] OR "smoking"[All Fields]) AND
("myocardial infarction"[MeSH Terms] OR "myocardial"
"infarction"[All Fields] OR "myocardial infarction"[All Fields])
```
Comparing $\chi_{lit}^2$ and mutual information measure (MIM)

An experiment was conducted to investigate and compare the effectiveness of $\chi_{lit}^2$ and MIM in predicting the known and unknown medical associations based on the concept co-occurrence in the literature. Because both measures follow very different theoretical approaches and their superiority tends to vary from one application to another [54], the experiments are required to assess the actual effectiveness of each measure in the context of the current work. Five known associations (Table 3.3) and five unknown associations were acquired from a medical expert (Table 3.4). The known associations were the associations that the expert considered to be true and acceptable with respect to his knowledge; the unknown associations were the otherwise. For each association, $\chi_{lit}^2$ and MIM scores were calculated ($p = 0.05$; $df = 1$; $a = 3.84$).

Table 3.3: Five known associations provided by a medical expert. Total Pubmed count: 19,572,400 (Query date: March 8, 2010)

<table>
<thead>
<tr>
<th>No.</th>
<th>Association Pairs</th>
<th>Plain Pubmed Query</th>
<th>$\chi_{lit}^2$</th>
<th>MIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>myocardial infarction smoking</td>
<td>24,025.10</td>
<td>5.31</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>cerebrovascular accident smoking</td>
<td>11,926.31</td>
<td>4.13</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>smoking malignant neoplasms</td>
<td>4,906.47</td>
<td>1.52</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>myocardial infarction mental depression</td>
<td>2,612.49</td>
<td>2.21</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>myocardial infarction atrial fibrillation</td>
<td>18,154.72</td>
<td>9.00</td>
<td></td>
</tr>
</tbody>
</table>

The performance of $\chi_{lit}^2$ and MIM was compared using three criteria: (1) recall rate, (2) minimum and maximum score comparison, and (3) Mann-Whitney U-Test.

1. Recall rate.

For each set of five known and five unknown associations, the percentage (%) of significant $\chi_{lit}^2$ and the percentage of significant MIM were calculated and compared. Table 3.3 shows that $\chi_{lit}^2$ identified all significant associations ($\chi_{lit}^2 > 3.84$) with 100% recall. A similar result was achieved by MIM (MIM>1.00).

In Table 3.4, $\chi_{lit}^2$ identified four associations (80% recall) which are statistically insignificant ($\chi_{lit}^2 \leq 3.84$) (i.e. the associations do not actually exist, or medically
CHAPTER 3: METHODOLOGY AND DESIGN

Table 3.4: Five unknown associations provided by a medical expert. Total Pubmed count: 19,572,400 (Query date: March 8, 2010)

<table>
<thead>
<tr>
<th>No.</th>
<th>Association Pairs</th>
<th>Plain Pubmed Query</th>
<th>$\chi^2_{lit}$</th>
<th>MIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>von willebrand factor intracardiac thrombus*</td>
<td></td>
<td>25.86</td>
<td>7.03</td>
</tr>
<tr>
<td>2.</td>
<td>smoking endothelial progenitor cell lev.†</td>
<td></td>
<td>0.33</td>
<td>1.14</td>
</tr>
<tr>
<td>3.</td>
<td>smoking human anti-murine antibody lev.†</td>
<td></td>
<td>0.38</td>
<td>0.0</td>
</tr>
<tr>
<td>4.</td>
<td>matrix metalloproteinases intracardiac thrombus*</td>
<td></td>
<td>1.30</td>
<td>0.0</td>
</tr>
<tr>
<td>5.</td>
<td>plaque rupture† freshwater fish†</td>
<td></td>
<td>1.47</td>
<td>0.00</td>
</tr>
</tbody>
</table>

* non-MeSH standard term
† non-standard term

invalid). MIM identified only three associations (60% recall) as insignificantly correlated ($MIM \leq 1$).

Based on this criterion, $\chi^2_{lit}$ and MIM seem to perform equally well, except for the association between von willebrand factor and intracardiac thrombus (Table 3.4). Although its $\chi^2_{lit}$ score is significant, its value is much lower when compared to the other $\chi^2_{lit}$ scores for the known associations. However, its MIM score is abnormally higher than some of the known association scores. This was anomalous given that the association between von willebrand factor and intracardiac thrombus was relatively unknown to the medical expert, hence was supposed to exhibit a much lower MIM score as compared to the known associations.

2. The minimum and maximum score comparison.

The second test assumes that in order to be a good correlation measure, the minimum score of $\chi^2_{lit}$ ($Min\chi^2_{lit}$) and the minimum MIM ($MinMIM$) score from the known association set must be greater than the maximum score of $\chi^2_{lit}$ and MIM ($Max\chi^2_{lit}$ or $MaxMIM$) of the unknown association set (Eq. 3.3.4 and Eq. 3.3.5).

$$Min\chi^2_{lit} > Max\chi^2_{lit} \quad (3.3.4)$$

$$MinMIM > MaxMIM \quad (3.3.5)$$

29
For the known associations (Table 3.3), $\min \chi^2_{\text{lit}}$ was 2,612.49 and $\min \text{MIM}$ was 1.52. For the unknown associations (Table 3.4), $\max \chi^2_{\text{lit}}$ was 25.86 and $\max \text{MIM}$ was 7.03. This result suggests that $\chi^2_{\text{lit}}$ performs better than MIM because $\min \chi^2_{\text{lit}} > \max \chi^2_{\text{lit}}$ whereas $\max \text{MIM} > \min \text{MIM}$.

3. **Mann-Whitney U-Test**.
Mann-Whitney U-test tests the statistical difference between the mean scores of two independent samples. Unlike the *Student’s t-test*, the U-test is a non-parametric test that works well with small samples (<20) that are not normally distributed [103].

Table 3.5: The comparison between $\chi^2_{\text{lit}}$ and MIM over two independent samples

<table>
<thead>
<tr>
<th>Sample</th>
<th>No.</th>
<th>Category</th>
<th>$\chi^2_{\text{lit}}$</th>
<th>MIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1.</td>
<td>Known</td>
<td>24,025.10</td>
<td>5.31</td>
</tr>
<tr>
<td></td>
<td>2.</td>
<td>Known</td>
<td>11,926.31</td>
<td>4.13</td>
</tr>
<tr>
<td></td>
<td>3.</td>
<td>Known</td>
<td>4,906.47</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td>4.</td>
<td>Known</td>
<td>2,612.49</td>
<td>2.21</td>
</tr>
<tr>
<td></td>
<td>5.</td>
<td>Known</td>
<td>18,154.72</td>
<td>9.00</td>
</tr>
<tr>
<td>II</td>
<td>6.</td>
<td>Unknown</td>
<td>25.86</td>
<td>7.03</td>
</tr>
<tr>
<td></td>
<td>7.</td>
<td>Unknown</td>
<td>0.33</td>
<td>1.14</td>
</tr>
<tr>
<td></td>
<td>8.</td>
<td>Unknown</td>
<td>0.38</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>9.</td>
<td>Unknown</td>
<td>1.30</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>10.</td>
<td>Unknown</td>
<td>1.47</td>
<td>0.0</td>
</tr>
</tbody>
</table>

In Table 3.5, this test was used to determine the statistical significance of the difference between the mean $\chi^2_{\text{lit}}$ and MIM scores of two independent samples (I and II). The first sample consisted of the known associations. The second sample consisted of the unknown associations. The effectiveness of the $\chi^2_{\text{lit}}$ and MIM measures in discriminating both groups was then compared.

If a measure produces the mean scores that are statistically different between the two samples, then it is considered effective and reliable for distinguishing the known and unknown associations.

For this test, the following hypotheses were formulated:

\[ H_0: \mu_{\text{Known}} = \mu_{\text{Unknown}} \]
\[ H_1: \mu_{\text{Known}} \neq \mu_{\text{Unknown}} \]

The null hypothesis ($H_0$) assumed no difference between the mean scores ($\chi^2_{\text{lit}}$ / MIM) of both samples. At 0.05 significance level ($\alpha = 0.05$), null hypothesis is
rejected if the obtained \( p \)-value (two-tailed) is less than 0.05.

Table 3.6: Mann-Whitney u-test results for \( \chi^2_{\text{lit}} \)

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
<td>0.009</td>
</tr>
<tr>
<td>Exact Sig. [2*(1-tailed Sig.)]</td>
<td>0.008</td>
</tr>
</tbody>
</table>

The mean \( \chi^2_{\text{lit}} \) of the known category was 12,325.02 and 5.89 for the unknown category. Results in Table 3.6 showed that the difference between mean score of \( \chi^2_{\text{lit}} \) for the known and unknown categories was statistically significant at 0.05 significance level (\( \text{Sig.} = 0.009; \text{Sig.} < \alpha \)). The null hypothesis was rejected. \( \chi^2_{\text{lit}} \) could reliably differentiate between the known and unknown associations.

Table 3.7: Mann-Whitney u-test results for MIM

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
<td>0.072</td>
</tr>
<tr>
<td>Exact Sig. [2*(1-tailed Sig.)]</td>
<td>0.095</td>
</tr>
</tbody>
</table>

U-test was also performed on the MIM (Table 3.7). The MIM mean score of the known category was 4.43 and 1.63 for the unknown category. At 0.05 significance level, the difference between the mean MIM for the known and unknown categories was not statistically significant (\( \text{Sig.} = 0.072; \text{Sig.} > \alpha \)). The null hypothesis was not rejected. MIM could not reliably differentiate between the known and unknown associations.

The results suggest that \( \chi^2_{\text{lit}} \) performs better than MIM in approximating the true correlations between the medical concepts in the literature. It could effectively distinguish the genuine, known medical associations from the hypothetical, unknown associations. This finding is in agreement with [54] who also concluded \( \chi^2_{\text{lit}} \) as a better correlation measure than MIM.

3.3.3 Functional novelty score

The functional novelty score is defined as the function of correlation strength of each component association involved in inferring the relationship between a pair of user hypotheses. Given the compliant rule \( x \leftrightarrow y \), its functional novelty is equivalent to the weakest correlation strength of \( A \leftrightarrow x \) and \( y \leftrightarrow D \) in literature. This is similar to
assuming that the total strength of a chain is no greater than its weakest link \[43\]. The correlation is measured using \(\chi^2_{\text{lit}}\).

**Definition 8.** Let \(\chi^2_{\text{lit}_A}\) and \(\chi^2_{\text{lit}_D}\) be the chi-square scores calculated for \(A \leftrightarrow x\) and \(y \leftrightarrow D\) from the literature, respectively. \(\text{Min}\chi^2\{x,y\}\) is the least of the two chi-square scores (Eq.3.3.6).

\[
\text{Min}\chi^2\{x,y\} = \text{Min}\{\chi^2_{\text{lit}_A},\chi^2_{\text{lit}_D}\}
\]  

(Eq.3.3.6)

Previously, Sebastian et al. [102] proposed the use of literature evaluation score \((\text{LES}_{mn})\) (see Eq.3.3.7) and aggregate literature evaluation score \((\text{ALES}_{xy})\) (see Eq.3.3.8) for measuring the functional novelty. It was assumed that the closer \(\text{ALES}_{xy}\) score to 1, the more functionally novel a rule becomes.

\[
\text{LES}_{mn} = \frac{2(n\text{SET}_{mn})}{n\text{SET}_m + n\text{SET}_n}
\]  

(Eq.3.3.7)

where,

- \(n\text{SET}_{mn}\): the number of articles in which the concept \(m\) and \(n\) co-occur,
- \(n\text{SET}_m\): the number of articles in which the concept \(m\) occurs,
- \(n\text{SET}_n\): the number of articles in which the concept \(n\) occurs.

\[
\text{ALES}_{xy} = \frac{\text{LES}_{Ax} + \text{LES}_{yD}}{2}
\]  

(Eq.3.3.8)

where,

- \(\text{LES}_{Ax}\): LES score of association \(A \leftrightarrow x\),
- \(\text{LES}_{yD}\): LES score of association \(y \leftrightarrow D\).

However, \(\text{LES}\) and \(\text{ALES}\) have serious limitations. First, they do not have the substantial theoretical support. Secondly, unlike \(\chi^2\), they have no threshold value to determine the statistical significance. Lastly, \(\text{ALES}_{xy}\) suffers from the inability to discriminate between the true and false inference. The comparison is given in the following section.
Chapter 3: Methodology and Design

Table 3.8: The comparison between LES and $\chi^2_{lil}$ on two medical associations. Total Pubmed count: 19,579,927. Query date: March 11, 2010.

<table>
<thead>
<tr>
<th>Assoc. 1</th>
<th>Assoc. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>x: Myocardial infarction</td>
<td>x: Smoking</td>
</tr>
<tr>
<td>y: Smoking</td>
<td>y: Endothelial progenitor cell levels</td>
</tr>
<tr>
<td>x count</td>
<td>162447</td>
</tr>
<tr>
<td>y count</td>
<td>152266</td>
</tr>
<tr>
<td>xy count</td>
<td>6730</td>
</tr>
<tr>
<td>LES</td>
<td>0.043</td>
</tr>
<tr>
<td>$\chi^2_{lil}$</td>
<td>24041.25</td>
</tr>
</tbody>
</table>

Comparing LES and ALES against $\chi^2_{lil}$ and Min$\chi^2$

LES and $\chi^2_{lil}$ were compared on the basis of two real medical associations provided by a medical expert: (1) myocardial infarction and smoking (the known association), and (2) smoking and endothelial progenitor cell levels (the unknown association). LES and $\chi^2_{lil}$ scores were calculated for each association. The result in Table 3.8 shows that both measures scored Assoc. 1 higher than Assoc. 2. This is reasonable considering that Assoc. 1 was more well-known than Assoc. 2. Nevertheless, using LES, one cannot precisely indicate which association is statistically significant. On the other hand, $\chi^2_{lil}$ score clearly indicates that Assoc. 1 was statistically significant ($\chi^2_{lil} > 3.84$).

Next, to demonstrate the limitation of ALES, one may consider two imaginary inference cases, each having imaginary scores (Table 3.9). In Case 1, it is assumed that no correlation exist between A and x in the literature ($LES_{Ax} = 0.00$), so no inference can actually be made between A and D. In Case 2, $LES_{Ax}$ is 0.40, so an inference between A and D is more likely than Case 1. However, both cases will eventually produce a similar ALES$_{xy}$ score. Hence ALES$_{xy}$ was not a reliable approximator for the true inference.

Table 3.9: A hypothetical example of the false and true inferences

<table>
<thead>
<tr>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$LES_{Ax}$</td>
<td>0.00</td>
</tr>
<tr>
<td>$LES_{yD}$</td>
<td>0.80</td>
</tr>
<tr>
<td>ALES$_{xy}$</td>
<td>0.40</td>
</tr>
</tbody>
</table>

33
3.4 KELAM Framework Design

Mining the novel rules is technically challenging because it requires a KDD framework that can adequately model the user’s domain knowledge [7]. Asking the user to thoroughly specify all his background knowledge for the data mining input is not a feasible approach, especially for the highly complex biomedical data mining tasks. Instead, it is much more practical if the domain knowledge can be automatically extracted from a collection of domain-specific literature [77] or ontology [86].

KELAM (Knowledge Extraction via Logical Association Mining) is the proposed domain-driven KDD framework for mining functionally novel rules (Figure 3.4). It extracts the domain knowledge from the literature (Pubmed), the ontology (UMLS1), and the user hypothesis.

During the data preparation (Stage 1), the user performs the data selection, attribute discretization, and semantic mapping with the help of the domain ontology. In Stage 2, the user specifies the basic data mining parameters including the \( \text{minsupp} \), \( \text{minconf} \) and \( \text{maxitem} \). The output rule set is then semantically filtered against the semantic relations in the ontology (in Stage 3). In Stage 4, rules that pass the filtering process are checked for the compliance with the existing domain knowledge in the

1\url{http://www.nlm.nih.gov/research/umls/}

Figure 3.4: KELAM framework
literature. Finally, in Stage 5, the user prescribes a pair of hypothesis terms, each of which is mapped to the standard biomedical concept in the ontology. By querying the literature, the rules are ranked according to their functional novelty scores and finally presented to the user.

