PERSONALISED ONTOLOGY LEARNING AND MINING FOR WEB INFORMATION GATHERING

By

Xiaohui Tao
B.IT.(Honours) QUT

August 2009

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY AT QUEENSLAND UNIVERSITY OF TECHNOLOGY BRISBANE, AUSTRALIA
Copyright © Xiaohui Tao. All rights reserved.

x.tao@qut.edu.au

Permission is herewith granted to Queensland University of Technology, Brisbane to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions. The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author’s written permission.
Dedicated to my wife Yunyan Liao for without her love and support this thesis would not have been possible.
Keywords

Abstract

Over the last decade, the rapid growth and adoption of the World Wide Web has further exacerbated user needs for efficient mechanisms for information and knowledge location, selection, and retrieval. How to gather useful and meaningful information from the Web becomes challenging to users. The capture of user information needs is key to delivering users’ desired information, and user profiles can help to capture information needs. However, effectively acquiring user profiles is difficult.

It is argued that if user background knowledge can be specified by ontologies, more accurate user profiles can be acquired and thus information needs can be captured effectively. Web users implicitly possess concept models that are obtained from their experience and education, and use the concept models in information gathering. Prior to this work, much research has attempted to use ontologies to specify user background knowledge and user concept models. However, these works have a drawback in that they cannot move beyond the subsumption of super- and sub-class structure to emphasising the specific semantic relations in a single computational model. This has also been a challenge for years in the knowledge engineering community. Thus, using ontologies to represent user concept models and to acquire user profiles remains an unsolved problem in personalised Web information gathering and knowledge engineering.

In this thesis, an ontology learning and mining model is proposed to acquire user profiles for personalised Web information gathering. The proposed computational model emphasises the specific is-a and part-of semantic relations in one
computational model. The world knowledge and users’ Local Instance Repositories are used to attempt to discover and specify user background knowledge. From a world knowledge base, personalised ontologies are constructed by adopting automatic or semi-automatic techniques to extract user interest concepts, focusing on user information needs. A multidimensional ontology mining method, Specificity and Exhaustivity, is also introduced in this thesis for analysing the user background knowledge discovered and specified in user personalised ontologies. The ontology learning and mining model is evaluated by comparing with human-based and state-of-the-art computational models in experiments, using a large, standard data set. The experimental results are promising for evaluation.

The proposed ontology learning and mining model in this thesis helps to develop a better understanding of user profile acquisition, thus providing better design of personalised Web information gathering systems. The contributions are increasingly significant, given both the rapid explosion of Web information in recent years and today’s accessibility to the Internet and the full text world.
## Contents

Keywords vii

Abstract ix

List of Figures xvii

List of Tables xviii

Terminology, Notation, and Abbreviations xix

Statement of Original Authorship xxiii

Acknowledgements xxv

1 Introduction 1

1.1 Introduction to the Study 1

1.2 Research Questions and Significance 5

1.3 Research Methods and Thesis Outline 7

1.4 Previously Published Papers 9

2 Literature Review 11

2.1 Web Information Gathering 11

2.1.1 Web Information Gathering Challenges 11

2.1.2 Keyword-based Techniques 13

2.1.3 Concept-based Techniques 16
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2</td>
<td>Web Personalisation</td>
<td>26</td>
</tr>
<tr>
<td>2.2.1</td>
<td>User Profiles</td>
<td>26</td>
</tr>
<tr>
<td>2.2.2</td>
<td>User Information Need Capture</td>
<td>30</td>
</tr>
<tr>
<td>2.3</td>
<td>Ontologies</td>
<td>34</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Ontology Definitions</td>
<td>34</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Ontology Learning</td>
<td>36</td>
</tr>
<tr>
<td>2.4</td>
<td>Summary and Conclusion</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>Ontology-based Personalised Web Information Gathering</td>
<td>43</td>
</tr>
<tr>
<td>3.1</td>
<td>Concept-based Web Information Gathering Framework</td>
<td>44</td>
</tr>
<tr>
<td>3.2</td>
<td>Summary</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>Preliminary Study</td>
<td>49</td>
</tr>
<tr>
<td>4.1</td>
<td>Design of the Study</td>
<td>49</td>
</tr>
<tr>
<td>4.2</td>
<td>Semantic Analysis of Topic</td>
<td>50</td>
</tr>
<tr>
<td>4.3</td>
<td>Acquiring User Profiles</td>
<td>51</td>
</tr>
<tr>
<td>4.4</td>
<td>Experiments and Results</td>
<td>53</td>
</tr>
<tr>
<td>4.5</td>
<td>Summary and Conclusion</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>Ontology Learning for User Background Knowledge</td>
<td>61</td>
</tr>
<tr>
<td>5.1</td>
<td>World Knowledge Base</td>
<td>61</td>
</tr>
<tr>
<td>5.1.1</td>
<td>World Knowledge Representation</td>
<td>62</td>
</tr>
<tr>
<td>5.1.2</td>
<td>World Knowledge Base Construction</td>
<td>63</td>
</tr>
<tr>
<td>5.1.3</td>
<td>World Knowledge Base Formalisation</td>
<td>77</td>
</tr>
<tr>
<td>5.2</td>
<td>Taxonomy Construction for Ontology Learning</td>
<td>81</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Semi-automatic Ontology Taxonomy Construction</td>
<td>84</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Automatic Taxonomy Construction</td>
<td>89</td>
</tr>
<tr>
<td>5.3</td>
<td>Ontology Formalisation</td>
<td>92</td>
</tr>
<tr>
<td>5.4</td>
<td>Summary and Conclusion</td>
<td>93</td>
</tr>
</tbody>
</table>
List of Figures

1.1 A User Concept Model ........................................ 4
1.2 Research Methodology and Thesis Structure ............... 8
3.1 Concept-based Web Information Gathering Framework .... 44
4.1 The Google Performance ...................................... 55
4.2 The Experiment Dataflow in the Preliminary Study ........ 57
4.3 The Experimental Results in Preliminary Study ............ 58
5.1 Raw Data in the MARC 21 Format of LCSH. .................. 65
5.2 An Authority Record in MARC 21 Data ...................... 65
5.3 Parsing Result of a MARC 21 Authority Record ............ 70
5.4 Subjects and Cross References .............................. 76
5.5 The Library of Congress Classification Web ................ 78
5.6 The World Knowledge Base ................................... 82
5.7 Ontology Learning Environment .............................. 85
5.8 A Constructed Ontology ...................................... 88
6.1 An Information Item in the QUT Library Catalogue ........ 101
6.2 Mappings of Subjects and Instances ....................... 103
6.3 Discovering Potentially Interesting Knowledge ............ 107
6.4 Interesting Concepts Discovery Phases ...................... 113
7.1 The Experiment Framework ................................. 122
7.2 Topic Distribution in RCV1 Corpus .......................... 125
7.3 A Sample Document in RCV1 Corpus ......................... 127
7.4 Word Distribution in RCV1 Corpus .......................... 129
7.5 A TREC-11 Filtering Track Topic ............................. 129

8.1 The 11SPR Experimental Results ............................ 148
8.2 The MAP and $F_1$ Measure Experimental Results .......... 151
8.3 Percentage Change in Topics (Ontology-I vs. Manual) ...... 158
8.4 Percentage Change in Topics (Ontology-II vs. Manual) ..... 158
8.5 Percentage Change in Details (Ontology-I vs. Auto) ........ 170
8.6 Percentage Change in Details (Ontology-II vs. Auto) ...... 171
8.7 Average Percentage Change (Ontology-I vs. Ontology-II) .. 177


List of Tables

5.1 Comparison with Taxonomies in Prior Works ...................... 63
5.2 The Reference of MARC 21 Authority Record Leaders ............ 67
5.3 Subject Identity and References ........................................ 72
5.4 Types of Subjects Referred by Variable Fields .................... 72

8.1 The Mean Average Precision Experimental Results ................. 150
8.2 The Average Percentage Change Results ............................... 151
8.3 The Student’s Paired T-Test Results ................................... 151
8.4 The Macro $F_1$ Measure Experimental Results .................... 154
8.5 The Micro $F_1$ Measure Experimental Results ...................... 155
8.6 Comparisons Between the Ontology-I Model and Others .......... 159
8.7 Comparison of the size of Ontology-I and Manual User Profiles
   (MAP Results) .......................................................... 160
8.8 Comparison of the size of Ontology-II and Manual User Profiles
   (MAP Results) .......................................................... 161
8.9 Comparisons Between the Ontology Models and the Semi-auto Model164
8.10 User Concept Model Specified in the Semi-auto Model for Topic 101166
8.11 Comparisons Between the Ontology Models and Auto Model ... 169
8.12 Comparison of the size of Ontology-I and Auto User Profiles (MAP
   Results) .............................................................. 171
8.13 Comparison of the size of Ontology-II and Auto User Profiles
   (MAP Results) .......................................................... 172
8.14 Comparisons Between the Ontology-I and Ontology-II Models .. 176
Terminology, Notation, and Abbreviations

Terminology

Document
Text documents consisting of terms.

Exhaustivity
The extent of semantic meaning covered by a subject that deals with the topic

Is-a
Relations describe the situation that the semantic extent referred by a hyponym is within that of its hypernym.

Local Instance Repository
A user’s personal information collection, such as user created and stored documents, browsed Web pages and compiled/received emails, etc.

Part-of
Relations define the relationship between a holonym subject denoting the whole and a meronym subject denoting a part of, or a member of, the whole.

Query
The data structure given by a user to information gathering systems for the expression of an information need.
Related-to Relations are for two topics related in some manner other than by hierarchy.

Specificity The focus of a subject’s semantic meaning on a given topic.

Topic The topic statement of a user information need.

World knowledge Commonsense knowledge acquired by people from experiences and education.

Notation

$LIR$ A user’s Local Instance Repository.

$O$ An ontology.

$r$ A semantic relation.

$\mathbb{R}$ A set of semantic relations, in which each element is a relation $r$.

$s$ A subject.

$\mathbb{S}$ A set of subjects, in which each element is a subject $s$.

$\mathbb{T}$ A subset of subject set $\mathbb{S}$.

$T$ A topic as the semantic meanings of an information need.

$WKB$ The world knowledge base consisting of $\mathbb{S}$ and $\mathbb{R}$.

Abbreviations

DDC Dewey Decimal Classification

IGS Information Gathering System

LCC Library of Congress Classification
<table>
<thead>
<tr>
<th>Abbr.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCSH</td>
<td>Library of Congress Subject Headings</td>
</tr>
<tr>
<td>LIR</td>
<td>Local Instance Repository</td>
</tr>
<tr>
<td>ODP</td>
<td>Open Directed Project</td>
</tr>
<tr>
<td>OLE</td>
<td>Ontology Learning Environment</td>
</tr>
<tr>
<td>QUT</td>
<td>Queensland University of Technology</td>
</tr>
<tr>
<td>RCV1</td>
<td>The Reuters Corpus Volume 1</td>
</tr>
<tr>
<td>TREC</td>
<td>Text REtrieval Conference</td>
</tr>
</tbody>
</table>
Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signed: ___________________________ Date: ___________________________
From the start of my doctoral program to the completion of my dissertation, I have gone through a long journey. Throughout that journey I received both direct and indirect support from my supervisors, colleagues, friends, and family, all of whom I would like to thank.

I would like to express my sincerest thanks to my supervisor, Associate Professor Yuefeng Li, for his very generous contribution of time, expertise, and guidance not only through my academic career but also through my life as a personal friend. I also thank my associate supervisors, Dr. Richi Nayak and Professor Ning Zhong (external), for their support and advice. Special thanks also go to Peter Bruza, Taizan Chan, Shlomo Geva, and Yue Xu, for their valuable comments and opinions about my research work.

Likewise, I owe gratitude to the Library of Congress and Queensland University of Technology Library, for authorising the use of MARC and catalogue records in my doctoral research. Also, I would like to thank the staff of the School of Information Technology and the Library at QUT. Specifically I would like to thank Mark Carry-Smith, Patrick Delaney, John King, Jon Peak, Alan Woodley, Sheng-Tang Wu, Wanzhong Yang, and Xujuan Zhou, for their best support, understanding, and greatly appreciated friendship throughout my PhD journey.

I would also like to acknowledge Jennifer Beale for her valuable assistance in proofreading and correcting the English of this dissertation.

Last, but definitely not least, words alone cannot express my thanks to Yunyan
Liao, my wife, for her love, encouragement, and support throughout this long and difficult journey. Without her constant support, I could never have completed this work.

Yours sincerely,

Xiaohui Tao
August 2009
Chapter 1

Introduction

1.1 Introduction to the Study

In recent decades, the amount of Web information has exploded rapidly. How to gather useful information from the Web has recently become a challenging issue for all Web users. Many information retrieval systems for Web information gathering have been developed to attempt to solve this problem, resulting in great achievements. However, there is still no complete solution to the challenge [33].

The current Web information gathering systems cannot satisfy Web search users, as they are mostly based on keyword-matching mechanisms and suffer from the problems of information mismatching and information overloading [110]. Information mismatching means valuable information is being missed in information gathering. This usually occurs when one search topic has different syntactic representations. For example, “data mining” and “knowledge discovery” refer to the same topic, discovering knowledge from raw data collections. However, by the keyword-matching mechanism, documents containing “knowledge discovery” may be missed if using the query “data mining” to search. The other problem, information overloading, usually occurs when one query has different semantic
meanings. A common example is the query “apple”, which may mean apples, the fruit, or iMac computers. By using the query “apple” for the information need “apple, the fruit”, the search results may be mixed with useless information, for example, that about iMac computers [109,110]. Thus, if user information needs could be better captured and interpreted, say, if it is clear that a user needs information about “apples, the fruit” but not “iMac computers”, more useful and meaningful information can be gathered for the user. Therefore, there exists a hypothesis that if user information needs can be captured and interpreted, more user useful and meaningful information can be gathered.

Capturing user information needs through a given query is extremely difficult. In most Web information gathering cases, users provide only short phrases in queries to express their information needs [191]. Also, Web users formulate their queries differently because of different personal perspectives, expertise, and terminological habits and vocabularies. These differences cause the difficulties in capturing user information needs. Thus, the capture of user information needs requires the understandings of users’ personal interests and preferences. User profiles are widely used, in personalised Web information gathering, for user information need capturing and user background knowledge understanding [88].

However, acquiring user profiles is difficult. A great challenge is how to distinguish the topic-relevant concepts from those that are non-relevant. One example is the topic “Economic espionage”, created by the TREC linguists*:

What is being done to counter economic espionage internationally?

which is narrated as:

1.1. Introduction to the Study

Documents which identify economic espionage cases and provide action(s) taken to reprimand offenders or terminate their behaviour are relevant. Economic espionage would encompass commercial, technical, industrial or corporate types of espionage. Documents about military or political espionage would be irrelevant.

For the topic, various relevant and non-relevant concepts may be manually specified based on the description and narrative; these are illustrated as Figure 1.1. An assumption can arise that Web users implicitly possess a concept model consisting of such relevant and non-relevant concepts obtained from their background knowledge, and use the model in information gathering [110, 203]. Although such user concept models cannot be proven in laboratories, they may be observed in daily life. Web users can easily determine whether or not a document is interesting to them when reading through the document content. Their judgements are supported by an implicit concept model like Figure 1.1, which Web users may not easily describe clearly and explicitly. If user concept models can be specified in user profiles, user information needs can be better captured, thus more useful, meaningful, and personalised information can be gathered for Web users.

However, such topic relevant and non-relevant concepts are difficult for computational systems to specify. The manual concept specification is an implicit process in the human mind and is difficult to simulate clearly. Thus, user profile acquisition is challenging in information systems.

Ontologies, as a formal description and specification of knowledge, are utilised by many researchers to represent user profiles. Li and Zhong [110] used interesting patterns discovered from personal text documents to learn ontologies for user profiles. Some groups like [55, 181, 182] learned personalised ontological user profiles adaptively from user browsing history through online portals to specify user background knowledge. However, the knowledge described in these ontologies is constructed, based on the structures in a subsumption manner of super-class and sub-class relations, which is unspecific and incomplete.

Emphasising the complete, specific semantic relations in one computational
model is difficult. The relationships held by a super-class and its sub-classes could be differentiated to various specific semantic relations. A terminological ontology developed in the 1990s, named WordNet, has specification of synonyms (related-to), hypernyms/hyponyms (is-a), holonyms/meronyms (part-of), troponyms, and entailments for the semantic relations existing amongst the synsets and senses [49]. Some researchers claimed that WordNet contributed to the improvement of their information gathering models [130, 131, 241]. However, some others reported that WordNet could not provide constant and valuable support to information gathering systems, and argued that the difficulty of semantic relations handling was one of the downside of using WordNet [212]. Hence, some works attempted to focus on only one specific and basic semantic relation, such as is-a by [21,23,167,178], part-of by [58,59,164,169], and related-to by [71,205]. However, for the basic semantic relations of is-a, part-of, and related-to, there has not been any research work that could emphasise them in one single computational model and evaluate their impact to the associated concepts. This is a challenging issue, and has not been successfully solved by existing knowledge work.
1.2 Research Questions and Significance

The previous section in this chapter demonstrates that the acquisition of user profiles is challenging in personalised Web information gathering, and the difficulties in such user profile acquisition are the extraction and specification of the topic-related concepts. These problems yields a demand for a holistic exploration of using ontologies to acquire user profiles effectively.

This thesis aims to address these problems by exploring an innovative approach that learns and mines personalised ontologies for acquiring user profiles. The exploration contributes to better designs of personalised Web information gathering systems, and assists Web users to more effectively find personalised information on a topic.

The research questions for this thesis study are outlined as follows:

1. How can user background knowledge on a topic be discovered effectively?

2. How can the specific and complete semantic relations existing in the concepts be specified clearly?

3. How can user profiles can be acquired to capture user information needs, according to the user background knowledge discovered and semantic relations specified?

In order to find answers to these questions, surveys of Web information gathering, Web personalisation, and ontologies are performed. Based on the survey results, scientific research is also performed to address the problems in user profile acquisition. In this research, general Web users with information needs are the user group in focus, the full text documents are the focused Web information, and the user profiles attempted to be acquired are the routing user profiles that are kept static in Web information gathering.

In this thesis, an ontology learning and mining model that answers the previous research questions and acquires user profiles using personalised ontologies is
Chapter 1. Introduction

proposed. In this attempt to discover the background knowledge of Web users, a world knowledge base and user Local Instance Repositories (LIRs) are used in the proposed model. The world knowledge base is a taxonomic specification of commonsense knowledge acquired by people through their experiences and education [238]. The user LIRs are the personal information collections of users, for example, user created and stored documents, browsed Web pages, and compiled/received emails. The information items in the LIRs have connection to the concepts specified in the world knowledge base. Personalised ontologies based on these are constructed in the proposed model by adopting automatic or semi-automatic ontology learning methods to discover the concepts relevant to user information needs. A multidimensional ontology mining method, Specificity and Exhaustivity, is also introduced in the proposed model for analysing the concepts specified in ontologies. The model emphasises the specific is-a and part-of semantic relations in one single computational model, and aims at effectively acquiring user profiles to capture user information needs for personalised Web information gathering. The ontology learning and mining model is evaluated by comparing the acquired user profiles with those acquired by the baselines, including manual, automatic, and semi-automatic user profile acquiring models. These evaluation results are reliable and promising for the proposed model.

The goal of the research in this thesis is to develop a better understanding of user profile acquisition. The findings of this study can improve the performance of personalised Web information gathering systems, and can thus provide better design of these systems. The findings also have the potential to help design personalised systems in other communities, such as information retrieval, information filtering, recommendation systems, and information systems. The contributions are original and increasingly significant, considering the rapid explosion of Web information in recent years and given today’s accessibility to the Internet, online digital libraries, and the full text world.
1.3 Research Methods and Thesis Outline

To ensure the success of the project, scientific method is the research methodology used in this thesis. Research methodologies provide detailed descriptions of the approaches taken in carrying out the research, such as the characteristics of data, data collection instruments, and the data collection process [53, 95]. Research methodologies accepted by the information systems and knowledge engineering communities have been undergoing continuous development in the last decade. Methods include case studies, field studies, action research, prototyping, and experimenting [22]. In information systems and knowledge engineering, research work that involves the development of robust mechanisms has to be evaluated by experiments in the classic science methodologies. Therefore, the scientific method, consisting of the iterating phases of problem definition, framework, preliminary study, model development, and evaluation, is chosen as the research methodology in this thesis. The chosen scientific method and its application are illustrated in Figure 1.2.

The rest of this thesis is outlined as follows:

Chapter 2 This chapter is a literature review of related disciplines covering Web information gathering, Web personalisation, and ontology learning and mining. The literature review pinpoints the limitations of existing techniques in Web information gathering, and suggests the course of possible solutions.

Chapter 3 In this chapter, a concept-based Web information gathering framework is presented that introduces the research hypothesis to the research problems and defines the assumptions and scopes of the research work conducted in this thesis.

Chapter 4 This chapter describes and discusses the preliminary study conducted for the hypothesis introduced in Chapter 3, aiming to evaluate the hypothesis before moving on to the model development phase.
Chapter 5 This chapter presents the personalised ontology learning for Web users. A world knowledge base is utilised for user background knowledge extraction. The focuses of the chapter are on the construction methodology of the world knowledge base and the automatic and semi-automatic user personalised ontology learning methods.

Chapter 6 This chapter presents the multidimensional ontology mining method, *Specificity* and *Exhaustivity*, aiming to discover the on-topic concepts from user LIRs. The interesting concepts, along with their associated semantic relations of *is-a* and *part-of*, are analysed for user background knowledge discovery.

Chapter 7 In this chapter, the evaluation methodology of the proposed ontology learning and mining model is discussed, including experiment hypotheses, experiment designs, and the implementation of the experimental models.

Chapter 8 This chapter presents the performance measuring methods used in
the evaluation experiments, the experimental results, and the related discussions.

Chapter 9 This chapter concludes the thesis by discussing the contributions and suggesting the future work extended from the thesis.

1.4 Previously Published Papers

Some of the results from the research work discussed in this thesis have been previously published in (or submitted to) international conferences and journals. These refereed papers are listed as follows:

1. X. Tao, Y. Li, and N. Zhong. A Personalized Ontology Model for Web Information Gathering. Under the second round review by *the IEEE Transactions on Knowledge and Data Engineering*, 2009.


Other published works on this research are also listed as follows:


Chapter 2

Literature Review

The aim of this literature review chapter is to set up the research questions and the related research methodology that are introduced in Chapter 3. The reviewed literature covers Web information gathering including related challenges and techniques, Web personalisation including user profile acquisition and user information need capture, and the ontology-related issues including definitions and learning and mining techniques.

2.1 Web Information Gathering

2.1.1 Web Information Gathering Challenges

Over the last decade, the rapid growth and adoption of the World Wide Web have further exacerbated user need for efficient mechanisms for information and knowledge location, selection and retrieval. Web information covers a wide range of topics and serves a broad spectrum of communities [4,33]. How to gather useful and meaningful information from the Web, however, becomes challenging to Web users. This challenging issue is referred by many researchers as Web information gathering [47,86,96,101].
The current Web information gathering systems suffer from the problems of information mismatching and overloading. The Web information gathering tasks are usually completed by the systems using keyword-based techniques. The keyword-based mechanism searches the Web by finding the documents with the specific terms or topics matched. This mechanism is used by many existing Web search systems, for example, Google∗ and Yahoo!†, for their Web information gathering. Huberman *et al.* [69] and Han and Chang [65] pointed out that by using keyword-based search techniques, the Web information gathering systems can access the information quickly; however, the gathered information may possibly contain much useless and meaningless information. This is particularly referred as the fundamental issue in Web information gathering: information mismatching and information overloading [107–110, 242]. Information mismatching refers to the problem of useful and meaningful information being missed out in information gathering, whereas information overloading refers to the problem of useless and meaningless information being gathered. Li and Zhong [107] argued that these fundamental problems are caused by the large volume of noisy and uncertain data existing on the Web and thus in the gathered information. Also argued by Han and Chang [65] and Broder [15], these problems are caused by the features posed by the Web, such as complexity and the dynamic nature of Web information. Effectiveness of Web information gathering is a difficult task for all Web information gathering systems.

In attempting to solve these fundamental problems, many researchers have aimed at gathering Web information with better effectiveness and efficiency for users. These researchers have moved information gathering from keyword-based methods to concept-based techniques in recent years. The journey is reviewed as follows.

∗http://www.google.com
†http://www.yahoo.com
2.1.2 Keyword-based Techniques

Keyword-based information gathering techniques are based on the feature vector of documents and queries. In order to determine if a document satisfies a user information need, information gathering systems extract the features of the document and compare these features to those of the given query. A well-known feature extraction technique is term frequency times inverse document frequency, usually denoted as $tf \times idf$ and calculated by:

$$w(t, d) = tf_t \times \log \left( \frac{|D|}{df_t} \right);$$  \hspace{1cm} (2.1)

where $w(t, d)$ is the weight indicating the likelihood that the term $t$ represents a feature of document $d$, $tf$ is the term frequency of $t$ in $d$, and $df$ is the number of documents in collection $D$ that contain $t$. With the $tf \times idf$, the more frequently a term occurs in a document and the less frequently it occurs in other documents, the more accurately the term represents the feature of the document \[125, 188\].

A document can then be represented by a vector of features; each one is a term associating with a weighting value calculated by techniques like $tf \times idf$. The feature vector of documents is represented as $\vec{d} = \{w_{1,d}, w_{2,d}, \ldots, w_{n,d}\}$, where $n$ is the total number of features representing $d$. These vectors are called the “feature vectors” of documents in information gathering \[6\].

The relevance of documents to given queries is determined by their similarities, whereas the similarity of documents and queries is measured by comparing their feature vectors. A query, as expressed by users for information needs, usually consists of a set of terms and thus can also be considered a document and represented by feature vectors $\vec{q} = \{w_{1,q}, w_{2,q}, \ldots, w_{n,q}\}$, where $q$ is a query and $i$ is the total number of features representing the query. The factors considered in the similarity measure are summarised by \[188\]:
1. the topic of information need is discussed in these documents at length;
2. these documents should deal with several aspects of the topic;
3. these documents have many terms pertaining to the topic;
4. authors express the concept referring to the topic in multiple unique ways.

One of the well-known similarity measure methods is Cosine similarity. The similarity measure methods are based on the feature vectors. When extracting the feature vector of documents, the term frequencies are affected by the length of the documents. Thus, the distance (similarity) values calculated are also influenced by document length [19]. The Cosine similarity biases the document length and focuses on the angle between the feature vectors of documents and queries. It is calculated by [6]:

$$\text{Cosine}(\vec{d}, \vec{q}) = \frac{\sum_{x \in \vec{d}, y \in \vec{q}} xy}{\sqrt{\sum_{x \in \vec{d}} x^2 \times \sum_{y \in \vec{q}} y^2}} \quad (2.2)$$

Cosine similarity normalises the documents before calculating the similarity.

Another state-of-the-art retrieval function widely used in Web information gathering is BM25. The BM25 method is based on the probabilistic retrieval framework, and ranks a set of documents based on the query terms appearing in the documents [10]:

$$\text{bm25}(q, d) = \sum_{t \in q} \log\left(\frac{N - f_d + 0.5}{f_d + 0.5}\right) \times \frac{(k_1 + 1) f_{(d,t)}}{K + f_{(d,t)}} \quad (2.3)$$

where $t$ indicates the terms occurred in query $q$; $N$ is the overall number of documents in the collection; $f_d$ is the frequency of documents that a term $t$ occurs in, and $f_{(d,t)}$ is the term frequency occurring in document $d$; $K$ is the result of Equation (2.4), where the constants of $k_1$ is set as 1.2 and $b$ as 0.75, $L_d$ is the length of document $d$ measured in bytes, and $AL$ is the average document length over the collection:

$$K = k_1((1 - b) + b \times \left(\frac{L_d}{AL}\right)) \quad (2.4)$$
The BM25 function is explicitly sensitive to document length, and is used by the Zettair search engine\textsuperscript{‡} for retrieving information from the Web. The pivoted model developed by Singhal et al. [184], that normalises the feature vectors by reducing the gap between the relevance and the retrieved probabilities, is another model similar to the BM25.

Keyword-based information gathering techniques reflect the nature of information gathering conducted by human users. These techniques can also be called statistical techniques because they capture the semantic relations between terms, based on the statistics of their co-occurrence in documents [79]. Typical models include Latent Semantic Analysis (LSA) [46], Hyperspace Analogue to Language (HAL) [121], Point-wise Mutual Information using Information Retrieval (PMI-IR) [205], and Non Latent Similarity (NLS) [18]. These models represent document collections by a multidimensional semantic space and terms by a vector in the semantic space. As discussed previously, the closer distance between feature vectors in the semantic space means higher semantic similarity of their representative documents and queries [79]. The keyword-based techniques that use semantic spaces reflect human performance in information gathering, as argued by Landauer and Dumais [91].

However, the information gathering systems that utilise keyword-based and statistical techniques were reported suboptimal in many cases. When the queries are overly specific with a just few terms, these systems have insufficient index terms to search. Consequently, some useful and meaningful information is missed in the gathered results [198]. These systems also cannot capture the type of semantic relations existing in the terms and documents, such as \textit{is-a}, \textit{part-of}, and \textit{related-to}. These relations are important, as they exist in many web sites that incorporate hierarchical categorisations, like Amazon\textsuperscript{§}, eBay\textsuperscript{¶} and Yahoo!. Failing to consider these semantic relations results in some document features

being missed out in the information gathering process [79]. Moreover, systems utilising the keyword-based and statistical techniques cannot distinguish various senses referred to by one term [105,109,110]. For example, the term “apple” may mean either apples, the fruit, or iMac computers. The keyword-based systems cannot distinguish the information about “apple”, the fruit, from that about “apple”, iMac computers. Consequently, useless and meaningless information is gathered and the information overloading problem occurs. In addition, the systems employing the keyword-based information gathering techniques cannot clarify different terms that have the same meanings. For example, if searching for “laptop”, the information containing “notebook” computers may be missed by these systems. As a result, useful and meaningful information is missed and the information mismatching problem occurs. These limitations, incorporated by keyword-based and statistical techniques, motivate the research performed by many groups, aiming to promote Web information gathering from keyword-based to concept-based and hence to improve the performance of information gathering systems.

2.1.3 Concept-based Techniques

The concept-based information gathering techniques use the semantic concepts extracted from documents and queries. Instead of matching the keyword features representing the documents and queries, the concept-based techniques attempt to compare the semantic concepts of documents to those of given queries. The similarity of documents to queries is determined by the matching level of their semantic concepts. The semantic concept representation and extraction are two typical issues in the concept-based techniques and are discussed in the following sections.
2.1. Web Information Gathering

Semantic Concept Representation

Semantic concepts have various representations. In some models, these concepts are represented by controlled lexicons defined in terminological ontologies, thesauruses, or dictionaries. In some other models, they are represented by subjects in domain ontologies, library classification systems, or categorisations. In some models using data mining techniques for concept extraction, semantic concepts are represented by patterns. The three representations given have different strengths and weaknesses.

Lexicon-based (or entity-based) representation is one of the common concept-based representation techniques. In this kind of representation, semantic concepts are represented by the controlled lexicons or vocabularies defined in terminological ontologies, thesauruses, or dictionaries. A typical representation is the synsets in WordNet, a terminological ontology. Each synset represents a unique concept that refers to a set of senses grouped by the semantic relation of synonyms. The senses in WordNet are the entities (or instances) of concepts. Different senses of a word could be in different synsets, and therefore in different semantic concepts. As well as synonyms, WordNet also has hypernyms/hyponyms, holonyms/meronyms, troponyms, and entailments defined for the semantic relations existing amongst the synsets and senses [49]. The models utilising WordNet for semantic concept representation include [17, 54, 70] and [87].

Alternatively from representing semantic concepts using terminological ontologies, Wang [215] represented semantic concepts using the terms in thesauruses. In his work, a thesaurus was developed based on Chinese Classification Thesaurus (CCT) and bibliographic data in China Machine-Readable Cataloging Record (MARC) format (CNMARC). The thesaurus was used to semantically annotate scientific and technical publications. Also using thesaurus for semantic concept representation are Scime and Kerschberg [171], Akrivas et al. [2] and others.

Online dictionaries are another important resource used for semantic concept representation in Web information gathering models, such as [128]. However,
Smith and Wilbur [185] argued that the definitions and materials found in dictionaries need to be refined with the knowledge discovered in the content of experts’ written documents, not the freely contributed Web documents.

The lexicon-based representation defines the semantic concepts in terms and lexicons that are easily understood by users. Because these are being controlled, they are also easily utilised by the computational systems. However, when extracting terms to represent concepts for information gathering, some noisy terms may also be extracted because of the term ambiguity problem. As a result, the information overloading problem may occur in gathering. Moreover, the lexicon-based representation relies largely on the quality of terminological ontologies, thesaurus, or dictionaries for definitions. However, the manual development of controlled lexicons or vocabularies (like WordNet) is usually costly [31]. The automatic development is efficient, however, in sacrificing the quality of definitions and semantic relation specifications. Consequently, the lexicon-based representation of semantic concepts was reported to be able to improve the information gathering performance in some works [79, 87, 119], but to be degrading the performance in other works [208, 211].

Many Web systems rely upon subject representation of semantic concepts for concept-based information gathering. In this kind of representation, semantic concepts are represented by subjects defined in knowledge bases or taxonomies, including domain ontologies, digital library systems, and online categorisation systems. In domain ontologies, domain knowledge is conceptualised and formally described in hierarchical structures [127]. The concepts in the hierarchical structure of domain ontologies are usually linked by the semantic relations of subsumption like super-class and sub-class. Each concept is associated with a label that best describes the concept terminologically. Typical information gathering systems utilising domain ontologies for concept representation include those developed by Lim et al. [114], by Navigli [139], and by Velardi et al. [209]. Domain ontologies contain expert knowledge: the concepts described and specified in the
ontologies are of high quality. However, expert knowledge acquisition is usually costly in both capitalisation and computation. Moreover, as aforementioned, the semantic concepts specified in many domain ontologies are structured only in the subsumption manner, rather than the more specific \textit{is-a}, \textit{part-of}, and \textit{related-to}, the ones developed or used by [55,74,84] and [242]. Some attempted to describe more specified relations, like [21,23,167,178] for \textit{is-a}, [58,59,164,169] for \textit{part-of}, and [71,205] for \textit{related-to} relations only. However, there has not been any research that could portray the basic \textit{is-a}, \textit{part-of}, and \textit{related-to} semantic relations in one single computational model for concept representation.

Also used for subject-based concept representation are the library systems, like Dewey Decimal Classification (DDC) used by [84,201,217], Library of Congress Classification (LCC) and Library of Congress Subject Headings (LCSH) [50], and the variants of these systems, such as the “China Library Classification Standard” used by [237] and the Alexandria Digital Library (ADL) used by [216]. These library systems are human intellectual endeavours that have been undergoing continuous revision and enrichment for over one hundred years. They represent the natural growth and distribution of human intellectual work that covers the comprehensive and exhaustive topics of world knowledge [26]. In these systems, the concepts are represented by the subjects defined by librarians and linguists manually. The concepts are constructed in taxonomic structure, originally designed for information retrieval from libraries. The concepts are linked by semantic relations, such as subsumption like \textit{super-class} and \textit{sub-class} in the DDC and LCC, and \textit{broader}, \textit{used-for}, and \textit{related-to} in the LCSH. The concepts in these library systems are well defined and refined by experts under a well-controlled process [26], and the concepts and structure are designed for information gathering originally. These are beneficial to the information gathering systems. However, the information gathering systems using library systems for concept representation largely rely upon the existing knowledge bases. The limitations of the library systems, for example, the focus on the United States more than on other regions
by the LCC and LCSH, would be incorporated by the information gathering systems that use them for concept representation.

The online categorisations are also widely relied upon by many information gathering systems for subject-based concept representation. The typical online categorisations used for concept representation include the Yahoo! categorisation used by [55] and Open Directory Project ‡ used by [28, 149]. In these categorisations, concepts are represented by categorisation subjects and organised in taxonomical structure. The instances referring to a concept are extracted from the Web documents under that categorisation, by using the keyword-based techniques for feature extraction, as discussed previously. However, the semantic relations linking the concepts in this representation are still only specified as super-class and sub-class. The nature of categorisations is in the subsumption manner of one containing another, but not the semantic is-a, part-of, and related-to relations. Thus, the semantic relations associated with the concepts in such representations are not in adequate details and specific levels. These problems weaken the quality of concept representation and thus the performance of information gathering systems.

Another semantic concept representation in Web information gathering systems is pattern-based representation. Representing concepts by individual terms can easily prompt semantic ambiguity problems, as the example of “apple” the fruit and “apple” computers discussed previously. Also, the term-based representation is inadequate for concept discrimination because single terms are not adequately specific [196]. Aiming to overcome these problems, a pattern-based concept representation is developed that uses multiple terms (e.g. phrases) to represent a single semantic concept. Phrases contain more content than any one of their containing terms. For example, “data mining” refers to a process that discovers knowledge from data. The combination of specific terms “data” and “mining” prevents the concept from the semantic ambiguity that may possibly be

‡http://www.dmoz.org
posed by either “data” or “mining”, such as mineral mining. Research representing concepts by patterns include Li and Zhong [102,107–111], Wu et al. [222–224], Dou et al. [44], Ruiz-Casado et al. [165,166], Borges and Levene [13], Cooley [34], and Cooley et al. [35]. However, pattern-based semantic concept representation poses some drawbacks. The concepts represented by patterns can have only subsumption specified for relations. Usually, the relations existing between patterns are specified by investigation of their containing terms [107–110, 222–224]. If more terms are added into a phrase, making the phrase more specific, the phrase becomes a sub-class concept of any concepts represented by the sub-phrases in it. Thus, “data mining” is a sub-class concept of “data” and also a sub-class concept of “mining”. Consequently, no specific semantic concepts like is-a and part-of can be specified and thus some semantic information may be missed in pattern-based concept representations. Another problem of pattern-based concept representation is caused by the length of patterns. The concepts can be adequately specific for discriminating one from others only if the patterns representing the concepts are long enough. However, if the patterns are too long, the patterns extracted from Web documents would be of low frequency and thus, cannot support the concept-based information gathering systems substantially [222]. Although the pattern-based concept representation poses such drawbacks, it is still one of the major concept representations in information gathering systems.

**Semantic Concept Extraction**

The techniques used for concept extraction from text documents include text classification techniques and Web content mining techniques, including association rules mining and pattern mining. These techniques are reviewed and discussed as follows.

**Text Classification**

Text classification aims to classify documents into categories. Due to the large volume of Web documents, the manual assessment of Web information is impos-
sible [60]. Based on the semantic content of Web documents, text classification techniques classify Web documents into categories automatically, and thus are capable of helping to assess Web information [24, 55, 56, 104, 134, 151, 231].

Text classification is the process of classifying an incoming stream of documents into categories by using the classifiers learned from the training samples [116]. Technical speaking, text classification is to assign a binary value to each pair of \((d_j, c_i) \in D \times C\), where \(D\) is a set of documents and \(C\) is a set of categories [172]. With a set of predefined categories, this is referred to as supervised text classification or predictive classification. The performance of text classification relies upon the accuracy of classifiers learned from training sets. In general, a training set is a set of labelled (positive and negative) samples, along with pre-defined categories [100, 231]. Based on the training set, the features that discriminate the positive samples from the negative samples are extracted. These features are then used as classifiers to classify incoming documents into the categories. Apparently, the accuracy rate of classifiers determines their capability of separating the incoming stream of documents, and thus the performance of text classification [52, 65, 100, 116]. Therefore, learning classifiers from the training sets is important in text classification. Typical existing techniques to learn classifiers include Rocchio [162], Naive Bayes (NB) [159], Dempster-Shafer [168], Support Vector Machines (SVMs) [76, 77], and the probabilistic approaches [36, 57, 80, 132, 144]. Sometimes there are not an optimal number of negative samples available but just positive and unlabelled samples [52]. This problem is referred to as semi-supervised (or partially supervised) text classification. The mainstream of semi-supervised classification techniques is completed by two steps: extracting negative samples from the unlabelled set first, and then building classifiers as supervised classification methods [100, 116, 233, 234], such as S-EM [117], PEBL [233], and Roc-SVM [100]. Alternatively, some research works attempted to extract more positive samples rather than negative samples from the unlabelled sets, for example [52]. The classifiers classifying documents into cate-
2.1. Web Information Gathering

categories are treated as the semantic concepts representing these categories. Hence, in concept-based Web information gathering, the process of learning classifiers is also a process of extracting the semantic concepts to represent the categories.

These classifier learning techniques can be categorised into different groups. Fung et al. [52] categorised them into two types: kernel-based classifiers and instance-based classifiers. Typical kernel-based classifier learning approaches include the Support Vector Machines (SVMs) [76,77] and regression models [172]. These approaches may incorrectly classify many negative samples from an unlabelled set into a positive set, thus causing the problem of information overloading in Web information gathering. Typical instance-based classification approaches include the $K$-Nearest Neighbor ($K$-NN) [39] and its variants, which do not relay upon the statistical distribution of training samples. However, the instance-based approaches are not capable of extracting highly accurate positive samples from the unlabelled set. Other research works, such as [55,56,151], have a different way of categorising the classifier learning techniques: document representations based classifiers, including SVMs and $K$-NN; and word probabilities based classifiers, including Naive Bayesian, decision trees [51,76] and neural networks used by [235]. These classifier learning techniques have different strengths and weaknesses, and should be chosen based upon the problems they are attempting to solve.

Text classification techniques are widely used in concept-based Web information gathering systems. Chaffee and Gauch [24] and Gauch et al. [55] described how text classification techniques are used for concept-based Web information gathering. Web users submit a topic associated with some specified concepts. The gathering agents then search for the Web documents that are referred to by the concepts. Sebastiani [172] outlined a list of tasks in Web information gathering to which text classification techniques may contribute: automatic indexing for Boolean information retrieval systems, document organisation (particularly in personal organisation or structuring of a corporate document base), text fil-
tering, word sense disambiguation, and hierarchical categorisation of web pages. Also, as specified by Meretakis et al. [134], the Web information gathering areas contributed to by text classification may include sorting emails, filtering junk emails, cataloguing news articles, providing relevance feedback, and reorganising large document collections. Text classification techniques have been utilised by [63, 68, 92, 123, 133] to classify Web documents into the best matching interest categories, based on their referring semantic concepts.

Text classification techniques utilised for concept-based Web information gathering, however, incorporate some limitations and weaknesses. Glover et al. [60] pointed out that the Web information gathering performance substantially relies on the accuracy of predefined categories. If the arbitration of a given category is wrong, the performance is degraded. Another challenging problem, referred to as “cold start”, occurs when there is an inadequate number of training samples available to learning classifiers. Also, as pointed out by Han and Chang [65], the concept-based Web information gathering systems rely on an assumption that the content of Web documents is adequate to make descriptions for classification. When the assumption is not true, using text classification techniques alone becomes unreliable for Web information gathering systems. The solution to this problem is to use high quality semantic concepts, as argued by Han and Chang [65], and to integrate both text classification and Web mining techniques.

**Web Content Mining**

Web content mining is an emerging field of applying knowledge discovery technology to Web data. Web content mining discovers knowledge from the content of Web documents, and attempts to understand the semantics of Web data [35, 88, 110, 115, 192]. Based on various Web data types, Web content mining can be categorised into Web text mining, Web multimedia data mining (e.g. image, audio, video), and Web structure mining [88, 192]. In this thesis, Web information is particularly referred to as the text documents existing on the Web. Thus, the term “Web content mining” here refers to “Web text content mining”,
the knowledge discovery from the content of Web text documents. Kosala and Blockeel [88] categorised Web content mining techniques into database views and information retrieval views. From the database view, the goal of Web content mining is to model the Web data so that Web information gathering may be performed based on concepts rather than on keywords. From the information retrieval view, the goal is to improve Web information gathering based on either inferred or solicited Web user profiles. With either view, Web content mining contributes significantly to Web information gathering.

Many techniques are utilised in Web content mining, including pattern mining, association rules mining, text classification and clustering, and data generalisation and summarisation [107, 109, 192]. Li and Zhong [107–110] and Wu et al. [222–224] represented semantic concepts by maximal patterns, sequential patterns, and closed sequential patterns, and attempted to discover these patterns for semantic concepts extracted from Web documents. Their experiments reported substantial improvements achieved by their proposed models, in comparison with the traditional Rocchio, Dempster-Shafer, and probabilistic models. Association rules mining extracts meaningful content from Web documents and discovers their underlying knowledge. Existing models using association rules mining include Li and Zhong [106], Li et al. [103], and Yang et al. [229,230], who used the granule techniques to discover association rules; Xu and Li [226–228] and Shaw et al. [175], who attempted to discover concise association rules; and Wu et al. [225], who discovered positive and negative association rules. Text classification is to classify a set of text documents based on their values in certain attributes (classifiers) [48], as discussed previously. Alternatively, text clustering is to group a set of text documents into unsupervised (non-predefined) classes based upon their features. These clustering techniques can also be called descriptive or unsupervised clustering; the main techniques include $K$-means [124] and hierarchical clustering [1]. Text clustering techniques were used by Desai and Spink [41] to extract concepts from Web documents for relevance assessment.
The techniques were also used by Godoy and Amandi [61, 62], Wei et al. [219], Zhou et al. [245], and Lee et al. [94] to extract the concepts of user interests for personalised Web information gathering. Also, Hung et al. [70], and Maedche and Zacharias [126] clustered Web documents using ontologies. Reinberger et al. [152] and Karoui et al. [78] used text clustering to extract hierarchical concepts for ontology learning. Some works, such as Dou et al. [44], attempted to integrate multiple Web content mining techniques for concept extraction. These works were claimed capable of extracting concepts from Web documents and improving the performance of Web information gathering. However, as pointed out by Li and Zhong [108, 109], the existing Web content mining techniques incorporate some limitations. The main problem is that these techniques are incapable of specifying the specific semantic relations (e.g. is-a and part-of) that exist in the concepts. Their concept extraction needs to be improved for more specific semantic relation specification, considering the fact that the current Web is nowadays moving toward the Semantic Web [8].

2.2 Web Personalisation

2.2.1 User Profiles

Web user profiles are widely used by Web information systems for user modelling and personalisation [88]. User profiles reflect the interests of users [177]. In terms of Web information gathering, user profiles are defined by Li and Zhong [110] as the interesting topics underlying user information needs. Hence, user profiles are used in Web information gathering to capture user information needs from the user submitted queries, in order to gather personalised Web information for users [55, 65, 110, 202].

Web user profiles are categorised by Li and Zhong [110] into two types: the data diagram and information diagram profiles (also called behaviour-based profiles and knowledge-based profiles by [136]). The data diagram profiles are usually
acquired by analysing a database or a set of transactions [55, 110, 136, 182, 197]. These kinds of user profiles aim to discover interesting registration data and user profile portfolios. The information diagram profiles are usually acquired by using manual techniques; such as questionnaires and interviews [136, 202], or by using information retrieval and machine-learning techniques [55, 145]. They aim to discover interesting topics for Web user information needs.

**User Profiles Representation**

User profiles have various representations. As defined by [177], user profiles are represented by a previously prepared collection of data reflecting user interests. In many approaches, this “collection of data” refers to a set of terms (or vector space of terms) that can be directly used to expand the queries submitted by users [2, 9, 36, 37, 136, 202, 218]. These term-based user profiles, however, may cause poor interpretation of user interests to the users, as pointed out by [109, 110]. Also, the term-based user profiles suffer from the problems introduced by the keyword-match techniques because many terms are usually ambiguous. Attempting to solve this problem, Li and Zhong [110] represented user profiles by patterns. However, the pattern-based user profiles also suffer from the problems of inadequate semantic relations specification and the dilemma of pattern length and pattern frequency, as discussed previously in Section 2.1.3 for pattern-based concept representation.

User profiles can also be represented by personalised ontologies. Gauch et al. [55, 56], Trajkova and Gauch [202], and Sieg et al. [181, 182] represented user profiles by a sub-taxonomy of a predefined hierarchy of concepts. The concepts existing in the taxonomy are associated with weights indicating the user-perceived interests in these concepts. This kind of user profiles describes user interests explicitly. The concepts specified in user profiles have clear definitions and extents. They are thus excellent for inferences performed to capture user information needs. However, clearly specifying user interests in ontologies is a difficult task,
especially for their semantic relations, such as is-a and part-of.

User profiles can also be represented by a training set of documents, as used in text classification [11,161]. User profiles (the training sets) consist of positive documents that contain user interest topics, and negative documents that contain ambiguous or paradoxical topics. This kind of user profiles describes user interests implicitly, and thus have great flexibility to be used with any concept extraction techniques. The drawback is that noise may be extracted from user profiles as well as meaningful and useful concepts. This may cause an information overloading problem in Web information gathering.

User Profile Acquisition

When acquiring user profiles, the content, life cycle, and applications need to be considered [170]. The content of user profiles is the description of user interests, as defined by Wasfi [218]. Although user interests are approximate and explicit, it was argued by [55,110,148] that they can be specified by using ontologies. The life cycle of user profiles refers to the period that the user profiles are valuable for Web information gathering. User profiles can be long-term or short-term. For instance, persistent and ephemeral user profiles were built by Sugiyama et al. [197], based on the long term and short term observation of user behaviour. Applications are also an important factor requiring consideration in user profile acquisition. User profiles are widely used in not only Web information gathering [55,110], but also personalised Web services [65], personalised recommendations [135,136], automatic Web sites modifications and organisation, and marketing research [243]. These factors considered in user profile acquisition also define the utilisation of user profiles for their contributing areas and period.

User profile acquisition techniques can be categorised into three groups: the interviewing, non-interviewing, and semi-interviewing techniques. The interviewing user profiles are entirely acquired using manual techniques; such as questionnaires, interviews, and user classified training sets. Trajkova and Gauch [202] ar-
argued that user profiles can be acquired explicitly by asking users questions. One typical model using user-interview profiles acquisition techniques is the TREC-11 Filtering Track model [161]. User profiles are represented by training sets in this model, and acquired by users manually. Users read training documents and assign positive or negative judgements to the documents against given topics. Based upon the assumption that users know their interests and preferences exactly, these training documents perfectly reflect users’ interests. However, this kind of user profile acquisition mechanism is costly. Web users have to invest a great deal of effort in reading the documents and providing their opinions and judgements. However, it is unlikely that Web users wish to burden themselves with answering questions or reading many training documents in order to elicit profiles [109,110].

The non-interviewing techniques do not involve users directly but ascertain user interests instead. Such user profiles are usually acquired by observing and mining knowledge from user activity and behaviour [110,148,176,192,197,202,218]. Typical models include the ontological user profiles acquired by [55,148,202] and [182]. These models acquire user profiles by using global categorisations such as Yahoo! categorisation and Online Directory Project. The machine-learning techniques are utilised to analyse the user-browsed Web documents, and classification techniques are used to classify the documents into the concepts specified in the global categorisation. As a result, the user profiles in these models are a sub-taxonomy of the global categorisations. However, because the categorisations used are not well-constructed ontologies, the user profiles acquired in these models cannot describe the specific semantic relations. Instead of classifying interesting documents into the supervised categorisations, Li and Zhong [109,110] used unsupervised methods to discover interesting patterns from the user-browsed Web documents, and illustrated the patterns to represent user profiles in ontologies. The model developed by [118] acquired user profiles adaptively, based on the content study of user queries and online browsing history. In order to acquire user
profiles, Chirită et al. [27] and Teevan et al. [199] extracted user interests from the collection of user desktop information such as text documents, emails, and cached Web pages. Makris et al. [129] comprised user profiles by a ranked local set of categories and then utilised Web pages to personalise search results for a user. These non-interviewing techniques, however, have a common limitation of ineffectiveness. Their user profiles usually contain much noise and uncertainties because of the use of automatic acquiring techniques.

With the aim of reducing user involvement and improve effectiveness, the semi-interviewing user profiles are acquired by semi-automated techniques. This kind of user profiles may be deemed as that acquired by the hybrid mechanism of interviewing and non-interviewing techniques. Rather than providing users with documents to read, some approaches annotate the documents first and attempt to seek user feedback for just the annotated concepts. Because annotating documents may generate noisy concepts, global knowledge bases are used by some user profile acquisition approaches. They extract potentially interesting concepts from the knowledge bases and then explicitly ask users for feedback. For example, by using a so-called Quickstep topic ontology, Middleton et al. [135,136] acquired user profiles from unobtrusively monitored behaviour and explicit relevance feedback. The limitation of semi-interviewing techniques is that they largely rely upon knowledge bases for user background knowledge specification.

## 2.2.2 User Information Need Capture

User information need analysis aims to extract the personal interests of users in information gathering. Web users come with different information needs when performing information gathering tasks. For the same search topic “New York”, the information interests of business travellers may be different from those of leisure travellers. Hence, analysing user information needs can help deliver users meaningful and useful information, according to their personal interests. User information need analysis is important in Web personalisation.
2.2. Web Personalisation

The techniques for user information need analysis can be categorised as global analysis and local analysis, based on the resources that the analyses rely on.

The global analysis techniques produce consistent and effective performance in user information need analysis. Such techniques use global knowledge bases to support the user interests analysis [36], including term clustering, Latent Semantic Index [240], and similarity thesauri [239]. The knowledge bases may be ontologies, thesauruses, and Web knowledge bases. Ontologies are the most common knowledge base used by these techniques. A typical one is WordNet [31, 49], which is a terminological and generic ontology. WordNet was used by Zhang et al. [241], Mandala et al. [130, 131], for user information need analysis, and these models had improved performance in information gathering. However, as Voorhees and Hou [212] reported that using WordNet could improve performance in some queries but not in others. Some other works, such as Andreasen et al. [3], Stojanovic [195], and Tran et al. [203], learned ontologies to interpret semantic meanings of user queries; and Bata et al. [5], Cimiano et al. [30], Lee et al. [93], Shamsfard et al. [174], and Espinasse [47], learned ontologies to annotate text documents. By the use of ontologies for user information need analysis, Web systems have achieved remarkable performance in personalised information gathering.

Dictionaries and thesauruses are also common global knowledge bases used by information gathering systems for information need analysis. A thesaurus in the economic and environment domain was manually constructed and used by Kristensen [89] for user information need analysis. In his work, the recall performance was improved; however, there was a reduction in the precision performance. A model called INSTRUCT used term-clustering statistics and morphological processing to analyse user needs from given queries [214]. Another model called CITE was developed by [128] to analyse user needs using a dictionary and the MeSH thesaurus. Dictionaries and thesauruses were also used by [2, 179, 180, 215, 216] to help analyse user needs in information gathering. These systems extracted the feature terms from user queries to represent the information needs, and then
suggested synonym terms based on the dictionaries and thesauruses for better information gathering.

Web knowledge bases nowadays are being used more and more frequently to analyse the semantic meanings of user information needs. Wikipedia as a free, multilingual Web encyclopedia is typical of them. Wikipedia has 12 million articles (2.7 million in English) written collaboratively by volunteers around the world [221]. Wikipedia was used by [45, 138, 158] to help understand the user interests underlying the queries. The online categorisations, such as the Yahoo! categorisation and the Open Directory Project, are also widely used by many Web information gathering systems to analyse the semantic meanings of user information needs [55, 151, 202]. These Web knowledge bases, however, have some limitations. The articles in Wikipedia are freely contributed by volunteers. Consequently, the knowledge extracted from Wikipedia may lack authority. The online categorisations have concepts categorised, but in simple subsumption structure only, not in specific semantic relations such as is-a and part-of. For semantic analysis of user information needs, better knowledge bases with specific semantic relations specified may be necessary.

In contrast to global knowledge bases, local user information need analysis largely relies upon information feedback from users or observations on the user behaviour. The local techniques for user information need analysis techniques include user relevance feedback, pseudo-relevance feedback, and user logs analysis [36].

Relevance feedback techniques capture user information needs based on the terms or documents that are explicitly fed back from users. Users are provided with a set of terms or documents, and asked to select the terms or documents that they are interested in. The selected terms and documents are then analysed by the systems in order to capture the user information needs, using the keyword-based or concept-based techniques discussed in Section 2.1. Because users give direct feedback, these kinds of techniques are usually effective in capturing user
information needs [190]. Many works, like the CUPID developed by Magen-
nis and Rijsbergen [128], the MUSCAT developed by Porter and Galpin [147],
and the context sensitive information retrieval by [176], used relevance feedback
techniques to improve performance in information gathering. The drawback of
relevance techniques is the cost of user time, as users may not like to burden
themselves with explicit feedback.

Pseudo-relevance feedback techniques are developed with the aim of incorpo-
rating the benefits from relevance feedback techniques and avoid the drawback of
them. Pseudo-relevance feedback techniques initialise a search first and assume
that the top-\(k\) returned documents are as relevant as the feedback explicitly pro-
vided by users. The features of the top-\(k\) documents are extracted and then used
to capture user information needs [122]. Many information gathering systems us-
ing pseudo-relevance feedback have been reported as to have achieved significant
improvements in their performance [20, 32, 94, 236]. Amongst these works, Lee et
al. [94] clustered the top-\(k\) documents to find dominant documents in order to
emphasise the core concepts in user interests. Instead of treating each of the top-
\(k\) documents as equally relevant, Collins-Thompson and Callan [32] re-sampled
the top-\(k\) documents retrieved in the initial search according to the relevance
values estimated by probabilities. As a result, a document is more relevant if it
is ranked higher. However, because of the pseudo techniques used, the top-\(k\) doc-
uments contain some noise. Systems using pseudo-relevance feedback techniques
are usually not as effective as those using explicit relevance feedback.

User log analysis techniques are based on observations of user behaviour. This
kind of techniques attempts to discover the correlations between user queries and
documents in user logs, and capture user information needs from the correlations.
A typical work was conducted by Cui et al. [36, 37], who used data mining and
probabilistic techniques to capture user needs. Beitzel et al. [7] classified Web
queries using user query logs with the aim of discovering the topical meanings
of user information needs. Alternatively, Sekine and Suzuki [173] analysed query
logs to discover user background knowledge. User logs are nontrivial resources that contain user personal information implicitly. However, user log analysis techniques rely on data mining or classification techniques for knowledge discovery. The discovered results sometimes contain noise and require further filtering.

2.3 Ontologies

2.3.1 Ontology Definitions

Ontologies are formal descriptions and explicit specifications of conceptualisation. Zhong and Hayazaki [244] defined that

- conceptualization means modeling some phenomenon in real world to form an abstract model that identifies the relevant concepts of that phenomenon;
- formal refers to the fact that the ontology should be machine readable, that is, an ontology provides a machine-processable semantics of information sources that can be communicated between different agents; explicit means that the type of concepts used and the constraints on their use are explicitly defined.

This definition is also commonly supported by [29, 40, 64, 109, 110, 135, 136, 202, 243]. Ontologies are an important technology in the semantic Web and Web information gathering. They serve for the semantic Web by providing a controlled vocabulary of concepts, each with explicitly defined and machine-processable semantics. Ontologies also provide a common understanding of topics for communication between systems and users, and enable Web-based knowledge processing, sharing, and reuse between applications [29, 243, 244]. Moreover, ontologies help define and interpret the semantic meaning of Web content, and enable intelligent agents to gather Web information for users in knowledge-based Web gathering [4, 29, 38, 127, 139, 187, 243, 244].

Depending on the types of stored knowledge, ontologies can be categorised into two types: domain ontologies and generic (terminological) ontologies [139, 187].
Domain ontologies specify expert classified concepts and form the core knowledge in particular domains. Thus, the content of domain ontologies needs to be updated regularly with the update of domain knowledge [139]. The size of domain ontologies vary, depending on the domains described. Domain ontologies are described by [187, 243, 244] as a set of domain terms generated from the abstract descriptions of domain knowledge and a set of domain knowledge referred by the terms. Domain ontologies provide the possibility to specify domain knowledge in the form of axioms for problem solving [194].

Generic and terminological ontologies store the lexical relations of concepts in natural languages. Terms are organised in bags of synonyms connected through various semantic relations [12]. The knowledge specified in generic ontologies is usually in large size and does not require regular updates [83, 143]. A well-known terminological ontology is WordNet [31, 49, 137], in which concepts are represented by lexicons linked by the semantic relations of synonyms, hyponyms, holonyms, and meronyms, and each lexicon refers to a set of senses. WordNet was utilised by Budanitsky and Hirst [17] to clarify the semantic relations between lexicons, and by Gangemi [54] to analyse the concepts of lexical taxonomies. Also, Hung et al. [70] utilised WordNet ontology for documents clustering, and Kornilakis et al. [87] used WordNet to support interactive concept map construction in information gathering. Generic and terminological ontologies may extend to domain ontologies when more specific concepts are added to the ontologies for a particular domain [142, 189].

Ontologies have been widely used by many groups to specify user background knowledge in personalised Web information gathering. Li and Zhong [109] used ontologies to describe the user conceptual level model: the so called “intelligent” part of the world knowledge model possessed by human beings. Li and Zhong [110] also used pattern recognition and association rules mining techniques to discover knowledge from Web content for ontology construction. Tran et al. [203] introduced an approach to translate keyword queries to the Description
Logics conjunctive queries and to specify user background knowledge in ontologies. Gauch et al. [55] learned personalised ontologies for individual users in order to specify their preferences and interests in Web information gathering. These works utilised ontologies to specify user background knowledge for personalised Web information gathering.

Ontologies usually consist of a set of concepts (also known as classes), a set of vocabularies (instances), semantic relations, and some inference and logic rules (axioms) for a general purpose or a particular domain [4, 38, 40, 109, 127, 139, 187]. The concepts are usually described and referred to by the terms in vocabularies [109]. The semantic relations typically include hierarchical and non-hierarchical relations. The hierarchical relations represent the human cognitive view of classification, the subsumption of super-class and sub-class, or the more specific part-of and is-a relations. The non-hierarchical relations can be associative (cause-effect) or equivalence (synonymy or related-to) relations [73]. Maedche and Staab [193] formally defined ontologies as a 4-tuple \( \langle C, R, I, A \rangle \), where \( C \) is a set of concepts, \( R \) is a set of relations, \( I \) is a set of instances, and \( A \) is a set of axioms. Maedche [127] had a slightly different 5-tuple \( \langle C, R, H^C, rel, A^O \rangle \) definition for ontologies, where

- \( C \) and \( R \) are two disjoint sets whose elements are concepts and relations, respectively;
- \( H^C \) is a taxonomic backbone and a directed relation \( H^C \subseteq C \times C \) called taxonomy. \( H^C(C_1, C_2) \) means that \( C_1 \) is a sub-concept of \( C_2 \);
- \( rel : R \rightarrow C \times C \) is a function that relates concepts non-taxonomy;
- \( A^O \) is a set of ontology axioms expressed in an appropriate logical language.

### 2.3.2 Ontology Learning

Ontology learning is the process of constructing ontologies. Zhong and Hayazaki [244] described a two-phase ontology learning approach: conceptual relationship anal-
2.3. Ontologies

Analysis and ontology prototype generation. The first phase is to compute the weights of instances in a corpus and generate a network-like concept space for semantic relation specification. The second phase treats instances as neurons (units) and takes the relationship between them as the unidirectional, weighted connection between neurons. Zhong [242] also extended the two phases to a multi-phase process of content collection, morphological analysis, text (domain) classification, generation of classification rules, and conceptual relationship analysis, as well as the generation, refinement, conceptual hierarchy, and management of ontologies. Alternatively, an ontology learning framework was proposed by Maedche and Staab [125,127]. The framework contains four main components: ontology engineering and management environment; data import and processing component; algorithm library component; and graphical user interface and management component. In the framework, ontologies are learned through four phases: concept import and reuse, concept extract, concept prune, and concept refine. The framework extends typical ontology engineering environments by using semi-automatic ontology learning tools with human intervention, and constructs ontologies adopting the paradigm of balanced cooperative modelling. Antonious [4] proposed an ontology learning approach that consists of eight steps: determine scope; consider reuse; enumerate terms; define taxonomy; define properties; define facets; define instances; and check for anomalies. The phases specified in these ontology learning methodologies can be iterated and backtracked to earlier steps at any point if necessary, in terms of practice and ontology engineering.

Ontology learning was accomplished manually by many works in the last century. Typical ontologies learnt by using such mechanism are WordNet [31,49,137] and its extensive models, such as Sensus [85] and HowNet [237]. The WordNet ontology contains over 150,000 words and 207,000 senses, developed by ontology engineers manually. The manual ontology learning mechanism is effective in terms of knowledge specification but time consuming and costly in terms of finance and computation. Hiring ontology engineers for expert knowledge is ex-
pensive, and using human-power is error-prone. The manual ontology learning mechanism poses limitations, and hence the automated mechanism is necessary.

Automated ontology learning is accomplished using the hierarchical collections of documents or thesaurus [29, 56, 151]. One example is the so-called reference ontology built by [55, 56]. This ontology was constructed based on the subject hierarchies and their associated Web pages in Yahoo!, Lycos†, and the Open Directory Project. The reference ontology was used for Web user profile acquisition, by mapping users’ personal interests to the subjects in reference ontology. Zhong [242] argued that thesauruses can be used as a background knowledge base for ontology learning. A typical example is the IntelliOnto [84], an ontology describing world knowledge by using a three-level taxonomy of subjects constructed on the basis of the Dewey Decimal Classification system. The instances in the ontology were learned from the information items stored in library catalogues. By using the knowledge bases, the synonyms and the wider and narrower senses of terms can be incorporated by the construction method for the specification of semantic relationships. These learning methods increase the efficiency of ontology learning. However, the effectiveness of constructed ontologies largely relies on that of the used knowledge bases.

Many other works tried to learn ontologies automatically by using data mining techniques. Zhong [242] proposed an approach for domain ontology learning using various data mining and natural-language understanding techniques. Web content mining techniques were used by Jiang and Tan [74] to discover knowledge from domain-specific text documents for ontology learning. Jin et al. [75] attempted to integrate data mining and information retrieval techniques to further enhance ontology learning. Doan et al. [42, 43] proposed a model called GLUE and used machine learning techniques to find similar concepts in different taxonomies. Dou et al. [44] proposed a framework to learn domain ontologies using pattern decomposition, clustering and classification, and association rules.

mining techniques. An ontology learning tool called OntoLearn was developed by Navigli et al. [139] to attempt to discover semantic relations among the concepts from Web documents. These works attempted to explore a new route to specify knowledge efficiently.

The semantic association between concepts stored in ontologies may be discovered by computing the conceptual similarity (or distance) between them in the conceptual space of ontologies [73]. Viewing the network of notes as a topography, two kinds of approaches have been developed to measure the conceptual similarity of two classes in ontologies: the node-based and edge-based approaches correspond to the information content approaches and the conceptual distance approaches, respectively. The node-based conceptual similarity methods measure the extent of information shared in common by the measured concept classes. These approaches are theoretical; their typical approaches are [153, 154]. The edge-based methods measure the distance (e.g. edge length) between the measured concept classes in ontologies. Edges refer to the links connecting any two nodes in the ontology structure. The more edges covered by the path when travelling from one concept node to another indicates the less similarity of two concepts. Compared to the node-based methods, these approaches are more intuitive and direct. The typical models are [81–83]. However, Jiang and Conrath [73] pointed out that the structure information of ontologies is ignored by the node-based (information content) approaches. For the edge-based (conceptual distance) methods, Richardson et al. [150,157] reported that they performed poorly when applied to the WordNet ontology. Although the structure information is considered by the edge-based (conceptual distance) methods, none of the existing methods takes into account the influences produced by the different semantic relations, *is-a*, *part-of*, and *related-to*, to the best of the candidate’s knowledge. Therefore, both kinds of approaches have limitations in measuring the conceptual similarity between concepts in ontologies.

In summary, these previously discussed works all suffer from the same prob-
lem: inadequate knowledge specification. They cover only a limited number of concepts and emphasise only the super-class and sub-class relations, not the specific is-a, part-of, and related-to semantic relations. Thus, a research gap exists in learning ontologies to specify user background knowledge and to emphasise the semantic relations of is-a, part-of, and related-to in a single knowledge model.

2.4 Summary and Conclusion

This chapter presented several key issues for this thesis.

The literature review introduced the challenges existing in the current Web information gathering systems, and described how the current works gather Web information for users. Prior to the work presented in this thesis, the challenges of information mismatching and overloading remained unsolved by current efforts. The literature review pointed out that the key to gathering meaningful and useful information for Web users is to improve the Web information gathering techniques from keyword-based to concept-based.

The literature review in this chapter also noted the issues in Web personalisation, focusing on Web user profiles and user information needs in personalised Web information gathering. The survey confirmed that the concept-based models using user background knowledge can help gather useful and meaningful information for Web users. However, the representation and acquisition of user profiles need to be improved for the effectiveness of user information need capture.

The literature review in this chapter also covered ontologies, including ontology learning and mining for Web information gathering. The literature review indicated that ontologies can provide a basis for the match of user information needs and the existing concepts and relations. This helps to acquire user profiles. User background knowledge can be specified by using personalised ontologies. However, the existing ontologies and ontology learning methods have limitations and need to be improved for more specific knowledge description and specification.
2.4. Summary and Conclusion

Promoting Web information gathering from keyword-based to concept-based requires the semantic understanding of user information needs. The personalised ontologies that specify user background knowledge can help capture user information needs. This thesis addresses the limitations of current models for these issues by proposing a novel ontology learning and mining model in Chapters 3, 4, 5, and 6, then evaluates the model against numerous existing personalised Web information gathering models using ontologies in Chapters 7 and 8.
Chapter 3

Ontology-based Personalised Web Information Gathering

The capture of user information needs can help deliver personalised and useful information to Web users. As discussed in Chapter 2, acquiring user profiles and specifying these in personalised ontologies can benefit the effectiveness of user information need capture. Existing models, however, have limitations in user information capture and user profile acquisition. To address these limitations, this thesis proposes a novel model for acquiring user profiles via personalised ontologies. Here, a concept-based Web information gathering framework is presented that introduces the research hypothesis, and defines the assumptions and scopes of the research performed in this thesis. The hypothesis is preliminarily tested in Chapter 4, developed in Chapter 5 and 6, and finally evaluated in Chapters 7 and 8.
Chapter 3. Ontology-based Personalised Web Information Gathering

3.1 Concept-based Web Information Gathering Framework

The research for this thesis was conducted under the assumptions and scopes specified by a so-called *concept-based Web information gathering framework*. The framework consists of four models: a user concept model, a user querying model, a computer model, and an ontology model, as illustrated in Figure 3.1. The user concept model is of a user’s background knowledge system. The querying model is a user’s expression of an information need in Web information gathering. The computer model is to capture the information need expressed in the querying model. The ontology model is produced by the computer model as an explicit representation of the implicit user concept model associated with the information need. The following paragraphs describe the relationships existing between these models, then explain the assumptions and scopes of the thesis research performed.