3.4.1 Stage 1: Semantic-based attribute discretization and mapping

Algorithm 1 shows the semantic-based discretization procedures.

**Algorithm 1** Semantic-based attribute discretization

**Input:**
- \( \alpha \) = attribute in the data set
- \( \Lambda = \) set of \( n \) data set attributes, \( \Lambda = \{\alpha_1, \alpha_2, \ldots, \alpha_n\} \)
- \( \pi = \) value interval defined for \( \alpha \) based on domain knowledge in ontology
- \( \Pi = \) set of \( \pi \)
- \( \pi_{\text{lower}} = \) lower bound value of interval \( \pi \)
- \( \pi_{\text{upper}} = \) upper bound value of interval \( \pi \)
- \( \phi = \) categorical value defined for \( \alpha \)
- \( \Phi = \) set of \( \phi \) for \( \alpha \)

**Output:**
- \( i = \) binary item
- \( v_i = \) valid value for \( i \)

**Procedures:**
for each \( \alpha \) in \( \Lambda \) do
  if \( \alpha \) is continuous attribute then
    Get \( m \) domain cut-off values
    Define \( m + 1 \) interval \( \Pi \)
    for each \( \pi \) in \( \Pi \) do
      Define \( i \), where \( v_i \in \{\pi_{\text{lower}}, \ldots, \pi_{\text{upper}}\} \)
    end for
  end if
  if \( \alpha \) is categorical attribute then
    Get \( \Phi \)
    for each \( \phi \) in \( \Phi \) do
      Define \( i \), where \( v_i = \phi \)
    end for
  end if
end for
Enriching the rules with the conceptual information from the ontologies helps bridge the semantic gap between the pattern representations and user interpretation [104]. First, each data set attribute is discretized. In general, the association rule algorithm only processes binary attributes called *items*. For a continuous attribute, the discretization process classifies the attribute’s values into $n$ intervals [105]. The final discretization outcome is a set of binary items having two discrete values, 0 or 1 (No or Yes, respectively). Value 1 indicates the item occurrence in a transaction while 0 indicating the otherwise.

In this work, the discretization was performed using the domain cut-off values found in the ontology and the other relevant domain knowledge sources (such as the published clinical guidelines). For instance, two cut-off values could be found in the UMLS Metathesaurus for discretizing the *Age* attribute: 6-12 years old for child and 13-18 years old for adolescent. Hence the discretization produces two binary items, [Age6-12=Yes] and [Age13-18=Yes].

For the categorical attributes, the binary item was constructed for each possible value of the categorical attributes. For an attribute having $m$ number of categorical values, $m$ items were generated. For instance, the attribute *Gender* produces two items, [Male=Yes] and [Female=Yes].

### Item mapping

Each binary item was subsequently mapped to the standard biomedical concept by manually searching for the most appropriate concept in the UMLS Metathesaurus. In the previous example, [Age6-12=Yes] was assigned with the concept *Child* (Age Group), and [Age13-18=Yes] was assigned with the concept *Adolescent* (Age Group). The item [Male=Yes] was assigned with the concept *Male Gender* (Organism Attribute). The term enclosed in the parentheses refers to the concept’s semantic type.

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of all binary items in data set. A semantic map was manually constructed that consisted of a vector of the semantic information for each $i_n$. A sample of the semantic map is shown in Table 3.10. A qualifier was included to provide the additional information that cannot be fully represented by the standard biomedical concept.
### Table 3.10: An example of semantic map contents

<table>
<thead>
<tr>
<th>Item</th>
<th>Concept</th>
<th>CUI</th>
<th>Semantic Type</th>
<th>Attribute</th>
<th>Qualf. (opt.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Male=Yes]</td>
<td>Male Gender</td>
<td>C0024554</td>
<td>Organism Attribute</td>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>[Age6-12=Yes]</td>
<td>Child</td>
<td>C0008059</td>
<td>Age Group</td>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>[hpt_stage1=Yes]</td>
<td>Hypertensive Disease</td>
<td>C0020538</td>
<td>Disease or Syndrome</td>
<td>SBP, DBP</td>
<td>stage 1</td>
</tr>
</tbody>
</table>

One of the main uses of the ontology for the domain-driven data mining is for disambiguating the meaning of the data attributes [86]. In the less ideal situation where an item cannot be precisely mapped to a specific concept in the ontology, the domain expert solved the ambiguity by manually assigning the appropriate term and semantic type to the item. This technique is similar to [86]. Algorithm 2 shows the algorithm for mapping items to the concepts.

#### 3.4.2 Stage 2: Generating the association rules with FP-Growth algorithm

To generate the association rules, the FP-Growth algorithm was chosen over the Apriori. Unlike the Apriori that requires an indefinite number of transaction scans, the FP-Growth encodes the data set into a highly compacted FP-Tree data structure and requires only two scans on the data set to generate all association rules. It avoids the expensive generate-and-test approach of the Apriori-like algorithms, hence more efficient. For certain types of transactional data sets, it outperformed the Apriori by several orders of magnitude [106] and proved to be efficient particularly in mining large and dense data sets [107] such as the biomedical data sets.

A Java implementation of the FP-Growth in the RapidMiner open-source data mining package was used [108]. Algorithm 3 shows the high-level algorithm for mining the association rules [106].

This work imposed a limit on the maximum number of items in the frequent itemsets, maxitem, to only two items in order to focus on mining the association rules that exactly represent a birelational association BC. This constraint is useful because simpler rules with the less number of antecedent and consequent items will be more comprehensible by the user as compared to the more complex rules. It also provides the mechanism for avoiding mining the low support rules as the support count of
Algorithm 2 Item-to-concept mapping using ontology

Input:
\( \alpha = \) attribute in the data set
\( \Lambda = \) set of \( n \) data set attributes, \( \Lambda = \{a_1, a_2, \ldots, a_n\} \)
\( i = \) binary item
\( I = \) set of \( n \) items in data set, \( I = \{i_1, i_2, \ldots, i_n\} \)
\( C = \) set of \( n \) biomedical concepts in ontology, \( C = \{c_1, c_2, \ldots, c_n\} \)
\( S = \) set of \( m \) semantic types defined over \( C \) in ontology, \( S = \{s_1, s_2, \ldots, s_m\} \)

Output:
\( c_i = \) biomedical concept for item \( i \), where \( c_i \in C \)
\( s_{c_i} = \) semantic type of \( c_i \), where \( s_{c_i} \in S \)
\( CUI = \) concept unique identifier
\( v = \) vector of \( \{i, c_i, CUI, s_i, \alpha\} \)
\( q = \) qualifier (optional)

Procedures:
for each \( i \) in \( I \) do
  Search for \( c_i \) in \( C \)
  if \( c_i \neq \emptyset \) then
    Get \( c_i \)
    Get CUI for \( c_i \)
    Get \( s_{c_i} \)
  else
    Get \( c_i \) from domain expert
    Set \( CUI = null \)
    Get \( s_{c_i} \) from domain expert
  end if
  Get \( \alpha \) of \( i \), \( \alpha \in \Lambda \)
  Get \( q \) for \( i \) (optional)
  Create \( v \)
end for

rules will usually decrease with the increased number of items in a rule.

3.4.3 Stage 3: Semantic-based filtering

Following Definition 5, a semantic-based filter was constructed that utilizes the information from the semantic map and UMLS Semantic Network. The components of the UMLS are described in detail in the Appendix section.

Example.

Consider two association rules.
\( R_1: [hpt\_stage1=Yes] \leftrightarrow [\text{Male}=\text{Yes}] \)
Algorithm 3 Association rule generation

**Input:**
- \( \text{minsupt} \) = minimum support threshold
- \( \text{minconf} \) = minimum confidence threshold
- \( \text{maxitem} \) = maximum number of items in an itemset
- \( I \) = set of \( n \) items in data set, \( I = \{i_1, i_2, \ldots, i_n\} \)

**Output:**
- \( \hat{I} \) = candidate frequent itemset generated from \( I \), \( \hat{I} \in I \)
- \( k \) = number of items in \( \hat{I} \)
- \( \text{supp}(\hat{I}) \) = support count of \( \hat{I} \)
- \( r \) = candidate association rule generated from \( \hat{I} \)
- \( R \) = set of \( m \) candidate association rules, \( R = \{r_1, r_2, \ldots, r_m\} \)
- \( \text{conf}(r) \) = confidence of \( r \)
- \( \Omega \) = final association rule set

**Procedures:**

1. Generate a set of \( \hat{I} \), \( \text{supp}(\hat{I}) \geq \text{minsupt} \), \( 1 < k \leq \text{maxitem} \)
2. For each \( \hat{I} \) do
3.   Generate \( R \)
4.     For each \( r \) in \( R \) do
5.       Compute \( \text{conf}(r) \)
6.       If \( \text{conf}(r) \geq \text{minconf} \) then
7.         Include \( r \) into \( \Omega \)
8.       End if
9.     End for
10. End for

\( R_2 \): [Male=Yes] \( \Leftrightarrow \) [Age6-12=Yes]

Referring to the information in the semantic map, a semantic type was extracted for each rule item to produce the following semantic associations:

\( R_1' \): Disease or Syndrome \(\rightarrow\) Organism Attribute

\( R_2' \): Organism Attribute \(\rightarrow\) Age Group

Each semantic type association was then compared against the UMLS Semantic Network. Rule \( R_1 \) was considered semantically valid because the relation between Disease or Syndrome and Organism Attribute types was defined in Semantic Network (Disease or Syndrome, associated_with, Organism Attribute). The content of the UMLS Semantic Network is described in Table 6.2 of the Appendix. \( R_2 \) was not semantically valid because the relation between Organism Attribute and Age Group was undefined.
Following this mechanism, the semantic-based rule filter excluded $R_2$ from the subsequent rule evaluation.

The semantic-based filter ensures that only the association rules which are compliant with the domain knowledge at the ontology level can proceed to the next level of the rule evaluation. This filtering technique is similar to [86] and [87]. However, it is simpler in that it does not explicitly calculate the item-relatedness of the rule antecedent and consequent using the taxonomic distance in the ontology. This is unnecessary within the scope of this work.

Note that the filter currently does not favor a certain semantic relation. In the future, it is possible to allow the user to specify a particular semantic relation he or she is interested in. For example, a domain expert aiming at discovering the etiological associations may focus on finding the rules whose items are related by the semantic relation *causes* in Semantic Network [109].

Algorithm 4 shows the filtering algorithm.

**Algorithm 4** Semantic-based rule filtering

Input:
- $\rho = \text{association rule } x \leftrightarrow y$
- $\Omega = \text{set of } n \text{ association rules, } \Omega = \{\rho_1, \rho_2, \ldots, \rho_n\}$
- $C = \text{set of biomedical concepts in ontology}$
- $S = \text{set of semantic types defined over } C \text{ in ontology}$
- $R = \text{set of semantic relations defined over } S$
- $c_x, c_y = \text{biomedical concept of item } x \text{ and } y \text{ respectively, } \{c_x, c_y\} \in C$
- $s_x, s_y = \text{semantic type of item } x \text{ and } y \text{ respectively, } \{s_x, s_y\} \in S$
- $r_{s_x \rightarrow s_y} = \text{semantic relation between } s_x \text{ and } s_y$

Output:
- $\Delta = \text{set of } m \text{ semantically valid association rules, } \Delta = \{\rho_1, \rho_2, \ldots, \rho_m\}$

Procedures:
- for each $\rho$ in $\Omega$ do
  - Get $x$ and $y$
  - Get $c_x$ and $c_y$
  - Get $s_x$ and $s_y$
  - if $r_{s_x \rightarrow s_y} \in R$ then
    - Include $\rho$ into $\Delta$
  - end if
- end for
3.4.4 Stage 4: Discovering the domain-knowledge compliant rules

In Stage 4, each rule that passed the filtering process was subsequently evaluated for its compliance with the existing medical knowledge in the literature. First, a standard biomedical concept was extracted for the rule antecedent/consequent from the semantic map file. A Pubmed query statement was subsequently constructed from these concepts to retrieve the frequency of their co-occurrence in the literature. $\chi^2_{lit}$ was computed to measure the correlation strength between the two concepts. Statistically significant $\chi^2_{lit}$ score of a rule, i.e. $\chi^2_{lit} > 3.84$ (see Def. 7), indicates that the rule is compliant and agreed by the domain expert.

3.4.5 Stage 5: Evaluating and ranking functionally novel rules

A pair of hypotheses was acquired from the user. The hypothesis terms were mapped to the standard biomedical vocabularies in the UMLS Metathesaurus, both representing $A$ and $D$. For each rule $x \leftrightarrow y$, $\chi^2_{lit}$ was calculated for the association between $A$ and $x$ and between $y$ and $D$ in order to produce $\text{Min}\chi^2$ score. The rules were then sorted based on $\text{Min}\chi^2$ in a descending order. The rules with $\text{Min}\chi^2 \geq 3.84$ were considered to be functionally novel.

3.5 Experimental Designs

The proposed experiments have two main objectives. The first objective is to find out whether incorporating more domain knowledge from the literature will increase the system’s ability to mine the domain knowledge-compliant rules. This objective is addressed in Experiment I. The second objective is to demonstrate the ability of the proposed measures to discover the functionally novel rules and whether they give the interesting results to the medical expert. This is addressed in Experiment II.

A common method to evaluate the effectiveness of many KDD systems is to discover the new knowledge that is compatible with the prior knowledge [94]. This method has the advantage of being able to assess the capability of the KDD system in discovering the potentially new knowledge by using the domain expert evaluation as the evaluation ‘gold standard’. It can also be applied to a wide variety of the available medical data sets. Several notable works [110, 76, 88, 26] employed this
method. Experiment I and II adopt the similar evaluation method by using the medical expert’s evaluation as the evaluation gold standard.

3.5.1 The medical expert

This section briefly describes the medical expert. At the time of the experiments, Dr. AFYY was a Cardiologist at the Department of a Cardiology at the Sarawak General Hospital (SGH), the Clinical Lead for the Cardiac Magnetic Resonance Imaging of the department, and the Secretary to the hospital’s Clinical Research Centre. He graduated from the University of Bristol (UK) in 1999 and completed a post-graduate training in General Medicine in 2004 (MRCP, UK). His areas of research interest include the Cardiovascular Imaging and Biomarker Research. He has participated in over 20 Phase I, II and III Clinical Trials and co-written several chapters in the National Medicine Use Survey 2005 and 2006 and the Annual Reports of the National Cardiovascular Disease Registry (NCVD-ACS and NCVD-PCI). He was a reviewer of the National Medical Research Registry. He has also published and reviewed articles published in several well-recognised international journals (the American Heart Hospital Journal, the European Journal of Radiology and the American Heart Journal) as well as in various international cardiology meetings (EuroPCR, World Congress of Cardiology, NHAM Annual Scientific Meeting).

3.5.2 Data sets descriptions

Two cardiovascular data sets were used for the experiments. The first data set, the Coronary Heart Disease Database (CHD_DB)\(^2\) is a synthetic cardiovascular data set curated by Suka et al. It is a replica of the original Framingham Heart Study data. From this database, a training data set containing 13,000 records was arbitrarily chosen. Table 3.11 shows a list of attributes in CHD_DB data set.

The second data set, ECHO_MSCT data set, is extracted from the real cardiovascular clinical databases with the help of the medical expert selected for this study and with the permission from a local cardiology centre (Table 3.12). The data set contains 543 records obtained from the echocardiography (ECHO) and multi-slice computed tomography scan (MSCT) examinations. The data set contains no missing values.