A Web information gathering task starts from a user information need. From observations, when users were in need of some information and began an information gathering task, they usually fell into one of the following cases:
1. they knew nothing about that information;

2. they had tried but failed to infer that information from what they already knew;

3. they might know something but were not sure, so they needed to confirm.

From the first case, an assumption is made that users hold a concept repository in their brains that stores the user background knowledge. Given this assumption, users can check in the repository to see if some information or knowledge is possessed or not. The second case raises another assumption: that the concepts stored in the knowledge repository may be linked to each other. Only with this assumption available can users perform inference tasks from what is known to what is unknown. The last case also raises an assumption that users hold an implicit confidence rate for the concepts stored in the knowledge repository, although they cannot express the confidence rate clearly. With this assumption raised, users know what information or knowledge they are certain of and what they are uncertain of. Based on these assumptions, although the mechanism of a human user’s brain-working in Web information gathering has not yet been clearly understood in laboratories, the following assumption can arise:

**Assumption 1.** *Users have a knowledge repository, in which:*

- the stored concepts are embedded in a taxonomic structure;

- the stored concepts are associated with implicit confidence rates.

Performing an information gathering task also means the process of gathering more information or knowledge to store in this user knowledge repository.

Based on Assumption 1, and calling a user’s implicit knowledge system a *concept model*, a user concept model can be formalised as:

**Definition 1.** *A user concept model is a 3-tuple \( U \coloneqq (K, \hat{B}, G) \), where*
Chapter 3. Ontology-based Personalised Web Information Gathering

- $\mathcal{K}$ is a non-empty set of pairs $\{(k, w_k)\}$, where $k$ is a concept possessed by the user and $w_k$ is the user’s confidence in $k$;

- $\hat{\mathcal{B}}$ is a taxonomic structure containing concepts and their relationships;

- $\mathcal{G}$ is a set of gaps $\{g_1, g_2, \ldots, g_i\}$ existing on $\hat{\mathcal{B}}$, in which each gap $g$ is one or more concepts that the user does not possess.

Note that the $:\approx$ is used in Definition 1 instead of $:=$, as this definition is given under Assumption 1, which is based on observations and cannot currently be proven in laboratories.

The information gathering tasks are performed by users when attempting to find the related concepts to fill the gaps $g$ on the $\hat{\mathcal{B}}$ of $\mathcal{U}$. The desired concepts are user information needs. When attempting to find the desired concepts, users express their information needs by short phrases in their own languages. The phrases consist of a set of terms, and are formulated in a certain data structure. In information gathering, these user-formulated data structures for information needs are called queries. Thus, the following assumption can arise:

**Assumption 2.** Queries are users’ expressions of information needs in their own languages.

Based on Assumption 2, a user query can be formalised as a querying model in the concept-based Web information gathering framework:

**Definition 2.** A user querying model $\mathcal{Q}$ is a set of terms $\{t \mid t \in \mathcal{L}_U\}$, in which elements are primitive units in the user’s language $\mathcal{L}$.

In order to distinguish the user querying model of an information need from the accurate concepts referred to by an information need, the latter is called a topic and denoted as $\mathcal{T}$ in this thesis.

Capturing user information needs means discovering the concepts related to the gaps in user models. Users do not possess the concepts referred to by the gaps $g \in \mathcal{G}$ in $\mathcal{U}$. As a result, they may have to describe their information needs
by using concepts they possess that associate with the gaps on the $\hat{B}$ of $\mathcal{U}$. Thus, information need capture can be understood as an inverse process of exploiting the unknown concepts referred to by the $g \in \mathcal{G}$ from user description $\mathcal{Q}$. However, tracing from a $\mathcal{Q}$ back to the concepts of $g \in \mathcal{G}$ is difficult. Queries are often small sets of terms and contain only limited information [72]. Users have different backgrounds, perspectives, terminological habits, and vocabulary. Consequently, there are many uncertainties existing in the information need capturing process.

A hypothesis thus arises that if user background knowledge can be discovered and user concept models can be represented, the concepts referred to by the gaps in the $\hat{B}$ can be discovered, and thus, user information needs can be captured effectively. Ontologies are the formal specification of knowledge. User background knowledge can be specified by using personalised ontologies, and these ontologies can be used to capture user information needs. This hypothesis is developed in the computer model, denoted by $\mathcal{C}$ in this concept-based Web information gathering framework. A personalised ontology is learned in the $\mathcal{C}$ to represent a user concept model $\mathcal{U}$, through a given querying model $\mathcal{Q}$. The ontology structure represents the taxonomy structure $\hat{B}$, and the concepts specified in the ontology represent the user background knowledge $\mathcal{K}$ in $\mathcal{U}$. Discovering the concepts associated with the gaps $g \in \mathcal{G}$ from the personalised ontology can then help to define the concepts referred by the $gs$ – in other words, topic $\mathcal{T}$ of the information need. The personalised ontology constructed for $\mathcal{T}$ is called the ontology model in this concept-based Web information gathering framework and is denoted by $\mathcal{O}(\mathcal{T})$.

Under the concept-based Web information gathering framework, developing the computer model becomes the motivation of the research performed in this thesis. The model being developed in this thesis, the ontology learning and mining model, aims to discover relevant and non-relevant concepts in order to acquire user profiles and capture user information needs effectively.
3.2 Summary

This chapter introduced the hypothesis to the research problem in this thesis.

As addressed in Chapter 1, this thesis aims to acquire user profiles by learning and mining personalised ontologies. In this chapter, a concept-based Web information gathering framework was presented. In the framework, a hypothesis was discussed for solving the research problem in this thesis, in which user personalised ontologies are learned to represent the user concept models, and user information needs are captured by specifying the gaps in user concept models. The research in this thesis is conducted to develop and evaluate this hypothesis, under the assumptions and scopes defined in the concept-based Web information gathering framework as well.

The ontology learning and mining model proposed in Chapters 5 and 6 develops the hypothesis, and Chapters 7 and 8 present the evaluation of the hypothesis. The proposed model acquires user profiles, and also allows user information needs to be captured effectively.
Chapter 4

Preliminary Study

In this chapter, a preliminary study is conducted to test the hypothesis introduced in Chapter 3 before moving on to the development phase. In the preliminary study, a method is introduced that acquires user profiles from the Web by using user concept models. The method investigates the given topics, constructs user concept models, and uses the constructed models to gather Web information for user profiles. The proposed method is evaluated through the experiments performed on a large, standard data set. The experimental results confirm that by using user concept models specifying user background knowledge, useful and meaningful Web information can be gathered. The hypothesis introduced in Chapter 3 is promising.

4.1 Design of the Study

The preliminary study aims to evaluate the hypothesis introduced in Chapter 3: user profiles can be acquired and user information needs can be captured effectively by extracting user background knowledge and specifying user concept models. Thus, the Web information gathering performance can be improved,
and the Web information gathering systems can be designed in concept-based rather than keyword-based. The preliminary study was conducted to assess the feasibility of the hypothesis before developing the hypothesis.

The user background knowledge was specified manually, and then used to acquire user profiles in this preliminary study. As a contribution to Web information gathering, the preliminary model was tested in evaluation experiments by using the acquired user profiles to gather Web information. If the Web information gathering system benefited from the acquired user profiles, the preliminary model was promising, and the feasibility of the hypothesis could also be proven.

The details of the preliminary study presented in the following sections include topic analysis, user profile acquisition, and evaluation.

4.2 Semantic Analysis of Topic

In order to capture a user information need, the concept space referred to by the information need, namely a topic and denoted as $\mathcal{T}$, is identified. Let $\mathcal{S}$ be a set of concepts in which each element $s$ is a subject and $s \in \mathcal{S}$. The concept space referred by a topic $\mathcal{T}$ can be described by two sets of positive subjects $S^+$ and negative subjects $S^-$. The positive subjects refer to the concepts that $\mathcal{T}$ can be best described and discriminated from others. The negative subjects refer to the concepts that may cause paradoxical or ambiguous interpretation of $\mathcal{T}$. Identifying the concept space referred by $\mathcal{T}$ is thus to extract the $S^+$ and $S^-$ of topic $\mathcal{T}$.

In this preliminary study, these positive and negative subjects are manually identified, based on the descriptions and the narratives provided by users for the given topic. Depending on the level of subjects supporting or against the given topic, the positive subjects and negative subjects are identified with a support value $sup(s, \mathcal{T})$, which is measured by:

$$sup(s, \mathcal{T}) = MB(\mathcal{T} | s) - MD(\mathcal{T} | s).$$ (4.1)
where $MB(T|s)$ is the belief (how strong $s$ is for $T$) and $MD(T|s)$ is the disbelief (how strong $s$ is against $T$) of subject $s$ to topic $T$. When $MB(T|s)$ is greater than $MD(T|s)$, $s$ supports $T$ and becomes a positive subject. In contrast, when $MB(T|s)$ is smaller than $MD(T|s)$, $s$ is against $T$ and becomes a negative subject. In the preliminary study, the $MB(T|s)$ and $MD(T|s)$ were specified by the user manually, and the range of $sup(s,T)$ values is $[-1,1]$. Based on these, the positive and negative subjects can be defined by:

$$
\begin{align*}
    s & \in S^+ \quad \text{if} \quad sup(s,T) > 0; \\
    s & \in S^- \quad \text{if} \quad sup(s,T) \leq 0.
\end{align*}
$$

(4.2)

Drawing a boundary line for the positive and negative subjects is difficult, because uncertainties may exist in these subject sets. The overlapping space between $S^+$ and $S^-$ is considered negative in this preliminary study. Therefore, the concept space referred by $T$ can be defined as:

$$
\text{space}(T) = S^+ - (S^+ \cap S^-).
$$

(4.3)

### 4.3 Acquiring User Profiles

User profiles in this preliminary study are represented by training sets, one of the common representations of user profiles in Web information gathering [110]. Usually, a training set consists of a subset of positive samples and a subset of negative samples. Thus, in terms of user profiles, the positive samples are a set of documents that contain the user background knowledge and thus help to capture user information needs; the negative samples are the documents that contain the concepts that are paradoxical and ambiguous to the information needs. The previously discussed positive subjects $S^+$ and negative subjects $S^-$ can be used to acquire the positive and negative samples for user profiles in Web information gathering.
In this preliminary study, the training sets are acquired from the Web through a Web search agent using the Google API search tool*. For a given topic, a set of queries is formulated based on the identified positive and negative subjects. Each $s \in S^+$ produces a query to retrieve a set of positive candidate documents, and each $s \in S^-$ produces a query for negative candidates. The level of training documents supporting or against the given topic depends on many factors: the precision performance of the search agent, the document’s index position in the returned list, and the support value of $s$ that produced the query to retrieve the document.

The precision performance of a Web search agent can be measured by observing the result gathered in a training round. A common Web information gathering performance measure is precision [6], which is calculated by:

$$\varphi_\kappa = \frac{|D^+|}{\kappa}$$

where $\varphi$ is the precision performance, $|D^+|$ is the number of relevant documents gathered when reaching the cutoff point $\kappa$, and $|D^+| \leq \kappa$. For example, if there are eight relevant documents in the cutoff 10, the precision performance of this agent is then 80%. Thus, higher precision performance means better capability of retrieving positive training documents for user profiles.

The support values are also affected by the document index positions in the returned list, retrieved by the Web search agent. Although the retrieving algorithm used by a Web search agent usually remains hidden from the public, one assumption is valid: the index position is evidence, from the search agent of the document’s relevance to the given topic, that the documents indexed towards the top of the returned list are more relevant.

Based on these, with Equation (4.1), the support value $sup$ of a document $d$
4.4 Experiments and Results

The model was evaluated by assessing the performance of a Web information gathering system that utilised the proposed model discussed in Section 4.2 and 4.3. In response to a given topic, two user profiles (training sets) were acquired by the proposed model and the benchmark model. These user profiles were used by the same system to capture user information needs and gather information to a given topic $T$ can be measured by:

$$sup(d, T) = \sum_{s \in S^+ \cup S^-} sup(d, s) \times sup(s, T); \quad (4.5)$$

where $sup(d, s)$ is the support value of $d$ to $s$, which is calculated by:

$$sup(d, s) = \beta \times \varphi(\kappa) \times \left(\frac{k-(D[d] \mod(k))+1}{k}\right); \quad (4.6)$$

where $\beta$ is a parameter value $[0|1]$, for the occurrence of $d$ in $D$. Thus, if $d$ does not occur in the $D$ gathered by using $s$, $sup(d, s) = 0$. $D[d]$ is the index of $d$ in the returned set $D$, which is determined by the Web search agent, as previously discussed, and $k$ is a static number of how many documents in each cutoff.

Because $s \in S^+$ gives positive $sup(s, T)$ values and $s \in S^-$ gives negative $sup(s, T)$ values, Equation (4.5) may finally give a training document a positive or negative value, depending on the related subjects. Thus, the final training documents representing the user profile can be extracted from the training sets, and defined as:

$$\begin{align*} 
D^+ &= \{d, |sup(d, T) > 0\} \\
D^- &= \{d, |sup(d, T) \leq 0\}. 
\end{align*} \quad (4.7)$$

where $D^+$ is the positive document set that contains the relevant concept of $T$; and $D^-$ is the negative document set that contains the paradoxical and ambiguous concepts of $T$. 

4.4 Experiments and Results
from the testing set. The performance of the system then indicated the quality of input user profiles because everything in the experiments remained the same, except the input profiles. By comparing the performance of the Web information gathering system using different user profiles, the proposed model was evaluated quantitatively.

The experiment design is briefly described as follows. The Web information gathering system was implemented based on Li and Zhong’s model (refer to Section 7.4 in Chapter 7 for details). Two experimental models were implemented:

**Manual User Profile Acquiring Model** The user profiles (training sets) were manually acquired by the TREC linguists who read each document and marked it as either positive or negative, according to the given topics [160]. The user background knowledge contained in the user profiles was checked and approved by the users. Thus, these user profiles may be deemed “perfect”. This model is shortened as the “Manual model” in the related discussions, and can be referred to Section 7.6.1 for detailed implementation;

**Semi-automatic User Profile Acquiring Model** The implementation of the user profile acquiring model, the “Semi-auto model”, was as introduced (see Sections 4.2 and 4.3). User concept models were constructed manually. The positive and negative subjects were identified manually, where the subjects are listed in Appendix B for details. The $MB(T|s)$ value was set one and $MD(T|s)$ zero for positive subjects, and $MB(T|s)$ was zero and $MD(T|s)$ one for negative subjects. Therefore, the $sup(s,T)$ of Equation (refequ-CF) was also the boundary value, set as one for all positive subjects and zero for all negative subjects. The user profiles were acquired from the Web based on the identified subjects. Google was chosen as the Web search agent because it is probably the most popular Web search engine nowadays. The performance achieved by Google was determined by using a training topic (“Economic espionage”) and manually measuring the precision of gathered results. The precision performance is plotted in Figure 4.1. At the first
Figure 4.1: The Google Performance
portion of cutoffs (top 30 documents), Google achieved high precision performance. However, the performance dropped quickly when the number of retrieved documents increased. The precision performance of Google affects the support value of training documents acquired by using Google, as discussed in Section 4.3.

The experiment dataflow is plotted as Figure 4.2, where for the same topics, two different user profiles were acquired and used by the same system to gather information from the RCV1 testing set. The results were then compared for evaluation.

The Reuters Corpus Volume 1 (RCV1) was used as the test bed in experiments. The RCV1 contains 806,791 documents, and was also the standard test bed used in TREC-11 2002. TREC-11 had topics designed by linguists and associated with the training sets and testing sets. The first fifteen of these topics (101–115) were used in the experiments. The detailed description and justifications of RCV1 and TREC topics can be referred to in Sections 7.3.2 and 7.3.3 in Chapter 7.

The performance of the Web information gathering system used in this experiment was measured by precision and recall, the modern quantitative measures of performance in information gathering evaluations [213]. The precision and recall experimental results are presented in the curves plotted by the precision at eleven standard recall levels (11SPR) [161, 204] in Figure 4.3. The detailed descriptions and justifications of precision, recall, and 11SPR can also be referred to in Section 8.1.1 in Chapter 8.

The Semi-auto model implemented in the experiments aims to preliminarily test the hypothesis of using personalised ontologies for user information need capture and user profile acquisition. As shown in Figure 4.3, the Web information gathering system using the Semi-auto user profiles outperformed that using the Manual user profiles. The Semi-auto model is promising and encouraging.

In the Semi-auto model, for a given topic, users first specified their background
Figure 4.2: The Experiment Dataflow in the Preliminary Study
Figure 4.3: The Experimental Results in Preliminary Study
knowledge and constructed the concept models manually against the topic. The concept models, constructed by positive and negative subjects, were used by the Web search agent to retrieve training documents from the Web. The Semi-auto user profiles were then acquired by filtering these retrieved Web documents. The advantage in the Semi-auto model is that the user-interested concepts were specified explicitly. By using these concepts to acquire user profiles, Web documents that were not only syntactically but also semantically relevant to the given topics were acquired. Another advantage of the Semi-auto model is that the training documents were retrieved from the Web. As a result, there were a total of 2775 Web documents (on average 185 per topic) retrieved from the Web by the Semi-auto model for the 15 experimental topics, in which 1398 documents are identified as positive and 1377 negative. Web information has great coverage of topics. Thus, using Web documents as the training sets benefited the topic coverage of user profiles acquired by the Semi-auto model. In addition, the non-binary support values assigned to the documents also benefited the Semi-auto model performance. The documents that were only partially relevant to the topics were then accurately judged, instead of roughly assigning either one for fully relevant or zero for non-relevance. The noise and uncertainty caused by partially relevant documents were then avoided. These advantages leveraged the Semi-auto model performance.

The Manual model acquired user profiles manually. The users read the training documents personally and judged the documents as positive or negative, according to the relevance or non-relevance of documents to the given topics. However, the Manual model suffered from a couple of problems. Although user reading checked and approved that the contents in the training documents were of interest, the concepts contained in the contents were not formally defined. Users had no problem extracting them manually from the documents when reading them. The computational models, however, were unable to extract the underlying concepts as well as human users can. Consequently, the performance of the
information gathering system using the Manual user profiles was weakened. Also, the number of training documents provided to TREC users to read was limited, and thus the topic coverage of user profiles in the Manual model was not as good as that of the Semi-auto model user profiles. For the 15 experimental topics, the Manual model acquired a total of 1054 documents (on average 70 per topic) for user profiles, where 354 are positive and 699 negative, much less than that in the Semi-auto model user profiles. In addition, the Manual model assigned binary values of positive and negative to the training documents. This might cause misjudgement when some documents incorporated only partial relevance to the topics. Consequently, compared with the advantages posed by the Semi-auto model, these weaknesses caused the Manual model to be overtaken by the Semi-auto model in the experiments.

4.5 Summary and Conclusion

The preliminary study presented in this chapter aims to evaluate the hypotheses presented in Chapter 3 before moving on to the development phase. In the study, a method was introduced to acquire training sets from the Web to represent user profiles. Based on the user-given topics, the user concept models were constructed manually. The positive and negative subjects in the concept models were specified, and their influences on user profiles acquisition were measured quantitatively. The training documents retrieved from the Web were filtered and re-ranked based on the positive and negative subjects specified in user concept models. The preliminary model was tested by experiment evaluation using the Reuters Corpus Volume 1 (RCV1) data set. The experimental results demonstrated that higher performance had been achieved by the Web information gathering system using the preliminary model. The preliminary model was promising and encouraging. With this successful result, the hypotheses presented in Chapter 3 are developed in Chapters 5 and 6, and finally evaluated in Chapters 7 and 8.
Chapter 5

Ontology Learning for User Background Knowledge

This chapter presents the methods for extracting user background knowledge and learning ontologies for user concept models. A global ontology, namely world knowledge base, is constructed first. It is utilised for user background knowledge extraction and personalised user ontologies construction. Two ontology learning methods, one semi-automatic and one automatic, are introduced. These use the world knowledge base to construct user personalised ontologies. This chapter focuses on the construction of the world knowledge base and ontology learning methods, and is the first phase in developing the hypothesis introduced in Chapter 3.

5.1 World Knowledge Base

World knowledge is the commonsense knowledge possessed by people and is acquired through their experience and education [238]. It plays an important role in information gathering: as stated by Nirenburg and Raskin [141],
world knowledge is necessary for lexical and referential disambiguation, including establishing coreference relations and resolving ellipsis as well as for establishing and maintaining connectivity of the discourse and adherence of the text to the text producer’s goal and plans.

A world knowledge base is a global ontology that formally describes and specifies world knowledge. With a world knowledge base, a user’s background knowledge is extracted, including concepts both relevant and non-relevant to user information needs. The world knowledge base is utilised by the ontology learning and mining model presented in this thesis.

5.1.1 World Knowledge Representation

Because it aims to extract user background knowledge, the world knowledge base needs to cover an exhaustive range of topics, since users may come from different backgrounds. The Library of Congress Subject Headings* (LCSH) system is ideal for world knowledge base construction. The LCSH system is a thesaurus developed for organising and retrieving information from a large volume of library collections. As a human intellectual endeavour, for over a hundred years the knowledge contained in the LCSH has undergone continuous revising and enriching. The LCSH system represents the natural growth and distribution of human intellectual work, and covers comprehensive and exhaustive topics of world knowledge [26]. In addition, the LCSH system is the most comprehensive non-specialised controlled vocabulary in English. In many respects, the system has become a de facto standard for subject cataloging and indexing, and is used not only as a major subject access tool in library catalogs but also as a means for enhancing subject access to knowledge management systems [26]. Hence, the LCSH provides an ideal knowledge resource in the construction of the world knowledge base.

As one of the the largest and most well-developed intellectual systems ever

constructed, the LCSH has many features. It covers all disciplines of human knowledge. The descriptors in LCSH are classified by professionals, and the classification quality is guaranteed by well-defined and continuously-refined cataloging rules. Compared with other classification/categorisation systems used as knowledge bases in previous works, such as the Library of Congress Classification (LCC) used by Frank and Paynter [50], the Dewey Decimal Classification (DDC) used by Wang and Lee [217], and the reference categorisation (RC) developed by Gauch et al. [55] using online categorisations, the LCSH system is superior, as shown in Table 5.1. The LCSH system covers more topics, and has more descriptors, a more specific structure, and more specific semantic relation specifications. These features make the LCSH system a superior descriptor for world knowledge, and an ideal knowledge base for research on knowledge engineering.

### 5.1.2 World Knowledge Base Construction

**MARC Forms of the LCSH**

The Library of Congress Subject Headings are stored in MARC 21 records for use in computational systems. MARC stands for *Machine-Readable Cataloging*, which is the standard formats for the representation and communication of bibliographic and related information in machine-readable form [113]. The MARC 21 records are in two types: bibliographic records and authority records [113]. Bibliographic records contain information about books, serials, sound recordings, and video recordings. They represent materials in a library’s collection. Authority records contain standardised and controlled forms for names, titles, and subjects,
Chapter 5. Ontology Learning for User Background Knowledge

for use in bibliographic records. The LCSH is specified from the authority records, and used to construct the world knowledge base.

Additionally, authority records provide authority control for the subjects and cross references in catalogs, and thus a linking framework for subjects. Authority control means establishing a recognised form for a subject and using that subject as an access point in a related bibliographic record. Hence, authority forms are used to achieve consistency amongst bibliographic records (materials in a library collection), and to organise the catalog to assist user information gathering in library collections. As the formulation of subjects in authority records is based on generally accepted cataloging and thesaurus-building conventions, the authority forms are also called the authorised, authoritative, or established form in some texts [112].

The raw MARC 21 authority records are stored in a sequential stream of data, as illustrated in Figure 5.1. In fact, the raw MARC 21 authority records provided by the Cataloging Distribution Service (CDS) in the Library of Congress are in a single 130MB file, containing only one data stream. By using the text processing technique of regular expression, the data stream can be separated for individual authority records. After text processing, there are 291,511 individual records specified. Figure 5.2 presents one of the authority records, specified from the part of records illustrated in Figure 5.1. However, as illustrated by the figures, the records are still in the MARC 21 format for computational systems, not for human users. The referring subjects and cross references still remain unclear. Thus, in order to construct the world knowledge base, the raw data authority records need to be parsed, and the meanings underlying the records need to be discovered. In the following subsection, how to parse the MARC 21 formatted authority records will be discussed.
Figure 5.1: Raw Data in the MARC 21 Format of LCSH.

Figure 5.2: An Authority Record in MARC 21 Data
Extraction of the MARC 21 Authority Records

The MARC 21 format of authority records consists of three main components: the leader, the directory, and the variable fields. The leader provides primary information required for processing an authority record. It is the first field in an authority record and has a fixed length of 24 characters. The characters in different positions have meanings for the context of leaders. These meanings are presented in Table 5.2 for the interpretation of the leaders in MARC 21 authority records [113]. The z in the character position 06 indicates that the record is an authority record. The obsolete and deleted records, indicated by d, o, s, and x in the 05 position, refer to non-existing concepts. These non-valuable records can be skipped, and only the valuable authority records indicated by a, c, or n in the 06 position need to be extracted. Also note that the character positions 07 and 08 are with the fixed value of “##” (where # denotes an empty space), the positions 10 and 11 are fixed with “22”, and the last six characters are fixed with “##4500”. Thus, by using the following text:

\[\cdots [a|c|n]z## \cdot 22 \cdots [n|o]##4500\]

with regular expression text processing techniques\(^\dagger\), the raw MARC 21 authority data can be parsed into individual authority records, where 4500 splits the stream data, and \([a|c|n]z\) ensures that only the valuable authority records are extracted. Consequently, each extracted authority record is like the one displayed in Figure 5.2.

Thus, as the first 24 characters form the record leader, for the record presented in Figure 5.2, its leader can be extracted and displayed as:

\[01061cz##2200313n##4500\]

\(^\dagger\)As the usage in regular expression techniques, “.” denotes any but one character, “[a|c|n]” means any one character of a, c, or n, and the same as “[n|o]”.\]
<table>
<thead>
<tr>
<th>Character Positions</th>
<th>References</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>00-04</td>
<td>Record length</td>
<td>a - Increase in encoding level; c - Corrected or revised; d - Deleted; n - New; o - Obsolete; s - Deleted and heading split into two or more headings; x - Deleted and heading replaced by another heading</td>
</tr>
<tr>
<td>05</td>
<td>Record status</td>
<td>n - New; o - Obsolete; s - Deleted and heading split into two or more headings; x - Deleted and heading replaced by another heading</td>
</tr>
<tr>
<td>06</td>
<td>Type of record</td>
<td>z - Authority data</td>
</tr>
<tr>
<td>07-08</td>
<td>Undefined character positions</td>
<td># - Undefined</td>
</tr>
<tr>
<td>09</td>
<td>Character coding scheme</td>
<td># - MARC-8; a - UCS/Unicode</td>
</tr>
<tr>
<td>10</td>
<td>Indicator count</td>
<td>2 - Number of character positions used for indicators</td>
</tr>
<tr>
<td>11</td>
<td>Subfield code length</td>
<td>2 - Number of character positions used for a subfield code</td>
</tr>
<tr>
<td>12-16</td>
<td>Base address of data</td>
<td>[number] - Length of Leader and Directory</td>
</tr>
<tr>
<td>17</td>
<td>Encoding level</td>
<td>n - Complete authority record; o - Incomplete authority record</td>
</tr>
<tr>
<td>18-19</td>
<td>Undefined character positions</td>
<td># - Undefined</td>
</tr>
<tr>
<td>20</td>
<td>Length of the length-of-field portion</td>
<td>4 - Number of characters in the length-of-field portion of a Directory entry</td>
</tr>
<tr>
<td>21</td>
<td>Length of the starting-character-position portion</td>
<td>5 - Number of characters in the starting-character-position portion of a Directory entry</td>
</tr>
<tr>
<td>22</td>
<td>Length of the implementation-defined portion</td>
<td>0 - Number of characters in the implementation-defined portion of a Directory entry</td>
</tr>
<tr>
<td>23</td>
<td>Undefined</td>
<td>0 - Undefined</td>
</tr>
</tbody>
</table>

Table 5.2: The Reference of MARC 21 Authority Record Leaders [113].
With the context explained in Table 5.2, one may see that this authority record contains 1061 characters, as indicated in the 00 to 04 character positions; and is a corrected or revised authority record, as indicated by the c and z in positions 05 and 06. Out of the entire 1061 characters, the leader and directory occupy 313 characters, as indicated in the positions 12 to 16. Finally, this authority record is a complete record, as referred by the n at the position 17. The semantic meanings of the leader in the authority record displayed in Figure 5.2 are interpreted.

The directory defines the structure and format information of an authority record. In leaders, the character positions 12 to 16 indicate the length of the leader and directory portion in the authority records. Because the directory immediately follows the leader and starts with the character position 24, by removing the first 24 characters from this portion, the directory of authority records can be extracted. For the authority record displayed in Figure 5.2, the length of leader and directory is 313 characters. Thus, with one hidden character at the end indicating the finish of the leader and directory portion, the directory can be extracted and displayed as:

```
001001300000 003000400013 005001700017 008004100034 010001700075 040001800092 053001100110 150002600121
450002300147 450002900170 450002700199 450002300226 450002400249 450002400273 450002600297 450002500323 450002700348 450002800375 550002300403 550002700426 550002900453 550003800482 670007800520 680014900598
```

The data elements in a directory can be divided into a series of entries, each with 12 character positions in length. Each entry is for a variable field (control or data) present in the authority record. Each directory entry is 12 character positions in length and contains three portions: the field tag (three character positions), the field length (four character positions), and the starting character
position (five character positions) [112,113]. The field tags are in a fixed length of three characters, indicating the tag number, such as 001 and 550. By using the highlighted entry in the above displayed directory 150002600121 as an example, the first portion 150 means the variable filed is of tag 150, the second portion 0026 means the length is 26 characters, and the last portion 00121 indicates that the starting character is counted from position 121 (position 0 is the first character after the directory portion). As a result, the referring variable field can be extracted and displayed as:

```
150 aBusiness intelligence
```

In these directory entries, the one for variable control fields (field tag 000) is specified first, followed by other variable fields arranged in ascending order of field tags [113].

By interpreting the references contained in the entries in the Leader, the semantic contents of authority records can be discovered. Figure 5.3 displays the interpretation of the raw data of the authority record displayed in Figure 5.2. Compared with the raw data, the interpreted authority record displayed in Figure 5.3 is more meaningful to human users.

**Interpretation of MARC Authority Forms**

In this section, the interpretation of variable fields in MARC 21 authority records is discussed, towards to constructing the world knowledge base from the Library of Congress Subject Headings system.

The subject information stored in MARC 21 authority records consists of three basic portions: headings, cross references, and notes of the authority record [112]:
<table>
<thead>
<tr>
<th>LEADER</th>
<th>01061cs 2200313n 4500</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIRECTORY</td>
<td>001001300000 0050006400013 0050017000017 0080004100034 010001700075 0400001800092 053001100110 1500026000121 450002300147 450002900170 450002700199 450023000226 450002400249 450002400273 450002600297 450002900232 450002700338 4500028000375 550002300403 550002700426 550002900453 550003800482 670007800520 680014900598</td>
</tr>
<tr>
<td>VARIABLE FIELDS</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>001</td>
</tr>
<tr>
<td>2</td>
<td>003</td>
</tr>
<tr>
<td>3</td>
<td>005</td>
</tr>
<tr>
<td>4</td>
<td>008</td>
</tr>
<tr>
<td>5</td>
<td>010</td>
</tr>
<tr>
<td>6</td>
<td>040</td>
</tr>
<tr>
<td>7</td>
<td>053</td>
</tr>
<tr>
<td>8</td>
<td>150</td>
</tr>
<tr>
<td>9</td>
<td>450</td>
</tr>
<tr>
<td>10</td>
<td>450</td>
</tr>
<tr>
<td>11</td>
<td>450</td>
</tr>
<tr>
<td>12</td>
<td>450</td>
</tr>
<tr>
<td>13</td>
<td>450</td>
</tr>
<tr>
<td>14</td>
<td>450</td>
</tr>
<tr>
<td>15</td>
<td>450</td>
</tr>
<tr>
<td>16</td>
<td>450</td>
</tr>
<tr>
<td>17</td>
<td>450</td>
</tr>
<tr>
<td>18</td>
<td>450</td>
</tr>
<tr>
<td>19</td>
<td>550</td>
</tr>
<tr>
<td>20</td>
<td>550</td>
</tr>
<tr>
<td>21</td>
<td>550</td>
</tr>
<tr>
<td>22</td>
<td>550</td>
</tr>
<tr>
<td>23</td>
<td>670</td>
</tr>
<tr>
<td>24</td>
<td>680</td>
</tr>
</tbody>
</table>

Figure 5.3: The parsing result of a MARC 21 authority record. Note that the index 1 to 24 for variable fields are added by the candidate for sake of explanation. They are not specified in authority records.
5.1. World Knowledge Base

- **Heading**: the standardised “authoritative” form of a name, subject, or title that is used for access points on bibliographic records.

- **Cross references**: references that direct a user from a variant form of subject to the authoritative form (called a *see* reference) or from one authoritative form to another authoritative form because they are related to one another (called a *see also* reference).

- **Notes**: notes that contain general information about standardised headings or more specialised information, such as citations for a consulted source in which information is either found or not found about a heading.

Thus, the interpretation of MARC 21 authority records means to specify the heading, cross references, and notes information from the MARC 21 authority records.

Variable fields contain the subject headings and the cross reference information of the subjects present in authority records. There are two types of variable fields in an authority record: variable control fields and variable data fields. The variable control fields are with 001, 003, 005, and 008 tags. In Figure 5.3, the variable fields listed from 1 to 4 are variable control fields, and the remains are variable data fields. While in these variable data fields, some are with 0XX tags (where \( X \in \{0 - 9\} \)). These 0XX variable data fields, together with the variable control fields, contain the standard numbers, classification numbers, and codes that are associated with the authority record. They do not contain descriptive information about the referring subjects by the authority records, and are used only to identify and retrieve records by matching specific criteria \([112,113]\).

Thus, the 0XX variable data fields and variable control fields can be skipped in the world knowledge base construction, unless bridging to other systems like Library of Congress Classification and Dewey Decimal Classification is required.

The variable data fields without the 0XX tags contain the headings, cross references, and notes information of authority records \([112,113]\). These fields are for the descriptive information of the referring subjects. In the authority record
Table 5.3: Subject Identity and References

<table>
<thead>
<tr>
<th>Code</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>0XX</td>
<td>Control information</td>
</tr>
<tr>
<td>1XX</td>
<td>Heading (authoritative and reference)</td>
</tr>
<tr>
<td>2XX</td>
<td>Complex see references</td>
</tr>
<tr>
<td>3XX</td>
<td>Complex see also references</td>
</tr>
<tr>
<td>4XX</td>
<td>See from tracings</td>
</tr>
<tr>
<td>5XX</td>
<td>See also from tracings</td>
</tr>
<tr>
<td>6XX</td>
<td>Reference notes, treatment, notes, etc</td>
</tr>
<tr>
<td>7XX</td>
<td>Heading linking entries</td>
</tr>
<tr>
<td>8XX</td>
<td>Alternative graphics</td>
</tr>
<tr>
<td>9XX</td>
<td>Reserved for local implementation</td>
</tr>
</tbody>
</table>

Table 5.4: Types of Subjects Referred by Variable Fields

<table>
<thead>
<tr>
<th>Code</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>X00</td>
<td>Personal names</td>
</tr>
<tr>
<td>X10</td>
<td>Corporate names</td>
</tr>
<tr>
<td>X11</td>
<td>Meeting names</td>
</tr>
<tr>
<td>X30</td>
<td>Uniform titles</td>
</tr>
<tr>
<td>X40</td>
<td>Bibliographic titles</td>
</tr>
<tr>
<td>X48</td>
<td>Chronological terms</td>
</tr>
<tr>
<td>X50</td>
<td>Topical terms</td>
</tr>
<tr>
<td>X51</td>
<td>Geographic names</td>
</tr>
<tr>
<td>X55</td>
<td>Genre/form terms</td>
</tr>
<tr>
<td>X80</td>
<td>General subdivision terms</td>
</tr>
<tr>
<td>X81</td>
<td>Geographic subdivision names</td>
</tr>
<tr>
<td>X82</td>
<td>Chronological subdivision terms</td>
</tr>
</tbody>
</table>

presented in Figure 5.3, the variable fields listed from 8 to 24 are the descriptive variable data fields. One may see that the information contained in these fields is much more meaningful, comparing to the variable control fields from 1 to 7. The world knowledge base in this thesis is constructed based on the knowledge specified in these descriptive variable data fields of authority records.

The tags of variable data fields are used to identify the subject of an authority record and the related cross references. These cross references link the subjects and thus form the structure of the world knowledge base. The semantics of these tags are present in Table 5.3 and 5.4, in which $X \in \{0 - 9\}$. A tag number is the combination of two entries, one from each of Table 5.3 and 5.4 respectively.
The entry from Table 5.3 refers to the function of the data (a subject heading or a cross reference) within the variable fields, and the entry from Table 5.4 refers to the type of subject described in the variable fields.

Subject headings have various types: names, titles, uniform titles, chronological terms, topics, as presented in Table 5.4. These types are outlined in detail as follows [112,113]:

- **Name heading**: a heading that is a personal, corporate, meeting, or jurisdiction (including geographic) name.

- **Title heading**: a heading contains the title by which an item or a series is identified for cataloging purposes and may be a uniform or conventional title, a page title of a work, or a series title.

- **Uniform title heading**: a heading consisting of the title by which an item or a series is identified for cataloging purposes when the title is not entered under a personal, corporate, meeting, or jurisdiction name in a name/title heading construction.

- **Chronological heading**: A heading consisting of a chronological subject term.

- **Topical heading**: a heading consisting of a topical subject term.

- **Genre/form heading**: a heading consisting of a genre/form subject term.

- **Subdivision heading**: A heading consisting of a general (topical or language), form, geographic, or chronological subject subdivision term. An extended subdivision heading contains more than one subject subdivision term.

The subject information is specified by variable data fields with tag “1XX”. The extraction of subjects from authority records can be explained using the previously discussed variable field again (the variable field No.8 in Figure 5.3):
The field tag is 150, the combination of “1XX” and “X50”. From Table 5.3, the function referred by code “1XX” is “heading”, meaning that the subject specified in this variable data field is the one referred by this authority record. Also from Table 5.4, the type of the subject is “Topical Terms”, as referred by code “X50”. This means that the referring subject is a topical subject. Thus, combining two entries together, the field tag “150” defines that the referring concept is a topical subject, and the label of the subject is specified as “Business intelligence”. By using the information displayed in Table 5.4, other types of subject headings can also be specified, such as “110” for the corporate subject headings, “130” for the uniform title subject headings, and “140” for the bibliographic subject headings. Based on this approach, the subject information referred by authority records can be extracted.

The cross references of subjects can be extracted from the variable data fields with tags “4XX”. The “4XX - see from tracings” function variables refer to the cross references. For the example of the tag “450” in Figure 5.3, say, the variable field No.9:

```
450 aBusiness espionage
```

the field tag “450” is the combination of function code “4XX” and type code “X50”. The code “4XX” indicates the function “See from tracing”, and the code “X50” indicates the type of “topical term”. Combining them together, tag “450” means that the referring concept is a topical subject named “Business espionage”. This subject is a Used for cross reference to “Business intelligence” that is referred by this authority record, as specified in the variable field with tag “150”. One authority record may have multiple Used for cross references. As shown in Figure 5.3, the displayed authority record has many “450” tags. The Used for references specified by the field tags with function “4XX” link the
subjects together. They construct part of the taxonomic structure in the world knowledge base.

The taxonomic and non-taxonomic structure of the world knowledge base is also constructed by the cross references specified by “5XX - See also from tracings” variable fields. These variable fields are designed to specify the **Broader term** and **Related to** references. The variable fields for **Broader term** and **Related to** references are discriminated by \( w\cdot a \) at the beginning (where \( \cdot \) denotes any but one character). If a variable data field starts with \( w\cdot a \), the associated code “5XX” refers to the **Broader term** references; otherwise, it refers to the **Related to** references. These references link the subjects together, and also construct part of the structure in the world knowledge base, where the **Broader term** are taxonomic relations and the **Related to** are non-taxonomic relations.

Back to the sample authority record displayed in Figure 5.3, the No.19, No.20, and No.21 variable fields are with “550” tags and the \( w\cdot a \) at the beginning of the data fields:

<table>
<thead>
<tr>
<th>550 wgaBusiness ethics</th>
</tr>
</thead>
<tbody>
<tr>
<td>550 wgaCompetition, Unfair</td>
</tr>
<tr>
<td>550 wgaIndustrial management</td>
</tr>
</tbody>
</table>

The code “5XX” refers to the function “See also from tracings” and “X50” refers to the type “topical terms”. With the “\( w\cdot a \)” in the beginning, these variable fields indicate that the associated topical subjects “Business ethics”, “Competition, Unfair”, and “Industrial management” are the **Broader term** references of “Business intelligence”, specified by the variable field with tag “150”. In contrast, the subject “Business intelligence” is of the **Narrower term** reference of these “550” and “\( w\cdot a \)” subjects. The hierarchically related subjects are linked by means of these reciprocal **Broader term** and **Narrower term** references. A subject is linked to the level immediately above it and the level immediately below it in the appropriate hierarchical structure [25]. This constructs the taxonomic
Figure 5.4: Subject “Business intelligence” and its cross references extracted from the MARC 21 authority records, where BT refers to “Broader term”, RT refers to “Related to”, and UF refers to “Used for”.

The non-taxonomic structure is constructed by the Related to references in the world knowledge base. Similar to the Used for references, there may be multiple Broader term and Related to references associated with an authority record.

The non-taxonomic structure is constructed by the Related to references in authority records. The No. 22 variable field in Figure 5.3 is also with the tag “550”, however, without \( w^* a \) in the beginning of data field:

| 550 aConfidential business information |

The variable field refers to the subject “Confidential business information”, and the subject is of Related to reference with “Business intelligence”, the subject specified by tag “150” and referred by the authority record. The Related to references construct the non-taxonomic structure of the world knowledge base. These taxonomic references are also important to the construction of world knowledge base.
By interpreting the semantic meanings of variable data fields, the subjects defined by authority records can be extracted, as well as their associated cross references. Figure 5.4 illustrates the subject extraction result from the authority record displayed in Figure 5.3, which is parsed from the raw data displayed in Figure 5.2 and 5.1. This authority record defines a subject labelled “Business intelligence” and the associated cross references, as displayed. Each cross reference refers to another subject, which is defined by another authority record and has its own cross references. The interpretation can also be confirmed by the visualised LCSH system “Classification Web” developed by the Library of Congress\(^\text{‡}\). Figure 5.5 presents the screenshot taken from the Library of Congress Classification Web for the subject “Business intelligence”.

By tracing the cross references, a backbone structure consisting of subjects and linked by cross references can be constructed for the world knowledge base. As the result of construction, the complete world knowledge base contains 491,250 subjects, in which 439,329 are topical subjects, 46,136 are geographic subjects, and 5785 are corporate subjects. These subjects are linked to each other by either taxonomic *Broader term/Narrower term* and *Used for* references, or non-taxonomic *Related to* references. In terms of the taxonomic structure, the backbone structure of the world knowledge base has a maximum depth of 37 levels (on average of 7.29 levels per subject path from leaf to root), far better than those constructed by the prior works presented previously in Table 5.1: the LCC knowledge base used by Frank and Paynter [50], the DDC used by Wang and Lee [217], and the RC used by Gauch *et al.* [55].

### 5.1.3 World Knowledge Base Formalisation

The world knowledge is constructed based on the subjects and cross references extracted from the Library of Congress Subject Heading system.

The primitive concept classes in the world knowledge base are subjects that

\(^\text{‡}\)The Library of Congress Classification Web, http://classificationweb.net/. Note the access is for subscribed users only.
Figure 5.5: Subject "Business intelligence" and its cross references visualised in the Library of Congress Classification Web.
are defined by the subjects specified by the authority records stored in MARC 21 data, the machine-readable form of the LCSH system. In this thesis, the subjects are formally defined:

**Definition 3.** Let $\mathbb{S}$ be the set of subjects, a subject $s \in \mathbb{S}$ is formalised as a 2-tuple $s := \langle \text{label}, \sigma \rangle$, where

- \textit{label} is the label of $s$ specified by the authority records in LCSH MARC 21 repository, and is denoted by \text{label}(s);

- $\sigma(s)$ is a signature mapping defining the cross references of $s$ that directly link to $s$, and $\sigma(s) \subseteq \mathbb{S}$.

Subjects in the world knowledge base are linked to each other by the semantic relations of \textit{is-a}, \textit{part-of}, and \textit{related-to}.

Formally, \textit{is-a} relations describe the situation that the semantic extent referred by a hyponym is within that of its hypernym: for example, a “car” is a “automobile”, and the “car” and “automobile” are on different levels of abstraction (or specificity). \textit{Is-a} relations are transitive and asymmetric. Transitivity means if subject $A$ is a subject $B$ and $B$ is a subject $C$, then $A$ is also a $C$. Asymmetry means if $A$ is a $B$, $B$ then cannot be an $A$: for example, the statement of ‘an automobile is a car” is false because not all automobiles are cars, like motorcycles.

Alternatively, \textit{part-of} relations define the relationship between a holonym subject denoting the whole and a meronym subject denoting a part of, or a member of, the whole: for example, a “wheel” is a part of a “car”. \textit{Part-of} relations also hold the transitivity and asymmetry properties. If $A$ is a part of $B$ and $B$ is a part of $C$, $A$ is also a part of $C$. If $A$ is a part of $B$ and $A \neq B$, $B$ is not a part of $A$.

\textit{Related-to} relations are for two topics related in some manner other than by hierarchy, such as “ships” and “boats”. The semantic meanings of the two topics may overlap to some extent. \textit{Related-to} relations hold the property of symmetry
but not transitivity. If $A$ is related to $B$, $B$ is also related to $A$. Related-to relations are not transitive, which means if $A$ is related to $B$ and $B$ is related to $C$, $A$ may not be necessarily related to $C$, if none of the semantic extents referred by $A$ and $C$ overlap.

The semantic relations in the world knowledge base are formally defined:

**Definition 4.** Let $R$ be the set of relations, a relation $r \in R$ is a 2-tuple $r := (\text{edge, type})$, where

- an edge connects two subjects that holds a type of relation;

- a type of relations is an element of \{is-a, part-of, related-to\}.

The semantic relations of is-a, part-of, and related-to are defined by the cross references clarified in the authority records in LCSH MARC 21 repository. There are three types of cross references defined in the LCSH system: *Broader term* (shortened as $BT$) and *Narrower term* (shortened as $NT$); *Used-for* (shortened as $UF$); and *Related to* (shortened as $RT$) [26]. The $BT$ and $NT$ references are for two subjects describing the same topic but in different abstract (or specific) levels [113]. These references define the is-a relations in the world knowledge base that link the associated pair of subjects. The $UF$ references in the LCSH system describe compound subjects and the subjects subdivided by others [113].

A *Used-for* reference is usually used in two different situations: to help describe an action, for example, “a fork is used for dining”; or to help describe an object, for example, “a wheel is used for a car”. In the these cases, the $UF$ references are in fact the part-of relations. When subject $s$ is used for an action, $s$ actually becomes a part of that action, like “using a fork when dining”; when $s_1$ is used for another subject $s_2$, $s_1$ becomes a part of $s_2$, like “a wheel is a part of a car”. Hence, the $UF$ references define the part-of relations in the world knowledge base.

The $RT$ references are for two subjects related in some manner other than by hierarchy, and are transformed into the related-to relations in the world knowledge base.
5.2. Taxonomy Construction for Ontology Learning

Finally, the world knowledge base is formally defined:

**Definition 5.** Let $WKB$ be a world knowledge base, which is a taxonomy constructed as a directed acyclic graph. $WKB$ consists of a set of subjects linked by their semantic relations, and can be formally defined as a 2-tuple $WKB := \langle S, R \rangle$, where

- $S$ is a set of subjects $S := \{s_1, s_2, \cdots, s_m\}$;
- $R$ is a set of semantic relations $R := \{r_1, r_2, \cdots, r_n\}$ linking the subjects in $S$.

The concept classes in the world knowledge base are defined by the subjects classified in the authority records in the LCSH MARC 21 repository, and the structure is constructed by the cross references defined in the authority records in LCSH MARC 21 repository. Figure 5.6 illustrates a part of the constructed world knowledge base, for the portion dealing with the subject “Business intelligence” that has been discussed throughout this chapter.

5.2 Taxonomy Construction for Ontology Learning

The personalised ontologies in this thesis represent the implicit concept models possessed by users. The ontologies also specify user background knowledge dealing with a given topic. Web users can easily make a decision if a document interests them or not, when they read through the document’s content, because Web users implicitly possess an established concept model based on their background knowledge, and use that model in Web information gathering [110]. Bearing in mind that ontologies are the formal descriptions and specifications of knowledge, if ontologies can be learned to represent user concept models, the semantic mean-
Figure 5.6: A portion of the world knowledge base dealing with the subject "Business intelligence."
ing of information needs can be captured effectively, and thus Web information gathering performance can be improved.

The world knowledge base contains a large volume of subjects and covers an exhaustive range of topics. Thus, the world knowledge base can work as a global ontology in user personalised ontology learning for user background knowledge extraction. For a given topic, three different sets of concepts may need to be extracted: *positive subjects* refer to the concepts that are interesting to the user with respect to the topic; *negative subjects* refer to the concepts that may make paradoxical or ambiguous interpretations of the topic, thus making it difficult to capture the information needs. The last set, *neutral subjects*, refers to the concepts that have no indication of either positive or negative subjects. Because the world knowledge base provides a large frame of concepts, the positive, negative, and neutral subjects can be extracted from it, along with their semantic relationships.

In order to clarify the likelihood of a subject being relevant to the given topic $T$ and interesting to the user, a support $sup(s, T)$ value can be assigned to subjects. The $sup(s, T)$ describes the subjects’ support level to the given topic, within the range of $[1, -1]$. $sup(s, T) = 1$ gives the evaluating subject $s$ the highest support to $T$, $sup(s, T) = -1$ gives the $s$ the lowest support to $T$, and $sup(s, T) = 0$ indicates that $s$ is on the boundary of neither positive nor negative.

These positive, negative, and neutral subjects are extracted from the world knowledge base along with their support values and relationships, and are used to construct personalised ontologies for Web users. Two methods, semi-automatic and automatic, are proposed to extract user background knowledge from the world knowledge base and to construct the personalised ontologies for users. They are introduced in the following sections.
5.2.1 Semi-automatic Ontology Taxonomy Construction

In this semi-automatic ontology learning method, the personalisation of ontologies is adopted through user-system interaction. To help the user-system interaction, a tool called *Ontology Learning Environment* (OLE) is developed to function as a graphic interface between users and the computer system. The OLE provides users with candidate subjects to identify for positives and negatives. The candidate subjects are extracted from the world knowledge base, according to user information needs. Figure 5.7 illustrates a screenshot of the OLE, generated in response to the user given topic “Economic espionage”.

The candidate subjects are presented in the OLE for users to select. The subjects listed on the top-left panel of OLE are the candidate positive subjects extracted from the *WKB*, organised in hierarchical form. Comparing the title of topic (“Economic espionage”) to the label of subjects (*label*(s)), the matched or partially matched subjects are retrieved. The three options located in the middle, “Most coverage”, “General coverage”, and “Least coverage”, determine the matching level defining whether a subject is to be retrieved or not. “Least coverage” is the most restricted option and for full matchings only. Thus, very few subjects can be considered potentially positive and extracted. “General coverage” covers full matchings and one term not-matchings; for example, subjects with label covering “economic” or “espionage” would be considered potentially positive and extracted. “Most coverage” is the most relaxed option and can be for two not-matching terms. However, the “Economic espionage” sample has only two terms in its option. In this case, selecting “Most coverage” gives the same results as selecting “General coverage”. The matching subjects are extracted as candidates for the user to select the positive subjects.

The subjects that directly or indirectly link to the matching candidates are also extracted for candidates. All these subjects are organised in hierarchical form and displayed to the user. This mechanism is to ensure that the candidate subjects not only syntactically match, but are also semantically relevant to the given
5.2. Taxonomy Construction for Ontology Learning

Figure 5.7: Ontology Learning Environment
topic. Note that the “Root” subjects in Figure 5.7 do not exist. The subjects on the first level, for example, “Industrial espionage”, are the most specific subjects extracted from the WKB. The subjects on the branches expanded from the first level subjects are thus the ancestors of these most specific subjects. These subjects are displayed on the top-left panel as candidates. The user selects the least positive subjects to highlight the path of positive subjects; for example, subjects on the path from “Industrial espionage” to “Crime”, and moves the highlighted subjects to the top-right panel. These are the positive subjects feedback from the user.

The negative candidates are extracted based on the user feedback positive subjects. The subjects on the top-right panel are the relevant subjects selected by the user from the top-left panel. These subjects, along with their descendents (more specific) subjects, are extracted and displayed on the bottom-left panel as the negative candidate subjects. Thus, the positive subjects at this stage are displayed together with the negative candidates, they will be discarded from the final negative subject set.

Negative subjects are those ambiguous to the topic. Those subjects linked to positive subjects but paradoxical to the topic have to be identified. The negative candidates are also organised in hierarchical form, and displayed on the bottom-left panel. The subjects on the first level are the ancestor (most abstract) subjects from the positive candidates, and the subjects on the branches expanded from the first level subjects are the descendant (more specific) subjects. The mechanism of displaying negative candidates in a manner from ancestor to descendents subjects, as well as displaying positive candidates from descendents to ancestor subjects, is to ensure that the candidates can have adequate coverage.

The user selected negative subjects are moved to the bottom-right panel, such as “Political ethics” and “Student ethics”, as well as the subjects on the path linking with them. As the positive candidates are a subset of negative candidates, it is possible that some user selected positive subjects may also be on the path with
the selected negative subjects; such as “Ethics”, “Crime”, “Commercial crimes” and “Competition, Unfair” in Figure 5.7. These positive subjects would not be collected for negative subjects. Thus, the subjects displayed on the bottom-right panel, not counting the already identified positive subjects, are the negative subjects feedback by the user.

The remaining subjects from the positive and negative candidates, that are not feedback from the user as either positive or negative subjects, become the subjects neutral to the given topic.

The positive, negative, and neutral subjects define the concept classes in the user personalised ontology, and the semantic relations linking these subjects construct the backbone structure of the ontology. The concepts contained in the personalised ontology consist of three sets:

- positive subjects relevant to the given topic and denoted by $S^+$. Their support values are the highest ($sup(s, T) = 1$ where $s \in S^+$) because they are selected manually by the user, thus, their positive values are approved by the user;

- negative subjects that are paradoxical or ambiguous to the topic and denoted by $S^-$. Their support values are the lowest negative one ($sup(s, T) = -1$ where $s \in S^-$), as they are also selected manually and approved by the user;

- neutral subjects that have no evidence belonging to either side (positive or negative) and denoted by $S^\emptyset$. Their support values are set as the boundary value zero ($sup(s, T) = 0$ where $s \in S^\emptyset$) for not being any site of positives or negatives.

An ontology is constructed in respect of the given topic, based on the user interaction with the OLE.

Figure 5.8 illustrates the ontology (partially) constructed in respect of the topic “Economic espionage”. Note that the semantic relations of is-a, part-of,
Figure 5.8: An ontology constructed for topic “Economic Espionage.” Note that this is only a part of the ontology due to space limit. The white nodes in the ontology are positive subjects, the dark nodes are the negative, and the gray nodes are the neutral subjects.
and related-to are not considered when users select the positive and negative subjects. However, these relations are extracted from the world knowledge base, as well as the candidate subjects. Thus, they also construct the ontology backbone structure with different semantic relations.

This constructed ontology is personalised because the user expresses personal preferences and interests when selecting the positive and negative subjects through the OLE. Therefore, if a user has a topic “New York” and plans for a business trip, the user would have different subjects selected and a different ontology structure, from those selected and constructed by a user planning for a leisure holiday in New York.

5.2.2 Automatic Taxonomy Construction

In the previous section, a semi-automatic ontology taxonomy construction method was introduced that learns personalised ontologies for users adopting user interaction through the OLE. However, the semi-automatic method has limits as it largely relies on user feedback. User involvement improves the effectiveness; however, this makes the method inefficient because Web users may not always like to burden themselves with providing feedback [109, 110]. To solve this, another ontology learning method is introduced here to construct the taxonomy for user ontologies automatically. The method extracts user background knowledge from the world knowledge base according to a given topic, and constructs the ontology taxonomy for the user automatically.

The user background knowledge is represented by the positive and negative subjects in this taxonomy construction method, identified according to the given topic. The title of the given topic $T$ is the starting point of information need capture, which is a set of terms, thus $T := \{t_1, t_2, \ldots, t_n\}$. By using these terms, an automatic syntax-matching mechanism can be used to extract the related subjects from the world knowledge base, along with their associated semantic relationships. The mechanism is presented in Algorithm 1.
5. Ontology Learning for User Background Knowledge

**Algorithm 1:** Automatic Ontology Taxonomy Construction

The support of a subject to the given topic, \( sup(s, T) \), is calculated by:

\[
sup(s, T) = \frac{|label(s) \cap T|}{n} \tag{5.1}
\]

where \( n \) is the size of term set of \( T \). The subjects are different from the semi-automated ontology learning method, and are not extracted by users manually but via an algorithm. Thus, the \( sup(s, T) \) values associated with the subjects can be more specific than only one or zero, because machine learning is more explicit compared with the decision making by human users. Any subjects with \( sup(s, T) > 0 \) are extracted as the positive subjects.

Based on these positive subjects, the negative subjects are extracted from the neighbourhood of positive subjects. As shown in Algorithm 1, \( s \mapsto s' \) and \( dis(s, s') \) are used for negative subject extraction. The \( s_1 \mapsto s_2 \) denotes a path existing between a positive subject \( s \) to another subject \( s' \) in the \( \mathcal{WKB} \). The \( dis(s, s') \) is the conceptual distance, measured by the number of subjects crossing over on the path \( s \mapsto s' \) \cite{81–83}. As argued by Khan et al. \cite{81–83}, the concepts with longer distance in an ontology have smaller similarity values. Thus, the subjects with greater conceptual distance \( dis(s, s') \) values and separated by longer distance are more different in semantics. Based on this argument, only the subjects with \( dis(s, s') \leq 3 \) are extracted in Algorithm 1. The subjects with more than that distance to a positive subject are considered no longer significant and
are ignored. This approach also promises the efficiency of Algorithm 1. These extracted subjects are close to the positive subjects based on their conceptual distance. However, no evidence currently exists that they may support the topic. To be discreet in user background knowledge specification, these subjects are categorised into the negative set temporarily, and their $\text{sup}(s, T)$ are set to negative. These specified subjects will be refined in the next chapter by using an ontology mining method that is based on their semantic relationships and the user’s local document collection.

Two sets of positive and negative subjects are extracted from the world knowledge base:

- positive subjects ($S^+$) that support the topic. Their support values are calculated by Equation (5.1);

- negative subjects ($S^-$) that currently have no evidence of supporting the topic $T$. Their support values are set as $(\text{sup}(s, T) = -1)$.

These positive and negative subjects define the classes in the constructed ontology taxonomy. The semantic relations of *is-a*, *part-of*, and *related-to* linking the positive and negative subjects are also extracted from the world knowledge base with the subjects, as described in Algorithm 1. These semantic relations construct the ontology taxonomy with different semantic relations.

Currently no neutral subjects are identified in the ontology learned by this automatic learning method. Also, as shown on Algorithm 1, different users may have the same positive and negative subject sets if they have the same topic. Hence, at this stage, the ontology taxonomy constructed by the automatic learning method is not yet personalised. In Chapter 6, the constructed ontology taxonomy will be refined for personalisation.
5.3 Ontology Formalisation

The constructed personalised ontologies aim to represent Web users’ implicit concept models $U$, as discussed in Chapter 3. In Section 5.2, the positive, negative, and neutral subjects for a given topic are extracted from the world knowledge base. The semantic relations existing between the subjects, such as is-a, part-of, and related-to, are also extracted, along with the subjects. These subjects and relations construct the classes and backbone structure of an ontology.

The personalised ontologies are formally defined as follows:

Definition 6. The structure of an ontology that describes and specifies topic $T$ is a graph consisting of a set of subject nodes. The structure can be formalised as a 3-tuple $O(T) := \langle S, \text{tax}^S, \text{rel}, \text{axioms} \rangle$, where

- $S$ is a set of subjects consisting of three subsets $S^+$, $S^-$, and $S^\aleph$, where $S^+$ is a set of positive subjects to $T$, $S^- \subseteq S$ is negative, and $S^\aleph \subseteq S$ is neutral;

- $\text{tax}^S$ is the taxonomic structure of $O(T)$, which is a noncyclic and directed graph $(S, E)$, where for each edge $e \in E$, $\text{type}(e) = \text{is-a}$ or $\text{part-of}$, and $\text{tax}(s_1 \rightarrow s_2) = \text{True}$, iff $(s_1 \rightarrow s_2) \in E$;

- $\text{rel}$ is a Boolean function defining the related-to relationship held by two subjects in $S$;

- $\text{axioms}$ are a set of functions, rules, and theorems that restrict the subjects and their relationships in $O(T)$.

The subjects $s \in S$ are associated with a support value $sup(s, T)$, indicating the support rate of $s$ to $T$. For the ontologies learned by using the automatic ontology learning method in Section 5.2.2, $S^\aleph$ is an empty set. The axioms that restrict the subjects and relationships will be discussed in Chapter 6, as the ontology mining methods for ontology personalisation.
5.4 Summary and Conclusion

Learning and mining ontologies to specify user background knowledge is a major objective in this thesis. This chapter presented the methods of extracting user background knowledge from a world knowledge base and for constructing user personalised ontologies. The detailed methods of world knowledge base construction and personalised ontology learning were presented, as the basis of achieving the aforementioned thesis objective.

The world knowledge base is constructed based on a library system. The Library of Congress Subject Headings is a library system that represents human intellectual endeavour and has been undergoing continuous revising and enrichment for over a hundred years. The subjects and associated semantic relations are extracted from the MARC 21 Authority records, the standard formats of LCSH in machine-readable form. Large volumes and a great range of topics are defined in the LCSH system and thus the constructed world knowledge base contains 491,250 topical, geographic, and corporate subjects. Also, various semantic relations associated with the subjects are extracted from the LCSH system and specified in the world knowledge base, including the is-a, part-of, and related-to relations. The semantic relations linking subjects construct the backbone structure of the world knowledge base, which is ideal for knowledge engineering researches and experiments.

The personalised ontologies are constructed, based on the user background knowledge extracted from the world knowledge base. Two ontology learning methods, semi-automatic or automatic, were introduced in this chapter. The semi-automatic ontology learning method relies on a tool called the Ontology Learning Environment to extract interesting subjects from the world knowledge base by user interaction, including the positive, negative, and neutral subjects related to the given topics. The automatic method extracts positive and negative subjects from the world knowledge base by measuring the conceptual distance between subjects and the given topics. Linking via associated semantic relations,
these subjects construct the user personalised ontologies, based on the structure defined by the world knowledge base.

However, the semi-automatic and automatic ontology learning methods have their limitations. While the semi-automatic ontology learning method benefits from the effectiveness achieved by users selecting the subjects of interest manually, it suffers from problems such as:

- users may miss some interesting subjects when selecting from a large set of candidates;
- the candidate subjects provided by the OLE may have inadequate coverage of possible subjects, as some semantically related but not syntactically related subjects may be missed.

The automatic ontology learning method does not require effort from users and thus has no user-prone errors. However, it also suffers from the second problem. The syntactic mechanism used by the automatic learning method cannot guarantee the semantic accuracy of extracted positive subjects. In addition, the ontologies constructed by using the automatic learning method contain only positive and negative subjects. They are rough compared to those constructed by the semi-automatic learning method, and also need more effort to refine the ontologies for personalisation.

Thus, the personalised ontologies constructed by using either the semi-automatic or the automatic ontology learning methods need to be refined. Their specification of user background knowledge also needs to be improved. In the next chapter, these issues will be addressed by introducing a multidimensional ontology mining method.
Chapter 6

Ontology Mining for Personalisation

This chapter introduces an ontology mining method that aims to refine and populate the taxonomy of ontologies constructed in Chapter 5 and discover more on-topic concepts from these ontologies. As discussed in Chapter 5, the taxonomy of ontologies constructed by using either the semi-automatic or automatic methods needs to be refined and populated with instances. In this chapter, a multidimensional ontology mining method, Specificity and Exhaustivity, is introduced to solve this problem, using the user Local Instance Repositories.

Ontology mining in this thesis refers to discovering and weighting the concepts in ontologies. In the multidimensional ontology mining method, specificity describes the focus of a subject’s semantic meaning on a given topic, whereas exhaustivity restricts the extent of semantic meaning covered by a subject that deals with the topic. This multidimensional method aims to investigate the concepts and the strength of associations between them in ontologies.
6.1 Specificity

The specificity (denoted spe) describes a subject’s semantic focus on a topic. A subject’s specificity has two focuses: the subject’s focus on its referring concepts and the subject’s focus on the given topic. They should be addressed separately. By calling the former semantic specificity and the latter topic specificity, the specificity and exhaustivity of subjects are defined and utilised for user background knowledge specification in the following sections.

6.1.1 Semantic Specificity

The semantic specificity refers to a subject’s focus on its referring concepts. The strength of such focus is influenced by the subject’s locality in the taxonomic structure of ontologies [203]. As stated in Definition 6 in this thesis, the taxonomic structure $\text{tax}^S$ of ontology $\mathcal{O}(T)$ is a graph linked by semantic relations. The subjects located at upper bound levels toward the root are more abstractive than the subjects at lower bound levels towards the “leaves”. The upper bound level subjects have more descendent subjects covered and thus more concepts referred, compared with the lower bound level subjects. Thus, in terms of a particular concept being referred to by both an upper bound and a lower bound subject, the latter has stronger focus because it has fewer concepts referred.

The concepts referred to by a child subject are more specific than those referred to by its parent subjects. The child subject’s semantic specificity is hence greater than that of its parent subjects. Different hierarchical relations, such as is-a and part-of, may have different contributions to the semantic specificity posed by subjects. Thus, the semantic specificity measure of a subject $s$ relies on the hierarchical semantic relations in the ontology structure. Because subjects have fixed locality on the $\text{tax}^S$ of $\mathcal{O}(T)$, semantic specificity can also be called absolute specificity, and denoted by $\text{spe}_a(s)$.

The semantic analysis of subject locality for measuring $\text{spe}_a$ is described in Algorithm 2. The $\text{isA}(s)$ and $\text{partOf}(s)$ are two functions in the algorithm.
The isA(s) returns the is-a child subjects in taxS (the subjects that directly link to s and hold is-a relationship to s). The partOf(s) returns the part-of child subjects in taxS (the subjects that directly link to s and hold part-of relationship to s). They satisfy isA(s) ⊆ σ(s) ⊂ S, partOf(s) ⊆ σ(s) ⊂ S, and isA(s) ∩ partOf(s) = ∅. Algorithm 2 is efficient, with the complexity of only $O(n)$, where $n = |S|$. It terminates eventually because the taxS is a directed acyclic graph, as defined in Definition 6.

As the taxS structure in ontology $\mathcal{O}(T)$ is a graphic taxonomy, the leaf subjects have no descendants. Thus, they have the strongest focus on their referring concepts and the highest semantic specificity $spe_a(s)$. By setting the $spe_a$ range as $(0,1]$ (greater than 0, less than or equal to 1), the leaf subjects have the strongest $spe_a(s)$ and full value 1, and the root subject of taxS has the weakest $spe_a(s)$ and the smallest value in $(0,1]$. Toward the root of taxS, the semantic specificity $spe_a(s)$ decreases for each level up. A coefficient $\theta$ is applied to $spe_a(s)$ analysis, defining the decreasing rate of semantic specificity for focus lost from lower bound toward upper bound levels in the taxS. ($\theta = 0.9$, meaning that the reducing rate is 10%, was used in the experiments conducted in this thesis.)

```
input: a personalised ontology $\mathcal{O}(T) := \langle tax^S, rel \rangle$; a coefficient $\theta$ between (0,1).
output: $spe_a(s)$ applied to specificity.
1 set $k = 1$, get the set of leaves $S_0$ from taxS, for ($s_0 \in S_0$) assign $spe_a(s_0) = k$;
2 get $S'$ which is the set of leaves in case that we remove the nodes $S_0$ and the related edges from taxS;
3 if ($S' == \emptyset$) then return;//the terminal condition;
4 foreach $s' \in S'$ do
5    if (isA($s'$) == $\emptyset$) then $spe_1_a(s) = k$;
6    else $spe_1_a(s) = \theta \times \min\{spe_a(s)|s \in isA(s')\}$;
7    if (partOf($s'$) == $\emptyset$) then $spe_2_a(s) = k$;
8    else $spe_2_a(s) = \sum_{s \in partOf(s') \text{ and } spe_a(s)}$;
9    $spe_a(s') = \min(spe_1_a(s), spe_2_a(s))$;
10 end
11 $k = k \times \theta$, $S_0 = S_0 \cup S'$, go to step 2.
Algorithm 2: Analysing semantic relations for semantic specificity
```

From the leaf subjects toward upper bound levels in the taxonomic structure taxS in ontology $\mathcal{O}(T)$, if a subject has is-a child subjects, it should not have
greater *semantic specificity* compared with any one of its *is-a* child subjects. In *is-a* relationships, a parent subject is the abstractive description of its child subjects to the referring concepts. However, the abstraction sacrifices the focus and specificity of the referring concepts. Thus, the $spe_a(s)$ value of a parent subject is defined as the smallest $spe_a(s)$ value of its *is-a* child subjects, applying the decreasing rate coefficient $\theta$.

If a subject has *part-of* child subjects, the *semantic specificity* of all *part-of* child subjects takes part of their parent subject’s *semantic specificity*. As a *part-of* relation, the concepts referred by a parent subject are the combination of that by its *part-of* child subjects. Therefore, a subject’s $spe_a$ is defined as the average $spe_a$ value of its *part-of* child subjects, applying the reducing rate $\theta$ coefficient.

The $spe_a$ values of a subject’s *is-a* and *part-of* child subjects should be addressed separately, if the subject has direct child subjects mixed with *is-a* and *part-of* child subjects. A *semantic specificity* value $spe1_a$ is first calculated for the *semantic specificity* inherited from the *is-a* child subjects, and then $spe2_a$ for the *semantic specificity* inherited from the *part-of* child subjects. The approaches to calculate $spe1_a$ and $spe2_a$ are the same as described previously. Following the principle that specificity decreases for the subjects located toward the upper bound levels, the smaller value of $spe1_a$ and $spe2_a$ is chosen and assigned to the parent subject for the final *semantic specificity*.