\(^2\)Available for use with permission from http://ichimura.ints.info.hiroshima-cu.ac.jp/chddb/
## Chapter 3: Methodology and Design

### Table 3.11: CHD_DB data attributes

<table>
<thead>
<tr>
<th>No.</th>
<th>Attribute</th>
<th>Description</th>
<th>Data type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID</td>
<td>Row Id.</td>
<td>Seq. values</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>CHD</td>
<td>Development of coronary heart disease</td>
<td>Categorical</td>
<td>0=non CHD case; 1=CHD case</td>
</tr>
<tr>
<td>3</td>
<td>Origin</td>
<td>National origin</td>
<td>Categorical</td>
<td>0=native-born; 1=foreign-born</td>
</tr>
<tr>
<td>4</td>
<td>Educate</td>
<td>Educational level</td>
<td>Categorical</td>
<td>0=grade school; 1=high school, not graduate; 2=high school, graduate; 3=college</td>
</tr>
<tr>
<td>5</td>
<td>Tobacco</td>
<td>Tobacco usage</td>
<td>Categorical</td>
<td>0=never; 1=stopped; 2=cigars or pipes; 3=cigarettes (&lt;20/day); 4=cigarettes (≥20/day)</td>
</tr>
<tr>
<td>6</td>
<td>Alcohol</td>
<td>Alcohol intake</td>
<td>Continuous</td>
<td>Unit in oz/mo</td>
</tr>
<tr>
<td>7</td>
<td>SBP</td>
<td>Systolic blood pressure</td>
<td>Continuous</td>
<td>Unit in mmHg</td>
</tr>
<tr>
<td>8</td>
<td>DBP</td>
<td>Diastolic blood pressure</td>
<td>Continuous</td>
<td>Unit in mmHg</td>
</tr>
<tr>
<td>9</td>
<td>TC</td>
<td>Cholesterol</td>
<td>Continuous</td>
<td>Unit in mg/dl</td>
</tr>
<tr>
<td>10</td>
<td>LVH</td>
<td>Left ventricular hypertrophy</td>
<td>Categorical</td>
<td>0=none; 1=definite or possible</td>
</tr>
</tbody>
</table>

### Table 3.12: ECHO_MSCT data attributes

<table>
<thead>
<tr>
<th>No.</th>
<th>Attribute</th>
<th>Description</th>
<th>Data type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sex</td>
<td>Patient’s gender</td>
<td>Categorical</td>
<td>Male; Female</td>
</tr>
<tr>
<td>2</td>
<td>Age</td>
<td>Patient’s age</td>
<td>Continuous</td>
<td>Unit in years</td>
</tr>
<tr>
<td>3</td>
<td>LMS, pLAD, mLAD, dLAD, pLCX, mLCX, dLCX, pRCA, mRCA, dRCA, PDA, PLV</td>
<td>Blockages of heart arteries</td>
<td>Continuous</td>
<td>Unit in %</td>
</tr>
<tr>
<td>4</td>
<td>Smoking</td>
<td>CVD risk factor</td>
<td>Categorical</td>
<td>True; False</td>
</tr>
<tr>
<td>5</td>
<td>Hypertension</td>
<td>CVD risk factor</td>
<td>Categorical</td>
<td>True; False</td>
</tr>
<tr>
<td>6</td>
<td>Diabetes mellitus</td>
<td>CDV risk factor</td>
<td>Categorical</td>
<td>True; False</td>
</tr>
<tr>
<td>7</td>
<td>Hypercholesterolaemia</td>
<td>CVD risk factor</td>
<td>Categorical</td>
<td>True; False</td>
</tr>
<tr>
<td>8</td>
<td>Family history of coronary heart disease</td>
<td>Fraction of blood pumped out of heart ventricles with each heart beat</td>
<td>Categorical</td>
<td>True; False</td>
</tr>
<tr>
<td>9</td>
<td>Ejection fraction</td>
<td></td>
<td>Categorical</td>
<td>True; False</td>
</tr>
</tbody>
</table>
CHAPTER 3: METHODOLOGY AND DESIGN

3.5.3 Design of Experiment I

Experiment I is designed to determine the usefulness of $\mu$, $\chi^2$, and $\chi^2_{lit}$ measures for discovering the association rules that are compliant with domain knowledge. The performance of these measures will be compared and evaluated based on the degree of agreement with the medical expert’s judgment. It is hypothesized that $\chi^2_{lit}$ would exhibit a higher recall rate than $\chi^2$ in discovering the domain knowledge-compliant rules because $\chi^2_{lit}$ explicitly incorporates the domain knowledge from Pubmed.

Experimental settings:

1. Given $\text{minsupp} = 0.0001$, $\text{minconf} = 0.00$, and $\text{maxitem} = 2$, all 2-item association rules are mined. The rationale behind selecting these support-confidence threshold values is explained in the next chapter. For each rule that passes the semantic-based filtering, its $\mu$ score is calculated.

2. If the number of rules $n \leq 25$, all rules are selected and ranked by $\mu$ in a descending order. If $n > 25$, all rules are first ranked by $\mu$ in a descending order and only the top-25 rules are selected.

3. $\chi^2$ and $\chi^2_{lit}$ score are computed for each selected rule.

4. The rules are then presented to the medical expert for an evaluation. $\mu$, $\chi^2$, and $\chi^2_{lit}$ scores are hidden from the evaluator to avoid the evaluation bias.

5. The medical expert manually labels the quality of each rule based on his knowledge and expertise. Unlike Ohsaki et al [76], this experiment does not use the interesting or not interesting labels because the meaning of these labels are imprecise and unclear. Instead, the label compliant (CP), contradictory (CT), and not sure (NS) are used. These labels more precisely indicate whether a rule was compliant with the existing knowledge, contradictory to it, or neither both, respectively.

6. The performance of $\mu$, $\chi^2$, and $\chi^2_{lit}$ is evaluated and compared as follows:

   (a) Rank rules in a descending order according to each measure.

   (b) Calculate Precision-at-fifteen (P@5), Precision-at-ten (P@10), and Precision-at-fifteen (P@15). These are the fraction of top-5, top-10, and top-15 rules labeled CP, respectively.
7. For each $\chi^2$ and $\chi^2_{lit}$ measures, Precision, Recall and $MC(CP)$ are calculated. Let $R_{CP} = \{r | r \in R \land L_{me}(r) = CP\}$ be a set of all rules labeled as CP by the medical expert (subsequently denoted as me), $R$ be the set of all rules to be evaluated, and $L_{me}(r)$ denoting the assigned rule label. Let $R_{stat} = \{r | r \in R \land \chi^2(r) \vee \chi^2_{lit}(r) > \alpha\}$ be a set of all statistically significant rules based on $\chi^2$ or $\chi^2_{lit}$, and $\alpha$ denoting the critical value. Let $R_{CP, stat} = \{r | r \in R_{CP} \land r \in R_{stat}\}$ be a set of all statistically significant rules labeled as CP. Each measure is evaluated using three evaluation metrics used in [76]:

$$\text{Precision} = \frac{R_{CP, stat}}{R_{stat}} \hspace{2cm} (3.5.1)$$

$$\text{Recall} = \frac{R_{CP, stat}}{R_{CP}} \hspace{2cm} (3.5.2)$$

$$MC(CP) = \frac{2(\text{Precision})(\text{Recall})}{\text{Precision} + \text{Recall}} \hspace{2cm} (3.5.3)$$

### 3.5.4 Design of Experiment II

In the second experiment, the usefulness of $\text{Min}\chi^2_{lit}$ for estimating and ranking rule functional novelty is investigated. Following results obtained in Experiment I (described in next chapter), the significant $\chi^2_{lit}$ score is used as the main indicator for the compliant rules. Because $\text{Min}\chi^2_{lit}$ selects the minimum $\chi^2_{lit}$ score of the associations $A \leftrightarrow x$ and $y \leftrightarrow D$, the higher $\text{Min}\chi^2_{lit}$ score of the rule $x \leftrightarrow y$, the higher the rule’s propensity to be functionally novel.

Four pairs of medical hypotheses are given by the medical expert based on his interest:

i smoking $\leftrightarrow$ endothelial progenitor cell levels (EPC lev.)

ii smoking $\leftrightarrow$ human anti murine antibody levels (HAMA lev.)

iii matrix metalloproteinase (MMP) $\leftrightarrow$ intracardiac thrombus

iv von willebrand factor (VWF) $\leftrightarrow$ intracardiac thrombus

Hypotheses [i, ii] are used for mining CHD_DB data set, and [iii, iv] for ECHO_MSCT data set. Experimental settings:
CHAPTER 3: METHODOLOGY AND DESIGN

1. Given $\minsupp = 0.0001$, $\minconf = 0.00$ and $\maxitem = 2$, the algorithm mines all 2-item association rules. $\chi^2_{li}$ score for each rule is computed.

2. The domain-knowledge compliant rules are selected ($\chi^2_{li} > 3.84$).

3. For each selected rule, $\text{Min} \chi^2$ score is calculated and ranked in a descending order. The top-10 rules and bottom-10 rules are selected for an analysis. This method aims at uncovering the clear distinction in the interestingness features between the top- and bottom-ranked rules. The same approach was adopted by [76] for the practical reason because it is not feasible for the medical expert to evaluate all rules due to the time constraint and the significant workloads involved.

4. For each rule-hypothesis pair, its component associations are extracted. For instance, given an inference $A \leftrightarrow x \leftrightarrow y \leftrightarrow D$, association $A \leftrightarrow x$ and $y \leftrightarrow D$ were extracted and presented to the medical expert.

5. The medical expert labels each of these component associations as compliant (CP), contradictory (CT), and not sure (NS).

6. The rules $x \leftrightarrow y$ which had both $A \leftrightarrow x$ and $y \leftrightarrow D$ labeled as CP are selected.

7. These rules are presented to the medical expert for the final evaluation. He is asked to judge if the rule $x \leftrightarrow y$ could potentially infer the association between hypothesis $A$ and $D$ in a meaningful sense. Four types of inference quality labels are used by the medical expert:

   (a) **YES**: all component associations are medically valid. $A$ and $D$ are likely to be related.

   (b) **PROBABLY**: only some component associations are medically valid. $A$ and $D$ are less likely to be related.

   (c) **NO, AS FAR AS I KNOW**: medical expert lacks sufficient background knowledge to make proper judgment.

   (d) **NOT SPECIFIC ENOUGH**: rule does not provide unambiguous and sufficient information to allow medical expert to make proper judgment.

The additional comments by the medical expert are also acquired.
8. Unlike the Experiment I, it is not possible to precisely determine the recall rate of $\text{Min}_{\chi^2_{\text{lit}}}^2$ due to the exclusion of the middle-ranked rules from the evaluation procedure. Hence the performance of $\text{Min}_{\chi^2_{\text{lit}}}^2$ has to be quantitatively determined by comparing the percentage of the number of candidate functionally novel rules in the top-10 and bottom-10 rules. The ranking is based on the $\text{Min}_{\chi^2_{\text{lit}}}^2$ score. The higher percentage of rules in top-10 that are labeled YES by the medical expert, the better the performance of $\text{Min}_{\chi^2_{\text{lit}}}^2$ in estimating a rule’s functional novelty. Finally, the quality of the candidate functionally novel rules will be qualitatively assessed by examining the medical expert’s comments, background knowledge, and the other relevant information in the literature.

3.6 Chapter Summary

The functional novelty is proposed as the new non-pairwise rule novelty criterion. An association rule is functionally novel if it establishes the previously unknown relation between a pair of user hypotheses. Unlike the traditional pairwise novelty, the functional novelty does not emphasize on the demonstration of the previously unknown combination of rule items.

The KELAM knowledge discovery framework incorporates domain intelligence from the ontology, literature and the domain expert. The two $\chi^2$-based measures being proposed utilize the domain knowledge in the literature. The first measure, $\chi^2_{\text{lit}}$, is a $\chi^2$ measure that is calibrated with the medical information from Pubmed for the purpose of estimating the domain knowledge-compliant rules. The second measure, $\text{Min}_{\chi^2_{\text{lit}}}^2$, estimates the functional novelty of a rule in the context of a pair of user hypotheses.

To provide the proof-of-concept, two experimental designs are proposed which involve the real cardiovascular data sets. The experiments involved a cardiologist who performs the expert evaluation on the results. The Experiment I compares and evaluates the usefulness of the $\mu$, $\chi^2$, and $\chi^2_{\text{lit}}$ measures for mining the domain knowledge-compliant rules based on the rules’ degree of agreement with the medical expert’s judgment. The Experiment II investigates the usefulness of the $\text{Min}_{\chi^2}^2$ for estimating and ranking rule functional novelty.
CHAPTER 4

Implementation and Results

4.1 Technological Architecture

*KELAM* was designed as a semi-automatic interactive knowledge discovery application [102]. It was implemented with Java Server Pages (JSP) technology, providing the Web accessibility. Figure 4.1 describes its technological architecture.

![Figure 4.1: KELAM technological architecture](image)
The data preprocessing stage was manually performed by the user. Using the DBMS and spreadsheet software, the target data set was manually converted from its native database format into the specific .dat and .aml data formats. The constructed semantic map and UMLS Semantic Network files were acquired in the form of .csv format.

There are two possible ways to access the Metathesaurus content. The first is to browse the contents using the UMLS MetamorphoSys program\(^1\). The second is to use the MetaMap Transfer (MMTx\(^2\)) which has the capability to map free texts to the Metathesaurus concepts.

The Data Mining subsystem consisted of two main modules: the *FP-Growth* algorithm module and the semantic-based rule filtering module. The Evaluation subsystem consisted of the hypothesis term mapping module for acquiring the user hypotheses and, using MMTx classes, automatically mapping them to the Metathesaurus concepts.

To access Pubmed citations and the article contents, the Entrez Programming Utilities (E-Utils\(^3\)) was integrated into the proposed application. E-Utils is a collection of Java classes that programmatically retrieve Pubmed contents outside the regular Pubmed Web interface. Three E-Utils packages were used: *EInfo* for retrieving the total number of Pubmed article count; *ESearch* for retrieving the citation count and the citation PMIDs (PubMed identifiers); and *EFetch* for retrieving the citations’ details (e.g. title, author list, abstract text, medical subject headings) based on the previously retrieved PMIDs.

![Figure 4.2: E-Utils data pipeline for real time access to Pubmed](image)

E-Utils allows the data pipelines to be constructed in various ways using its component classes. Figure 4.2 shows the data pipeline arrangement of *ESearch* and *EFetch*. *ESearch* returned Pubmed citation count, PMID list, WebEnv and QueryKey in

---

1\(^{http://www.nlm.nih.gov/pubs/factsheets/umlsmetamorph.html}\)
2\(^{http://ii.nlm.nih.gov/MMTx.shtml}\)
3\(^{http://eutils.ncbi.nlm.nih.gov/}\)
response to a query. WebEnv and QueryKey are the cookie-like temporary variables generated by Pubmed servers which act as a reference to the previously retrieved PMID list. These variables were useful as they enable EFetch to retrieve the details of Pubmed citations without having to regenerate the PMID list, thus reducing retrieval time.

The Rule Visualization subsystem was designed for two purposes: (1) to visualize the data mining outputs; and (2) to provide the user with the context-rich environment for the rule evaluation. It consisted of the rule sorting module for ranking the association rules from the most functionally novel to the least novel; the literature module for retrieving and displaying the Pubmed citation titles, authors, journal names, and abstract texts; and the database module for retrieving the relevant medical records pertinent to each rule from the underlying clinical database.

4.1.1 Data preparation

Attribute selection

The data set was manually selected and extracted by the user using the appropriate DBMS data exporting tools into a spreadsheet file. Since the medical data may reside in various types of database management systems (DBMS), it is necessary to create the standard data input format that KELAM can process, i.e. .dat and .aml.

Attribute discretization

Each selected data attribute was manually discretized and categorized into binary items. Discretization enables data mining algorithms to operate more efficiently. It reduces the complexity of processing multiple values in continuous attributes by grouping the values into a number of \( n \) intervals \[105\]. For discretizing continuous attributes, the domain cut-off values were first acquired from the Metathesaurus. If no suitable cut-off value is available from the ontology, it is acquired from the relevant medical literature or the medical expert.