In summary, the *semantic specificity* of subjects is measured based on the investigation of subject locality in the taxonomic structure $taxS$ of $O(T)$. In particular, the influence of locality comes from the subject’s taxonomic semantic (*is-a* and *part-of*) relationships with neighbours. Investigations on influences of the concept locality in ontologies have been conducted by many prior works; such as Tran et al. [203]. However, the existing works do not emphasise the strength of semantic *is-a* and *part-of* relations. The *semantic specificity* method in this thesis solves this problem and emphasises the *is-a* and *part-of* semantic relations.
6.1.2 Topic Specificity

The *topic specificity* refers to the focus of subjects on a given topic. The prior ontology learning research does not often take into account the problems that the ontologies aim to solve, such as [4, 127, 137, 193]. As pointed out by Noy [142], ontologies are to share knowledge among different applications for problem solving. Ontology construction and utilisation should be considered under the constraint of the problems intended to solve. In this section, *topic specificity* is introduced to value the strength of subjects focused on user given topics. The method deals with the personalised Web information gathering problem.

The *topic specificity* measures the focus of subjects on the given topic, which refers to a user’s personal interests and information needs. The ontology constructed in Chapter 5 aims to discover interesting concepts for personalised Web information gathering. Business travellers can expect to have personalised results gathered for searching the same topic of “New York” that are different from those gathered by leisure travellers. To business travellers, the concepts associated with “leisure holiday in New York” have a different focus from their information needs, in comparison with that associated with “business trip in New York”. The *topic specificity* scales the strength of subjects regarding user information needs.

The interesting concepts can be discovered from a user’s personal information collections, such as user stored documents, browsed Web pages, and compiled/received emails [123]. These documents have content-related descriptors associated with the concepts specified in external knowledge bases [38]; for example, the metadata tags in XML, RDF, OWL, DAML, and XHTML documents citing the concepts in knowledge bases. This kind of documents with semantic meta-data becomes more and more popular on the Web today, and are argued to be the mainstream of semantic Web documents [4, 127, 206]. In this thesis, such personal information collected is called a user’s Local Instance Repository (LIR), and each document is an instance in the LIR. Because of the specified content-related descriptors, these instances can be used to populate the ontology
taxonomy constructed for the user, and the user’s interests can be also discovered from his (her) LIR.

For the sake of experiments, in this thesis the information items in library catalogues are used as the semantic Web documents. A user’s LIR is simulated by a collection of user-visited information items in library catalogues. As previously discussed in Chapter 5, the LCSH system has a thesaurus that contains the content-related descriptors (subjects) in controlled vocabularies. Corresponding to the descriptors in LCSH, the catalogues of library collections also contain associated descriptive information of library stored books and documents. Figure 6.1 displays a sample instance in the LIRs used in this thesis, an information item in the catalogue describing a book stored in the Queensland University of Technology (QUT) Library*. The descriptive information, such as the title and table of contents, are provided by the author, and the summary is provided by specialist librarians. This descriptive information is expert-classified and trustworthy, and thus can be recognised as the extensive knowledge resource. A list of content-based descriptors (subjects) is also cited on the bottom of Figure 6.1, indexed by their focuses on the item’s content. These subjects provide a bridging connection for the extensive concepts and the concepts specified in the world knowledge base. The ontology mining method is to discover interesting concepts for a user, from both the user’s LIR and personalised ontology.

By calling an information item in LIRs an instance, the relationship between a user’s personalised ontology and LIR can be explored. Firstly, the strength of an instance to a citing subject needs to be clarified. As mentioned previously, the subjects cited by an instance are indexed by their focuses on the content of the instance. Many subjects cited by one instance would thus cause subject specificity loss, as one subject deals only with a part of the instance content. In contrast, the connection held by a subject and an instance is strengthened if fewer subjects are cited by the instance, and the subject is ranked on the top of

6.1. Specificity

Figure 6.1: An Information Item in the QUT Library Catalogue

<table>
<thead>
<tr>
<th>Author</th>
<th>Landsman, Mark, 1966-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Dictatorship and demand: the politics of consumerism in East Germany / Mark Landsman</td>
</tr>
<tr>
<td>Published</td>
<td>Cambridge, MA: Harvard University Press, 2005</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ITEM IDEN</th>
<th>CALL NO</th>
<th>STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carseldine</td>
<td>306.30943109045 1</td>
<td>IN LIBRARY</td>
</tr>
</tbody>
</table>

Table of Contents

<table>
<thead>
<tr>
<th></th>
<th>Production and consumption: establishing priorities</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The contest begins: the currency reform, the Berlin blockade, and the introduction of the HO</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>The planned and the unplanned: consumer supply and provisioning crisis</td>
<td>74</td>
</tr>
<tr>
<td>3</td>
<td>The rise, decline, and afterlife of the new course</td>
<td>115</td>
</tr>
<tr>
<td>4</td>
<td>Demand research and the relations between trade and industry</td>
<td>149</td>
</tr>
<tr>
<td>5</td>
<td>Crisis revisited: the main economic task, and the building of the Berlin Wall</td>
<td>173</td>
</tr>
</tbody>
</table>

Description xii, 296 p.; 24 cm.

Series Harvard historical studies ; 147

ISBN 087401969X (alk. paper)

Bibliography Includes bibliographical references (p. 223-287) and index.

Summary "Based on research in recently opened East German state and party archives, this book depicts a regime caught between competing pressures. While East Germany's leaders followed a Soviet model, which fetishized productivity in heavy industry and prioritized the production of capital goods over consumer goods; they nevertheless had to contend with the growing allure of consumer abundance in West Germany. The usual difficulties associated with satisfying consumer demand in a socialist economy acquired a uniquely heightened political urgency, as millions of East Germans red across the open border. "A new vision of the East-West conflict emerges, one fought as much with washing machines, televisions, and high fashion as with political propaganda, espionage, and nuclear weapons. Dictatorship and Demand deepens our understanding of the Cold War."--BOOK JACKET.

Subject Consumption (Economics) -- Germany (East)  
Socialism -- Germany (East)  
Germany (East) -- Economic conditions  
Germany (East) -- Politics and government.
the subject citing list. Hence, the strength of an instance \( i \) to a subject \( s \) can be calculated by:

\[
str(i, s) = \frac{1}{\text{priority}(s, i) \times n(i)};
\]  

(6.1)

where \( n(i) \) is the number of subjects on the citing list of instance \( i \), \( \text{priority}(s, i) \) is the index (starting from one) of \( s \) on the list cited by \( i \). The \( str(i, s) \) aims to measure the reference strength between instances and subjects.

With the strength value of instances to subjects determined, the relationship between the instances in a user’s LIR and the subjects in the personalised ontology can be defined. Let \( \Omega = \{i_1, i_2, \ldots, i_k\} \) be a finite and nonempty set of instances in the LIR, \( f(i, s) \) defines the existing relationship between an instance \( i \) and a subject \( s \):

\[
f(i, s) = \begin{cases} 
\text{True} & \text{if } str(i, s) \geq \text{min}_str; \\
\text{False} & \text{otherwise.}
\end{cases}
\]  

(6.2)

where \( \text{min}_str \) is the minimal \( str \) value for filtering out the noisy pairs. Given an \( i \in \Omega \), a set of subjects can be extracted from \( \mathcal{S} \) by using the following mapping:

\[
\eta : \Omega \rightarrow 2^\mathcal{S}, \quad \eta(i) = \{s \in \mathcal{S} | f(i, s) = \text{True}\}.
\]  

(6.3)

The mapping function \( \eta(i) \) describes the subjects cited by an instance \( i \). In order to classify instances, the reverse mapping \( \eta^{-1} \) of \( \eta \) can also be defined:

\[
\eta^{-1} : \mathcal{S} \rightarrow 2^\Omega, \quad \eta^{-1}(s) = \{i \in \Omega | f(i, s) = \text{True}\}.
\]  

(6.4)

The mappings \( \eta \) and \( \eta^{-1} \) reveal the relationships between instances and subjects. Each instance maps to a set of subjects in \( \mathcal{S} \), and each subject is cited by a set of instances in \( \Omega \). Each pair, \( (i, s) \), is associated with a strength value defined by Equation (6.1). Figure 6.2 presents a sample mapping related to the topic “Business intelligence”.

From Definition 6, it can be understood that a user’s personalised ontology contains a set of positive subjects, a set of negative subjects, and a set of neutral
Specificity

Figure 6.2: Mappings of Subjects and Instances

subjects, pertaining to a given topic. Based on the mapping of Equation (6.3), if an instance maps to only positive subjects, the instance fully supports the semantic of the given topic; if it maps to only negative subjects, it is strongly against the semantic of the given topic. Hence, the strength of an instance supporting or against a given topic $T$ can be measured by utilizing the mappings of Equation (6.3) and the instance-subject strength value in Equation (6.1):

$$str(i, T) = \sum_{s \in \eta(i)} str(i, s) \times sup(s, T).$$  

Recall back to the discussions in Chapter 5, where positive subjects have $sup(s, T) > 0$, negative subjects have $sup(s, T) < 0$, and neutral subjects have $sup(s, T) = 0$. The value of $str(i, T)$ could be negative if the more and stronger negative subjects are cited by an instance $i$. In that case, the concepts referred by instance $i$ are against topic $T$. The value of $str(i, T)$ could also be zero, if the subjects cited by $i$ are neutral subjects only, or the cited negative subjects have exactly the same strength as that of cited positive subjects. In this case, the concepts referred to by $i$ have no evidence of supporting or being against $T$. Finally, if $str(i, T) > 0$, the cited positive subjects must have strong support for $T$, and thus the concepts referred by instance $i$ support the topic.

The topic specificity of subjects is measured based on the instance-topic
strength of citing instances. With respect to the absolute specificity, the topic specificity can also be called relative specificity, denoted by $spe_r(s, T, LIR)$. Because the concepts referred by instances are specified by the cited subjects, a subject’s $spe_r(s, T, LIR)$ can be calculated by:

$$spe_r(s, T, LIR) = \sum_{i \in \eta^{-1}(s)} str(i, T).$$

(6.6)

Since the $str(i, T)$ from Equation (6.5) could be positive or negative, the value of $spe_r(s, T, LIR)$ could also be positive or negative as well. The topic specificity of subjects is based on the referring concepts of subjects, and not on the selection of users (in the semi-automatic ontology learning method introduced in Chapter 5.2.1) or the measure of the syntactic mechanism (in the automatic ontology learning model introduced in Chapter 5.2.2).

As discussed previously, the specificity describes a subject’s semantic focus. Thus, a subject’s focus on its referring concepts and on the given topic must both be counted. Therefore, the final specificity of a subject is composed of semantic and topic specificity values, and calculated by:

$$spe(s, T) = spe_a(s) \times spe_r(s, T, LIR).$$

(6.7)

The $spe_a(s)$ is scaled by investigating the subject locality, and the $spe_r(s, T, LIR)$ is measured by assessing its citing instances in LIRs. As a result of Equation (6.7), the subjects that are located towards the lower bound levels in the ontology and with more positive citing instances would have greater specificity values.

### 6.2 Exhaustivity

The exhaustivity (denoted $exh$) of a subject refers to the extent of concepts dealt with by the subject, in respect to a given topic. The extent of on-topic concepts referred by a subject extends if the subject has more positive descendants to the
topic. In contrast, if the subject has more negative descendants, the extent of on-topic concepts referred by the subject shrinks.

Because the extent is defined for on-topic concepts, *exhaustivity* needs to take the user interests into account. A subject in the personalised ontology for a business traveller should have different *exhaustivity* value from that for a leisure traveller. Hence, by defining $vol(s)$ as a set of direct and indirect descendants of subject $s$ (including $s$ and all its *is-a* and *part-of* child subjects), where the elements are determined by:

$$vol(s) = \{s'|s' \in S, \exists \text{ a path in } \mathcal{E} \text{ from } s' \text{ to } s\} \quad (6.8)$$

a subject’s *exhaustivity* is measured by aggregating the *topic specificity* of all subjects appearing in its $vol(s)$:

$$exh(s, T) = \sum_{s' \in vol(s)} \sum_{i \in \eta^{-1}(s')} str(i, T) \times spe_r(s', T, LIR). \quad (6.9)$$

Note that in Equation (6.9), the *exhaustivity* relies on the *semantic specificity*, as the *exhaustivity* refers to the extent of on-topic interesting concepts. Thus, if more positive subjects with higher *specificity* values are in the $vol(s)$ of a subject, the referring on-topic and interesting concepts would be extended and the subject’s *exhaustivity* value increases. In contrast, if more negative subjects are in the $vol(s)$, the negative proportion in the referring concepts becomes bigger and the subject’s *exhaustivity* value decreases. The constraints of *specificity* and *exhaustivity* in ontologies will be further investigated later in this chapter.

Subjects are considered on-topic and interesting to the user only if the subjects’ *specificity* and *exhaustivity* are of positive values. Thus, the subject sets of $S^+, S^-$ and $S^\#$, originally identified by the user in Definition 6, can be refined
after ontology mining for the *specificity* and *exhaustivity* of subjects:

\[
\begin{align*}
S^+ &= \{s | (\text{spe}(s, T) > 0), (\text{exh}(s, T) > 0), s \in \mathcal{S}\}; \\
S^- &= \{s | (\text{spe}(s, T) < 0), (\text{exh}(s, T) < 0), s \in \mathcal{S}\}; \\
\mathcal{S}^\text{\#} &= \{s | s \in (\mathcal{S} - (\mathcal{S}^+ \cup \mathcal{S}^-))\}.
\end{align*}
\]

6.3 Interesting Concepts Discovery

According to a given topic, the positive, negative, and neutral subjects are initially extracted in the ontology learning phase. These subjects are refined in the previous section based on semantic analysis. However, there may still be some potentially interesting concepts that are overlooked in previous phases. In this section, a method is presented that aims to discover such potentially interesting concepts from the negative subject set \( \mathcal{S}^- \) and neutral subject set \( \mathcal{S}^\text{\#} \). The method further refines the constructed personalised ontologies.

The potentially interesting concepts are discovered from the user’s LIR, based on the citation of subjects to instances. First introduced is the cover set of a subject \( \text{coverset}(s) \) that refers to the extent of instances in an LIR citing \( s \). \( \text{coverset}(s) \) is defined based on the mappings of Equation (6.3) and (6.4) by:

\[
\text{coverset}(s) = \eta^{-1}(s).
\]

The cover set \( \text{coverset}(s) \) aims at defining the related-to subjects of \( s \) (the subjects that directly link to \( s \) and hold related-to relationship with \( s \)). If \( \text{coverset}(s_1) \cap \text{coverset}(s_2) \neq \emptyset \), \( s_1 \) and \( s_2 \) have concepts overlapped. One may then conclude that they are related to each other more or less. Figure 6.3 illustrates this semantic discovery by using \( \text{coverset} \). In the figure, \( s_3 \) and \( s_4 \) are relevant to \( s_1 \), but \( s_2 \) is not. Assume that subject \( s_1 \) in Figure 6.3 belongs to the positive set, \( s_2, s_3, \) and \( s_4 \) belong to the negative set (or neutral set); it can be said that \( s_3 \) and \( s_4 \) are also interesting to the user because they are relevant to
positive $s_1$, although they are classified in the negative set $S^-$ originally. Based on that, the underlying interesting subjects that were overlooked in the previous phases can be determined from $S^-$ and $S^\aleph$.

The interest level of these newly discovered subjects can be measured according to the size of their overlapping concepts with the positive subjects. These subjects from $S^-$ or $S^\aleph$ become interesting because they hold related-to relationships with the positive subjects in $S^+$. Thus, these positive subjects have the authority to determine the interest level of the newly discovered interesting subjects. A subject is more interesting if it has more related-to positive subjects and these related-to positive subjects are more on-topic of $T$. Based on these, let

$$\hat{S}(s) = \{s'| s' \in S^+, \text{coverset}(s') \cap \text{coverset}(s) \neq \emptyset\}; \quad (6.14)$$

the interest level of a $s \in S^- \cup S^\aleph$ can be calculated by:

$$\text{interest}(s, T) = \frac{\sum_{s' \in \hat{S}(s)} \text{conf}(s' \rightarrow s) \times \text{sup}(s', T)}{|\hat{S}(s)|}; \quad (6.15)$$
where \( \text{sup}(s', T) \) could be either the specificity \( \text{spe}(s', T) \) from Equation (6.7) or the exhaustivity \( \text{exh}(s', T) \) from Equation (6.9), depending on the specificity or exhaustivity preference of the system, as long as the preference is consistent. The \( \text{conf}(s' \rightarrow s) \) is the confidence of \( s \) received from the positive subject \( s' \) and calculated by:

\[
\text{conf}(s' \rightarrow s) = \frac{|\text{coverset}(s') \cap \text{coverset}(s)|}{|\text{coverset}(s')|}.
\] (6.16)

In order to prune the noisy and weak findings in the discovered interesting subjects, a minimum interest level should be applied to the method. A subject \( s \in S^- \cup S^w \) can be recognised as interesting to the user only if its interest level is greater than the minimal requirement. Because the discovered subjects rely on their related-to positive subjects, these positive subjects also have the authority to determine the minimum interest level:

\[
\text{min}_\text{interest} = \alpha \times \frac{\sum_{s \in S^+} \text{sup}(s, T)}{|S^+|};
\] (6.17)

where \( \alpha \) is a parameter for adjusting the minimum interest level. (Based on the experiments conducted and discussed in Chapter 7 and 8, \( \alpha = 1.5 \) delivers the best performance to the experimental model.)

With the interest level defined, the support value \( \text{sup}(s, T) \) of discovered interesting subjects can be calculated. This makes these subjects able to be used consistently with other subjects in \( O(T) \). The \( \text{sup}(s, T) \) takes count of the specificity (or exhaustivity) and the citing instances. Thus, for a discovered interesting subject with interest level greater than \( \text{min}_\text{interest} \), its support value to \( T \) is calculated by:

\[
\text{sup}(s, T) = \frac{\sum_{s' \in \hat{S}(s)} \text{conf}(s' \rightarrow s) \times \text{sup}(s', T)}{|\hat{S}(s)|}.
\] (6.18)

The newly discovered interesting subjects have not yet counted into \( S^+ \) at this
stage. This is because $\hat{S}(s) \subseteq S^+$, according to Equation (6.14), $\{\text{sup}(s', T) | s' \in \hat{S}(s)\}$ is fixed. This guarantees that Equation (6.18) would not fall into recursive deadlock in computation.

With their $\text{sup}(s, T)$ values associated, the underlying interesting subjects discovered from $S^-$ and $S^\aleph$ can be finally added into $S^+$ by:

\[
S^+ = S^+ \cup \{s | s \in S^- \cup S^\aleph, \text{interest}(s, T) \geq \text{min}_\text{interest}\}; \quad (6.19)
\]

\[
S^- = S^- - \{s | s \in S^-, \text{interest}(s, T) \geq \text{min}_\text{interest}\}; \quad (6.20)
\]

\[
S^\aleph = S^\aleph - \{s | s \in S^\aleph, \text{interest}(s, T) \geq \text{min}_\text{interest}\}. \quad (6.21)
\]

The personalised ontology, constructed in Chapter 5 and refined early in this Chapter by a multidimensional method using Specificity and Exhaustivity, is then further refined for personalisation.

### 6.4 Theorems for Ontology Restriction

A few theorems are now introduced, based on the subject analysis of multidimensional specificity and exhaustivity:

**Theorem 6.1.** A leaf subject in an ontology has the same value of specificity and exhaustivity.

**Proof 1.** As $s$ is a leaf subject, from Equation (6.8), we have $\text{vol}(s) = \{s\}$, from Eq (6.9), we have

\[
exh(s, T) = \sum_{s' \in \text{vol}(s)} \sum_{i \in \eta^{-1}(s')} \text{str}(i, T) \times \text{spe}_a(s', T)
\]

\[
= \text{spe}_a(s', T) \times \sum_{i \in \eta^{-1}(s)} \text{str}(i, T)
\]

\[
= \text{spe}_a(s', T) \times \text{spe}_r(s, T, \text{LIR})
\]

\[
= \text{spe}(s, T) \quad \square
\]
Theorem 6.2. Let \( s_1, s_2 \) be two subjects in \( \mathcal{O}(T) \), \( s_1 \in \text{vol}(s_2) \), and \( \eta^{-1}(s_1) = \eta^{-1}(s_2) \), we always have

\[
spe(s_1, T) \geq spe(s_2, T).
\]

Proof 2. From Equations (6.6) and (6.7), we have:

\[
spe(s_1, T) - spe(s_2, T) = spe_a(s_1) \times spe_r(s_1, T, \text{LIR}) - spe_a(s_2) \times spe_r(s_2, T, \text{LIR})
\]

\[
= spe_a(s_1) \times \sum_{i \in \eta^{-1}(s_1)} str(i, T) - spe_a(s_2) \times \sum_{i \in \eta^{-1}(s_2)} str(i, T)
\]

\[
=(spe_a(s_1) - spe_a(s_2)) \times \sum_{i \in \eta^{-1}(s_1)} str(i, T)
\]

\[\therefore \text{There exists a path from } s_1 \text{ to } s_2 : s_1 \rightarrow s' \rightarrow \cdots \rightarrow s'' \rightarrow s_2;\]

From Algorithm 2, we have \( spe_a(s_1) \geq spe_a(s'), \cdots, spe_a(s'') \geq spe_a(s_2); \)

\[\therefore spe_a(s_1) \geq spe_a(s_2) \text{ and } spe(s_1, T) - spe(s_2, T) \geq 0. \]

Theorem 6.3. Let \( s_1, s_2 \) be two subjects in \( \mathcal{O}(T) \), and \( s_1 \in \text{vol}(s_2) \).

1. If \( \text{vol}(s_2) \subseteq S^+ \), we always have \( exh(s_1, T) \leq exh(s_2, T) \);

2. If \( \text{vol}(s_2) \subseteq S^- \), we always have \( exh(s_1, T) \geq exh(s_2, T) \).
6.5. Ontology Learning and Mining Model

Proof 3. From Equation (6.9), we have:

\[
\begin{align*}
\text{exh}(s_2, T) - \text{exh}(s_1, T) & = \sum_{s' \in \text{vol}(s_2)} \sum_{i \in \eta^{-1}(s')} \text{str}(i, T) \times \text{spe}_a(s', T) - \sum_{s'' \in \text{vol}(s_1)} \sum_{i \in \eta^{-1}(s'')} \text{str}(i, T) \times \text{spe}_a(s'', T) \\
& = \sum_{s''' \in (\text{vol}(s_2) - \text{vol}(s_1))} \sum_{i \in \eta^{-1}(s''')} \text{str}(i, T) \times \text{spe}_a(s''', T) \\
& = \sum_{s''' \in (\text{vol}(s_2) - \text{vol}(s_1))} \text{spe}(s''', T) \\
\end{align*}
\]

\[\vdash \text{From Equation (6.10), for } \forall s''' \in \text{vol}(s_2) \text{ and } \text{vol}(s_2) \subseteq S^+ \Rightarrow \text{spe}(s''', T) > 0\]

\[\vdash \text{exh}(s_2, T) - \text{exh}(s_1, T) \geq 0; \text{ Analogically, from Equation (6.11), for }\]

\[\forall s''' \in \text{vol}(s_2) \text{ and } \text{vol}(s_2) \subseteq S^- \Rightarrow \text{spe}(s''', T) < 0\]

\[\vdash \text{exh}(s_2, T) - \text{exh}(s_1, T) \leq 0, \text{ if } \text{vol}(s_2) \subseteq S^-. \]

These theorems restrict the utilisation of specificity and exhaustivity in ontology mining. Theorem 6.1 describes the leaf subjects in terms of specificity and exhaustivity. Theorem 6.2 guarantees that a subject must be more specific than any one at a higher level in the ontology, if they hold the same strengths to a topic. Theorem 6.3 constrains the influence of positive and negative subjects to exhaustivity. Based on these theorems, the definitions of specificity and exhaustivity are suitable for ontology mining. A subject in ontologies may be highly exhaustive but not specific, in respect to a topic. Similarly, a subject may be highly specific but deal with only a limited semantic extent referred by a topic.

6.5 Ontology Learning and Mining Model

The ontology learning and mining model proposed here and in Chapter 5 learns a user’s concept model and develops the hypothetical computer model introduced in the concept-based Web information gathering framework and presented in Chap-
In respect to a user information need, the ontology learning and mining model learns a personalised ontology to represent a user’s concept model. It also specifies and scales the concepts in the ontology regarding the user information need. The ontology learning and mining model is formalised as:

**Definition 7.** The ontology learning and mining model $C$ is a 3-tuple $C := \langle WKB, LIR, F \rangle$, where

- $WKB$ is a world knowledge base that frames a user’s background knowledge;
- $LIR$ is a user’s local instance repository, in which the elements cite the knowledge in $WKB$;
- $F$ is a set of functions, inferences, algorithms, and theorems that learn and mine an ontology for a user using $WKB$ and $LIR$.

To represent a user’s concept model $U$, an ontology is constructed based on the $WKB$ and personalised using the user $LIR$, co-responding to a querying model $Q$ for a $g \in G$ describing the user information need. The ontology model represents the user’s concept model $U$. The concepts $K$ in $U$ are represented by $S$, in which the subjects in $S^+$ are relevant and $S^-$ are non-relevant to the $T$ representing an information need $g \in G$. The weight $w_k$ for a concept $k$ in $K$ is reproduced by $\text{sup}(s, T)$ for the subjects in $S$. The $\hat{B}$ in $U$ is constructed by $R$, $\text{tax}^S$ and $\text{rel}$ in $O(T)$. The user concept model $U$ is represented by the ontology $O(T)$.

Figure 6.4 presents the process of interesting concepts discovery in the ontology learning and mining model. In respect to a given topic, by using the semi-automatic ontology learning method introduced in Section 5.2.1 of Chapter 5, three sets of positive, negative, and neutral subjects are extracted from the world knowledge base. This is presented as Phase One in Sub-Figure (A). In this chapter, these subjects are first refined by the multidimensional ontology mining method using Specificity and Exhaustivity, and the is-a and part-of relations are investigated for the knowledge specification. Based on the content of user LIR, the noisy subjects in the positive set are filtered, and user overlooked positive
Figure 6.4: The phases of interesting concepts discovery, where (A) is for the ontology learning and mining model using the semi-automatic learning method, and (B) is for using the automatic learning method.
subjects are discovered from the negative and neutral sets and added into the positive set. The subjects are refined by Equations (6.10), (6.11), and (6.12) in Section 6.2. This process is described as Phase Two in Sub-Figure (A). Finally, in Section 6.3, more positive subjects are discovered from the negative and neutral sets, based on the investigation of related-to relationships held by the subjects. This is presented as Phase Three in Sub-Figure (A). The positive, negative, and neutral subjects in the ontology are refined and scaled, and the user background knowledge is specified.

Sub-Figure (B) in Figure 6.4 illustrates the process of knowledge discovery using the automatic ontology learning method introduced in Section 5.2.2 of Chapter 5. As displayed in Sub-Figures (A) and (B), the difference between automatic and semi-automatic learning methods is that initially the neutral subject set is empty in Phase One. The neutral subjects are acquired in Phase Two. The positive, negative, and neutral subjects are refined and scaled in Phase Two and Three for user background knowledge specification.

6.6 Summary and Conclusion

In this chapter, a multidimensional ontology mining method was introduced. The method aims to refine the personalised ontologies learned and discussed in Chapter 5. Two dimensions were introduced to investigate the concepts defined and specified in ontologies: specificity refers to the semantic focus of subjects on a particular topic, and exhaustivity refers to the semantic extent of subjects that deals with a topic. The subjects in ontologies may be of great exhaustivity but poor specificity, or of great specificity but poor exhaustivity, with respect to the given topic. The specificity of subjects consists of two parts: the semantic specificity specifying the focus of referring concepts, and the topic specificity specifying the focus of the given topic. An algorithm was presented in this chapter to evaluate the semantic specificity of subjects based on their locality in ontologies and the investigation of their associated semantic relations; such as is-a and part-of re-
lations. In addition, a method is also presented to measure the *topic specificity* by using user LIRs. The positive, negative, and neutral subjects extracted in the personalised ontology learning phase present in Chapter 5 are thus refined based on their *specificity* or *exhaustivity* values.

The ontology mining method introduced in this chapter also aims to discover more interesting and on-topic concepts from the ontologies. The features of positive subjects are extracted from their referring instances in user LIRs. Also referring to the same instances, the underlying interesting subjects from the original negative and neutral sets can be discovered and added into the positive set. Their support values to the given topic are evaluated, based on their referring positive features. The personalised ontologies are hence further refined for personalisation, with more interesting and on-topic subjects discovered.

In this chapter, a set of theorems was introduced to define the relationships between *specificity* and *exhaustivity*, and to restrict the utilisation of *specificity* and *exhaustivity* in ontology mining.

In addition, the ontology learning and mining model was formalised in this chapter according to the concept-based Web information gathering framework discussed in Chapter 3. The ontology learning and mining model validates the hypothesis introduced in Chapter 3 that aimed to solve the research problem of specifying user background knowledge in ontologies to capture user information needs for Web information gathering. The experimental evaluation of the ontology learning and mining model will be presented and discussed in Chapter 7 and 8.
Chapter 7

Evaluation Methodology

The preceding chapters introduced the ontology learning and mining model for acquiring user profiles for in Web information gathering, and the following chapter evaluates the introduced model. This chapter addresses the design issues of experiments for evaluating the proposed ontology learning and mining model through environment, data set, topics, and dataflow. It then describes the implementation of the ontology learning and mining model and the baseline models in experiments. This evaluation methodology bridges the gap between the preceding method chapters and the following results and discussion.

7.1 Experiment Hypotheses

The proposed ontology learning and mining model aims to acquire user profiles for personalised Web information gathering. Hence, the existing user profile acquiring models are the baseline models in the experiments for evaluating the proposed model. Based on the survey conducted and discussed in Chapter 2, two main hypotheses were established in the design of evaluation experiments. The validity of the proposed model can be proven if:
1. **the accuracy of user profiles acquired by the proposed model can be better than that of the user profiles acquired by the state-of-the-art computational models; and**

2. **the accuracy of user profiles acquired by the proposed model can approximate that of the user profiles acquired by the human-based user profile acquiring models.**

These experiment hypotheses drove the design of the experiments and are discussed in detail as follows.

The baseline models in the experiments were selected from the results of the survey of user profile acquisition techniques, as discussed previously in Chapter 2. The user profile acquiring models can be categorised into three groups: interviewing, non-interviewing, and pseudo-relevance feedback. In the evaluation experiments, the user profiles acquired by the proposed model demanded to compare with those acquired by the typical models, representing the interviewing, non-interviewing, and pseudo-relevance feedback mechanisms respectively.

The interviewing user profile acquiring models use human efforts. Users are provided with a set of questions to answer, or a set of documents to read and judge for relevance or non-relevance to the information needs. These models are human based and represent explicit human effort in user profile acquisition and information need capture. A typical model is that of user profiles acquired and used in the TREC-11 Filtering Track*. In this model, linguists who created the topics read a set of training documents and provided judgements of positive or negative to them against given topics [161]. Because the topics and user profiles are created and acquired by the same users manually, these user profiles perfectly reflect the user concept models for the topics, under an assumption:

**Assumption 3.** *Users know their information needs perfectly.*

Also under another assumption:

Assumption 4. *Human brains work better than computational models.*

these manually acquired and perfect user profiles can only be approximated, not outperformed. Therefore, the user profiles acquired manually in the TREC-11 Filtering Track were selected as the objective baseline in the evaluation experiments. If the accuracy of the user profiles acquired by the proposed model was close to that of the manual user profiles, the approximation of the former to the latter could be proven. This proven approximation to the human-based models could also prove the efficiency of the proposed model because the proposed ontology learning and mining model is a computational model, compared with the interviewing user profile acquiring models.

The non-interviewing user profile acquisition techniques do not require human efforts from users. Instead, they observe and discover the topic-relevant concepts from user activities and behaviours [202]. The typical models of these implicit techniques include the OBIWAN model proposed by Gauch et al. [55, 56, 202] and the ontology-based user profiles proposed by Sieg et al. [181, 182]. Similar to the proposed ontology learning and mining model, these state-of-the-art models use ontologies to represent user profiles. However, these existing ontology-based user profile acquiring models have different mechanisms from that proposed in the ontology learning and mining model. Thus, these models were selected as the baselines in the experiments. If the user profiles acquired by the proposed model had better accuracy than that of the user profiles acquired by these baseline models, the effectiveness of the proposed model could be proven, in terms of comparison with the non-interviewing user profile acquiring models.

Another experiment hypothesis intended to be tested by comparison with the OBIWAN model [55, 56, 202] and the ontology-based user profiles model [181, 182] is for the semantic relations specification. While using ontologies to represent user profiles, the ontologies in these compared models are constructed in a subsumption structure of *super-class* and *sub-class* relations. In contrast, the proposed ontology learning and mining model in this thesis emphasises the specific seman-
tic relations of *is-a*, *part-of*, and *related-to*, and also evaluates their impacts on the associated concepts. Thus, by comparing the proposed model with the OBI-WAN and the ontology-based user profiles models, the benefit from the specific and complete semantic relation specification to the information gathering systems can be evaluated. If the proposed model outperformed the OBIWAN and the ontology-based user profiles models, the validity of emphasising *is-a*, *part-of*, and *related-to* relations could thus be proven.

The pseudo-relevance feedback profiles are generated by semi-automatic techniques, different from the interviewing and non-interviewing mechanisms. The pseudo-relevance feedback techniques assume a certain number of top documents on an initially extracted list as the positive information feedback from a user. The topic relevant concepts are then discovered from these documents. One typical approach of these techniques is the preliminary model introduced in Chapter 4, which specifies user concept models manually and acquires user profiles using the concept models. The manually specified concept models, including relevant and non-relevant concepts to the topics, are supposed to be more accurate than the concepts discovered from the pseudo-relevant documents, also under Assumption 4 discussed previously. Hence, this preliminary study model was also selected as a baseline model in the experiments for the representative of the pseudo-relevance feedback user profile acquiring models. If the user profiles acquired by the proposed model had better accuracy than that of the user profiles acquired by this baseline model, the effectiveness of the proposed model could be proven, compared with the pseudo-relevance feedback user profile acquiring models.

In the evaluation experiments, the validity of previously discussed experiment hypotheses would be tested by comparing the user profiles acquired by the proposed model with the profiles acquired by these baseline models. If the hypotheses could be confirmed, the validity of the proposed ontology learning and mining model could also be proven by the evaluation experiments.
7.2 Experiment Framework

The user information needs in these evaluation experiments are described and represented by user profiles. Such representation was introduced by [11] and further improved by Li and Zhong [110]. The same representation was also used in the participating models in the Text REtrieval Conference Filtering Track series. In the experiments in this thesis, user profiles were represented by training sets, consisting of a subset of positive samples $D^+$ that contain the on-topic concepts, and a subset of negative samples $D^-$ that contain the concepts that may confuse the topic interpretation. Each sample was a document $d$ holding a support value $\text{support}(d)$ to the given topic. Based on this representation, the baseline models in the experiments were selected carefully.

In the fields of Web information gathering, a common batch-style experiment is developed for the comparison of different models. The experiment is to select a collection of documents (testing set) and a set of topics associated with relevance judgements, and then measure the performance of each experimental model [186]. Because this thesis work investigated the hypothesis of using ontologies to acquire user profiles and benefit Web information gathering, the experiment framework was designed following this common batch-style.

The comparison of user profiles acquired by the proposed and baseline models was conducted in the experiment framework illustrated in Figure 7.1. Four models were implemented for the evaluation, according to the experiment hypotheses:

- **Ontology Model** that represented the proposed ontology learning and mining model, in which two versions were implemented: the **Ontology-I model** according to the automated and the **Ontology-II model** according to the semi-automated ontology learning methods presented in Chapter 5. The ontology mining method introduced in Chapter 6 was used by both the Ontology-I and -II models;

- **Manual User Profile Acquiring Model** that represented the typical
Chapter 7. Evaluation Methodology

Figure 7.1: The Experiment Framework
human-based interviewing user profile acquiring models. It is shortened as the “Manual model” in the related discussions;

- **Semi-automatic User Profile Acquiring Model** that implemented the model developed in the preliminary study, and represented the pseudo-relevance feedback user profile acquisition techniques. It is shortened as the “Semi-auto model” in the related discussions;

- **Automatic User Profile Acquiring Model** that represented the typical non-interviewing user profile acquisition techniques, including the models developed by Gauch *et al.* [55, 56, 202] and by Sieg *et al.* [181, 182]. It is shortened as the “Auto model” in the related discussions.

The topics went into the user profile acquiring models, and different user profiles were acquired. The user profiles were used by a common system to gather information from the testing set. The performance of the information gathering system was then determined by the input of user profiles, given the incoming user profiles as the only difference to the system. Based on that, the accuracy of user profiles could be measured by measuring the performance achieved by the information gathering system using the profiles.

The details of the experiment design, including the experimental environment, the common Web information gathering system, and the implementation of experimental user profile acquiring models, are described as follows.

## 7.3 Experimental Environment

Because it is difficult to predict what background Web users may come from, an environment covering a large range of topics was demanded for the evaluation experiments in the field of Web information gathering. In this thesis, the evaluation experiments were performed using the environment set up by the TREC-11 Filtering Track in 2002.
7.3.1 TREC-11 Filtering Track

The Text REtrieval Conference (TREC) aims to support research within the information retrieval community. The TREC series are co-sponsored by the National Institute of Standards and Technology (NIST)† and the United States Department of Defense‡. The TREC provides the infrastructure for large-scale evaluation of text retrieval methodologies. Its main objectives include to encourage research in information retrieval based on large text collections, and to increase the availability of appropriate evaluation techniques for use by industry and academia. Since 1992, for each annual TREC, NIST provides a test set of documents and questions. These TREC test collections, topics, and evaluation software are available to the retrieval research community, so organisations can evaluate their own retrieval systems at any time [200,210].

The TREC-11 2002 Filtering Track aimed to evaluate the information gathering methods using user profiles for separating relevant and non-relevant documents in an incoming stream. In the TREC-11 2002 Filtering Track, user profiles were represented by training sets consisting of positive and negative documents. The TREC Filtering Track argued that the information gathering performance can be improved by using user profiles, and evaluated information gathering methods based only on the quality of the retrieved document set [161]. According to the experiment design discussed previously, the TREC-11 2002 Filtering Track provided a perfect experimental environment for the evaluation experiments in this thesis.

7.3.2 Experimental Data Set

The TREC-11 2002 Filtering Track used the Reuters Corpus Volume 1 (RCV1) corpus provided by Reuters for research purposes [155]. The RCV1 corpus is large data sets of XML (Extensible Markup Language) documents with great topic cov-

---

7.3. Experimental Environment

Figure 7.2: Topic Distribution in RCV1 Corpus [155]

ereage. Reuters is the largest international text and television news agency. Every day, Reuters produces about 11,000 stories in 23 languages by its journalists. Stories are both distributed in real time and made available via online databases and other archival products. Produced by Reuters, the RCV1 is an archive of 806,791 documents drawn from one of those online databases for research purposes. The RCV1 consists of all and only stories in English and was produced between August 20, 1996 and August 19, 1997. The documents are distributed in the form of 365 zip files, one per day, in approximately 3.7Gb [155]. The CRV1 corpus is split into two different sets, one for training and one for testing: the first six weeks’ items in RCV1, 20 August through 30 September 1996, are taken as the training set, and the remainder makes up the testing set. As a result, the training set contains 23,307 documents, and the testing set contains 783,484 documents. Distributed by Reuters to 520 groups, RCV1 is widely used in many areas to support substantial research advances [99, 210].

The RCV1 corpus has many advantages over other data sets that are popularly used in experiments in information retrieval and gathering: such as Reuters-
21578 collection [97] and OHSUMED [67, 231]. The number of documents contained in the RCV1 corpus is 35 times that of the Reuters-21578 and double that of the OHSUMED documents (at 348,566 documents) [99, 163]. While useful, OHSUMED does not contain the full texts of documents. Also, the RCV1 corpus covers a large range of topics. The RCV1 topics are categorised manually by the Reuters’ editors. Figure 7.2 illustrates the distribution of topics in the RCV1 corpus [155]. One may see that the “Corporate/Industrial” category has the largest volume of more than 350,000 stories (documents), whereas “War, Civil war”, “Crime, Law enforcement”, and “Capacity/Facilities” have the smallest volume of less than 50,000 stories in each. In contrast, OHSUMED is focused specifically on the medical domain only, and thus has a limited number of topics. These advantages make the RCV1 corpus the best choice for the evaluation experiments in this thesis.

The text documents in the RCV1 corpus have been processed by substantial verification and validation of the content; removed of spurious or duplicated documents; normalisation of dateline and byline formats; and addition of copyright statements. These documents have been formatted using a consistent XML schema of MewsML§, which is the extensive use of descriptive metadata, largely accepted by the Web intelligence community as one of the highly potential types of Web documents in the future [140].

A sample document in the RCV1 data set is illustrated in Figure 7.3. Each RCV1 document has a \(<\texttt{newsitem}>\) field for identification, where the \texttt{itemid} is a unique number for identification, the \texttt{date} is the time the story was produced, and the \texttt{xml:lang} indicates the language of the document. For the document displayed in Figure 7.3, its \texttt{itemid} is “128275”, it was produced in “1996-10-18” (which also means that it is a document in the testing set), and the document is in English. Each document has a title marked by the tag \(<\texttt{title}>\), a headline by the tag \(<\texttt{headline}>\), and a dateline by \(<\texttt{dateline}>\). The main content of the document

§http://www.newsml.org
ISRAEL Blast kills two pro Israel militiamen Lebanon

Blast kills two pro Israel militiamen Lebanon

JERUSALEM 1996 10 18

roadside charge blew killed two pro Israeli South Lebanon Army SLA
militiamen wounded two Jewish state s south Lebanon occupation zone Friday
Israeli security sources

incident occurred eastern sector Israel s self declared security zone

Israel carved 15 nine mile wide occupation zone withdrew bulk 1982 Lebanon
invasion force 1985

Pro Iranian Hizbollah guerrillas fighting oust Israel zone

April Israel launched air sea naval bombardment Lebanon retaliation
guerrilla rocket attacks northern Israel 200 Lebanese civilians killed
onslaught Israelis killed

fighting ended U S brokered ceasefire forbidding sides fire civilian areas

Figure 7.3: A Sample Document in RCV1 Corpus
is framed in a distinct <text> field and paragraphed in several <p> fields. In the experiments in this thesis, only the text in the title and main content is used as the document content. The information in the headlines usually duplicates that in the titles. The information in the datelines is about regions and times. This information is out of the research scope in this thesis, and is thus discarded in the text preprocessing of the data set. Also discarded is the paragraph structure in the RCV1 documents. This thesis focuses only on the semantics of document contents, not the structure. Hence, the <p> tags are discarded and the text in different paragraphs is treated the same in the experiments.

The text preprocessing of the RCV1 corpus includes stopword removal and word stemming. The RCV1 documents vary from a few hundred to several thousand words in length [99,163]. Figure 7.4 illustrates the word distributions in the RCV1 corpus. With the aim of reducing the dimensionality and complexity of the feature vectors representing the documents, the stopword removal and word stemming techniques were used in the text preprocessing of RCV1 corpus. During the stopword removal phase, the commonly occurring words, such as “to”, “or”, “and”, “of”, “the”, “a”, are removed from the documents. Word stemming is recommended by many researches in information retrieval and Web information gathering communities [76]. The terms with a common stem usually have the same semantic meanings, for example, “connect”, “connected”, “connecting”, “connection”, and “connections”. In the word stemming phase, these words are conflated into a single stem “connect” by removing the various suffixes of -ed, -ing, -ion, and -ions. The Porter stemmer algorithm [146] was used, as it is widely used by many text mining works. After the stopword removal and word stemming, 40% to 50% of the total number of words can be filtered out in text preprocessing [66].
Figure 7.4: The Word Distribution in RCV1 Corpus [156]

Figure 7.5: A TREC-11 Filtering Track Topic
7.3.3 Experimental Topics

In the experiments conducted in this thesis, the topics created by and used for TREC-11 2002 Filtering Track were chosen for the experiments. The TREC distinguishes between user information needs and search queries: statements of user information needs are called topics; data structures given by users to a retrieval system are called queries, which are generated to describe the topics (user information needs). The TREC test collections provide a wide range of topics for experiments: each one has a clear statement of what criteria makes a document relevant [210]. The topic statements consist of four sections: an identifier, a title, a description, and a narrative. Figure 7.5 displays a sample topic in TREC-11 Filtering Track, one of the topics used in the experiments in this thesis. In the experiments, the titles of topics were used as the querying models $Q$, based on the assumption that in real world users often have only small numbers of terms in their queries [72]. In these querying models, the longest queries have five terms, the shortest queries have two terms, with an average of three terms only. These titles of TREC topics are listed in Appendix A for reference.

Two distinct types of topics were created for the TREC-11 Filtering Track. The first set of 50 topics, covering a wide range, was created by the NIST assessors manually using the standard topic development protocol. Each NIST assessor came up with some candidate topics that were created based on his or her own interests. The assessor then searched the RCV1 corpus to estimate the approximate number of relevant documents that corresponded to each candidate topic. The final set of topics was then selected by the NIST TREC team from among these candidates, based on the estimated number of relevant documents in RCV1 and balancing the load across assessors. For the topics in this set, the assessors who created the topic statements were also the same people who performed the relevant assessments for these topics. The second set of 50 topics was created automatically based on the intersection of Reuters category cate-
Because of their natural bias, the second set of topics seem to be more appropriate to test methods for classification and categorisation, with the first set more appropriate to test methods for capturing user information needs and acquiring user profiles. Because the proposed ontology learning and mining model in this thesis aims to capture user information needs for Web information gathering, the first set of 50 manually created topics was used in the evaluation experiments.

The 50 experimental topics ensure the stability and validity of the evaluation results. In the experiments, it was assumed that each topic came from an individual user. Thus, the 50 topics were coming from 50 different users, and as a result, the experiments could cover a large range of topics. The Web information gathering system then learned a personalised ontology according to each topic, to specify the user’s background knowledge and capture the information need. This mechanism is more effective than using subjects for experiments because these topics are carefully created and selected by the NIST TREC team, and have associated RCV1 training and testing sets [161,210]. In this case, the experiments can be well controlled and the evaluation result is valid. Also, as reported and suggested by Buckley and Voorhees [16], 50 topics are substantial to make a benchmark for stable evaluations in information gathering experiments. Thus, the 50 topics used in the experiments ensure the high stability of evaluation results for the thesis.

### 7.4 Web Information Gathering System

An information gathering system (IGS) was implemented for common use by all experimental models. The IGS is an implementation of a model developed by [110] that uses user profiles for Web information gathering. The input support values associated with the documents in user profiles affect the IGS’s performance sensitively. The [110] model is chosen: not only is it verified better than the *Rocchio* and *Dempster-Shafer* models, but it is also extensible in using support
values of training documents for Web information gathering.

The IGS first uses the training set to evaluate weights for a set of selected terms $T$. After text pre-processing of stopword removal and word stemming, a positive document $d$ becomes a pattern that consists of a set of term frequency pairs $d = \{(t_1, f_1), (t_2, f_2), \ldots, (t_k, f_k)\}$, where $f_i$ is $t_i$’s term frequency in $d$. The semantic space referred by $d$ is represented by its normal form $\beta(d)$, which satisfies $\beta(d) = \{(t_1, w_1), (t_2, w_2), \ldots, (t_k, w_k)\}$, where $w_i$ ($i = 1, \ldots, k$) are the weight distribution of terms and $w_i = f_i / \sum_{j=1}^{k} f_j$.

A probability function on $T$ can be derived based on the normal forms of positive documents and their supports for all $t \in T$:

$$pr_\beta(t) = \sum_{d \in D^+, (t, w) \in \beta(d)} support(d) \times w. \quad (7.1)$$

The testing documents can be indexed by $weight(d)$, which is calculated using the probability function $pr_\beta$:

$$weight(d) = \sum_{t \in T} pr_\beta(t) \times \tau(t, d); \quad (7.2)$$

where $\tau(t, d) = 1$ if $t \in d$; otherwise $\tau(t, d) = 0$.

Attempting to clarify the semantic ambiguity from $D^-$, a set of negative documents $ND$ is selected firstly from $D^-$, which satisfies $ND = \{d' \in D^- | weight(d') \geq \min_{d \in D^+} weight(d)\}$. The supports or normal forms of positive documents $d$ are also updated in the following situations: (i) if $\exists d' \in ND$, and $d \subseteq d'$, the support is adjusted by $support(d) = \frac{1}{\mu} \times support(d)$, where $\mu = 8$ in our experiments; otherwise, (ii) if $d \cap d' \neq \emptyset$, instead of updating $support(d)$, its normal form $\beta(d)$ is adjusted for all $(t, w) \in \beta(d)$ and $t \in d'$ by $w = \frac{w}{\mu}$, and for the rest $(t, w) \in \beta(d)$ and $t \notin d'$ by:

$$w = w + w \times \frac{\mu - 1}{\mu} \times \frac{s_{offering}}{base}$$

where $s_{offering} = \sum_{(t, w) \in \beta(d), t \in d'} w$ and $base = \sum_{(t, w) \in \beta(d), t \notin d'} w$. The probabil-
ity function Equation (7.1) and then the weight function Equation (7.2) can be updated based on the changes of the supports and normal forms.

In summary, the input to the Web information gathering system is the user profiles consisting of a set of training documents $D = \{d \mid d \in D^+ \cup D^-\}$ in which each document is associated with a support value $support(d)$ to the given topic. The experimental user profile acquiring models, including the proposed Ontology models and the baseline models, would match this requirement in their user profiles.

### 7.5 Ontology Models

In this section, the implementations of the ontology learning and mining model proposed in this thesis are presented, including semi-automatic and automatic Ontology models.

In the experiments, each topic was treated as an individual user with an information need, a large number of subjects representing different interests are needed when attempting to evaluate the proposed model in an environment covering a wide range of topics. However, it is unrealistic to obtain a group of subject participants holding such a large range of topics in their personal interests. Thus, it was assumed in the experiments that each of the 50 topics came from an individual user with a personal information need, and the experiments attempted to learn the user’s personalised ontology in order to acquire the user profile. As illustrated in Figure 7.1 and required by the IGS, the input to the implemented Ontology models (including both semi-automatic and automatic models) was a topic, and the output was a training set consisting of positive documents ($D^+$) and negative documents ($D^-$). Each document was associated with a $support(d)$ value indicating its support rate to the topic. These training documents with $support(d)$ values were the user profile corresponding to the given topic that describes the associated user background knowledge and helps capture the information need.

Before introducing the implemented models, the world knowledge base and
user LIRs that were commonly used by both semi-automatic and automatic Ontology models are discussed.

7.5.1 World Knowledge Base

As previously discussed in Chapter 5, the global ontology, so called world knowledge base $\mathcal{WKB}$ in this thesis, was implemented and constructed based on the LCSH system. The LCSH authority records distributed by the Library of Congress were a single file of 130MB in MARC (MAchine-Readable Cataloging) 21 format, which is sequential raw data compiled in a machine-readable form. After data pre-process using regular expression techniques, the MARC 21 authority records were translated to human-readable text and organised in the SQL database in a size of about 750MB. Theoretically, the LCSH authority records consist of subjects for personal names, corporate names, meeting names, uniform titles, bibliographic titles, topical terms, and geographic names. In order to make the Ontology models run more efficiently, only the topical, corporate, and geographic subjects were kept in the world knowledge base, as they have covered most topics in daily life. Eventually, the constructed $\mathcal{WKB}$ contained 491,250 subjects covering a wide range of topics.

The semantic relations in the world knowledge base were transformed from the references specified in the LCSH. The Broader/Narrower, Used-for, and Related-to references (represented by “450 |w | a”, “450” and “550” in the MARC 21 authority records, respectively) cross referencing the subjects were also extracted to define the semantic relations of is-a, part-of, and related-to in the $\mathcal{WKB}$ respectively. The BT and NT references are for two subjects describing the same topic but in different levels of abstraction (or specificity) [113]. These references defined the is-a relations in the world knowledge base. The Used-for references are usually used in two situations: to help describe an action, for example, “a turner is used for cooking”; or to help describe an object, for example, “a wheel is used for a car”. It is assumed in this thesis that in these cases, they are the
7.5. Ontology Models

part-of relations. When object A is used for an action, A actually becomes a part of that action, like “using a turner in cooking”; when A is used for object B, A becomes a part of B, like “a wheel is a part of a car”. Hence, the Used-for references in the LCSH system defined the part-of relations in the world knowledge base. The RT references are for two subjects related in some manner other than by hierarchy. They defined the related-to relations in the world knowledge base. The subjects in the implemented world knowledge base are linked by these three types of semantic relations.

7.5.2 Local Instance Repository

In the implementation, a user’s local instance repository was collected through searching the subject catalogue of the Queensland University of Technology (QUT) Library by using the given topic, as previously discussed in Chapter 6. The QUT library catalogue stores a large volume of information, summarising over four hundred thousand information items. The catalogue was distributed by the QUT library as a 138MB text file containing information for 448,590 items, and used in the experiments as the corpus for user LIR extraction. All of this information can be accessed through QUT library’s Web site (http://www.library.qut.edu.au/) and is available to the public.

Before use in the experiments, the catalogue information was also pre-processed by using text processing techniques such as stopword removal, word stemming, and term grouping. Librarians and authors have assigned title, table of content, summary, and a list of subjects to each information item in the catalogue. In order to simplify the experiments, only the abstracted information (title, table of content, summary) was used to represent an instance in LIRs. Each information item cites a list of subjects defined in the LCSH system for the semantic content. Therefore, treating each information item in the catalogue as an instance, as discussed in Chapter 6, each instance cites a set of subjects in the constructed

\[\text{This figure is for the collection in QUT library prior to 2007.}\]
world knowledge base. On average, there are about 2.06 subjects cited by each instance. For each one of the 50 experimental topics and thus each one of the 50 corresponding users, the user’s LIR was extracted from this catalogue data set. As a result, there were about 1111 instances existing in one LIR on average.

7.5.3 Model I: Semi-automatic Ontology Model

This model is the implementation of the ontology learning and mining model using the semi-automatic ontology learning method discussed in Chapter 5.

A user’s personalised ontology was constructed for a given topic by user interaction, as described in Section 5.2.1 of Chapter 5. The user roles were played by the candidate of this thesis. Based on the description and narrative of each experimental topic (as shown in Figure 7.5), the user selected positive and negative subjects from the world knowledge base, through Ontology Learning Environment, and used the subjects to construct a personalised ontology. While in the experiments, on average each constructed ontology contained about 16 positive and 23 negative subjects. These subjects were connected by the is-a, part-of, and related-to semantic relations, as defined by the cross references in the LCSH system and structured in the world knowledge base.

For each topic, the ontology mining method was also performed on the constructed ontology and the user’s LIR to discover interesting concepts, as described in Chapter 6. This implementation model appreciated specificity more than exhaustivity, in terms of ontology mining. The semantic relations of is-a and part-of, were thus considered in the ontology mining phase for interesting concepts discovery. For the coefficient $\theta$ in Algorithm 2 discussed in Chapter 6, some preliminary tests had been conducted for various values (0.5, 0.7, 0.8 and 0.9), and as a result of that, $\theta = 0.9$, meaning that the rate for specificity decreased for each level up in tax$^S$ is 10%, giving the Ontology model the best performance. Thus, the coefficient $\theta$ was set as 0.9 for Algorithm 2 utilised in the experiments in this thesis. The $\alpha$, a parameter for adjusting the minimum interestingness level
for interesting subjects discovery (in Equation (6.17) introduced in Section 6.3 of Chapter 6, was set as $\alpha = 1.5$, which also gave the best performance in the preliminary tests.

7.5.4 Model II: Automatic Ontology Model

This model is the implementation of the proposed model using the automatic ontology learning method discussed in Chapter 5.

Once the world knowledge base and an LIR were ready, an initialised ontology was learned first, as described in Section 5.2.2 in Chapter 6. This implementation appreciated specificity rather than exhaustivity, because accurately capturing user information needs is the top priority.

The ontology mining phase taken by this automatic Ontology model was the same as that taken by the semi-automatic Ontology model, as described previously.

7.5.5 Weighting the Training Documents

For both the semi-automatic and automatic Ontology Models, a document $d$ in the user profiles was acquired from an instance $i$ in the user’s LIR. The document’s associated support value $support(d)$ was measured by:

$$support(d_i) = str(i, T) \times \sum_{s \in \eta(i)} spe(s, T)$$

where $s \in S$ is a subject in the user’s personalised ontology $O(T)$, $str(i, T)$ is defined by Equation (6.5) and $spe(s, T)$ by Equation (6.7). While conducting the experiments, various parameters for classifying $support(d)$ to positive or negative were investigated. However, because the constructed ontologies were personalised and focused on a wide range of topics, there was no universal parameter existing for all topics. Therefore, the parameter was set as $support(d) = 0$, following the nature of positive and negative defined and discussed in this thesis. Thus, in this
Ontology model, the documents with \( \text{support}(d) > 0 \) went to the positive set \( D^+ \), and the ones with a negative \( \text{support}(d) \leq 0 \) went to the negative set \( D^- \).

### 7.6 Baseline Models

#### 7.6.1 Manual User Profile Acquiring Model

The Manual User Profile Acquiring Model (Manual model) demonstrates the interviewing user profile acquisition mechanisms, in which the acquired user profiles reflect user concept models perfectly. As previously mentioned, the RCV1 data set used in TREC-11 Filtering Track aims to evaluate the methods of persistent user profiles for separating relevant and non-relevant documents in an incoming stream: the TREC linguists in NIST separated the RCV1 set into training sets and testing sets for the topics designed by the TREC linguists [161]. These training sets were used as the user profiles in the Manual model in the experiments, as they were manually acquired by the TREC linguists who created the topics, and thus best reflected users’ interests in these topics.

The concepts contained in the content of Manual training documents represent the user interests in the experimental topics perfectly. The 50 topics used in the experiments are the topics designed in TREC-11 Filtering Track. They are designed by linguists manually, and associated with positive and negative training documents from the RCV1 data set [161].

In the topic design phase, each TREC linguist came to NIST with a set of candidate topics based on his or her own interests. For each candidate topic, the TREC linguist estimated the approximate number of relevant documents by searching the RCV1 data set using the NIST’s search system, which was a statistic-based ranking information retrieval engine. The NIST TREC team selected the final set of topics from among these candidate topics based on the estimated number of relevant documents and balancing the load across the TREC linguists.
The training sets associated with the topics were acquired through two phases: the retrieval phase and fusion phase, aiming at providing more accurate relevance judgements for the training documents. In the retrieval phase, extensive searches using multiple retrieval and classification systems were conducted at NIST for each topic. This process included two to seven rounds. After each round, relevant information was used as feedback to improve the search queries used for the next round. The process continued until no more relevant documents were found or five rounds had passed (some topics had more than five rounds due to glitches in the feedback system) [161].

Based on the relevant documents found in the retrieval phase, the author of each topic was given five document sets to judge for the topic in the fusion phase. Each document set consisted of about 100 documents, chosen from the relevant documents found in the retrieval phase. The author read each one of them and marked the document as positive or negative for relevance or non-relevance to the topic. The combined set of judged and marked documents were used as the training data for that topic [210]. Thus, the Manual training sets perfectly reflected the users’ interests in the experimental topics, as the topics were created by the same author who performed the relevance assessments for that topic.

The Manual training documents associated with the topics were used as the user profiles in the Manual model in this thesis’ experiments. Against a given topic, each document in the training set is associated with “positive” or “negative” for relevance or non-relevance to the topic. If a document $d$ is marked “positive”, it is a positive document in the user profile and $support(d) = \frac{1}{|D^+|}$; otherwise, it is a negative document and $support(d) = 0$. These positive and negative documents then form a Manual user profile for the given topic.

### 7.6.2 Automatic User Profile Acquiring Model

This experimental model demonstrates the non-interviewing user profile acquisition techniques, in particular the Gauch et al. OBWAN model [55, 56, 202] and
the Sieg et al. ontological user profile model [181, 182]. In these models, a user’s interests and preferences are represented by a set of weighted positive subjects that are learned from the user’s browsing history. The subjects are constructed in an ontology that contains and specifies the semantic relations in the subsumption of $super$-$class$ and $sub$-$class$ manner. The user profiles are acquired based on these positive subjects.

In this experimental model, the sets of positive subjects corresponding to given topics are manually fed back by the user. The user feedback process and results were exactly the same, as the positive subjects were extracted in the Ontology-I model, through the Ontology Learning Environment and from the world knowledge base. Different from the Ontology-I model, there were no specific $is$-$a$, $part$-$of$, and $related$-$to$ semantic relations considered, and no ontology mining phase (the ones introduced in Chapter 6) performed in the Auto model. The positive subjects were equally weighted as one, because there was no evidence to show which positive subjects users preferred more than others.

The training sets in this Auto model were extracted through searching the subject catalogue of the QUT library, the same process in the Ontology models for user LIRs. However, in this model a document’s $support(d)$ value was determined by the number of positive subjects cited by $d$. Because the positive subjects were equally weighted, more positive subjects cited by $d$ would make the document semantically closer to the user interests, and thus strengthened its $support(d)$ value.

There was no negative training set generated by this model, as they were not required by the OBWAN model [55] and the Sieg et al.’s ontological user profile model [182].

### 7.6.3 Semi-automatic User Profile Acquiring Model

The Semi-automatic User Profile Acquiring Model (Semi-auto model) is an implementation of the preliminary study model, presented and discussed in Chapter 4.
In the Semi-auto model, user concept models were specified by users manually. The positive and negative subjects were first identified by users manually (the thesis candidate played the role of the users). The identified subjects can be found in Appendix B for details. The $MB(T|s)$ value was set one and $MD(T|s)$ zero for positive subjects, and $MB(T|s)$ was zero and $MD(T|s)$ one for negative subjects. Therefore, the $sup(s, T)$ of Equation (4.1) in Chapter 4) was also the boundary value, as one for all positive subjects and zero for all negative subjects. In accordance with the semantic analysis of user concept models, the user profiles were acquired from the Web using a Web search agent.

For each given topic $T$, its referring concept space $space(T)$ was specified, including the positive subjects $S^+$ and negative subjects $S^-$. The positive and negative subjects were extracted from the description and narrative provided in the topic, as shown in Figure 7.5 and described in Section 7.3.3. Also identified by users were support values $sup(s, T)$ of the positive and negative subjects, regarding the given topic. The positive and negative subjects with their specified support values constructed the user concept model describing the given topic.

The $S^+$ and $S^-$ subjects were then used to acquire the positive and negative documents for user profiles in Web information gathering. Each $s \in S^+$ produced a query for retrieving a set of positive candidate documents, and each $s \in S^-$ produced a query for negative candidates. The support value $support(d)$ of acquired documents was determined by the performance of the Web search agent, the document’s index position in the returned list, and also the support value of $s$ that produced the query to retrieve the document. This support value of training documents was calculated, the same as the $sup(d, T)$ by Equations (4.5) and (4.6) presented in Chapter 4. Finally, the training sets acquired for user profiles are refined based on the training documents’ support values, using the Equation (4.7) introduced in Chapter 4.

Google was chosen as the Web search agent in the experiments, the same as that in the preliminary study. The performance achieved by Google was de-
determined using a training topic (Topic 101 as displayed in Figure 7.5) and by manually measuring the precision of gathering results. The precision results measured at the different cutoff points were plotted in Figure 4.1 in Chapter 4. At the first portion of cutoffs (top 30 documents), Google achieved high precision performance. However, the performance dropped quickly when the number of retrieved documents increased. The precision performance of Google influenced the support value of training documents acquired by using Google, as discussed previously.

The implementation of the Semi-auto model can also be referred back to Chapter 4 for detailed descriptions.

7.7 Summary

This chapter addressed the design issues of evaluation experiments for the proposed ontology learning and mining model. The experiment hypotheses were first discussed, followed by the experiment framework and environment. The implementation details of the ontology learning and mining model and the baseline models were also addressed in this chapter for evaluation methodology. The proposed ontology learning and mining model was evaluated by comparing the user profiles acquired by the model to those acquired by the human-based and state-of-the-art computational models. The comparisons were performed based on the experiments using a common Web information gathering system and the standard data set and topics created by the TREC, which is a widely accepted platform in information gathering evaluations. The accuracy of user profiles were measured by measuring the performance achieved by the Web information gathering system using the profiles. The validity of proposed ontology learning and mining model could then be evaluated based on the comparisons of acquired user profiles. The evaluation methodology designed in this chapter was carried through to the evaluation experiments, and the related results are presented and discussed in Chapter 8.
Chapter 8

Results and Discussions

The experiments conducted in this thesis were designed to evaluate the proposed ontology learning and mining model by comparing the user profiles acquired by the proposed model to those acquired by the baseline models, as outlined in Chapter 7. The user profiles learned by the proposed model and the baseline models were used by the common system for Web information gathering. The experiments used a standard data set and topics. The performance achieved by the Web information gathering system evaluated the accuracy of user profiles, and thus the effectiveness of the models that acquired the user profiles.

The experiment hypotheses are that the implementation of the proposed model (namely the Ontology-I and Ontology-II models) can (i) achieve the same performance as (or close to) that of the Manual model, and (ii) outperform the Semi-auto and Auto models, as discussed in Chapter 7. In this chapter, the experimental results and their related discussion are presented, against the experiment hypotheses discussed in Chapter 7, for the evaluation of the ontology learning and mining model proposed in Chapters 5 and 6.
8.1 Performance Measures

The performance of the experimental models is measured by three methods: the precision averages at eleven standard recall levels (11SPR), the mean average precision (MAP), and the $F_\beta$ Measure. These are all based on precision and recall, the standard, modern method of information gathering evaluations [6,16].

8.1.1 Precision and Recall

*Precision* and *Recall* are two standard quantitative measures of the performance achieved by information retrieval models [213]. Precision indicates the capacity of a system to retrieve only the relevant information items, whereas recall indicates the capacity of a system to retrieve all the relevant information items. They are calculated by [81, 83, 187]:

\[
\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}} \tag{8.1}
\]

\[
\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents in the collection}} \tag{8.2}
\]

An ideal information gathering model is to deliver information with the highest rates of both precision and recall to users. However, in reality, information gathering models may not be able to retrieve all the relevant items from a collection, especially when the collection is large. Thus, the recall ratio is one of the principal factors measuring the performance of a system: it denotes the rate of relevant information items gathered in a given situation. The other principal factor, precision, indicates an information gathering model’s ability to avoid retrieving irrelevant information items. This factor denotes the rate of unwanted items being withheld in a given situation. Often, when the precision performance of a system is improved, the recall rate is degraded; when the recall performance is improved, the precision rate is degraded [81, 183]. It is difficult for a system to achieve the ideal performance with both highest precision and highest recall.
8.1.2 Effectiveness Measuring Methods

Precision and recall are set-based measures and suitable for evaluating the quality of an unordered set of gathered documents. Attempting to facilitate computing average performance and evaluate information gathering models over a set of $N$ topics, the precision values at each individual topic can be interpolated to a set of standard recall level (0 to 1 in increments of 0.1). The mechanism of interpolating precision at standard recall level $\lambda \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$ is to use the maximum precision obtained for the topic for any actual recall level greater or equal to $\lambda$. The interpolated precision values are then plotted to a curve to show the performance achieved by the information gathering model [161, 204]. This measure is so-called Precision at 11 standard recall levels and shortened as 11SPR in the TREC.

For a set of experimental topics, an 11SPR value is computed by summing the interpolated precisions at the specified recall cutoff and then dividing by the number of topics:

$$\frac{\sum_{\lambda=1}^{N} \text{precision}_\lambda}{N}.$$  \hfill (8.3)

The 11SPR measure is proved suitable for information gathering and has become one of the most common methods in information gathering evaluations [161,204].

The mean average precision (MAP) over all relevant documents is a stable measure and a discriminating choice in information gathering evaluations. The average precision value is a single value measure that reflects the experimental model’s performance over all relevant documents. For each topic, rather than being an average of the precision at standard recall levels, the MAP measure is the mean of the precision values obtained after each relevant document is retrieved. The MAP value for a set of experimental topics is then the mean of the average precision values of each of the individual topics in the experiments. Different from the 11SPR measure, the MAP reflects the performance in a non-interpolated recall-precision curve [204]. As reported by Buckley and Voorhees [16], the MAP measure is a stable information gathering measuring method, recommended for
general-purpose information gathering evaluations.

The $F_\beta$ measure, also widely used in information retrieval and Web information gathering [98,99,207], is calculated by:

$$F_\beta = \frac{(\beta^2 + 1) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$$  \hspace{1cm} (8.4)

where $\beta$ is a parameter balancing precision and recall, depending on the precision or recall preferred by the system. When the value of $\beta = 0.5$, precision is weighted by the system to be twice as much as recall. When $\beta = 2$, recall is weighted as twice as much as precision. When $\beta = 1$, recall and precision are evenly weighted, and the $F_\beta$ measure corresponds to the harmonic mean and becomes the commonly used $F_1$ measure [99]:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Because precision and recall are equally important in Web information gathering in this thesis, the $F_1$ measure was used in the experiments for effectiveness measuring. Furthermore, the macroaverage and microaverage $F_1$ measures were used for detailed investigation on the effectiveness across the experimental topics, where the macro-$F_1$ measures the unweighed mean of effectiveness and micro-$F_1$ measures the effectiveness computed from the sum of results. The macro-$F_1$ measure averages the precision and recall and then calculates the $F_1$ measure for each experimental topic. The micro-$F_1$ measure calculates the $F_1$ measure for each returned result and then averages the $F_1$ values. The greater $F_1$ values indicate better effectiveness.

### 8.1.3 Statistical Significance Tests

In scientific research, statistical significance tests play an important role to evaluate the reliability of experiment results. They allow researchers to detect significant improvements. The proposed computational models need to prove that they
8.1. Performance Measures

truly achieve the designated goals rather than by chance only [14, 186]. In this thesis, two statistical significance tests were used for evaluations, the *Percentage change in performance* and *Student’s Paired T-Test*, where the former was to measure the difference in the mean of measuring metric, and the latter was to compute the probability that the result values occurred by chance.

The *percentage change in performance* is a traditional statistical method used to compute the difference between two sets of results. It is also a method commonly used in information gathering and knowledge management for evaluations [32, 120, 220]. The percentage change in performance is calculated by:

\[
\%\text{Change} = \frac{V_{\text{Ontology}} - V_{\text{Competitor}}}{V_{\text{Competitor}}} \times 100\%; \quad (8.5)
\]

and the average \(\%\text{Change}\) is calculated by:

\[
\text{avg } \%\text{Change} = \frac{\sum_{i=1}^{N} \%\text{Change}_i}{N}; \quad (8.6)
\]

where \(N\) is the number of experimental topics, and \(V\) is the result achieved by an experimental model for topic \(i\). The larger \(\%\text{Change}\) value indicates a more significant improvement achieved by the proposed model.