Chapter 4: Implementation and Results

Item-to-concept mapping

Each resulting binary item was mapped to a medical concept name. The medical concept names were searched from the Metathesaurus using the MMTx or UMLS Metamorphosys. Because the Metathesaurus consists of various types of biomedical vocabularies, determining which vocabulary to be used is important because no single vocabulary can meet all of the healthcare information needs [112]. Since all Pubmed literature are indexed using the MeSH terms, the MeSH medical concept names were preferred in order to increase the relevance of the retrieved documents. Where no matching MeSH concept exists, concepts from the other medical vocabularies were used.

The outcome of this mapping procedure is the semantic map file (.csv). Figure 4.3 shows an example of the semantic map file containing data in the following order of values: item name, medical concept name, concept unique identifier, semantic type, original attribute name.

<table>
<thead>
<tr>
<th>Item name</th>
<th>Medical concept name</th>
<th>Unique identifier</th>
<th>Semantic type</th>
<th>Original attribute name</th>
</tr>
</thead>
<tbody>
<tr>
<td>coronary heart disease</td>
<td>Disease or Syndrome, CHD</td>
<td>CO010068</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cholesterol low</td>
<td>Disease or Syndrome, CHD</td>
<td>C0860920</td>
<td></td>
<td></td>
</tr>
<tr>
<td>borderline cholesterol</td>
<td>Disease or Syndrome, CHD</td>
<td>C0694540</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laboratory or Test Result, CHD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cholesterol levels raised</td>
<td>Disease or Syndrome, CHD</td>
<td>C0948569</td>
<td></td>
<td></td>
</tr>
<tr>
<td>normal blood pressure</td>
<td>Disease or Syndrome, CHD</td>
<td>C022105</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prehypertension, C1696708</td>
<td>Disease or Syndrome, CHD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypertensive disease</td>
<td>Disease or Syndrome, CHD</td>
<td>C0020538</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypertensive crisis</td>
<td>Disease or Syndrome, CHD</td>
<td>C0020546</td>
<td></td>
<td></td>
</tr>
<tr>
<td>left ventricular hypertrophy, C0149721</td>
<td>Disease or Syndrome, LVH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>elementary school education level, C1552029</td>
<td>Idea or concept, Educate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high school</td>
<td>Population Group, Educate</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.3: Semantic map file (.csv): an example

Data set format conversion

The target data set was converted into a pair of .dat and .aml files. The first consists of the binary data values (Figure 4.4). The second file is an XML-based description of each binary item associated with the data values in .dat file (Figure 4.5).

The primary objective of the association rule mining is to detect the co-occurrences of items in a data set [3]. The occurrence (presence) and non-occurrence (absence) of items in an itemset are most commonly represented in binary form [3, 113, 78, 13, 114, 115, 116, 117]. Consequently, a binomial (binary) model of attribute is applied in this
study. Nominal values "0", "false", or "no" have been used equally to represent the absence of item, and values "1", "true", or "yes" to represent the presence of item. This study chooses binomial values "yes" and "no" on the preference basis. Even though a few examples of the use of non-binomial attribute value formats have been reported in the literature, e.g. the use of granules to represent attribute values [118], the case of spatial association rule mining for image databases [119], and the use of vertical data format [117], the binomial model remains the most widely-used for mining the
4.1.2 Data mining

**FP-Growth module**

To perform data mining, KELAM accepted four input files: a semantic map file (.csv), a binary data value file (.dat), an attribute file (.aml) and a UMLS Semantic Network file (.csv). It allowed the user to specify the three main association rule mining parameters: the minsupp, minconf, and maxitem.

**Rule semantic filter module**

The association rules generated using the FP-Growth algorithm were semantically filtered. The filter used the information from the semantic map and the UMLS Semantic Network files as the basis for filtering. For each rule, the semantic types of rule items were acquired from the semantic map file and their relation was checked against the list of semantic relations defined in the UMLS Semantic Network. The filtering process was performed at the rule-level instead of at the itemset-level due to the non-symmetric property of the semantic relations.

4.1.3 User hypothesis

**Hypothesis standardization module**

The user entered a pair of hypotheses into KELAM. Each hypothesis term was mapped to a standard Metathesaurus concept using the MetaMap Transfer (MMTx) application. The standardization helps to disambiguate the meaning of the term and increased the relevance of the retrieved Pubmed literature.

Because the MMTx’s mapping precision was reportedly low [121], the mapping result for each hypothesis term needs to be displayed to and confirmed by the user. In case where no appropriate medical concept name existed, the user prescribed the most suitable medical concept name.
4.1.4 Rule scoring

Functional novelty module

The functional novelty module performed two main functions: (1) identifying the domain knowledge-compliant rules, and (2) scoring the rule functional novelty. For each association rule that was generated and semantically filtered by the data mining module, $\mu$, $\chi^2$, and $\chi^2_{lit}$ scores were calculated. For each compliant rule, $\text{Min}\chi^2$ score was calculated.

4.1.5 Rule ranking and visualization

Rule sorting module

The aim of this module is to allow the user to identify the interesting rules easily and quickly [82]. The rules were ranked based on the $\text{Min}\chi^2$ score in a descending order. The most functionally interesting rules were displayed on top of the rule set so that they could be quickly noticed and examined by the user. The least interesting rules were displayed at the bottom. To enhance the interactivity, the user had the options to sort the rules based on the $\mu$, $\chi^2$, and $\chi^2_{lit}$ scores.

4.1.6 Literature details visualization module

Incorporating the domain knowledge from the literature helps to contextualize the user’s rule evaluation process. Contextualization is an important aspect of the frequent pattern analysis and visualization [122, 18]. Contextualization helps the user better understand the meaning and significance of each rule within the context of the most current domain knowledge. It also helps the user assess the validity of the discovered rule and eventually determine the extent to which it can reliably mediate the user-prescribed hypotheses.

The details of literature that support each rule were presented to the user, including the titles, authors, journal titles, and publication years. When the user selected a rule, a list of literature supporting the associations $x \leftrightarrow y$, $A \leftrightarrow x$, and $y \leftrightarrow D$ was displayed. Upon selecting each article, the corresponding abstract text was displayed. Providing an access to the abstract text is important as it allows the user to investigate
and determine the actual relevance of each article.

Two parameters were set for retrieving the literature details: (1) *maximum number of retrieved articles*; (2) *article sorting criteria*. The decision to limit the number of the retrieved articles was reasonable. It was not feasible to visualize all supporting articles as their number may easily reach hundreds or even thousands. The articles were by default sorted according to the most recent publication year.

### 4.1.7 Database access module

KELAM provided the connectivity with the underlying medical records stored in a relational database. It allowed the user to query the subsets of patient records that contributed to the occurrence of each rule. The tight connectivity with the analysed database is deemed as a desirable feature for an effective knowledge discovery tool and is necessary for creating a context-rich environment for the rule evaluation.

### 4.2 Description of KELAM Prototype

This section provides the detailed descriptions of the working prototype system implemented based on the proposed technological architecture. Using the feature classification, this section also provides a comparison between the prototype and the other comparable prototype systems. The comparison is useful for distinguishing KELAM from the other existing systems.

#### 4.2.1 KELAM prototype system

Following KELAM’s technological architecture, the prototype implementation is divided into several modules: the input module, data mining module, evaluation module, and visualization module.

**Input module**

The input module allows the user to specify two input data: the target data set and the semantic map file. Figure 4.6 shows the system’s user interface.

The target data set consists of two separate files: .dat and .aml files. The ‘Dataset file’
CHAPTER 4: IMPLEMENTATION AND RESULTS

Figure 4.6: Data set and semantic map specification

Figure 4.7: Normal mode setting

option specifies .dat source file, the ‘Attribute file’ option specifies the .aml source file, and the ‘Mapping file’ option specifies the semantic map source file. The ‘Minimum
Support’ and ‘Minimum Confidence’ parameters can be adjusted by the user to suit the discovery goals. The ‘Max. Retrieved Literature’ parameter allows the user to determine the maximum number of Pubmed citations that can be retrieved in support of each rule. The user can choose between two display modes: the ‘Normal Mode’ and the ‘Expert Mode’. The former displays only a few basic rule quality indicators (see Figure 4.7). The latter includes all the available rule quality indicators, such as the Support, Confidence, Min$\chi^2$, etc. Alternatively, the user may independently select different indicators as needed.

Data mining and rule sorting module

After specifying the appropriate knowledge discovery settings, the user enters a pair of medical hypothesis terms. In case where several target data sets have been previously uploaded onto the system, the user can choose which data set to be used for mining the rules. See Figure 4.8.

For each hypothesis term, the system displays a list of recommended UMLS standard medical concepts. This functionality is achieved using the MMTx classes. For example in Figure 4.9, 17 concepts were found but no exact match was found for the term high-sensitivity c-reactive protein.

Figure 4.8: Hypothesis term entry

Figure 4.9: The first hypothesis term mapping
CHAPTER 4: IMPLEMENTATION AND RESULTS

In contrast, Figure 4.10 shows four concepts closely associated with a hypothesis term. The first concept best matched the second hypothesis term *matrix metalloproteinase*.

![Figure 4.10: The second hypothesis term mapping](image)

The data mining module produces a set of rules that may potentially connect both hypothesis terms (Figure 4.11). For each rule, the values of the selected rule quality indicators, e.g., Min$\chi^2$, are shown (Figure 4.12). The user then may sort the rule set based on these different indicators to understand the rules from various point of views.

![Figure 4.11: Displayed rule set](image)

![Figure 4.12: Rule quality indicators](image)
CHAPTER 4: IMPLEMENTATION AND RESULTS

Literature visualization module

Pubmed citations that support a rule can be retrieved and displayed by clicking on each rule. The parameter ‘Max. Retrieved Literature’ determines the maximum number of citations to be displayed. Figure 4.13 shows a set of citations retrieved for the rule \textit{hypercholesterolaemia }\leftrightarrow \textit{hypertension}.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4_13.png}
\caption{Pubmed citations retrieved for rule \textit{hypercholesterolaemia }\leftrightarrow \textit{hypertension}}
\end{figure}

For each citation, its \textit{title}, \textit{author}, \textit{journal}, \textit{year}, and \textit{PMID} details are accessible. Citation’s full abstract text is displayed by clicking its title. For example, Figure 4.14 shows the full abstract text for an article entitled ‘\textit{Obesity a risk factor for coronary events in Japanese patients with hypercholesterolaemia on low-dose simvastatin therapy}’. 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4_14.png}
\caption{Full abstract of a Pubmed literature}
\end{figure}
CHAPTER 4: IMPLEMENTATION AND RESULTS

Database access module

Using this module, the user queries patient records from the underlying database which contribute to the occurrence of each rule. Figure 4.15 shows a subset of the patient records corresponding to the rule \( \text{hypercholesterolaemia} \leftrightarrow \text{hypertension} \). For this subset of records, the descriptive statistics are computed and visualized to describe the demographic characteristics that are clinically important, e.g. race and gender. In this example, there is an almost equal proportion of the male and female patients who had the co-occurrence of \( \text{hypercholesterolaemia} \) and \( \text{hypertension} \). Interestingly, more than half of the patients were Chinese (65.85%), followed by a a smaller percentage of Malay patients (17.50%). The goal is to help the user interpret a rule in a clinically meaningful way and provide the more holistic view of the association rule.

Related Patient Record

<table>
<thead>
<tr>
<th>No.</th>
<th>ID</th>
<th>Sex</th>
<th>Race</th>
<th>Weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>103</td>
<td>Female</td>
<td>Chinese</td>
<td>85</td>
</tr>
<tr>
<td>2</td>
<td>105</td>
<td>Male</td>
<td>Iban</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>106</td>
<td>Female</td>
<td>Chinese</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>107</td>
<td>Female</td>
<td>Chinese</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>108</td>
<td>Female</td>
<td>Iban</td>
<td>72</td>
</tr>
<tr>
<td>6</td>
<td>109</td>
<td>Male</td>
<td>Chinese</td>
<td>69</td>
</tr>
<tr>
<td>7</td>
<td>110</td>
<td>Female</td>
<td>Chinese</td>
<td>80</td>
</tr>
</tbody>
</table>

Number of patients found: 120

![Fig 4.15: Patient records corresponding to rule \( \text{hypercholesterolaemia} \leftrightarrow \text{hypertension} \)](image)

A drill-down data analysis can be performed on each patient record. Figure 4.16 shows the results of the echocardiography and multi-slice computed tomography (MSCT) examinations of a one patient. The medication data can also be displayed to track down the patient’s treatment history.
CHAPTER 4: IMPLEMENTATION AND RESULTS

4.2.2 Comparison with other prototypes

There were 43 knowledge discovery systems in 1999 as surveyed by Goebel and Gruenwald [123]. This number has continued to increase in the past one decade. This section discusses the key differences between KELAM and the other data mining prototypes. It compares KELAM with two relevant and recent knowledge discovery research prototypes: GEMINI and ARROWSMITH.

GEMINI is a research prototype built recently at the National University of Singapore for mining and ranking the interestingness of rules within an equivalent rule group [68]. Its binary version is available free for download [61]. GEMINI represents a category of KDD-based hypothesis discovery systems. The other similar prototype systems, such as IAS Visualization system [82], are excluded due to their inaccessibility or obsoleteness. This comparison utilizes the feature classification groups introduced in [123] which include the general characteristics (architecture, operating systems), the database connectivity (data sources, database connectivity, model, attributes, queries), and the data mining characteristics (discovery tasks, discovery methodology, human interaction). The additional comparison feature, domain knowledge source, is also included. Not all the criteria under each characteristic group are included in this comparison. Some criteria are irrelevant to the assessment of KELAM. The others lack the required information.

The comparison between KELAM and GEMINI is shown in Table 4.1. The similarities

---

**Table 4.1:** Feature comparison of GEMINI and KELAM

<table>
<thead>
<tr>
<th>Arch.</th>
<th>OS</th>
<th>Data</th>
<th>DB Access</th>
<th>Model</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>C/S</td>
<td>U</td>
<td>D</td>
<td>W</td>
<td>TF</td>
</tr>
<tr>
<td>GEMINI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KELAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mining Tasks</td>
<td>Method</td>
<td>Interaction</td>
<td>DK Source</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>Cla</td>
<td>A</td>
<td>Vis</td>
<td>RI</td>
<td>G</td>
</tr>
<tr>
<td>GEMINI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KELAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

can be observed between KELAM and GEMINI in terms of their data source format (ASCII file), data model (one table), data mining methodology (rule induction), and human computer interaction model (human guided). The differences are observed in the system architecture (standalone versus client-server system), the operating systems (UNIX-based OS, Windows versus Windows), the database connection (online versus offline), the attribute types (categorical versus continuous), the database queries (allowed versus not allowed), and the supported data mining tasks (link analysis only versus a combination of link analysis, classification, and visualization).

The most contrasting difference is in the domain knowledge source category in which KELAM incorporates the comprehensive domain knowledge sources (literature and ontology), whereas GEMINI involves none.

The second comparison is made against the ARROWSMITH [5], which belongs to the category of literature-based hypothesis discovery systems. ARROWSMITH is a research prototype built by the University of Illinois at Chicago [48] based on Swanson’s literature discovery model. Its purpose is to find the disjoint yet complementary relationships among biomedical entities in Pubmed.

Using the ARROWSMITH, a user first selects a set of ‘A-Literature’ by searching keyword A. In a similar way, the user then selects another set of ‘C-Literature’ by searching keyword C. A set of ‘B-terms’ is generated, containing the terms which occur separately in the ‘A-Literature’ as well as in the ‘C-Literature’ but which do not occur together with ‘A’ and ‘C’ within the same literature. The terms are ranked

based on their relevance. They can be further filtered based on their semantic groups. Because ARROWSMITH does not have most of the features generally found in a data mining system, a comparison based on the above feature classification is not suitable. Instead, the comparison focuses on the domain knowledge source’s breadth (i.e. how many types of sources are involved in the system) and depth (i.e. how much details of the domain knowledge is provided by the system to the user).

Table 4.2: The domain knowledge usage comparison of ARROWSMITH and KELAM

<table>
<thead>
<tr>
<th>Domain Knowledge Utilization</th>
<th>Breadth</th>
<th>Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ont.</td>
<td>Lit.</td>
</tr>
<tr>
<td>ARROWSMITH</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>KELAM</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

Although KELAM and ARROWSMITH seem to possess a similar breadth of domain knowledge sources, KELAM involves the more in-depth domain knowledge by incorporating the semantic relations in the ontology for the rule filtering. In contrast, ARROWSMITH does not take the semantic relations into account when calculating the probability of B-terms in linking the disjoint terms [46].

4.3 Experiment I

4.3.1 CHD_DB

Data preparation

Prior to data mining, the following data attributes were excluded from the original data set:

1. ID
   Reason: not meaningful.

2. ORIGIN
   Reason: values were only relevant in the context of the original US-based
CHAPTER 4: IMPLEMENTATION AND RESULTS

Framingham study.

3. ALCOHOL

Reason: not of the medical expert’s interest.

The continuous attributes were discretized into binary items using cut-off values obtained from the domain knowledge sources (Table 4.3).

Table 4.3: CHD_DB: Discretization of continuous data attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Item</th>
<th>Cut-off Values</th>
<th>Domain Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>chol_low=yes</td>
<td>&lt; 200 mg/dl</td>
<td>American Heart Association</td>
</tr>
<tr>
<td></td>
<td>chol_border=yes</td>
<td>200 – 239 mg/dl</td>
<td>Heart Association</td>
</tr>
<tr>
<td></td>
<td>chol_raised=yes</td>
<td>≥ 240 mg/dl</td>
<td>Guideline</td>
</tr>
<tr>
<td>SBP, DBP</td>
<td>bp_normal=yes</td>
<td>SBP &lt; 120 AND DBP &lt; 80</td>
<td>1. American Heart Association</td>
</tr>
<tr>
<td></td>
<td>hpt_pre=yes</td>
<td>SBP = 120 – 139 OR DBP = 80 – 89</td>
<td>Guideline</td>
</tr>
<tr>
<td></td>
<td>hpt_stage1=yes</td>
<td>SBP = 140 – 159 OR DBP = 90 – 99</td>
<td>2. Chobanian et al.</td>
</tr>
<tr>
<td></td>
<td>hpt_stage2=yes</td>
<td>SBP = 160 – 180 OR DBP = 100 – 110</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hpt_crisis=yes</td>
<td>SBP &gt; 180 OR DBP &gt; 110</td>
<td></td>
</tr>
</tbody>
</table>

The remaining categorical data attributes were itemized according to their categorical values (Table 4.4). Each resulting item was subsequently mapped to a Metathesaurus concept.

Table 4.4: CHD_DB: Items derived from categorical attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Generated Items</th>
<th>Corresponding Attribute Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD</td>
<td>chd=yes</td>
<td>1</td>
</tr>
<tr>
<td>LVH</td>
<td>hypertrophy_left=yes</td>
<td>1</td>
</tr>
<tr>
<td>Educate</td>
<td>edu_grade=yes</td>
<td>0</td>
</tr>
<tr>
<td>Educate</td>
<td>edu_high_not_graduate=yes</td>
<td>1</td>
</tr>
<tr>
<td>Educate</td>
<td>edu_high_graduate=yes</td>
<td>2</td>
</tr>
<tr>
<td>Educate</td>
<td>edu_college=yes</td>
<td>3</td>
</tr>
<tr>
<td>Tobacco</td>
<td>smoker=yes</td>
<td>2,3,4</td>
</tr>
<tr>
<td>Tobacco</td>
<td>ex_smoker=yes</td>
<td>1</td>
</tr>
<tr>
<td>Tobacco</td>
<td>never_smoke=yes</td>
<td>0</td>
</tr>
</tbody>
</table>

The contents of the resulting semantic map are shown in Figure 4.17 (from the left to the right column: item names; medical concept names; CUIs; semantic types; original attribute names, qualifier). From the semantic map, the system can automatically check if a qualifier exists for an item. If a qualifier exists, the qualifier term was appended.
to the Pubmed query statement. For instance, the query statement for \textit{hpt\_stage1=yes} would be formulated as \textit{hypertensive+disease+stage+1}. The query statement for \textit{chd=yes} the query statement would be \textit{coronary+heart+disease} (no qualifier).

![Table](image)

**Figure 4.17:** CHD_DB: item-to-concept mapping results (.csv)

### Data mining

This section first justifies the selected \textit{minsupp} and \textit{minconf} threshold setting. There has been no optimal \textit{minsupp} and \textit{minconf} threshold values that apply to all situations. The research to solve this problem is still ongoing [16]. From the domain-driven perspective, we can select threshold values that best satisfy the specific requirements of the domain and of the target data set. Since the objective of this work is to mine as many rules as possible without leaving out the significant rules, \textit{minsupp} = 0.0001 and \textit{minconf} = 0.00 thresholds were chosen based on the following rationales:

1. CHD_DB data set consists of 13,000 transactions. At the minimum, an association rule must satisfy 1.3 transactions \((0.0001 \times 13,000 = 1.3)\). Such low-support rules will most probably happen by chance. They are medically insignificant as they only apply to one patient/clinical instance. The chosen setting ensures that all medically significant rules will not be left out from the subsequent rule evaluation processes.

2. ECHO_MSCT data set consists of 543 transactions. At the minimum, an association rule must satisfy 0.0543 transaction \((0.0001 \times 543 = 0.0534)\). This
setting ensures that virtually all rules will be generated.

3. A key weakness of confidence score is its inability to reliably measure true implications [78]. The high confidence score does not guarantee the true correlation between the rule antecedent and consequent. We chose minconf = 0.00 so as not to exclude the true correlation rules that may happen to reside at a relatively low confidence region.

Figure 4.18 shows the number of rules generated by FP-Growth algorithm for both data sets at various minsupp levels (minconf=0.00 and maxitem=2). The setting minsupp at 0.0001 removed more than half of the rules generated from the CHD_DB and ECHO_MSCT data sets (544→228; 312→124, respectively), thus greatly reducing the number of rules to be evaluated.

The Data Mining module successfully generated 228 association rules. Of these, only 168 rules passed the rule semantic filtering ($\frac{60}{228} \times 100\% = 23.32\%$ reduction). It was then discovered that the rule set actually contained redundant pairs of rules, i.e. two rules having similar items but positioned differently in terms of its antecedent and consequent. For example, rule [chd=yes]⇔[smoker=yes] and rule [smoker=yes]⇔[chd=yes] were redundant. Both rules essentially reports the same co-occurrence of items [chd=yes] and [smoker=yes] although their confidence scores are different. Because the association rules do not necessarily imply the causal relationships between the items, such redundancy can be safely eliminated by arbitrarily removing one of the rules. Because the medical expert was also informed
that an association rule only indicated the co-occurrence of items without asserting the causal relations, the removal of redundant rules would not cause any significant evaluation bias. Eighty three rules were removed, leaving only 85 non-redundant rules ($\frac{83}{168} \times 100\% = 49.41\%$ reduction). The remaining rules were sorted and ranked in a descending order according to the $\mu$ score. Because the total number of rules exceeds 25 rules, only the top 25 rules were selected to be evaluated by the medical expert.

**Results**

Table 4.5 presents the evaluation results. Nineteen rules were labeled as Compliant (indicated by the black dots). These are rules that the medical expert confidently believed as true according to his medical knowledge. The remaining six rules were neither marked as Contradictory or Not-Sure due to the insufficient data and clinical parameters for making judgment (see the medical expert’s comments). Because the experiment’s objective is to assess the extent to which $\mu$, $\chi^2$, and $\chi^2_{lit}$ scores can predict the compliant rules, the results remain valid. The rules were subsequently sorted in a descending order according to $\mu$, $\chi^2$, and $\chi^2_{lit}$ measures. P@5, P@10, and P@15 of each measure were then calculated and the results are shown in Table 4.6.

Among all the three measures, $\mu$ demonstrated the best ranking precision for detecting the compliant rules (in bold). A comparison between $\chi^2$ and $\chi^2_{lit}$ showed that the precision of $\chi^2_{lit}$ increased with the increase in the number of rules being evaluated. On the contrary, the $\chi^2$'s precision gradually decreased as more rules are evaluated. This result suggests that $\chi^2_{lit}$ was likely to have a higher recall rate than $\chi^2$.

![Figure 4.19: CHD_DB: evaluation results](image)

The qualitative analysis results are shown in Figure 4.19 following the method used in [76]. CP denotes the rules labeled as Compliant by the medical expert ($R_{CP}$); N denotes the otherwise. The white cells indicate the statistically significant rules ($R_{stat}$); the black cells indicate the otherwise. The degree of agreement between the medical
expert’s evaluation and the machine’s evaluation was calculated by observing the number of white cells that matched CP (i.e. $R_{CP,Stat}$). The recall rates of $\chi^2$ and $\chi^2_{lit}$ were calculated and compared.

The result showed that $\chi^2_{lit}$ demonstrated a higher degree of agreement with the
CHAPTER 4: IMPLEMENTATION AND RESULTS

Table 4.6: CHD_DB: performance comparison between $\mu$, $\chi^2$, and $\chi^2_{lit}$

<table>
<thead>
<tr>
<th></th>
<th>$\mu$</th>
<th>$\chi^2$</th>
<th>$\chi^2_{lit}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P@5$</td>
<td>1.00</td>
<td>0.80</td>
<td>0.60</td>
</tr>
<tr>
<td>$P@10$</td>
<td>0.90</td>
<td>0.60</td>
<td>0.70</td>
</tr>
<tr>
<td>$P@15$</td>
<td>0.93</td>
<td>0.73</td>
<td>0.80</td>
</tr>
</tbody>
</table>

medical expert’s evaluation compared to $\chi^2$. When measured across the 25 rules, $\chi^2_{lit}$ demonstrated a higher precision (0.79 vs 0.69), a higher recall (0.79 vs 0.47), and a higher MC(CP) (0.79 vs 0.56) than $\chi^2$.

This result suggests that a domain-driven measure ($\chi^2_{lit}$) more effectively detect the domain knowledge-compliant rules than the traditional data-driven measure ($\chi^2$).

4.3.2 ECHO_MSCT

Data preparation

A total of 543 records with non-missing values were extracted from the ECHO_MSCT data base. The continuous attributes were discretized into binary items using cut-off values obtained from the domain knowledge sources (Table 4.7).

Table 4.7: ECHO_MSCT: Discretization of the continuous data attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Item</th>
<th>Cut-off values</th>
<th>Domain sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>age19-44yrs=yes</td>
<td>age19-44</td>
<td>1. UMLS Metathesaurus</td>
</tr>
<tr>
<td></td>
<td>age45-64yrs=yes</td>
<td>age45-64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>age65-79yrs=yes</td>
<td>age65-79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>age80andover=yes</td>
<td>age80≤</td>
<td></td>
</tr>
<tr>
<td>Ejection fraction</td>
<td>hf=yes</td>
<td>&lt; 40</td>
<td>2. American Heart Association Guideline</td>
</tr>
</tbody>
</table>

Table 4.8 shows the remaining categorical attributes tranformed into items according to their categorical values. Figure 4.20 shows the resulting semantic map.

Data mining

At $\text{minsupp} = 0.0001$ and $\text{minconf} = 0.00$, the data mining module produced 124 association rules. Of these, 86 rules passed the semantic filtering (38/124 * 100% = 30.65 reduction). Forty-three redundant rules were removed, leaving 43
Table 4.8: MSCT_ECHO: Items derived from categorical attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Item</th>
<th>Corresponding values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>male=yes</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>female=yes</td>
<td>Female</td>
</tr>
<tr>
<td>LMS, pLAD, mLAD</td>
<td>ca=yes</td>
<td>50-74=True OR 75-100=True for any one of the artery.</td>
</tr>
<tr>
<td>dLAD, pLCX, mLAD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dLCX, pRCA, mRCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dRCA, PDA, PLV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td>smoking=yes</td>
<td>True</td>
</tr>
<tr>
<td>Hypertension</td>
<td>hypertension=yes</td>
<td>True</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>dm=yes</td>
<td>True</td>
</tr>
<tr>
<td>Hypercholesterolaemia</td>
<td>hypercholesterolaemia=yes</td>
<td>True</td>
</tr>
<tr>
<td>Family history of CAD</td>
<td>ca_fam_history=yes</td>
<td>True</td>
</tr>
</tbody>
</table>

Figure 4.20: ECHO_MSCT: item-to-concept mapping results (.csv)

non-redundant rules (43/86×100% = 50% reduction). All 43 rules were sorted and ranked in a descending order by μ.

Results

Because the total number of rules exceeded 25 rules, the top 25 rules were selected to be evaluated by the medical expert. The medical expert labeled all rules as Compliant.

The rules were then sorted in a descending order according to each measure. P@5, P@10, and P@15 scores were calculated and shown in Table 4.9.

No difference in precision was observed across all the three measures. Consequently, the variation in the performance had to be determined by the precision and recall analysis over all 25 rules. Table 4.10 shows the complete evaluation results. Figure
Table 4.9: ECHO_MSCT: performance comparison between $\mu$, $\chi^2$, and $\chi^2_{lit}$

<table>
<thead>
<tr>
<th></th>
<th>$\mu$</th>
<th>$\chi^2$</th>
<th>$\chi^2_{lit}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P@5$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$P@10$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$P@15$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Figure 4.21: ECHO_MSCT evaluation results

4.4 Experiment II

The results obtained in Experiment I demonstrated $\chi^2_{lit}$’s a high recall rate. Consequently, Experiment II used $\chi^2_{lit}$ as the basis for determining the compliant rules.

The usefulness of $Min\chi^2$ for measuring the functional novelty was investigated by mining the rules from the CHD_DB and MSCT_ECHO data sets, guided by four pairs of hypotheses from the medical expert. Four result sets are presented in Table 4.11, Table 4.12, Table 4.13, and Table 4.14. Each table includes top-10 and bottom-10 rules sorted by $Min\chi^2$. The $\chi^2_{lit,ax}$ measures the correlation strength of association between the hypothesis $A$ and rule antecedent $x$. $\chi^2_{lit,yd}$ was calculated for the rule consequent $y$ and hypothesis $D$. $CP_{Ax}$ and $CP_{yD}$ refer to the medical expert’s evaluation on the compliance of $A \iff x$ and $y \iff D$ with the domain knowledge.

Six candidate functionally novel rules (rule no. 2,3,4,6,7,8) were found among the top-10 rules (Table 4.11). Six candidate rules were also found among the top-10 rules in Table 4.12 (rule no.4,5,6,8,9,10). In both Table 4.13 and Table 4.14, only one candidate rule was found in the top-10 rules (Rule no.5 and no.9, respectively). In all result sets, no candidate functionally novel was found among the bottom-10 rules.

All six candidate rules in Table 4.11 and Table 4.12 were found to be the same.