The *Student’s Paired T-Test* is also a common statistical method used to compare two sets of results for significance [14, 231]. A typical *null hypothesis* in the Student’s Paired T-Test is that no practical difference exists in two compared models. When two tests produce highly different significance levels (substantially low *p-value* value, usually set as \(<0.05\)), the *null hypothesis* can be rejected, and the significant improvement achieved by one model over the other can be proven. In contrast, when two tests produce almost equivalent significance levels (high *p-value*, usually set as \(>0.1\)), there is little or no practical difference between two compared models. Although the Student’s Paired T-Test has an assumption of using the normal distribution in its *null hypothesis*, it is argued by Smucker *et al.* [186] that the Student’s Paired T-Test largely agrees with the bootstrap
and randomisation tests in terms of information retrieval evaluations, as they are likely to draw the same conclusions regarding the statistical significance of their results. Thus, in this thesis, the Student’s Paired T-Test was used with the percentage change in performance method for evaluations.

8.2 Experimental Results

8.2.1 11SPR Results

The experimental 11SPR results are plotted in Figure 8.1, where the higher values indicate better performance. The 11SPR curves demonstrate that the Ontology-I, Ontology-II, and Manual models have almost the same achievement in their performances. At the recall level 0.0, 0.1, and 0.2, the Ontology-II model has the same performance as the Manual model, and the Ontology-I model has slightly
lower performance compared to the Ontology-II and Manual models. At recall level 0.3, both Ontology-I and II have slightly lower performance than the Manual model. At recall level 0.4, the Ontology-I and II have the same performance as that of the Manual. After recall level 0.4, the Ontology-II model has almost the same performance as the Manual model in all remaining recall levels, and the Ontology-I model outperforms both the Ontology-II and Manual models.

In terms of the performance achieved by the Semi-auto and Auto models, the Semi-auto model outperforms the Auto model, but does not perform as well as the Ontology-I, Ontology-II, and Manual models. The Auto model has the poorest performance in all five experimental models, as it is only at recall level 0.9 that the Auto model achieves the same performance as that by the Ontology-II, Manual, and Semi-auto models.

For overall 11SPR performance achieved by the five experimental models, the Ontology-I is the best model, followed by the Manual model, and then the Ontology-II model. The Semi-auto model outperforms only the Auto model, which itself has the poorest performance in all five experimental models. Thus, in terms of the 11SPR performance, the experimental hypotheses is evaluated, that the Ontology models (i) can achieve the same performance as (or close performance to) that of the Manual model, and (ii) can outperform the Semi-auto and Auto models. The ontology learning and mining model proposed in this thesis is promising.

8.2.2 MAP Results

The detailed MAP results are presented in Table 8.1, and the MAP results of five experimental models are plotted in Figure 8.2 for comparison. Based on the average MAP results, the Manual model has the best performance, followed by the Ontology-I model, and then the Ontology-II and Semi-auto models. The Auto
### Table 8.1: The Mean Average Precision Experimental Results

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Topic Detection</th>
<th>Auto</th>
<th>Semi-auto</th>
<th>Man</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.832624562</td>
<td>0.682659709</td>
<td>0.718420359</td>
<td>0.722781523</td>
<td>0.681552234</td>
</tr>
<tr>
<td>2</td>
<td>0.674207979</td>
<td>0.674207979</td>
<td>0.674207979</td>
<td>0.674207979</td>
<td>0.674207979</td>
</tr>
<tr>
<td>3</td>
<td>0.726203521</td>
<td>0.682659709</td>
<td>0.718420359</td>
<td>0.722781523</td>
<td>0.681552234</td>
</tr>
<tr>
<td>4</td>
<td>0.686559975</td>
<td>0.686559975</td>
<td>0.686559975</td>
<td>0.686559975</td>
<td>0.686559975</td>
</tr>
<tr>
<td>5</td>
<td>0.748671982</td>
<td>0.682659709</td>
<td>0.718420359</td>
<td>0.722781523</td>
<td>0.681552234</td>
</tr>
<tr>
<td>6</td>
<td>0.612204195</td>
<td>0.674207979</td>
<td>0.674207979</td>
<td>0.674207979</td>
<td>0.674207979</td>
</tr>
</tbody>
</table>

8. Results and Discussions

On the basis of the experimental results, it can be observed that the semi-auto-topic detection approach yields the highest mean average precision across all the experiments. This suggests that semi-auto-topic detection is more effective in identifying topics in scientific documents compared to the other approaches. Further analysis is required to understand the specific reasons behind this observation.

---

**Note:** The table values are placeholders and should be replaced with actual experimental results. The structure and format of the table are designed to facilitate the presentation of comprehensive experimental data in a clear and organized manner.
8.2. Experimental Results

Figure 8.2: The MAP and $F_1$ Measure Experimental Results

<table>
<thead>
<tr>
<th></th>
<th>Ontology-I</th>
<th>Ontology-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>7.66%</td>
<td>9.25%</td>
</tr>
<tr>
<td>Macro-FM</td>
<td>7.00%</td>
<td>8.57%</td>
</tr>
<tr>
<td>Micro-FM</td>
<td>6.69%</td>
<td>8.28%</td>
</tr>
</tbody>
</table>

Table 8.2: The Average Percentage Change Results

<table>
<thead>
<tr>
<th></th>
<th>Ontology-I</th>
<th>Ontology-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.8823</td>
<td>0.0261</td>
</tr>
<tr>
<td>Macro-FM</td>
<td>0.5512</td>
<td>0.0060</td>
</tr>
<tr>
<td>Micro-FM</td>
<td>0.5195</td>
<td>0.0053</td>
</tr>
</tbody>
</table>

Table 8.3: The Student’s Paired T-Test Results
model has the lowest MAP performance achieved in the five models. However, as shown in Figure 8.2, the differences between the experimental models are not significant. Thus, the statistical tests were performed for significance and reliability on the detailed results presented in Table 8.1. The percentage change results can be found in Table 8.2 and the Student’s Paired T-Test results can be found in Table 8.3.

According to the average percentage change results, compared with the Auto model, the Ontology-I model has achieved 20.42% improvement and the Ontology-II has achieved 16.46%. These improvements are significant. The significance is also confirmed by the Student’s T-Test results presented in Table 8.3, in which the $p$-value produced by the Ontology-I and Auto comparison is only 0.0002, and for the Ontology-II and Auto comparison is only 0.0064. As discussed in Section 8.1.3, when two models produce substantially low $p$-value ($<0.05$), the null hypothesis (that no difference exists in two comparing models) can be rejected, and the significant improvement achieved by one model over the other can be proven. The $p$-values produced by the Ontology-I vs. Auto and the Ontology-II vs. Auto comparisons are far less than the boundary value of 0.05, and therefore the improvements achieved by the Ontology models over the Auto model can be proven significant.

Compared with the Semi-auto model, the Ontology-I model has also achieved 9.25% improvement and the Ontology-II has only achieved 3.87%, as shown in Table 8.2. Although these figures, especially the percentage change achieved by the Ontology-II model, are not obviously significant, the Student’s Paired T-Test results report that the improvements are significant. According to the T-Test results presented in Table 8.3, the $p$-value produced by the Ontology-I and Semi-auto comparison is only 0.0261, and for the Ontology-II and Semi-auto comparison is only 0.0209. Apparently, the $p$-values produced by the Ontology-I and Semi-auto comparison and the Ontology-II and Semi-auto comparison are much smaller than 0.05, and the significant improvement achieved by the Ontology
models over the Semi-auto model can still be proven.

Finally, compared with the Manual model, the Ontology models have also achieved some improvements. As shown in Table 8.2, the Ontology-I model has made a 7.66% improvement from the Manual model, and the Ontology-II model 3.31%. However, according to the T-Test results produced by the Ontology-I vs. Manual comparison ($p$-value=0.8823) and Ontology-II vs. Manual comparison ($p$-value=0.4842), the $p$-values are substantially higher and much greater than the boundary value 0.05. The T-Test results indicate that there is no (or little) practical difference existing between the Ontology models and the Manual model. Therefore, both the percentage change and T-Test results confirm that the Ontology models can achieve the same performance as (or close to) the Manual model.

Based on the MAP results, the experiment hypotheses are evaluated, that the Ontology models can (i) achieve the same performance as (or close performance to) that of the Manual model, and (ii) outperform the Semi-auto and Auto models. The ontology learning and mining model proposed in this thesis is promising.

### 8.2.3 $F_1$ Measure Results

The illustration comparison of the $F_1$ Measure performance achieved by five experimental models is plotted in Figure 8.2; the detailed macro-$F_1$ Measure results are presented in Table 8.4, and the micro-$F_1$ Measure results are in Table 8.5. Both the macro-$F_1$ and micro-$F_1$ Measure results have the same report: that the Ontology-I model has the best performance, followed by the Manual model, the Ontology-II model, and then the Semi-auto model, and finally the Auto model. These are as similar as those reported by the 11SPR and MAP results. The statistic tests were also performed on the $F_1$ Measure results presented in Tables 8.4
Chapter 8. Results and Discussions
154

Topic
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125

Manual
0.733317629
0.728481648
0.359996605
0.644068861
0.554752893
0.232358154
0.22967502
0.17940681
0.450757768
0.217572106
0.108189944
0.193969771
0.315167435
0.412803766
0.506325933
0.632009787
0.361150987
0.11139913
0.409717102
0.672906687
0.471247673
0.449294739
0.184104698
0.236125231
0.465337871

Auto
0.653577842
0.506427367
0.187343629
0.588022425
0.408933859
0.26305838
0.227839005
0.160283774
0.625929559
0.278660702
0.093826064
0.095277841
0.338040299
0.423929733
0.373307068
0.557115745
0.313291943
0.213732636
0.307610239
0.529265611
0.361671718
0.461626286
0.161154624
0.411969866
0.443575553

Semi-auto
Ontology-I
0.652367463
0.597521727
0.528653811
0.563335088
0.346627676
0.388045164
0.646578031
0.628422292
0.569622322
0.583989948
0.256256226
0.280845594
0.206951478
0.227827053
0.149581832
0.159029142
0.653081576
0.661814205
0.156336062
0.280089208
0.10066714
0.131831982
0.194685346
0.201461082
0.212757975
0.352576131
0.427032703
0.43713724
0.5516033
0.536620539
0.511799167
0.57550365
0.37406275
0.334487223
0.176867792
0.220594931
0.249207517
0.290103619
0.655657483
0.665584461
0.464731064
0.402553677
0.433783502
0.472813328
0.171502427
0.169482079
0.386266808
0.356506512
0.474357462
0.453823076
Average (101 - 150)

Ontology-II
0.645111707
0.531855944
0.360205946
0.64651699
0.57338162
0.247903193
0.201774372
0.149578854
0.655536361
0.196181484
0.091209155
0.212743913
0.253934357
0.429368786
0.54859822
0.600170763
0.377487176
0.197088938
0.299585923
0.673432876
0.481574382
0.413670018
0.179384099
0.37043716
0.483177963

Topic
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150

Manual
0.772321103
0.48297926
0.330627346
0.336828277
0.169328675
0.614954074
0.117420303
0.265731137
0.453809669
0.627277259
0.306692732
0.137654166
0.405605098
0.247177182
0.41706272
0.51731016
0.314422579
0.130592815
0.462479773
0.143834809
0.613144111
0.330273138
0.769955125
0.213142289
0.334403532
0.387503312

Auto
0.641486961
0.517806477
0.354183622
0.305202586
0.149621066
0.548589626
0.140748687
0.176438539
0.397749307
0.507395242
0.426828479
0.133800901
0.318908431
0.265296977
0.50481279
0.384779598
0.204033286
0.086403152
0.351678578
0.091624632
0.707798642
0.28664968
0.75571895
0.14925503
0.375440258
0.355354465

Table 8.4: The Macro F1 Measure Experimental Results

Semi-auto
0.689031977
0.504997587
0.308943383
0.357833496
0.203709873
0.627870499
0.170896141
0.244726432
0.335639906
0.524142503
0.309007949
0.134212691
0.292890965
0.286030334
0.404558628
0.442954215
0.325737224
0.097345871
0.421748096
0.083761975
0.659128133
0.327087831
0.747859021
0.351979318
0.391587705
0.375894413

Ontology-I
0.65275019
0.486732809
0.345884907
0.35401697
0.165930011
0.601032085
0.169657465
0.263202308
0.420673867
0.497307986
0.402663032
0.135431714
0.37631675
0.292211199
0.496412931
0.39874777
0.362154695
0.138021803
0.423254861
0.087147337
0.720554146
0.371949764
0.729115778
0.424134722
0.415351319
0.394053107

Ontology-II
0.641507864
0.519187168
0.313387217
0.335527103
0.216561843
0.638825352
0.167387809
0.264047277
0.349681866
0.545634107
0.328913819
0.13152871
0.369071038
0.280504673
0.413940433
0.478922511
0.330706832
0.113242969
0.46250416
0.092165055
0.671789023
0.345291102
0.7558595
0.389898934
0.315873275
0.385837397


## 8.2. Experimental Results

<table>
<thead>
<tr>
<th>Topic</th>
<th>Manual</th>
<th>Auto</th>
<th>Semi-auto</th>
<th>Ontology-I</th>
<th>Ontology-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>0.66625927</td>
<td>0.59341781</td>
<td>0.592045083</td>
<td>0.541501979</td>
<td>0.580368611</td>
</tr>
<tr>
<td>102</td>
<td>0.67119781</td>
<td>0.46954164</td>
<td>0.512324227</td>
<td>0.49466787</td>
<td>0.581396536</td>
</tr>
<tr>
<td>103</td>
<td>0.324215622</td>
<td>0.174609414</td>
<td>0.315111494</td>
<td>0.34583373</td>
<td>0.32244787</td>
</tr>
<tr>
<td>104</td>
<td>0.58511518</td>
<td>0.54350904</td>
<td>0.593824651</td>
<td>0.581765543</td>
<td>0.591773603</td>
</tr>
<tr>
<td>105</td>
<td>0.509154105</td>
<td>0.375278787</td>
<td>0.520709026</td>
<td>0.53364592</td>
<td>0.52383283</td>
</tr>
<tr>
<td>106</td>
<td>0.22229861</td>
<td>0.2449396632</td>
<td>0.239411588</td>
<td>0.259619948</td>
<td>0.230748346</td>
</tr>
<tr>
<td>107</td>
<td>0.20613899</td>
<td>0.209349393</td>
<td>0.18858697</td>
<td>0.209618787</td>
<td>0.186732544</td>
</tr>
<tr>
<td>108</td>
<td>0.167580289</td>
<td>0.15071789</td>
<td>0.140968741</td>
<td>0.140938018</td>
<td>0.14072162</td>
</tr>
<tr>
<td>109</td>
<td>0.42049801</td>
<td>0.579087388</td>
<td>0.308603333</td>
<td>0.601138344</td>
<td>0.412596815</td>
</tr>
<tr>
<td>110</td>
<td>0.201896468</td>
<td>0.256709533</td>
<td>0.145908435</td>
<td>0.256816821</td>
<td>0.18151512</td>
</tr>
<tr>
<td>111</td>
<td>0.10169594</td>
<td>0.091248715</td>
<td>0.095922901</td>
<td>0.12644561</td>
<td>0.087459869</td>
</tr>
<tr>
<td>112</td>
<td>0.179967024</td>
<td>0.095580209</td>
<td>0.178725302</td>
<td>0.184001196</td>
<td>0.19515309</td>
</tr>
<tr>
<td>113</td>
<td>0.286666854</td>
<td>0.314115105</td>
<td>0.19459678</td>
<td>0.325711855</td>
<td>0.23426848</td>
</tr>
<tr>
<td>114</td>
<td>0.373157876</td>
<td>0.391276891</td>
<td>0.391524691</td>
<td>0.399262504</td>
<td>0.395181187</td>
</tr>
<tr>
<td>115</td>
<td>0.45224763</td>
<td>0.337003033</td>
<td>0.49324563</td>
<td>0.489124677</td>
<td>0.49297612</td>
</tr>
<tr>
<td>116</td>
<td>0.37787956</td>
<td>0.512915411</td>
<td>0.466356909</td>
<td>0.526533062</td>
<td>0.546993237</td>
</tr>
<tr>
<td>117</td>
<td>0.330693877</td>
<td>0.292643429</td>
<td>0.344040704</td>
<td>0.309367195</td>
<td>0.346581236</td>
</tr>
<tr>
<td>118</td>
<td>0.10771323</td>
<td>0.202542247</td>
<td>0.167907097</td>
<td>0.208347464</td>
<td>0.186600943</td>
</tr>
<tr>
<td>119</td>
<td>0.380343555</td>
<td>0.28971278</td>
<td>0.235509061</td>
<td>0.27262897</td>
<td>0.28103036</td>
</tr>
<tr>
<td>120</td>
<td>0.614788402</td>
<td>0.481893874</td>
<td>0.590079085</td>
<td>0.601877141</td>
<td>0.608801309</td>
</tr>
<tr>
<td>121</td>
<td>0.410362759</td>
<td>0.3296963</td>
<td>0.41208737</td>
<td>0.36081822</td>
<td>0.428551619</td>
</tr>
<tr>
<td>122</td>
<td>0.401738781</td>
<td>0.418542187</td>
<td>0.397304733</td>
<td>0.427027532</td>
<td>0.381088531</td>
</tr>
<tr>
<td>123</td>
<td>0.17216309</td>
<td>0.153442684</td>
<td>0.161197265</td>
<td>0.161778265</td>
<td>0.168622812</td>
</tr>
<tr>
<td>124</td>
<td>0.228631165</td>
<td>0.379909313</td>
<td>0.355636206</td>
<td>0.336281589</td>
<td>0.34346656</td>
</tr>
<tr>
<td>125</td>
<td>0.422953271</td>
<td>0.401201411</td>
<td>0.419694306</td>
<td>0.404229091</td>
<td>0.428584241</td>
</tr>
</tbody>
</table>

| Average (101 - 150) | 0.355942711 | 0.328750061 | 0.345822145 | 0.362188892 | 0.354828583 |

### Table 8.5: The Micro F<sub>1</sub> Measure Experimental Results
and 8.5, in order to evaluate the significance and reliability of the experimental results. The percentage change results are presented in Table 8.2 and the Student’s Paired T-Test results in Table 8.3, together with the MAP statistic test results.

According to the average percentage change results, compared with the Auto model, the Ontology-I model has 18.41% performance improvement in macro-$F_1$ and 16.93% improvement in micro-$F_1$ Measure results, and the Ontology-II model has 14.65% improvement in macro-$F_1$ and 13.35% improvement in micro-$F_1$ results. These significant improvements are confirmed by the Student’s T-Test results presented in Table 8.3. The $p$-values produced by the Ontology-I and Auto comparison are only 0.0001 in both the macro-$F_1$ and micro-$F_1$ Measure results, and by the Ontology-II and Auto comparison are only 0.0059 in macro-$F_1$ and 0.0071 in micro-$F_1$ Measure results. These $p$-values are much smaller than the boundary value of 0.05. The null hypothesis of no difference existing between two models is rejected, and the significant improvement achieved by the Ontology models over the Auto model is proven. Thus, based on the statistic test results, the Ontology models have significantly outperformed the Auto model in $F_1$ Measure results.

Compared with the Semi-auto model, the Ontology-I model has also 8.57% performance improvement in macro-$F_1$ and 8.28% improvement in micro-$F_1$ results, and the Ontology-II model has only 3.74% improvement in macro-$F_1$ and 3.61% improvement in micro-$F_1$ results. Similar as the indication from the MAP results, though these figures are not obviously significant, the Student’s Paired T-Test results argue that these improvements are significant. As presented in Table 8.3, the $p$-values produced by the Ontology-I and Semi-auto comparison are only 0.0060 in macro-$F_1$ and 0.0053 in micro-$F_1$ Measure results. In terms of the Ontology-II and Semi-auto comparison, the $p$-values produced are also only 0.0092 in macro-$F_1$ and 0.0082 in micro-$F_1$ results. These $p$-values are much smaller than 0.05, the significance boundary. Hence, the significant improvement
achieved by the Ontology models from the Semi-auto model is proven by the Student’s Paired T-Test. The Ontology models are confirmed better than the Semi-auto model significantly, in terms of the $F_1$ Measure results.

In terms of the comparison with the Manual model, the Ontology models have also made some improvements. As shown in Table 8.2, the Ontology-I model has improved from the Manual model 7.00% in macro-$F_1$ and 6.69% in micro-$F_1$ Measure performance. The Ontology-II model has improved by 2.69% in macro-$F_1$ and 2.55% in micro-$F_1$ Measure performance. Though the percentage change results report such improvements, the Student’s Paired T-Test results argue that there is no practical difference between the Ontology models and the Manual model. The Ontology-I and Manual comparison produces the $p$-value of 0.5512 in terms of macro-$F_1$ and 0.5159 in terms of micro-$F_1$ Measure results. The Ontology-II and Manual comparison produces the $p$-values as 0.8620 in macro-$F_1$ and 0.8958 in micro-$F_1$ Measure results. These $p$-values are substantially high and much greater than the significance boundary value of 0.05. Therefore, the Student’s T-Test argues that no practical difference exists between the Ontology models and the Manual model and the null hypothesis stands. Hence, the statistic tests, including the percentage change and T-Test, confirm that the Ontology models have the same performance as the Manual model in $F_1$ Measure experimental results.

Based on the $F_1$ Measure results, the experiment hypotheses are evaluated, that the Ontology models (i) can achieve the same performance as (or close to) the perfect Manual model, and (ii) can outperform the Semi-auto and Auto models. The ontology learning and mining model proposed in this thesis is encouraging.
Figure 8.3: Percentage Change in Topics (Ontology-I vs. Manual)

Figure 8.4: Percentage Change in Topics (Ontology-II vs. Manual)
8.3 Discussion

8.3.1 Ontology Models vs. Manual Model

The experiments performed on the Ontology models and the Manual model aim to evaluate the proposed computational model. This evaluation is conducted by comparing the user profiles acquired by the Ontology-I and Ontology-II models to those acquired by the Manual model, in which the concepts are specified and proven by users manually. According to the experimental results presented in Section 8.2, the Ontology-I and Ontology-II models have achieved the same performance as that of the Manual model in the experiments.

The experimental results indicate that the MAP, macro-FM, and micro-FM experimental results largely agree with each other. Table 8.6 presents the comparisons between the Ontology models and others, based on the number of topics that the Ontology models won, lost, and tied in the experiments. For each pair of comparisons, whether the Ontology model is better than, worse than, or equal to the Manual model is compared with a predefined fuzziness value. The fuzziness value, introduced by Buckley and Voorhees [16], is a value that if the percentage change made by two scores is smaller than the fuzziness value, the two scores are deemed equivalent. In this discussion, the fuzziness value is set as 5%, the same as set in [16] for information retrieval experiments. Thus, any percentage change values within 5% of one another are deemed as equal. For the 50 experimental topics, the results give the number of topics that the Ontology model won, lost, or tied in the comparisons with other models. For an example in Table 8.6, the Ontology-I model is better than the Manual model in 22 topics, worse in 17 topics, and equal to it in 11 topics in terms of MAP performance. Based on Table 8.6,
Chapter 8. Results and Discussions

Average number of documents in user profiles acquired by

<table>
<thead>
<tr>
<th></th>
<th>Ontology-I</th>
<th>Manual</th>
<th>Proportional difference (Ontology-I/Manual)</th>
</tr>
</thead>
<tbody>
<tr>
<td>For topics that Ontology-I won (22)</td>
<td>1348</td>
<td>49</td>
<td>28</td>
</tr>
<tr>
<td>For topics that Ontology-I lost (17)</td>
<td>1095</td>
<td>65</td>
<td>17</td>
</tr>
<tr>
<td>For all topics (50)</td>
<td>1111</td>
<td>54</td>
<td>21</td>
</tr>
</tbody>
</table>

Table 8.7: Comparison of the size of Ontology-I and Manual User Profiles (MAP Results)

one may see that the numbers of topics in which the Ontology models are better than, worse than, and equal to the Manual model are very similar on the MAP, macro-$F_1$ and micro-$F_1$ Measures results. This finding is confirmed by the topic distribution of percentage change results plotted in Figures 8.3 and 8.4. In these figures, in most of the topics the percentage change values calculated based on the MAP, macro-$F_1$, and micro-$F_1$ results are coincidental. These results largely agree with each other for the experimental models’ performance.

The user profiles produced by the Ontology models have better user background knowledge coverage than that produced by the Manual model. However, the Manual user profiles have the better specification. In the investigation into the experimental results, it is found that the proportional difference of the training set sizes has influence on the performance of models. This is reported by the figures in Table 8.7, which presents the comparison of the size of the Ontology-I and Manual user profiles, in terms of the MAP performance. The proportional difference is calculated by the average number of documents in the Ontology-I user profiles divided by that of the Manual user profiles. Because the MAP, macro-$F_1$ and micro-$F_1$ Measures results largely agree with each other, the related discussions use only the MAP performance for explanation, for the sake of simplicity.

In the topics where the Ontology models outperformed the Manual model, the numbers of training documents contained in the Ontology user profiles and Manual user profiles have large proportional difference. For the 22 topics in which the Ontology-I outperformed the Manual model in MAP results (as presented in
8.3. Discussion

Table 8.8: Comparison of the size of Ontology-II and Manual User Profiles (MAP Results)

<table>
<thead>
<tr>
<th></th>
<th>Average number of documents in user profiles acquired by</th>
<th>Proportional difference (Ontology-II/Manual)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ontology-II</td>
<td>Manual</td>
</tr>
<tr>
<td>For topics that Ontology-II won (14)</td>
<td>7848</td>
<td>44</td>
</tr>
<tr>
<td>For topics that Ontology-II lost (19)</td>
<td>6423</td>
<td>60</td>
</tr>
<tr>
<td>For all topics (50)</td>
<td>7610</td>
<td>54</td>
</tr>
</tbody>
</table>

Table 8.6 and plotted in Figure 8.3), the average number of training documents in the Ontology-I user profiles is 1348. This is about 28 times the documents in the Manual user profiles, which average only 49 documents.

On the other hand, for the 17 topics that the Ontology-I lost in comparison with the Manual model, the size of training sets representing the Ontology-I user profiles has relatively small proportional difference compared with that of the Manual user profiles. The average number is 1095 for the documents in Ontology-I user profiles and 65 for the documents in Manual user profiles. The Ontology-I average number is only about 17 times the average number in the Manual model, much smaller than 27, the proportional difference made in the topics by which the Ontology-I won the Manual model.

In the overall 50 topics, the proportional difference between the Ontology-I and Manual user profiles is in the middle, considering only Ontology-I winning and losing topics. The average number is 1111 for the documents in Ontology-I user profiles and 54 for the documents in Manual profiles; the proportional difference is 21 times.

The finding in the investigation on Table 8.7 can also be confirmed by the information in Table 8.8, which presents the comparisons of the size of training sets representing the Ontology-II and Manual user profiles in MAP performance. As shown on the table, for the 14 topics in which the Ontology-II outperformed the Manual model, the average size of Ontology-II training sets is 7848, which is 178 times 44, the average size of Manual training sets, whereas for the 19 topics by which the Ontology-II lost, the average size of Ontology-II profiles is 6423, only
107 times the Manual average size of 60. For the overall 50 topics, the average size of Ontology-II profiles is 7610 and the Manual profile is 54. The proportional difference is 141 times, again in the middle, considering only winning and losing topics. The comparison of the Ontology-II and Manual user profiles confirms the finding in Table 8.7 for the Ontology-I and Manual user profiles comparison.

Based on these comparisons, it can be seen that the number of training documents in the Ontology user profiles influences the performance of Ontology models. The influence is caused by the user background knowledge extracted and specified in the Ontology models and the Manual model. In the Ontology models, the user background knowledge was extracted from the world knowledge base implemented according to the LCSH system. The world knowledge base has excellent coverage of topics in the world, containing 439,329 topical subjects, 46,136 geographic subjects, and 5785 corporate subjects. Using the world knowledge base, the Ontology models can have less chance of missing relevant subjects when extracting the user background knowledge. In the Ontology-I model, the computational model first extracted the potential relevant subjects from the world knowledge base, and users selected positive and negative subjects from them. As a result, the Ontology-I model has an average of 1111 documents in their user profiles. In those topics where the Ontology-I performed well, the average size of training sets is even as large as 1348. In the Ontology-II model, the computational model took care of the entire process and used data mining techniques for non-interesting knowledge filtering and new interesting knowledge discovery. The Ontology-II model, as a result, has a large amount of training documents in user profiles (on average 7610 for overall topics and 7848 for topics performed well). The large number of training documents extracted by the Ontology models ensures that more user background knowledge was extracted and specified.

In the Manual model, the user background knowledge was specified manually by users. As previously discussed in Section 7.6.1 in Chapter 7, the training documents for each topic were obtained in two steps. The Manual linguists brought
up a topic, and first searched the RCV1 data set using the NIST’s PRISE search engine to retrieve a set of potentially relevant documents. The author of the topic then read the retrieved documents and judged them as positive or negative for relevance or non-relevance of each document to the topic. This procedure ensures that the training documents were judged accurately, however, in the trading of the user background knowledge coverage. Firstly, the number of documents retrieved from the RCV1 and provided to the Manual linguists to read were limited (54 on average). Secondly, only “positive” or “negative” could be chosen when the Manual linguists read a document. This restricted the judgements on binary values. In case of only a part of the content in a document being relevant, some user background knowledge would be missed if the document was judged “negative”. If the document was judged “positive”, some noisy concepts would be obtained in the user profiles. Consequently, the Manual model has limited user background knowledge coverage and poor knowledge presentation, which weakened the performance of the Manual model.

However, the user background knowledge contained in the Manual user profiles was proven by the users manually, because of the acquiring procedure. This is why the Manual model performed well in comparison with the Ontology models, especially in the beginning of the recall levels, as plotted in Figure 8.1. When the recall level increases, the performance of Ontology models drops more slowly, especially that of the Ontology-I model, compared with that of the Manual model. As discussed previously, users manually selected the positive and negative subjects in the Ontology-I model. This procedure maintains a relatively high accuracy rate of extracted user background knowledge, while acquiring user profiles with large concepts coverage. However, the Manual model still performed as the best in some measuring schemes that prefer precision performance to recall, such as average MAP shown in Table 8.1.

Another downside to the Manual user profiles is that the user background knowledge contained in the Manual user profiles is well formatted for human
On ontology-I vs. Semi-auto Ontology-II vs. Semi-auto

<table>
<thead>
<tr>
<th></th>
<th>Ontology-I vs. Semi-auto</th>
<th>Ontology-II vs. Semi-auto</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>(24, 10, 16)</td>
<td>(19, 5, 26)</td>
</tr>
<tr>
<td>Macro-FM</td>
<td>(24, 8, 18)</td>
<td>(16, 4, 30)</td>
</tr>
<tr>
<td>Micro-FM</td>
<td>(24, 7, 19)</td>
<td>(15, 4, 31)</td>
</tr>
</tbody>
</table>

Table 8.9: Comparisons Between the Ontology Models and the Semi-auto Model

beings to understand, but not for computers. As previously discussed, the Manual user profiles were acquired by the TREC linguists reading and judging each training document manually against the topics. The TREC linguists, being the authors who created the topics, perfectly understood their information needs and what they were looking for in the training documents. However, the TREC linguists, as ordinary Web users, still could not formally specify their background knowledge while acquiring the user profiles. The concepts contained in the Manual user profiles are implicit and difficult for computational models to understand.

The Ontology models, on the other hand, have the extracted user background knowledge formally specified. The interesting concepts were explicitly extracted from the world knowledge base and discovered from the user LIRs. In the experiments, on average there were 16 positive and 23 negative subjects manually extracted for each topic in the Ontology-I model, and 2315 subjects automatically extracted in the Ontology-II model. These subjects were constructed in ontology form, and linked by the semantic relations of \( is-a \), \( part-of \), and \( related-to \). Because of the mathematic formalisations, the ontology mining method introduced in Chapter 6 could perform and more interesting concepts could be discovered effectively. Thus, the user background knowledge contained in the Ontology user profiles is formally specified and ideal for use by computational models. This partially contributes to the superior performance achieved by the Ontology models, compared with that of the Manual model.

### 8.3.2 Ontology Models vs. Semi-auto Model

This experiment aims to evaluate the proposed ontology learning and mining model by comparing the user profiles acquired by the Ontology-I and Ontology-
II models to those acquired by the Semi-auto model, the implementation of the preliminary model introduced in Chapter 4.

The Semi-auto model in the experiments is the implementation of the preliminary model introduced in Chapter 4, as previously mentioned in Section 7.6.3. The preliminary study aimed to evaluate the hypothesis of using user concept models for Web information gathering. In the preliminary model (and thus the Semi-auto model), according to a given topic, users first specified their concept model manually. The concept models, represented by positive and negative subjects, were used by a Web search agent to retrieve training documents from the Web. The user profiles were then acquired by filtering and re-ranking the retrieved documents. In this evaluation experiment, the designed experiment environment was exactly the same as that discussed in Chapter 4 and for evaluating the preliminary model, except for 50 topics instead of 15 in the preliminary study.

The ontology learning and mining model developed in this thesis discovers and specifies user background knowledge automatically, which is superior to the preliminary model (Semi-auto model). The experimental results using all 50 topics, including the 11SPR, MAP, macro-$F_1$, and micro-$F_1$ results as previously discussed in Section 8.2, have confirmed that the Ontology-I and Ontology-II models outperformed the Semi-auto model substantially and significantly. The final developed model (Ontology models) is superior to the preliminary model (Semi-auto model), and thus the developed computational model is promising in terms of research methodology.

The user background knowledge specified in the Semi-auto model was not formalised, as that specified in the Ontology models. The concepts specified in the Semi-auto model were not supported by any knowledge base. There were neither formal definitions nor standard and consistent representations of concepts that the users could rely on. For instance, for the topic “Economic espionage” with identification number 101 presented in Figure 7.5 and discussed in Chapter 7, based on the associated description and narrative, the user concept model was
Table 8.10: User Concept Model Specified in the Semi-auto Model for Topic 101

<table>
<thead>
<tr>
<th>Positive Subject</th>
<th>Negative Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic espionage</td>
<td>Military espionage</td>
</tr>
<tr>
<td>Commercial espionage</td>
<td>Political espionage</td>
</tr>
<tr>
<td>Technical espionage</td>
<td></td>
</tr>
<tr>
<td>Industrial espionage</td>
<td></td>
</tr>
</tbody>
</table>

specified as the terms presented in Table 8.10. Consequently, the user background knowledge specified in the Semi-auto model was represented by free terms, not a controlled vocabulary and thesaurus, and the semantic relations existing in the specified concepts also remained implicit.

The utilisation of the world knowledge base leverages the Ontology models over the Semi-auto model. The world knowledge base was constructed according to the MARC 21 authority records in LCSH system. Almost 500,000 subjects were specified in the knowledge base, including topical, geographic, and corporate subjects. Also specified in the world knowledge base were the semantic relations linking the subjects, including is-a, part-of, and related-to. The world knowledge base as a global ontology provided clear and formal definitions to concepts specification in the Ontology models, as well as standard and consistent concept representations. Based on the world knowledge base, the Ontology-I model extracted the topic relevant and non-relevant concepts through the OLT semi-automatically and the Ontology-II extracted automatically. The Ontology models then constructed the extracted concepts into user personalised ontologies, as illustrated in Figure 5.8 in Chapter 5 for Topic 101. The concepts were well defined and specified in the ontologies. This benefited the Ontology models and made them superior to the Semi-auto model.