Table 4.10: MSCT_ECHO: Medical expert’s evaluation results

<table>
<thead>
<tr>
<th>No.</th>
<th>Rules</th>
<th>$\mu$</th>
<th>$\chi^2$</th>
<th>$\chi^2_{Lik}$</th>
<th>Comm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>[hypercholesterolaemia=yes] $\leftrightarrow$ [hypertension=yes]</td>
<td>0.424</td>
<td>70.634</td>
<td>107205.330</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>[age45-64yrs=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.400</td>
<td>4.727</td>
<td>16317.050</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>[age45-64yrs=yes] $\leftrightarrow$ [hypertension=yes]</td>
<td>0.400</td>
<td>1.059</td>
<td>57783.970</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>[hypertension=yes] $\leftrightarrow$ [female=yes]</td>
<td>0.372</td>
<td>6.767</td>
<td>27023.750</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>[ca=yes] $\leftrightarrow$ [female=yes]</td>
<td>0.363</td>
<td>23.619</td>
<td>29289.910</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>[male=yes] $\leftrightarrow$ [hypertension=yes]</td>
<td>0.296</td>
<td>6.767</td>
<td>33668.120</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>[female=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.235</td>
<td>0.027</td>
<td>13774.770</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>[hypertension=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.211</td>
<td>46.133</td>
<td>25812.780</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>[dm=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.208</td>
<td>38.588</td>
<td>15996.610</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>[dm=yes] $\leftrightarrow$ [hypertension=yes]</td>
<td>0.197</td>
<td>0.055</td>
<td>207658.140</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>[ca=yes] $\leftrightarrow$ [age19-44yrs=yes]</td>
<td>0.185</td>
<td>1.671</td>
<td>41467.960</td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>[dm=yes] $\leftrightarrow$ [age19-44yrs=yes]</td>
<td>0.172</td>
<td>0.767</td>
<td>15244.470</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>[ca=yes] $\leftrightarrow$ [age19-44yrs=yes]</td>
<td>0.161</td>
<td>0.672</td>
<td>80081.150</td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>[female=yes] $\leftrightarrow$ [dm=yes]</td>
<td>0.099</td>
<td>23.619</td>
<td>54465.510</td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td>[ca=yes] $\leftrightarrow$ [dm=yes]</td>
<td>0.088</td>
<td>2.788</td>
<td>25907.510</td>
<td></td>
</tr>
<tr>
<td>16.</td>
<td>[ca=yes] $\leftrightarrow$ [age19-44yrs=yes]</td>
<td>0.079</td>
<td>0.400</td>
<td>116890.840</td>
<td></td>
</tr>
<tr>
<td>17.</td>
<td>[ca=yes] $\leftrightarrow$ [age65-79yrs=yes]</td>
<td>0.074</td>
<td>7.136</td>
<td>81832.350</td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td>[male=yes] $\leftrightarrow$ [dm=yes]</td>
<td>0.065</td>
<td>14.674</td>
<td>56002.300</td>
<td></td>
</tr>
<tr>
<td>19.</td>
<td>[ca=yes] $\leftrightarrow$ [male=yes]</td>
<td>0.055</td>
<td>18.351</td>
<td>9724.550</td>
<td></td>
</tr>
<tr>
<td>20.</td>
<td>[hypertension=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.054</td>
<td>3.078</td>
<td>4657.880</td>
<td></td>
</tr>
</tbody>
</table>

rules, though ranked differently. This suggests that the medical expert’s knowledge goal might have changed as his hypotheses changed \[90\]. In effect, the perceived functional novelty of the rules might have been modified by the changing knowledge goals. This modification then manifested in the form of different rankings of the same rules.
Table 4.11: Functional novelty evaluation obtained by mining CHD_DB data set for smoking↔endothelial progenitor cell levels

<table>
<thead>
<tr>
<th>Rank (top)</th>
<th>Association Rules</th>
<th>(Min\chi^2)</th>
<th>(\chi^2_{\text{lin}})</th>
<th>(CP_{\text{Ax}})</th>
<th>(\chi^2_{\text{lin},D})</th>
<th>(CP_{D})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>[chol_low=yes] ↔ [chd=yes]</td>
<td>283.63</td>
<td>32760.4</td>
<td>283.63</td>
<td>Yes</td>
<td>283.63</td>
</tr>
<tr>
<td>2.</td>
<td>[hypertrophy_left=yes] ↔ [chd=yes]</td>
<td>283.63</td>
<td>1423.23</td>
<td>Yes</td>
<td>283.63</td>
<td>Yes</td>
</tr>
<tr>
<td>3.</td>
<td>[cholRaised=yes] ↔ [chd=yes]</td>
<td>283.63</td>
<td>1836.06</td>
<td>Yes</td>
<td>283.63</td>
<td>Yes</td>
</tr>
<tr>
<td>4.</td>
<td>[cholBorder=yes] ↔ [chd=yes]</td>
<td>283.63</td>
<td>4680.42</td>
<td>Yes</td>
<td>283.63</td>
<td>Yes</td>
</tr>
<tr>
<td>5.</td>
<td>[edu_high_not_graduate=yes] ↔ [chd=yes]</td>
<td>283.63</td>
<td>13033.15</td>
<td>Yes</td>
<td>283.63</td>
<td>Yes</td>
</tr>
<tr>
<td>6.</td>
<td>[hpt_stage1=yes] ↔ [chd=yes]</td>
<td>283.63</td>
<td>2145.46</td>
<td>Yes</td>
<td>283.63</td>
<td>Yes</td>
</tr>
<tr>
<td>7.</td>
<td>[hpt_stage2=yes] ↔ [chd=yes]</td>
<td>283.63</td>
<td>2342</td>
<td>Yes</td>
<td>283.63</td>
<td>Yes</td>
</tr>
<tr>
<td>8.</td>
<td>[smoker=yes] ↔ [chd=yes]</td>
<td>283.63</td>
<td>311701.7</td>
<td>Yes</td>
<td>283.63</td>
<td>Yes</td>
</tr>
<tr>
<td>9.</td>
<td>[never_smoke=yes] ↔ [chd=yes]</td>
<td>283.63</td>
<td>46082.72</td>
<td>Yes</td>
<td>283.63</td>
<td>Yes</td>
</tr>
<tr>
<td>10.</td>
<td>[edu_high_graduate=yes] ↔ [chd=yes]</td>
<td>273.84</td>
<td>273.84</td>
<td>283.63</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>[cholRaised=yes] ↔ [hpt_stage1=yes]</td>
<td>0.37</td>
<td>1836.06</td>
<td>Yes</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>[smoker=yes] ↔ [hpt_stage1=yes]</td>
<td>0.37</td>
<td>311701.7</td>
<td>Yes</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>[chd=yes] ↔ [cholRaised=yes]</td>
<td>0.12</td>
<td>85334.93</td>
<td>Yes</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>[hpt_crisis=yes] ↔ [chd=yes]</td>
<td>0.12</td>
<td>85334.93</td>
<td>Yes</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td>[hpt_crisis=yes] ↔ [cholRaised=yes]</td>
<td>0.12</td>
<td>2145.46</td>
<td>Yes</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>16.</td>
<td>[hpt_stage2=yes] ↔ [cholRaised=yes]</td>
<td>0.12</td>
<td>2342</td>
<td>Yes</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>17.</td>
<td>[never_smoke=yes] ↔ [cholRaised=yes]</td>
<td>0.12</td>
<td>46082.72</td>
<td>Yes</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td>[chd=yes] ↔ [cholBorder=yes]</td>
<td>0.07</td>
<td>85334.93</td>
<td>Yes</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>19.</td>
<td>[hpt_stage2=yes] ↔ [cholBorder=yes]</td>
<td>0.07</td>
<td>2342</td>
<td>Yes</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>20.</td>
<td>[hpt_stage1=yes] ↔ [cholBorder=yes]</td>
<td>0.07</td>
<td>2145.46</td>
<td>Yes</td>
<td>0.07</td>
<td></td>
</tr>
</tbody>
</table>

Rank (bottom)
Table 4.12: Functional novelty evaluation obtained by mining CHD_DB data set for smoking→human anti murine antibody levels

<table>
<thead>
<tr>
<th>Rank (top)</th>
<th>Association Rules</th>
<th>$\text{Min} \chi^2$</th>
<th>$\hat{\chi}_{\text{ill},A}$</th>
<th>$CP_{A_S}$</th>
<th>$\hat{\chi}_{\text{ill},D}$</th>
<th>$CP_{D}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>[chd=yes]↔[edu_high_not_graduate=yes]</td>
<td>491.36</td>
<td>85335.00</td>
<td>Yes</td>
<td>491.36</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>[chd=yes]↔[edu_high_graduate=yes]</td>
<td>302.11</td>
<td>85335.00</td>
<td>Yes</td>
<td>302.11</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>[chol_low=yes]↔[chd=yes]</td>
<td>31.00</td>
<td>32760.00</td>
<td></td>
<td>31.00</td>
<td>Yes</td>
</tr>
<tr>
<td>4.</td>
<td>[hypertrophy_left=yes]↔[chd=yes]</td>
<td>31.00</td>
<td>1423.20</td>
<td>Yes</td>
<td>31.00</td>
<td>Yes</td>
</tr>
<tr>
<td>5.</td>
<td>[chol_raised=yes]↔[chd=yes]</td>
<td>31.00</td>
<td>1836.10</td>
<td>Yes</td>
<td>31.00</td>
<td>Yes</td>
</tr>
<tr>
<td>6.</td>
<td>[chol_border=yes]↔[chd=yes]</td>
<td>31.00</td>
<td>4680.40</td>
<td>Yes</td>
<td>31.00</td>
<td>Yes</td>
</tr>
<tr>
<td>7.</td>
<td>[edu_high_not_graduate=yes]↔[chd=yes]</td>
<td>31.00</td>
<td>13033.00</td>
<td></td>
<td>31.00</td>
<td>Yes</td>
</tr>
<tr>
<td>8.</td>
<td>[hpt_stage1=yes]↔[chd=yes]</td>
<td>31.00</td>
<td>2145.50</td>
<td>Yes</td>
<td>31.00</td>
<td>Yes</td>
</tr>
<tr>
<td>9.</td>
<td>[hpt_stage2=yes]↔[chd=yes]</td>
<td>31.00</td>
<td>2342.00</td>
<td>Yes</td>
<td>31.00</td>
<td>Yes</td>
</tr>
<tr>
<td>10.</td>
<td>[smoker=yes]↔[chd=yes]</td>
<td>31.00</td>
<td>311702.00</td>
<td>Yes</td>
<td>31.00</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.16</td>
<td>2145.50</td>
<td>Yes</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.04</td>
<td>85335.00</td>
<td>Yes</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.04</td>
<td>4680.40</td>
<td>Yes</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.04</td>
<td>1836.10</td>
<td>Yes</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.04</td>
<td>311702.00</td>
<td>Yes</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.03</td>
<td>85335.00</td>
<td>Yes</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.03</td>
<td>4680.40</td>
<td>Yes</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.03</td>
<td>1836.10</td>
<td>Yes</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.03</td>
<td>273.84</td>
<td>Yes</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.03</td>
<td>311702.00</td>
<td>Yes</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>

Rank (bottom)
Table 4.13: Functional novelty evaluation obtained by mining ECHO_MSCT data set for *matrix metalloproteinase→intracardiac thrombus*

<table>
<thead>
<tr>
<th>Rank (top)</th>
<th>Association Rules</th>
<th>$\text{Min} \chi^2$</th>
<th>$\hat{\lambda}_{hit, A}$</th>
<th>$CP_A$</th>
<th>$\hat{\lambda}_{hit, D}$</th>
<th>$CP_D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>[ca=yes] ↔ [age65-79yrs=yes]</td>
<td>440.68</td>
<td>440.68</td>
<td>Yes</td>
<td>793.36</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>[ca=yes] ↔ [age45-64yrs=yes]</td>
<td>440.68</td>
<td>440.68</td>
<td>Yes</td>
<td>504.54</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>[ca=yes] ↔ [male=yes]</td>
<td>344.57</td>
<td>440.68</td>
<td>Yes</td>
<td>344.57</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>[ca=yes] ↔ [female=yes]</td>
<td>267.45</td>
<td>440.68</td>
<td>Yes</td>
<td>267.45</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>[hypertension=yes] ↔ [ca=yes]</td>
<td>108.74</td>
<td>108.74</td>
<td>Yes</td>
<td>381.60</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>[hypertension=yes] ↔ [age65-79yrs=yes]</td>
<td>108.74</td>
<td>108.74</td>
<td>Yes</td>
<td>793.36</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>[hypertension=yes] ↔ [age45-64yrs=yes]</td>
<td>108.74</td>
<td>108.74</td>
<td>Yes</td>
<td>504.54</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>[hypertension=yes] ↔ [age19-44yrs=yes]</td>
<td>108.74</td>
<td>108.74</td>
<td>Yes</td>
<td>901.79</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>[hypertension=yes] ↔ [male=yes]</td>
<td>108.74</td>
<td>108.74</td>
<td>Yes</td>
<td>344.57</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>[hypertension=yes] ↔ [female=yes]</td>
<td>108.74</td>
<td>108.74</td>
<td>Yes</td>
<td>267.45</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>[dm=yes] ↔ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>1.24</td>
<td>Yes</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>[age45-64yrs=yes] ↔ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>31.72</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>[age65-79yrs=yes] ↔ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>0.38</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>[age19-44yrs=yes] ↔ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>210.04</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>[female=yes] ↔ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>12.58</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>[ca_fam_history=yes] ↔ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>25.98</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>[male=yes] ↔ [dm=yes]</td>
<td>0.01</td>
<td>0.01</td>
<td>2.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>[male=yes] ↔ [ca=yes]</td>
<td>0.01</td>
<td>0.01</td>
<td>381.6</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>[male=yes] ↔ [hypertension=yes]</td>
<td>0.01</td>
<td>0.01</td>
<td>51.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>[male=yes] ↔ [hypercholesterolaemia=yes]</td>
<td>0.01</td>
<td>0.01</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rank (bottom)
Table 4.14: Functional novelty evaluation obtained by mining ECHO_MSCT data set for von willebrand factor $\rightarrow$ intracardiac thrombus

<table>
<thead>
<tr>
<th>Rank (top)</th>
<th>Association Rules</th>
<th>$\min \chi^2$</th>
<th>$\chi^2_{H_0}$</th>
<th>$CP_{AX}$</th>
<th>$\chi^2_{H_0D}$</th>
<th>$CP_{yD}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>[ca=yes] $\leftrightarrow$ [age65-79yrs=yes]</td>
<td>746.01</td>
<td>746.01</td>
<td>Yes</td>
<td>793.36</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>[ca=yes] $\leftrightarrow$ [age45-64yrs=yes]</td>
<td>504.54</td>
<td>746.01</td>
<td>Yes</td>
<td>504.54</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>[dm=yes] $\leftrightarrow$ [age65-79yrs=yes]</td>
<td>482.88</td>
<td>482.88</td>
<td>Yes</td>
<td>793.36</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>[dm=yes] $\leftrightarrow$ [age45-64yrs=yes]</td>
<td>482.88</td>
<td>482.88</td>
<td>Yes</td>
<td>504.54</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>[age65-79yrs=yes] $\leftrightarrow$ [ca=yes]</td>
<td>381.60</td>
<td>1017.10</td>
<td>Yes</td>
<td>381.60</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>[age45-64yrs=yes] $\leftrightarrow$ [ca=yes]</td>
<td>381.60</td>
<td>1252.70</td>
<td>Yes</td>
<td>381.60</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>[male=yes] $\leftrightarrow$ [ca=yes]</td>
<td>381.60</td>
<td>709.39</td>
<td>Yes</td>
<td>381.60</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>[female=yes] $\leftrightarrow$ [ca=yes]</td>
<td>381.60</td>
<td>799.90</td>
<td>Yes</td>
<td>381.60</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>[dm=yes] $\leftrightarrow$ [ca=yes]</td>
<td>381.60</td>
<td>482.88</td>
<td>Yes</td>
<td>381.60</td>
<td>Yes</td>
</tr>
<tr>
<td>10.</td>
<td>[dm=yes] $\leftrightarrow$ [male=yes]</td>
<td>344.57</td>
<td>482.88</td>
<td>Yes</td>
<td>344.57</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>[hypertension=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>341.78</td>
<td>Yes</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>[ca=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>746.01</td>
<td>Yes</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>[dm=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>482.88</td>
<td>Yes</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>[age45-64yrs=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>1252.70</td>
<td>Yes</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>[age65-79yrs=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>1017.10</td>
<td>Yes</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>[male=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>709.39</td>
<td>Yes</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>[age19-44yrs=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>1144.80</td>
<td>Yes</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>[female=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>799.90</td>
<td>Yes</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>[ca_fam_history=yes] $\leftrightarrow$ [hypercholesterolaemia=yes]</td>
<td>0.15</td>
<td>210.76</td>
<td>Yes</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

Rank (bottom)
The $\text{Min} \chi^2$ appears to be useful for estimating and ranking rules’ functional novelty. This is evidenced by the consistent observations in which the candidate functionally novel rules always appeared among the top-10 rules and never among the bottom-10 rules. But consistent with Experiment I, $\text{Min} \chi^2$ had a low precision. It achieved 0.6 (6/10) precision in Table 4.11 and Table 4.12 and 0.1 (1/10) precision in Table 4.13 and Table 4.14. Its actual recall rate was not calculated due to the exclusion of the middle-ranked rules in this evaluation.

At the final evaluation step, all candidate functionally novel rules were selected and carefully analyzed by the medical expert. Rule no.8 in Table 4.11 and rule no.10 in Table 4.12 were excluded because the association $\text{Smoking} \leftrightarrow [\text{smoker=yes}]$ was considered obvious and not meaningful to the user.

Table 4.15 presents the results. Of 12 rules being evaluated, 2 rules were labeled Yes (16.67%), 2 rules were Probably (16.67%), 6 rules were No, as far as I know (50.00%), and 2 rules were Not specific enough (16.67%).

**Functionally novel rules: no.12 and no.2**

Inference no.12 and no.2 were labeled as Yes by the medical expert.

1. **Inference no.12**

Inference no.12 was the most interesting discovery according to the medical expert. First, each association which constituted the inference was medically valid and acceptable as it was supported by the existing body of medical knowledge. Next, it was widely accepted that $\text{diabetes mellitus}$ was a key risk factor of $\text{coronary arteriosclerosis}$ ([ca=yes]). It was also known that $\text{Von Willebrand factor}$ was associated with $\text{diabetes mellitus}$ ([dm=yes]). $\text{Coronary arteriosclerosis}$ was known to be associated with $\text{intracardiac thrombus}$. These associations implied an indirect relationship between $\text{Von Willebrand factor}$ and $\text{intracardiac thrombus}$.

The mechanism by which $\text{Von Willebrand factor}$ and $\text{intracardiac thrombus}$ were related was relatively unknown in the literature. A Pubmed search for ‘$\text{von willebrand factor AND intracardiac thrombus}$’ returned only 5 citations, and a Pubmed search for ‘$\text{Von Willebrand factor AND diabetes mellitus AND coronary arteriosclerosis AND intracardiac thrombus}$’ returned 0 citation.
Table 4.15: Medical expert’s final evaluation on the functional novel rules

<table>
<thead>
<tr>
<th>No.</th>
<th>A</th>
<th>rule ( x \leftrightarrow y ) (support/confidence)</th>
<th>D</th>
<th>Label</th>
<th>Comments</th>
<th>Pubmed count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Smoking</td>
<td>[hypertrophy_left=yes] ( \leftrightarrow ) [chd=yes] (0.079/0.745)</td>
<td>EPC lev.</td>
<td>No, as far as I know</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Smoking</td>
<td>[chol_raised=yes] ( \leftrightarrow ) [chd=yes] (0.292/0.594)</td>
<td>EPC lev.</td>
<td>Yes</td>
<td>'border is not a good term as in borderline male, borderline female, etc.</td>
<td>0</td>
</tr>
<tr>
<td>3.</td>
<td>Smoking</td>
<td>[chol_border=yes] ( \leftrightarrow ) [chd=yes] (0.144/0.448)</td>
<td>EPC lev.</td>
<td>Not specific enough</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Smoking</td>
<td>[hpt_stage1=yes] ( \leftrightarrow ) [chd=yes] (0.106/0.398)</td>
<td>EPC lev.</td>
<td>No, as far as I know</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Smoking</td>
<td>[hpt_stage2=yes] ( \leftrightarrow ) [chd=yes] (0.177/0.545)</td>
<td>EPC lev.</td>
<td>No, as far as I know</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Smoking</td>
<td>[hypertrophy_left=yes] ( \leftrightarrow ) [chd=yes] (0.079/0.745)</td>
<td>HAMA lev.</td>
<td>No, as far as I know</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>Smoking</td>
<td>[chol_raised=yes] ( \leftrightarrow ) [chd=yes] (0.292/0.594)</td>
<td>HAMA lev.</td>
<td>Probably</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Smoking</td>
<td>[chol_border=yes] ( \leftrightarrow ) [chd=yes] (0.144/0.448)</td>
<td>HAMA lev.</td>
<td>Not specific enough</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>Smoking</td>
<td>[hpt_stage1=yes] ( \leftrightarrow ) [chd=yes] (0.106/0.398)</td>
<td>HAMA lev.</td>
<td>No, as far as I know</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>Smoking</td>
<td>[hpt_stage2=yes] ( \leftrightarrow ) [chd=yes] (0.177/0.545)</td>
<td>HAMA lev.</td>
<td>No, as far as I know</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>MMP</td>
<td>[hypertension=yes] ( \leftrightarrow ) [ca=yes] (0.145/0.305)</td>
<td>Intracardiac thrombus</td>
<td>Probably</td>
<td>Still under research, as HAMA levels is surrogate for EPC levels</td>
<td>0</td>
</tr>
<tr>
<td>12.</td>
<td>VWF</td>
<td>[dm=yes] ( \leftrightarrow ) [ca=yes] (0.057/0.375)</td>
<td>Intracardiac thrombus</td>
<td>Yes</td>
<td>Still under research</td>
<td>0</td>
</tr>
</tbody>
</table>
The rule \([\text{dm=yes}]\leftrightarrow[\text{ca=yes}]\) is functionally novel to the medical expert because it was unknown that diabetes mellitus and coronary arteriosclerosis could mediate the relationship between Von Willebrand factor and intracardiac thrombus. How likely is this? Interestingly, from the five citations retrieved, a research article by Lip et al \([125]\) seemed to indicate a potential association between Von Willebrand factor and intracardiac thrombus amid through a different biological mechanisms.

2. **Inference no.2**

Inference no.2 was interesting to the medical expert because the relationship between smoking and endothelial progenitor cell levels was previously unknown. A Pubmed search for ’smoking AND endothelial progenitor cell levels’ returned 0 citation. But it was well known that smoking would increase the blood cholesterol level \([\text{cholRaised=yes}]\). It was also known that the raised cholesterol level would increase the risk of incurring coronary heart disease \([\text{chd=yes}]\) and that endothelial progenitor cell level was instrumental for predicting the coronary heart disease outcomes.

However, a Pubmed search for ’smoking AND cholesterol levels raised AND coronary heart disease AND endothelial progenitor cell levels’ returned 0 citation. Consequently, the rule \([\text{cholRaised=yes}]\leftrightarrow[\text{chd=yes}]\) was considered as functionally novel as it might suggest a probable association between the smoking habit and EPC levels. To the medical expert, the finding represents a potential area for the future medical research.

**Other rules**

Inference no.11 was labeled *Probably*. It was known that hypertension increased the likelihood of coronary arteriosclerosis \([\text{ca=yes}]\). Coronary arteriosclerosis was closely associated with intracardiac thrombus. The medical expert, however, was not completely sure of the association between matrix metalloproteinase and hypertension.

Inferences no.1,4,5,9,10 were labeled as ‘Still under research ..’ because the medical expert lacked the personal background knowledge to make a solid judgment on their truth or falsity. The inference no.3 and no.8 were not specific enough for a proper evaluation to be done due to the ambiguity on the meaning of the borderline cholesterol (see the comments).
4.5 Discussions

4.5.1 Discussions on results obtained in Experiment I

$\chi^2_{lit}$ consistently demonstrated the higher usefulness than $\chi^2$ in recalling the domain knowledge-compliant rules. $\chi^2_{lit}$'s high recall value is more important than its lack of precision because it is crucial to mine as many compliant rules as possible within $\text{minsupp}$ and $\text{minconf}$ constraints prior to the functional novelty evaluation to avoid losing the important rules. This finding confirms the initial intuition that incorporating more domain knowledge would increase the recall rate for the compliant rules.

The fact that $\mu$ constantly outperformed $\chi^2$ and $\chi^2_{lit}$ in precision at the top fifteen rules is in agreement with the well-accepted data mining assumption that the rules with high support and high confidence tend to be compliant with the existing knowledge. However, it is difficult to use $\mu$ in practice for two reasons. Firstly, $\mu$ score does not have a threshold value that can be used to determine the statistically significant association. Secondly, $\mu$ is a composite measure of support and confidence. But since the optimal $\text{minsupp}$ and $\text{minconf}$ thresholds for data mining are still difficult to determine [16], it also becomes difficult to determine the optimal threshold of $\mu$.

4.5.2 Discussions on results obtained in Experiment II

The results obtained from the second experiment show the contrast between the functional novelty and the traditional pairwise novelty. For example, from the pairwise perspective, rule [dm=yes] $\leftrightarrow$ [ca=yes] is not novel because diabetes mellitus is known as a common risk factor of coronary arteriosclerosis. However, the rule is functionally novel because the experimental results suggest its potential in mediating a previously unknown association between VMF and intracardiac thrombus.

Min$\chi^2$ appeared to be useful for estimating and ranking the rule functional novelty. All functionally novel rules appeared among the top-10 rules. However, like $\chi^2_{lit}$, it suffered from the poor precision and resulted in a larger number of rules that need to be evaluated by the user. The lack of precision also resulted in the less accurate rule ranking, e.g. rule no. 1 in Table 4.11.

Min$\chi^2$'s lack of precision could be rectified by increasing the relevance of the
literature query. This study has only used a simple Pubmed query statement in which the keywords were simply appended with AND operator. The problem is that some concepts, e.g. Adult, Smoking, and Female Gender, tend to have very high citation counts. These are common patient demographic attributes collected in most medical studies and do not necessarily imply any meaningful representation of the main subjects of studies. Hence, in such cases, the high citation counts can be dangerously misleading. One way to address the problem is to focus the keyword search on a particular section of the abstract texts, for instance Results or Conclusions section. This may help distinguish the true correlations from the false ones. Alternatively, one could also restrict the query specifically to the title or major MeSH heading fields. A few works [69, 59] showed that focusing the search on Pubmed’s MeSH major descriptors and utilizing the UMLS semantic relations may increase the relevance of the literature retrieval results. Pratt and Yestigen-Yildiz [42] limited the search for a concept co-occurrence in the article titles only. Demner-Fushman and Lin [126] demonstrated that a combination of the semantic-based and statistical techniques may increase the relevance of literature in response to the evidence-based clinical questions. These are beyond the scope of the current study.

The results also suggest that the medical expert’s interest seemed to be correlated with specific semantic types. A careful observation reveals that the majority of rules in Table 4.15 comprise of items that correspond to the Disease or Syndrome, Finding, or Laboratory and Test Results types. The other semantic types such as Age Group, Organism Attribute, or Population Group appear to be less significant to the user. The reason behind such tendency is worth studying in the future.

4.6 Chapter Summary

KELAM, a semi-automatic knowledge discovery architecture was proposed and implemented as a Web-based application. The FP-Growth class provided an efficient algorithm implementation for mining the association rules. The UMLS Metathesaurus and the Semantic Network provided the highly structured domain knowledge needed for implementing the item mapping and the semantic-based rule filtering. The content of Pubmed literature was remotely accessed by the application through the Entrez Utilities package to provide access to the vast amount
CHAPTER 4: IMPLEMENTATION AND RESULTS

of the unstructured medical domain knowledge useful for the rule functional novelty evaluation.

Two experiments involving an expert cardiologist and two cardiovascular data sets (CHD_DB and MSCT_ECHO) were conducted. Experiment I demonstrated that $\chi_{lit}^2$ exhibited the higher recall for the domain knowledge-compliant rules in comparison to $\chi^2$ when measured over 25 rules. Based on the four real medical hypotheses provided by the expert, Experiment II aimed at evaluating the functional novelty of the association rules generated from both data sets. All the rules considered to be functionally novel by the medical expert appeared among the top-10 rules. $\text{Min} \chi^2$ seemed to be a useful measure for automatically measuring and ranking the functional novelty.
Chapter 5

Conclusions

This work presents a study on the functional novelty as a new association rule interestingness criterion. It is hypothesized that a functionally novel rule is interesting to the user because it mediates the relationship between a pair of hypotheses whose relationship is previously unknown. A novel domain-driven KDD framework is proposed that (a) mines the association rules from the target data set with the minimum support and minimum confidence thresholds and (b) evaluates the functional novelty of each rule in the context of the most current domain knowledge synthesized from the literature, ontology, and user hypotheses. A proof-by-construction approach is adopted to test the research hypothesis. A software prototype, KELAM, is developed and its findings are evaluated in medical domain by an expert cardiologist from a local cardiac treatment center.

5.1 Recapitulation

The functional novelty is proposed as a non-pairwise rule novelty criterion. The novelty is not determined by the presence of the unknown combinations of rule items but by the rule’s ability to mediate the previously unknown relationships between pairs of user hypotheses. The functional novelty has the obvious advantages over the pairwise novelty. It allows a novel rule to be discovered even from among the common, high-support and high-confidence association rules. It reduces the rare item problem. It also increases user acceptance of the data mining output.

The most suitable knowledge discovery paradigm for evaluating the rule functional novelty is not known previously. This study proposes a domain-driven data mining
(D³M) paradigm. Its emphasis on incorporating the most up-to-date domain knowledge into the knowledge discovery process is likely to increase the reliability and validity of the rule novelty evaluation. Medical field has a wide array of domain knowledge sources. To ensure comprehensiveness, three knowledge sources were selected: (a) literature, (b) ontology, and (c) the medical expert’s background knowledge. The medical literature is accessible via Pubmed. The medical ontology is available via the UMLS Metathesaurus. The medical expert’s specific interest and background knowledge are incorporated into the rule evaluation in the form of a user hypothesis pair.

The association rules are generated using the FP-Growth algorithm. Minimum support (0.0001) and minimum confidence (0.00) thresholds are set to generate most of the rules. The maximum itemset is set at 2 to limit the analysis on the rules that have 2-itemset.

The rule evaluation is composed of three stages:

   Each item in the target data set is mapped to a medical concept in the UMLS Metathesaurus. This in turn determines the semantic type of each rule item. The rules are then filtered against the content of the UMLS Semantic Network. Semantically valid rules are the rules that have their antecedent and consequent items semantically related in the Semantic Network file. The invalid rules are excluded from the subsequent evaluation as they are assumed not to be useful to the user. The filtering method can reduce the number of rules to be evaluated by up to 30%.

2. Rule compliance evaluation.
   Rules that pass the filtering process are further evaluated for their compliance with the domain knowledge in the literature, determined by the $\chi^2_{lit}$ score. Compliance increases the acceptability of the final rule set outcome. To do this, a pair of the corresponding medical concepts are extracted for each rule antecedent and consequent. An online query is formulated to get the total number of term co-occurrence of both concepts in Pubmed. Using the number of term occurrence, the $\chi^2_{lit}$ score is calculated. Rules that have a $\chi^2_{lit}$ score satisfying the critical value threshold are considered compliant to the domain knowledge.
3. Functional novelty evaluation.

The medical expert prescribes a pair of medical hypothesis terms whose relationship is unknown and interesting to be explored. Each hypothesis term is mapped to the UMLS medical concepts. For each domain knowledge-compliant rule generated in the previous stage, the functional novelty is evaluated based on the rule’s strength of correlation with the expert’s hypotheses. The correlation strength is approximated by $\text{Min}\chi^2$ score.

Two experiments were conducted involving two real world cardiovascular data sets. A cardiologist supplies hypotheses to be tested and evaluates the knowledge discovery results. $\chi^2_{\text{lit}}$ demonstrates the higher recall rate in detecting the domain knowledge-compliant rules in comparison to the traditional $\chi^2$. Based on the medical expert’s feedback, $\text{Min}\chi^2$ is found to be useful for estimating the functional novelty of rules although it suffers from the lack of precision.

An interesting finding from the experiment shows how a common rule ($\text{diabetes mellitus} \leftrightarrow \text{coronary arteriosclerosis}$) seems to mediate a rare association between a biomarker called Von Willebrand Factor and the occurrence of intracardiac thrombus. The medical expert notes that the association is potential and interesting to be studied in the future.

5.2 Significance and Contributions

On the outset, the extreme imbalance has been observed between numerous published data mining algorithms on one side and the lack of the actionable KDD frameworks on the other side. This has motivated the major proponents of the domain-driven data mining [24][127] to advocate a paradigm shift from data mining to the knowledge discovery as originally envisioned by Fayyad et al [2]. Instead of focusing on the development of the new, exotic data mining algorithms, this paradigm shift has resulted in the increasing emphasis on producing the novel and effective KDD frameworks and workflows that capitalize on the existing well-proven data mining algorithms.

The important aspect of this new wave of research is the enhancement of the post-mining interestingness evaluation through the introduction of techniques capable of automatically incorporating the ubiquitous domain intelligence for
evaluating the quality of rules generated by existing data mining algorithms [127]. For example, Mostafa et al [128] envisioned that a fruitful research area in the future will be one in which the associations that are discovered from the clinical data warehouses will be validated by the evidence in the literature.

Within this context, the work presented in this dissertation offers two contributions:

1. The functional novelty as a useful criterion for determining the interestingness of 2-item association rules. The experiments involving the real medical data sets have given the evidence that functionally novel rules are highly interesting to the medical expert.

2. The novel KDD framework that automatically acquires the ubiquitous domain knowledge from the literature, ontology, and the user at a post-mining phase. The framework is technically interesting because the existing data mining algorithms such as the FP-Growth are naïve towards the semantics of a specific domain. They do not normally produce patterns or rules that are semantically enriched so as to enable the automatic inference by a knowledge-based system. Therefore, additional processings are required to make these patterns more actionable and understandable.

The further implications of these contributions are:

1. By emphasizing the discovery of domain knowledge-compliant rules, the proposed method is likely to produce medically acceptable rules.

2. Assuming the optimal minsupp – minconf threshold values for finding interesting rules remain unknown, the proposed method may alleviate the computational burden caused by the rare item problem. It allows a data miner to set relatively high minsupp and minconf thresholds while maintaining the possibility for discovering the novel rules.

Finally, on a wider data mining research landscape, works in finding useful hypothesis linkages via the interesting rules is important as it is developing in parallel with the increasing interest in the link mining research, e.g. [22]. This trend can be interpreted as representing the increasing realization among the researchers and scientists that a substantial amount of new and valuable knowledge remains hidden among the already known knowledge.
5.3 Limitations

The limitations of this study are:

1. This study evaluates the interestingness of the 2-item association rules. It does not cover the analysis on the >2-item rules. The critics of this study may content that the rare item problem is less severe in the case of 2-item rules even if the \textit{minsupp} and \textit{minconf} thresholds are set to be very low. This is fine because the main focus of the study is to provide the initial proof of concept for the functional novelty’s usefulness in evaluating an association rule.

2. The quality of a Pubmed query is determined by the choice of search terms and the inherent limitations of Pubmed’s search facility [129]. Formulating the optimal queries that precisely represent the user’s information needs is a classic problem in information retrieval research [130, 131, 132, 133]. Choosing overly-specific and obscure search terms has the tendency to produce too few articles. Using overly-general terms may produce higher number but less relevant articles. For example, the search on Pubmed for "smoking AND cholesterol levels raised" returned 278 articles and the search for "smoking and increased cholesterol levels" returned 1,873 articles (results as per December 5, 2011) although "raised" and "increased" seem to suggest the same meaning, at least in the layman term.

Query expansion has long been proposed to address retrieval problems associated with the presence of synonyms. WordNet links English nouns, verbs, adjectives, and adverbs to sets of synonyms that are in turn linked through semantic relations that determine word definitions [134, 135]. This way, the more specific concepts inherit information from their more general concepts, thus enabling the automatic expansion of user queries [136]. Less-detailed user queries could be expanded and their retrieval effectiveness enhanced when WordNet was used [137].

In comparison to generic thesauri, query expansion is likely to yield more relevant results when domain-specific thesauri are used as the basis for the expansion [138, 139]. As a biomedical thesaurus, UMLS Metathesaurus’ primary objective is to facilitate different ways of saying the same things between humans and computer programs in the biomedical domain [140].
Chapter 5: Conclusions

is achieved by clustering synonymous terms from more than 140 biomedical source vocabularies (e.g. MeSH, SNOMED CT, ICD) into a concept [141]. Of these synonyms, a preferred term is chosen to represent the concept. The content of Metathesaurus is carefully crafted by human experts to effectively represent users’ information needs [142].

Unfortunately, there has been no universal agreement whether Metathesaurus-based query expansion actually benefits information retrieval. Contradictory results have been reported concerning the information retrieval performance following applications of Metathesaurus-based query expansions [143, 144]. However, there is a common finding found in both these studies. Aronson and Rindflesch [143] observed that the query expansion is only effective when it can be used with MeSH-indexed documents and Hersh et al [144] reported that MeSH terms led to a 10% improvement in information retrieval performance. It appears, therefore, that the improvement in information retrieval is not due to the query expansion mechanism but due primarily to the unique connection between the UMLS Metathesaurus terms (which are mapped to the user’s query) and the MeSH terms with which the biomedical literature is indexed. This makes sense considering that the MeSH vocabulary is a subset of the Metathesaurus.

Therefore, based on these considerations, the most reasonable approach adopted in this study is to manually search and carefully select the most appropriate Metathesaurus term for representing each attribute in the target data set. No query expansion is applied. For example, the term “cholesterol levels raised” is derived directly from the UMLS vocabulary (concept ID C0848569). See Figure 4.17 in the previous chapter. Using these terms, Pubmed queries are subsequently formulated. This approach capitalizes on the apparent correspondence between Metathesaurus terms and MeSH index terms assigned to each Pubmed article to produce highly relevant query results.

3. The ability to map the rule items and hypothesis terms to the standard medical concepts is limited by the breadth of the UMLS Metathesaurus’ contents. The experience during the data preparation stage shows that the domain cut-off values often need to be acquired from non-ontological resources such as the American Heart Association Guideline.
Chapter 5: Conclusions

4. This study does not aim to address the efficiency problem of the association rule mining algorithm. The data mining efficiency is limited by the efficiency of $FP$-$Growth$ implementation.

5. This study only involves one medical expert due to time constraint and the substantial effort required for evaluating the rules. But this limitation is compensated by selecting the medical expert who had excellent expertise and substantial research experience. For comparison, Ohsaki et al [76] only involved two medical experts in their experiments.

5.4 Future Works

The future works in the following areas should be considered:

1. Evaluating the functional novelty of $\geq$2-item rules is important as it is when the rare item problem becomes more severe. The problem is challenging because different rule evaluation techniques will be required to deal with the higher number of rule antecedents and consequents. On the other hand, one should be aware that increasing the size of the rule itemset will increase the difficulty of the rule evaluation faced by the medical experts and will eventually reduce the rules’ usefulness.

2. The relevance of the current semantic-based filtering results can be enhanced by creating a discriminative filtering mechanism based on different types of semantic relations defined in the UMLS Semantic Network. For instance, a user investigating the causal relations between a hypothesis pair may find an association rule that implies a causal relationship between its antecedent and consequent to be quite interesting. Hence the future system’s filtering can include only rules whose items are associated via the relation $causes$ or $produces$.

3. Finding a coherent chain of inference has recently been recognized as an important research topic in data mining [22]. Many relations can possibly exist between two information objects. Unless they possess a high degree of coherence, these relations can be misleading. Instead of relying on $\chi^2_{it}$ and $\min \chi^2$, one may investigate the use of other established term analysis
Chapter 5: Conclusions

...techniques, such as term frequency-inverse document frequency (TF-IDF) or co-word clustering.

4. In medicine, it is often equally important to analyze the relationships between negative co-occurrences. Mining the association rules that include the negative co-occurrences of items goes beyond the capability of the traditional association rules algorithms. The future works in this area is scientifically and practically relevant and will eventually increase the usefulness of the existing medical knowledge discovery systems.

5.5 Conclusions

In conclusion, this dissertation addresses an important research question of whether the novelty of association rules can be evaluated beyond the use of the traditional pairwise novelty criterion. This is successfully answered through the proposal of a new non-pairwise novelty criterion called functional novelty. By designing and implementing a novel medical knowledge discovery framework, this study demonstrates how functionally novel association rules can be discovered from the real cardiovascular data sets. The new framework successfully incorporated the medical domain knowledge from the literature, ontology, and user hypotheses into the post-mining rule evaluation process. Based on the medical expert’s feedback, the proposed evaluation technique and measures proved to be useful for estimating the candidate functionally novel rules. The outcome of this work is expected to become the first step towards medical knowledge discovery systems that can effectively aid the medical researchers in rapidly testing and validating the potential medical hypotheses at the initial stages of the medical discovery process.
Appendix

6.1 RapidMiner

RapidMiner is an open-source data mining software package that offers a diverse range of data mining algorithms. These algorithms are implemented in the form of a set of ‘operators’ arranged under an operator tree that models the knowledge discovery process [108]. The RapidMiner Community Edition is downloaded for free from its website. Because RapidMiner provides a Java class library of data mining algorithms, it can be easily integrated into the Java-based applications.

In this work, the association rules were generated in two steps via two RapidMiner operators. First, the FPGrowth operator generated frequent itemsets based on the minsupp and maxitem thresholds. These itemsets were subsequently passed to the AssociationRuleGenerator operator to construct association rules that satisfied the minconf threshold.

6.2 Ontology

6.2.1 Unified Medical Language System (UMLS)

The Unified Medical Language System is a collection of medical ontologies made available for public use by the United States National Library of Medicine (NLM) [1]. The two ontological components of the UMLS used in this study are the Metathesaurus and the Semantic Network.

Metathesaurus is a comprehensive collection of the standard medical vocabularies containing the medical concepts and their naming variations as well as the semantic relationships among them. Each concept is uniquely identified by the Concept Unique Identifier (CUI) code and belongs to a specific semantic type. Figure 6.1 shows the Metathesaurus’ output where Diabetes is the standard medical concept for diabetes illness. The CUI C0011847 is a unique identifier of the concept and Disease or Syndrome indicates its semantic type. Its two naming variants, Diabetes and Diabete, are derived from two different medical vocabulary sources: the Metathesaurus Names and the Psychological Index Terms.

The Semantic Network defines 135 semantic types for categorizing all concepts in the Metathesaurus. Fifty-four relations exist among these semantic types. The Semantic Network arranges the semantic types and their relations in a parent-child hierarchical model. In Figure 6.2, the semantic type Biologic Function has two children types: Physiologic Function and Pathologic Function. Subsequently, for each child type more types can be derived.
Figure 6.3 shows how semantic relations ‘manages’, ‘treats’, ‘disrupts’, ‘complicates’, ‘interacts_with’, and ‘prevents’ are derived from the parent relation ‘affects’.

**Figure 6.3:** A partial depiction of the hierarchies among the semantic relations in the UMLS Semantic Network

Figure 6.4 partially depicts relations among some semantic types that exist in the Semantic Network.

**Figure 6.4:** A partial depiction of the semantic relations among several semantic types

Table 6.1 shows how, for example, semantic type Disease or Syndrome and Organism Attribute are semantically related via the relation associated_with.

In this study, Metathesaurus is used to standardise medical terms found in the user hypotheses, data set attributes, and literatures. The standardisation helps solve the ambiguities that often exist between two different medical terms having a similar meaning. For example, the term High Blood Pressure and Hypertension can be standardised into a concept Hypertensive Disease (Disease or Syndrome). Hence the computer can reason that both terms actually have the same meaning.
Table 6.1: A partial depiction of the UMLS Semantic Network content

...  
Disease or Syndrome, associated_with, Organism Attribute  
Disease or Syndrome, associated_with, Clinical Attribute  
Disease or Syndrome, isa, Biologic Function  
...  
Organism Function, affects, Human  
...  

Finally, the Semantic Network was used as the basis for the semantic-based rule filtering. Because the Semantic Network defines the relations that may exist and may not exist between two semantic types, the filter can exclude the illogical association rules from the final knowledge discovery output.

MetaMap Transfer (MMTx)

MetaMap automatically maps biomedical texts to the standard concepts in the UMLS Metathesaurus. MMTx, a Java-implementation of MetaMap, is integrated into KELAM. It is available free-of-charge by the National Centre of Biomedical Communications (NCBI)².

MMTx’s mapping process involves the initial parsing of texts into the simple noun phrases, generating the lexical variants for each phrase, retrieving and evaluating a set of mapping candidates, and producing one or more final mapping results. For more complete explanations on its algorithm, refer to [145].

An early analysis on MetaMap’s performance against human’s in identifying the biomedical concepts from texts [42] demonstrated that MetaMap was excellent in identifying the common biomedical concepts from free texts, with a 93.3% recall and a 84.5% precision. It was noted that MetaMap’s high recall depends on the coverage of the UMLS concept vocabularies. As these vocabularies continue to expand over the years, it is plausible to expect its recall rate to gradually increase.

The performance of MMTx has also been studied and compared with the performance of MetaMap. Hliaoutakis et al [121] observed that MMTx’s over-generation of the term variants led to its low precision. Its usage of Metathesaurus terms for the term extraction also may reduce its accuracy in retrieving MEDLINE literature which are

indexed differently using the MeSH terms. Divita et al. concluded that, overall MetaMap mapped free texts to the UMLS concepts better than MMTx did. However, MMTx has the advantage over MetaMap as it can be easily integrated into another Java application. For this reason, this study chose to use MMTx instead of MetaMap.

### 6.3 Pubmed Literature

Pubmed is maintained by NCBI as a public Web search engine for accessing MEDLINE database. MEDLINE database is the NLM’s main bibliographic databases that contains biomedical journal citations and abstracts encompassing medicine, nursing, dentistry, veterinary medicine, health care system, and preclinical practices. Currently MEDLINE indexes 5,200 journals published in the U.S. and 80 journals from other countries. MEDLINE is the largest database component offered by Pubmed.

![Figure 6.5: An example of Pubmed citation record](image)

Each citation indexed by MEDLINE is assigned with one or more Medical Subject

---

Headings (MeSH). MeSH is a controlled vocabulary thesaurus developed at the NLM. Human indexers assign the MeSH term(s) to each MEDLINE citation to indicate its relevant subject(s). To determine the major focus of an article, an indexer may assign one or more MeSH terms as the Major Topic headings. Hence, MeSH can be used to determine the objective relevance of a citation for a particular topic and is useful for narrowing the focus of the literature search.

Figure 6.5 above shows a sample citation from a search result for ‘congestive heart failure’. PMID 18666366 (PMID stands for Pubmed Identifier) uniquely identifies this citation. It also displays the article’s titles, abstract texts, a list of authors, journal titles, date of publication, Medical Subject Headings (MeSH), and other essential information. Although useful, having the literature citations in this type of format is not suitable for computational manipulation intended by this work.

Figure 6.6 presents the details of the same citation but annotated with the eXtensible Markup Language (XML) tags by Pubmed. Having the information in this machine-readable form makes it easy to be manipulated by a computer application. Each piece of the citation information can be automatically extracted by referring to the relevant XML tags.

\[\text{http://www.nlm.nih.gov/mesh/}\]
6.3.1 Entrez Utilities

To automatically retrieve the Pubmed search results in an XML format, KELAM uses the Entrez Programming Utilities (E-utilities). E-utilities are developed by the NCBI for the Java environment and consist of eight utilities providing various functionalities. Publicly accessible in the form of Java classes, these utilities post HyperText Transfer Protocol (HTTP) requests to the NCBI server by constructing the appropriate URL strings. The XML output is similar to Figure 6.6. This work employs three utilities:

1. **EInfo** retrieves the latest total Pubmed citation count. Sample HTTP request: 

2. **ESearch** retrieves a list of citation PMIDs corresponding to Pubmed search terms. Sample HTTP request: 

3. **EFetch** retrieves citation details based on a list of citation PMIDs previously retrieved with ESearch. Sample HTTP request: 

6.4 Java

Java was chosen as the application development environment for two reasons. Firstly, through its Java Virtual Machine, the application becomes highly portable across different computing platforms. KELAM had been successfully tested on the Windows XP, Windows Vista, and Linux RedHat. Secondly, the RapidMiner’s FP-Growth implementation and the MMTx are all available in the form of Java classes. Implementing KELAM as a Java application makes it easy to integrate with these software components. KELAM is implemented as a Web-based application with Java Server Pages (JSP) technology, supported by the Microsoft SQL Server database.

---

8http://www.java.com/en/
References


REFERENCES


REFERENCES


REFERENCES