The user profiles in the Ontology models was more accurate and complete than those in the Semi-auto model. After the relevant concepts were extracted, as discussed previously, the concepts were further enriched and filtered by using the ontology mining method discussed in Chapter 6. This procedure filtered some noisy and uncertain subjects from the extracted user background knowledge. In
addition, more interesting concepts were discovered from the user LIRs. Thus, the user background knowledge contained in the final Ontology user profiles became more accurate and complete. In contrast, the Semi-auto model did not have the procedure of knowledge filtering and enriching. The subjects specified by the users were immediately used to acquire user profiles. The queries used by Google for user profile acquisition were formulated from the specified subject terms, as presented in Table 8.10. Because free terms were used in the user concept model specification, the terminological ambiguity could not be avoided, and as a result, noise existed in the specified subjects. The Semi-auto model had no filtering procedure to prune this noise. Moreover, because these subjects were used immediately for Web search, no discovery of interesting concepts occurred in the Semi-auto model, as it had in the Ontology models. Therefore, the user profiles acquired by the Semi-auto model were not as accurate as those acquired by the Ontology models.

The training documents in the Ontology-I and Ontology-II user profiles were also of higher quality, compared with that in the Semi-auto user profiles. The training sets representing the Semi-auto user profiles were retrieved from the Web. The benefit of such a procedure is that Web information covers a wide range of topics and serves a broad spectrum of communities [33]. No matter how uncommon or unusual the topic is, people can always find related information from the Web. However, the Web information has a large proportion of noisy data. When retrieving documents from the Web for user profiles, some noisy information was also retrieved by the Semi-auto model as well. Considering that the Semi-auto model had no filter procedure for the specified interesting concepts, the Semi-auto user profiles had more chances to obtain such noisy information in their training documents. Also, because of retrieving Web documents for acquiring user profiles, the quality of Semi-auto user profiles largely relied on the chosen Web search agent. The search agent employed by the Semi-auto model was Google, the Web search engine commonly used by many Web users. However,
by doing so, the Semi-auto model had no control over the search methods or algorithms but relied on Google completely. The quality of Web information acquired by the Semi-auto model for user profiles was therefore poorly controlled.

The Ontology models extracted the training documents from the user Local Instance Repositories (LIRs) for user profiles. The user LIR is the collection of a user’s personal information items, such as user stored documents, browsed Web pages, and compiled/received emails. These information items have content-related descriptors associated with the user background knowledge. In the experiments, the user LIRs were simulated by using the QUT library catalogue. The LIRs have content-related descriptors assigned to items, according to the subjects specified in the world knowledge base. Because of such content-related descriptors connecting the documents to the user background knowledge, the noise issue was largely controlled in the Ontology models. Also, considering that the Ontology model had a filtering procedure conducted in interesting concepts discovery and specification, fewer noise was obtained when retrieving documents from the user LIRs for user profiles. Consequently, in the Ontology models, the quality issue of training documents representing user profiles was better controlled, compared to the Semi-auto model. This leveraged the Ontology models and made them better than the Semi-auto model.

8.3.3 Ontology Models vs. Auto Model

The experiments performed on the Ontology-I, Ontology-II, and Auto models aimed to evaluate the computational model proposed in this thesis by comparing the user profiles extracted and specified by the Ontology models to those acquired the Auto model. The Auto model is implemented to demonstrate the non-interviewing user profile acquisition approaches, in particular the Gauch et al. OBWAN model [55] and the Sieg et al. ontological user profile model [182]. In the same way as the ontology learning and mining model proposed in this thesis, these models also utilise ontologies for user background knowledge spec-
8.3. Discussion

Table 8.11: Comparisons Between the Ontology Models and Auto Model

<table>
<thead>
<tr>
<th></th>
<th>Ontology-I vs. Auto</th>
<th>Ontology-II vs. Auto</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>(27, 6, 17)</td>
<td>(28, 14, 8)</td>
</tr>
<tr>
<td>Macro-FM</td>
<td>(27, 6, 17)</td>
<td>(26, 14, 10)</td>
</tr>
<tr>
<td>Micro-FM</td>
<td>(27, 6, 17)</td>
<td>(25, 13, 12)</td>
</tr>
</tbody>
</table>

Table 8.11: Comparisons Between the Ontology Models and Auto Model

ification, and contribute to personalised Web information gathering. However, in these models, user background knowledge is represented by a set of weighted positive subjects. The subjects are constructed in user personalised ontologies. The semantic relations specified in the ontologies and linking the subjects are the subsumption manner of super-class and sub-class. These models were summarised and implemented as the Auto model in the evaluation experiments in this thesis.

According to the experimental results presented in Section 8.2, the Ontology-I and Ontology-II models have significantly outperformed the Auto model. The experimental results of using all 50 topics, including the 11SPR, MAP, macro-$F_1$ and micro-$F_1$ Measure results, have confirmed that the Ontology-I and Ontology-II models have made substantial and significant improvement from that of the Auto model. The ontology model proposed in this thesis is more promising than the ontological user profile models developed by Gauch et al. [55] and Sieg et al. [182].

The experimental results indicate that the MAP, macro-$F_1$, and micro-$F_1$ Measure results largely agree with each other, in the same way that the Ontology and Manual models discussed in Section 8.3.1 do. Table 8.11 presents the comparisons between the Ontology models and the Auto model, based on the number of topics that the Ontology models won, lost, and tied in the experiments. Once again, the fuzziness value is set as 5%. From the table, one may see that the numbers of topics in which the Ontology models are better than, worse than, and equal to the Manual model are very similar, based on the results measured by different methods of MAP, macro-$F_1$, and micro-$F_1$. This can also be confirmed by the topic distribution of percentage change results plotted in
Figures 8.5 and 8.6. In these figures, in most of the topics the percentage change values calculated based on the MAP, macro-$F_1$ and micro-$F_1$ results are coincidental. These results largely agree with each other for the experimental models’ performance.

The user profiles acquired by the Ontology models had better user background knowledge coverage than that acquired by the Auto model. As discussed in Chapter 7.6.2, the Auto model retrieved the training documents from the same data set as that used by the Ontology models, the users’ Local Instance Repositories simulated by the QUT library catalogue. However, in the investigation, it is found that the information gathering performance had a connection with the size of training sets representing user profiles. This is shown in Table 8.12, which presents the comparisons of the average number of documents in the Ontology-I and Manual user profiles, in terms of the MAP performance. Because the MAP, macro-$F_1$ and micro-$F_1$ Measures results largely agree with each other, once again
this discussion uses only the MAP performance for explanation, for the sake of simplicity.

In the 27 topics in which the Ontology-I model outperformed the Auto model, the average numbers of training documents in the Ontology-I user profiles is 893, which is about 4.2 times 213, the average number of documents in the Auto user profiles. In the six topics that the Ontology-I lost in comparison with the Auto model, the average size of training sets representing the Auto user profiles is 1518, which is much bigger compared to 213. In contrast, the average size of training
Table 8.13: Comparison of the size of Ontology-II and Auto User Profiles (MAP Results)

<table>
<thead>
<tr>
<th>Topics that Ontology-II won (28)</th>
<th>Ontology-II</th>
<th>Auto</th>
<th>Proportional difference (Ontology-II/Auto)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topics that Ontology-II lost (14)</td>
<td>6252</td>
<td>451</td>
<td>13.9</td>
</tr>
<tr>
<td>All topics</td>
<td>7610</td>
<td>436</td>
<td>17.5</td>
</tr>
<tr>
<td></td>
<td>7960</td>
<td>237</td>
<td>33.6</td>
</tr>
</tbody>
</table>

sets representing the Ontology-I user profiles is 1729, almost the same as the
Auto average size. In the overall 50 topics, the proportional difference between
the Ontology-I and Auto user profiles is in the middle of that considering only
the Ontology-I winning and losing topics. The average number is 1111 for the
documents in Ontology-I user profiles and 436 for the Auto user profiles, and the
proportional difference is only 2.5 times.

The finding can also be confirmed by the training set size comparison between
the Ontology-II and Auto models, which is presented in Table 8.13. As shown
on the table, for the 28 topics in which the Ontology-II model outperformed
the Auto model, the average size of Ontology-II training sets is 7960, which is
33.6 times 237, the size of Auto training sets, whereas for the 14 topics that
the Ontology-II lost, the average size of Ontology-II profiles is 6252, only 13.9
times the average size of 451 in the Auto user profiles. For the overall 50 topics,
the average size of Ontology-II profiles is 7610, with 436 for the Manual user
profiles. The proportional difference is 17.5 times, again in the middle range
of those considering only the winning and losing topic. The comparison of the
Ontology-II profiles and Auto profiles confirms the finding in Table 8.12 for the
Ontology-I and Auto user profiles comparison.

Based on these comparisons, one may see that the number of training docu-
ments in the Ontology user profiles contributes to the improvement made by the
Ontology models from the Auto model. In the investigation, it is found that this
contribution was caused by the user background knowledge extracted and spec-
ified in the Ontology models, which was more accurate and had better coverage
than that in the Auto model.

The Ontology models used both positive and negative subjects for their user background knowledge specification, which makes the knowledge specification much more accurate than that of the Auto model using positive subjects only. Because the training documents for user profiles were acquired by using the specified subjects, the Ontology user profiles had negative training documents that the Auto user profiles did not have. Thus, the Ontology user profiles had more documents in their training sets.

The Ontology models and the Auto model extracted the relevant concepts from the same world knowledge base, as discussed in Chapter 7.6.2. In addition, the Auto model used exactly the same positive subject sets as that used by the Ontology-I model. However, the Ontology models, including Ontology-I and Ontology-II, specified user background knowledge not only in positive subjects but also in negative subjects. Many achievements have been reported by using both positive and negative samples to learn classifiers in the data mining and text classification communities [52,65,100,116,233,234]. Negative subjects thus helped the Ontology models to clarify the specification of user background knowledge in the experiments.

In the next phase after positive and negative subjects extraction, an ontology mining method was performed by the Ontology models, in which the negative subjects were used to filter the extracted positive concepts, as discussed in Chapter 6. Thus, the user background knowledge specified by the Ontology models was more accurate than that by the Auto model.

Also, the information gathering system used in the experiments, as discussed in Chapter 7.4, was designed to use training documents in Web information gathering. The same as the Rocchio [162] and Dempster-Shafer [90] models, the information gathering system used in the experiments was sensitive regarding the positive and negative training documents. Hence, having both positive and negative subjects and training sets significantly benefits the performance of the
Ontology models. In contrast, when the Ontology models did not have a sufficient number of negative subjects and training documents present, this advantage was weakened and the performance went down. As shown in the second row in Table 8.12, Ontology-I model lost in the comparison with the Auto model when the Ontology-I training sets had insufficient negative documents present and almost the same size as that of the Auto training sets.

Based on these, it can be concluded that the more accurate user background knowledge specification contributes to the superior performance of the Ontology models over that of the Auto model.

The specification of is-a, part-of, and related-to semantic relations also contributes to the high accuracy level of user background knowledge specified in the Ontology models. The concepts stored in the personalised ontologies constructed in the Ontology models were specified by is-a, part-of, and related-to semantic relations. They were more specific than the super-class and sub-class used in the ontologies constructed in the Auto model. Because of such specific semantic relations, the ontology mining method, as discussed in Chapter 6, was able to perform on the constructed personalised ontologies. The influence of subjects on each other was clarified in the Ontology models. Such influence was counted for the support value of subjects to the given topics, as well as the users’ personal interests discovered from the user Local Instance Repositories. The Auto model did not consider the specific difference within the super-class and sub-class of subjects, and had no ontology mining method performed to investigate such specific relations in the ontologies. Instead of that, the Auto model valued all positive subjects as one initially, and increased the support value of a subject when it was cited by more instances (documents) in the user’s Local Instance Repository. This procedure counted users’ personal interests, but failed to investigate the influence of semantic relations on the support value of subjects to the topics. Hence, the user background knowledge specified in the Ontology models was more accurate than that in the Auto model.
The user profiles acquired by the Ontology models had better coverage than those acquired by the Auto model. In the Ontology models, the user background knowledge was further enriched after extraction, which was completed by using the ontology mining method discussed in Chapter 6. During the procedure of knowledge enrichment, more interesting subjects were discovered from user Local Instance Repositories and added into the user background knowledge. The Auto model, in contrast, did not have the procedure of user background knowledge enrichment. Considering that the training documents representing user profiles were acquired by using subjects, the Ontology training sets have larger sizes than the Auto user profiles. As a result, the Ontology models performed better when the Ontology user profiles and Auto profiles have a large proportional difference in their training set sizes, as displayed in the first row in Table 8.12 and Table 8.13. Thus, the user background knowledge contained in the Ontology user profiles was more complete than that contained in the Auto user profiles.

8.3.4 Ontology-I Model vs. Ontology-II Model

The experiments performed on the Ontology-I and Ontology-II models aimed to evaluate the user profiles acquired by the semi-automatic and automatic Ontology models, as proposed in this thesis.

The ontology-I model was the implementation of the ontology learning and mining model using the semi-automatic ontology learning method, as discussed in Chapter 5. In this model, users’ personalised ontologies were constructed according to the given topics through the OLE. The candidate positive and negative subjects were extracted from the world knowledge base first, and then users selected the positive and negative manually, based on their judgements of the candidate subjects.

In the Ontology-II model, the implementation using the automatic ontology learning method extracted relevant subjects from the world knowledge base using the syntax-matching mechanism first, and relied on the ontology mining method
discussed in Chapter 6 to filter the noisy subjects and discover more interesting subjects. The experimental results, as the 11SPR results have shown in Figure 8.1, demonstrate that the Ontology-I and Ontology-II models have almost the same performance before recall level 0.5. After that, the Ontology-I model outperformed the Ontology-II model and has better precision results. The detailed MAP, macro-$F_1$, and micro-$F_1$ Measure results are presented in Tables 8.1, 8.4, and 8.5 respectively. All of these results have the same report, that the semi-automatic Ontology-I model has achieved higher performance in comparison with the automatic Ontology-II model in experiments.

The experimental results indicate that the MAP, $macro-F_1$, and $micro-F_1$ Measure results largely agree with each other, similar to those of the experiments on Ontology models vs. the Manual model and on Ontology models vs. the Auto model. Table 8.14 presents the comparisons between the Ontology-I and Ontology-II models, based on the number of topics that the Ontology-I models won, lost, and tied, in comparison with Ontology-II in the experiments. Again, the fuzziness value is set as 5%. The numbers of topics in which the Ontology-I model is better than, worse than, and equal to the Ontology-II model are very similar, based on the results measured by different methods of MAP, macro-$F_1$, and micro-$F_1$. These results largely agree with each other for the Ontology models’ performance.

The user profiles acquired by the Ontology-I user profiles had a better accuracy rate, compared with those acquired by the Ontology-II user profiles. The Ontology-I model had relatively better performance achieved in MAP results in comparison with that in $macro-F_1$ and $micro-F_1$ results, when compared to the Ontology-II model. This finding is visualised in Figure 8.7, the plotted average

<table>
<thead>
<tr>
<th></th>
<th>Ontology-I vs. Ontology-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>(20, 13, 17)</td>
</tr>
<tr>
<td>Macro-FM</td>
<td>(21, 13, 16)</td>
</tr>
<tr>
<td>Micro-FM</td>
<td>(20, 12, 18)</td>
</tr>
</tbody>
</table>

Table 8.14: Comparisons Between the Ontology-I and Ontology-II Models
percentage change comparisons between the Ontology-I and Ontology-II models. The improvement made by the Ontology-I over the Ontology-II model in MAP performance is almost double that made in the macro-$F_1$ and micro-$F_1$ performances. While the $F_1$ Measure balances the importance of precision and recall, the MAP appreciates precision more than recall in Web information gathering. Thus, the higher achievement in MAP performance than in $F_1$ Measure made by the Ontology-I model indicates that the user background knowledge specified in the Ontology-I user profiles had better accuracy than that in the Ontology-II user profiles.

In the Ontology-I model, users manually selected the positive and negative subjects. The number of specified subjects was limited as users read and selected the subjects carefully through the Ontology Learning Environment. As a result, the Ontology-I model had 39 subjects selected on average per topic, including 16 positive and 23 negative subjects. Such numbers of subjects were easy to control, and thus this procedure maintained a relatively high accuracy rate of extracted user background knowledge. This is also confirmed by the 11SPR performance
plotted in Figure 8.1. At the first six recall levels (0.0 to 0.5), the Ontology-II model achieved a similar performance to that of the Ontology-I model. However, when the recall level increases and more gathered Web documents are under assessment, the Ontology-II’s 11SPR performance decreases and is eventually lost in comparison with the Ontology-I model. In the Ontology-II model, however, the user background knowledge was extracted entirely automatically from the world knowledge base. As a result, the Ontology-II model had 2315 subjects extracted for each topic on average. For such a large number of subjects, it is difficult to avoid uncertainties. Thus, many noisy subjects were extracted as well as the useful and meaningful subjects, because of automatic extraction. Although the ontology mining method (as discussed in Chapter 6) was performed in the Ontology-II model, the issue could not be controlled as well as that in the Ontology-I model. Thus, when the recall level increases, this disadvantage of Ontology-II user profiles becomes more significant, and the Ontology-II performance decreases.

8.4 Conclusion

In this chapter, the experimental results were presented and discussed for the evaluation of ontology learning and mining model proposed in this thesis. The experiments were performed by comparing the information gathering performance achieved by using the proposed model with that achieved by using the human-based and state-of-the-art computational user profile acquiring models. According to the experimental results, the Ontology-I and Ontology-II models, the different implementations of the proposed model, achieved close performance to the human-based model and outperformed the state-of-the-art computational models. Therefore, the experimental hypotheses introduced in Chapter 7 are proven to be correct. The ontology learning and mining model proposed in this thesis is thus evaluated promisingly, for the ability to acquire user profiles and capture user information needs effectively.
Chapter 9

Conclusions and Future Work

9.1 Ontology Learning and Mining Model

Over the last decade, the rapid growth and adoption of the World Wide Web has further exacerbated user needs for efficient mechanisms for information and knowledge location, selection, and retrieval. Web information covers a wide range of topics and serves a broad spectrum of communities. However, how to gather useful and meaningful information from the Web has become challenging to Web users.

The current Web information gathering systems cannot satisfy Web users, as they are mostly based on keyword-matching mechanisms and suffer from the problems of information mismatching and information overloading [110]. Usually, Web users provide only short phrases in queries to express their information needs [191]. Also, Web users formulate their queries differently because of their personal perspectives, expertise, terminological habits, and vocabularies. If user information needs can be better captured and interpreted, more useful and meaningful information can be delivered to users and better Web information gathering performance can then be achieved.
Web users implicitly possess a concept model obtained from their background knowledge and use that model in information gathering [110]. They can easily determine whether or not a document is interesting to them when reading through the document content, although they may be unable to express the reason explicitly. Thus, a hypothesis arises that if this user concept model can be rebuilt, user information needs can be captured accurately, and thus more meaningful and personalised Web information can be gathered for users. Ontologies, as a formal description and specification of knowledge, are utilised by many researches to represent user concept models. However, few investigations have been performed on using ontologies to capture user information needs in Web information gathering.

In this thesis, an ontology learning and mining model is proposed that aims to simulate user concept models for personalised Web information gathering. The model is proposed under the assumptions and scopes defined by the concept-based Web information gathering framework in Chapter 3. The framework aims to use user background knowledge to improve Web information gathering performance. It consists of a user concept model, a querying model, a computer model, and finally an ontology model. The computer model is implemented by the ontology learning and mining model proposed in this thesis, and the ontology model is the personalised ontologies constructed for user concept models.

The ontology learning and mining model attempts to effectively acquire user profiles to capture user information needs. Two ontology learning methods, automatic and semi-automatic, are proposed in the model to learn personalised ontologies for users (Chapter 5). Based on the Library of Congress Subject Headings, which is a library system that represents human intellectual endeavour and has been undergoing continuous revising and enriching for over a hundred years, a world knowledge base is constructed to extract the topic relevant subjects for personalised ontology learning. The constructed user personalised ontologies are further investigated using an ontology mining method, Specificity and Exhaustivity, presented in Chapter 6. The aim is to discover more interesting and on-topic
subsubjects from the users’ LIRs, which are users’ personal collections of information items. The interesting subjects, along with their associated semantic relations of is-a and part-of, are analysed for user background knowledge specification. Based on the user background knowledge, the user profiles are acquired and information needs are captured effectively.

The ontology learning and mining model is evaluated by comparing the acquired user profiles with those acquired by the baseline models in experiments, as designed in Chapter 7. A large, standard data set was used in the experiments, and the experimental results were measured by using the modern and standard methods widely used in information gathering evaluations. By using the user profiles acquired by the proposed ontology learning and mining model, the Web information gathering system performed closely to that using the profiles acquired by a manual model, and significantly outperformed that which used the profiles acquired by other baseline models (Chapter 8). The ontology learning and mining model proposed in this thesis is promising and capable of specifying user background knowledge and capturing user information needs for Web information gathering.

9.2 Contributions

This thesis makes a number of contributions to knowledge engineering and Web information gathering research.

An important and challenging issue in knowledge engineering is to emphasise the specific semantic relations in one single computational model. Existing mathematic models formalise either subsumption relations only, such as super-class and sub-class by [55,74,84,242], or part-of only, such as [58,59,164,169], or related-to only, such as [71,205]. Few of them consider various specific semantic relations like is-a, part-of, and related-to, together in one framework. However, in the real world various semantic relations exist together. They are not isolated from each other. Thus, specifying various semantic relations in one single model
for investigation is important in knowledge engineering. This thesis presents a computational model that emphasises various semantic relations of *is-a, part-of*, and *related-to* in one single framework. The influence produced by various semantic relations is investigated and measured quantitatively. This work develops an explorative model for the design of new models in knowledge engineering, and explores a possible solution to the aforementioned important and challenging issue. It is a new contribution to knowledge engineering.

The focus and extent of concepts in ontologies have not yet been fully investigated. Formalising the focus and extent of concepts is important in knowledge engineering. If the relationship between the concept focus and concept extent can be specified and the influence they have on each other can be clarified, the concepts in ontologies can be better defined, and the utilisation of concepts can be more accurate and appropriate. In this thesis, two concepts, *specificity* and *exhaustivity*, are introduced; they formalise the focus and extent of concepts respectively. When the *specificity* of a concept is strong, the *exhaustivity* of the concept becomes limited; similarly, when the *exhaustivity* of a concept becomes large, the *specificity* of the concept is weak. An ontology mining method is also proposed in this thesis to measure the *specificity* and *exhaustivity* of concepts for concepts analysis in ontologies. This research explores novel schemes for concept investigation in ontologies, and is also a new theoretical contribution to knowledge engineering.

User profiles are largely used in web personalisation, but existing user profile acquisition techniques are either ineffective or inefficient. The interviewing user profile acquisition techniques, like that used by TREC-11 Filtering Track [161], are computationally costly although effective. The user profiles acquired by the non-interviewing techniques, such as by [55, 148, 202] and [182], show lack of accuracy. This thesis proposes an approach that acquires user profiles effectively. The user personalised ontologies are first constructed using the world knowledge base. The user profiles are acquired from the user LIRs, by using the user person-
alised ontologies. The experiments demonstrate that the acquired profiles have the same quality as that of the interviewing TREC user profiles, and are better than that acquired by the baseline non-interviewing techniques. This concept-based approach using personalised ontologies is a novel exploration of user profile acquisition, and provides a new benchmark for other researches. It is a new contribution to personalised Web information gathering.

The global knowledge bases are commonly used in ontology learning, but few are adequate. The global knowledge bases may be used to learn ontologies in multiple domains; therefore, the basic requirement of a global knowledge base is the large coverage of topics. However, many of the knowledge bases used in ontology learning cover only a small volume of topics, like that used by [84]. Aiming to learn ontologies, the global knowledge bases also need to define concepts formally, and specify various semantic relations existing in concepts. Many global knowledge bases, for example, those used for ontology learning by [55,74,84,242], may have concepts defined but various semantic relations not specified. Ontologies are also for knowledge sharing by different applications. Thus, the global knowledge bases need to be constructed by reliable methods, either by experts manually or by evaluated computational methods. Global knowledge bases like those used by [45,138,158] have adequate topic coverage, but are contributed by volunteers in an uncontrolled manner. It can hardly be said that their contained knowledge is reliable unless proven.

This thesis constructs a world knowledge base out of the LCSH system for ontology learning, which covers a great range (topical, geographic, and corporate) and has a huge volume (491,250) of topics. The world knowledge base also specifies various semantic relations in details, including is-a, part-of, and related-to relationships. As a human intellectual endeavour, the LCSH has been undergoing continuous manual revising and enriching for over a hundred years by linguists and librarians. The subjects in LCSH are classified by professionals, and the classification quality is guaranteed by well-defined and continuously-refined cat-
aloging rules [26]. Therefore, the world knowledge base constructed in this thesis is also reliable and the quality is guaranteed. This work provides an ideal world knowledge base for knowledge models developed by other scientific researches, and is a practical contribution to knowledge engineering.

The contributions claimed by this thesis work are under an exception that the users’ LIRs are the collections of information items compiled in the formats applicable to the Semantic Web because of the use of the library catalogue to simulate the LIRs. The Semantic Web has inter-operability standards for both the syntactic form of documents and the semantic content [232,243,244]. Thus, the Semantic Web documents have content-related descriptors associating with the concepts specified in external knowledge bases [38]; for example, the metadata tags in XML, RDF, OWL, DAML, and XHTML documents citing the concepts in knowledge bases. The findings of this thesis may be inapplicable to the Web documents that do not have such content-related descriptors specified. However, given more and more Semantic Web documents being available online and the recognition of Semantic Web as the future of the Web [4,127,206], the contributions made by this thesis work are valuable and increasingly significant.

9.3 Future Work

Given the above conclusions, a few avenues of research have arisen and will be pursued in the future work that extends from this thesis.

The user profile acquisition is extendable from routing user profiles to adaptive user profiles. User profile acquisition can be routing or adaptive, depending on the short-term or long-term period in which the user profiles are valuable for Web information gathering [197]. The user profiles acquired by the ontology learning and mining model in this thesis are routing for short-term user profiles. They do not consider the adaptive change of user interests during a temporal frame. It will be interesting to investigate the adaptive change of user interests in a long-term period and to measure its influence on Web information gathering performance.
The research work for acquiring user profiles conducted in this thesis can be extended from routing to adaptive, in order to investigate user information need capture, considering both short-term and long-term user interests, by using the world knowledge base and user LIRs.

The specification of various semantic relations is also extendable for further investigation. The LCSH system provides knowledge engineering researches with an ideal environment consisting of various semantic relations of is-a, part-of, and related-to. In this thesis, only the subjects with is-a and part-of relationships are extracted from the world knowledge base and investigated. The related-to relationships existing amongst subjects are identified based on their referring instances, but not the specifications in the world knowledge base. The specifications of subjects and semantic relations in the world knowledge base have therefore not yet been thoroughly investigated. Further emphasising the is-a, part-of, and related-to relations is a new challenge and the course that will be pursued in future work.

The visualisation of user profiles is also a potential work that can be extended from this thesis work.

9.4 Overall Conclusion

The major finding of this thesis is that user profiles can be acquired to capture user information needs effectively by using personalised ontologies. The finding is based on evaluation experiments which model Web users’ possessed concept models in Web information gathering.
# Appendix A

## TREC Topics in Experiments

<table>
<thead>
<tr>
<th>ID</th>
<th>Title</th>
<th>ID</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Economic espionage</td>
<td>126</td>
<td>Nuclear plants U.S.</td>
</tr>
<tr>
<td>102</td>
<td>Convicts, repeat offenders</td>
<td>127</td>
<td>U.S. automobile seat belt</td>
</tr>
<tr>
<td>103</td>
<td>Ferry Boat sinkings</td>
<td>128</td>
<td>Child labor laws</td>
</tr>
<tr>
<td>104</td>
<td>Rescue of kidnapped children</td>
<td>129</td>
<td>Problems illegal aliens U.S.</td>
</tr>
<tr>
<td>105</td>
<td>Sport Utility Vehicles U.S.</td>
<td>130</td>
<td>College tuition planning</td>
</tr>
<tr>
<td>106</td>
<td>Government supported school vouchers</td>
<td>131</td>
<td>Television U.S. children</td>
</tr>
<tr>
<td>107</td>
<td>Tourism Great Britain</td>
<td>132</td>
<td>Friendly fire deaths</td>
</tr>
<tr>
<td>108</td>
<td>Harmful weight-loss drugs</td>
<td>133</td>
<td>Anti-rejection transplant drugs</td>
</tr>
<tr>
<td>109</td>
<td>Child custody cases</td>
<td>134</td>
<td>Crime Statistics Great Britain</td>
</tr>
<tr>
<td>110</td>
<td>Terrorism Middle East tourism</td>
<td>135</td>
<td>WTO trade debates</td>
</tr>
<tr>
<td>111</td>
<td>Telemarketing practices U.S.</td>
<td>136</td>
<td>Substance abuse crime</td>
</tr>
<tr>
<td>112</td>
<td>School bus accidents</td>
<td>137</td>
<td>Sea turtle deaths</td>
</tr>
<tr>
<td>113</td>
<td>Ford foreign ventures</td>
<td>138</td>
<td>Creutzfeldt-Jakob, mad cow disease</td>
</tr>
<tr>
<td>114</td>
<td>Effects of global warming</td>
<td>139</td>
<td>Pig organ transplants</td>
</tr>
<tr>
<td>115</td>
<td>Indian casino laws</td>
<td>140</td>
<td>Computer simulation</td>
</tr>
<tr>
<td>116</td>
<td>Archaeology discoveries</td>
<td>141</td>
<td>Environment National Park</td>
</tr>
<tr>
<td>117</td>
<td>Organ transplants in the UK</td>
<td>142</td>
<td>Illiteracy Arab Africa</td>
</tr>
<tr>
<td>118</td>
<td>Progress in treatment of schizophrenia</td>
<td>143</td>
<td>Improving aircraft safety</td>
</tr>
<tr>
<td>119</td>
<td>U.S. gas prices</td>
<td>144</td>
<td>Mountain climbing deaths</td>
</tr>
<tr>
<td>120</td>
<td>Deaths mining accidents</td>
<td>145</td>
<td>Airline passenger disruptions</td>
</tr>
<tr>
<td>121</td>
<td>China Pakistan nuclear missile</td>
<td>146</td>
<td>Germ warfare</td>
</tr>
<tr>
<td>122</td>
<td>Symptoms Parkinson’s disease</td>
<td>147</td>
<td>Natural gas vehicles</td>
</tr>
<tr>
<td>123</td>
<td>Newspaper circulation decline</td>
<td>148</td>
<td>North American Free Trade Agreement</td>
</tr>
<tr>
<td>124</td>
<td>Aborigine health</td>
<td>149</td>
<td>Aid to handicapped people</td>
</tr>
<tr>
<td>125</td>
<td>Scottish Independence</td>
<td>150</td>
<td>Drive-by shootings</td>
</tr>
</tbody>
</table>
101 Economic espionage

**Description** What is being done to counter economic espionage internationally?

**Narrative** Documents which identify economic espionage cases and provide action(s) taken to reprimand offenders or terminate their behavior are relevant. Economic espionage would encompass commercial, technical, industrial or corporate types of espionage. Documents about military or political espionage would be irrelevant.

102 Convicts, repeat offenders

**Description** Search for information pertaining to crimes committed by people who have been previously convicted and later released or paroled from prison.

**Narrative** Relevant documents are those which cite actual crimes committed by “repeat offenders” or ex-convicts. Documents which only generally discuss the topic or efforts to prevent its occurrence with no specific cases cited are irrelevant.

103 Ferry Boat sinkings

**Description** Documents will report on any sinkings of Ferry Boats throughout the world.

**Narrative** Documents that identify any instances where a ferry boat has sunk or capsized are relevant; only boats identified as ferries should be considered relevant.

104 Rescue of kidnapped children

**Description** Identify a kidnapping of a child or children when the child or children have been rescued or released.
Narrative  Documents discussing abducted or kidnapped children are relevant. Documents referring to abuse of children without reference to kidnapping or abduction are irrelevant. Cases of kidnapping where some children are murdered or not found while others are rescued are relevant.

105  Sport Utility Vehicles U.S.

Description  Find documents that will illustrate the phenomenal growth in the number of SUV’s owned by Americans, and concerns about their safety and environmental impact.

Narrative  Documents that discuss the growth in ownership of Sport Utility Vehicles in the United States are relevant. Documents including sales reports and projections by manufacturers are relevant. Documents about Consumer groups identification of potential problems would be relevant. Documents about light trucks are not relevant.

106  Government supported school vouchers

Description  Research documents on the pros/cons of government supported school vouchers for private or religious schools.

Narrative  Documents containing statements by elected officials, civic groups or clergy on the use of public funds in support of private and religious schools for tuition, books, building maintenance and busing are relevant. Documents that include state or local ballot initiatives and the result on the use of public monies toward this end are relevant. Documents about lawsuits addressing this subject regardless of the court level are relevant.

107  Tourism Great Britain

Description  Retrieve documents pertaining to tourism into Great Britain and the efforts being undertaken to increase it.
Narrative Documents about Scotland, Wales and only Northern Ireland are relevant as well as documents about many offshore islands which may be mentioned without specifically being identified as part of Great Britain.

108 Harmful weight-loss drugs

Description Identify medicines used for obesity or weight-loss that have harmful side effects.

Narrative Relevant documents will show specific, harmful side effects.

109 Child custody cases

Description Research reports on child custody cases.

Narrative Relevant documents concentrate on custody cases between blood relatives such as parents or, grandparents and parents. Children being held in custody by police or social services due to family problems are irrelevant.

110 Terrorism Middle East tourism

Description Relevant documents directly correlate terrorism with its effect on tourism in the Middle East. Documents reflecting either terrorism or tourism in the area but not associating the effect of one or the other are irrelevant.

111 Telemarketing practices U.S.

Description Find documents which reflect telemarketing practices in the U.S. which are intrusive or deceptive and any efforts to control or regulate against them.

Narrative Telemarketing practices found to be abusive, intrusive, evasive, deceptive, fraudulent, or in any way unwanted by persons contacted are relevant. Only such practices in the U.S. are relevant. All efforts to halt these practices, including lawsuits, legislation or regulation are also relevant.

112 School bus accidents
Description Identify any documents noting school bus accidents that resulted in the death of a student.

Narrative Relevant documents will identify any instances where a school bus accident has resulted in the death of a student. Documents specifying location and number of deaths are relevant. Buses carrying children on school sponsored trips are relevant.

113  Ford foreign ventures

Description Track joint ventures, partnerships and cooperative alliances between the Ford Motor Co. and foreign entities.

Narrative Current, intact ventures are the only ones relevant. Ventures planned for, hoped for or being explored, as well as past ventures which have broken up are irrelevant. Ford Motor Co., and at least one other foreign entity must be named in the document. Units of Ford which are involved in alliances with foreign entities are relevant.

114  Effects of global warming

Description Evidence of effects of global warming or the greenhouse effect on climate and environment.

Narrative Only articles that describe actual changes due to global warming or the greenhouse effect are relevant. Current evidence that points to future effects is relevant.

115  Indian casino laws

Description Research the state laws regarding the construction, operation, and distribution of profits of the gambling casinos on U.S. Indian Reservations.

Narrative Documents that show laws and ballot initiatives pertaining to the operation of gambling casinos on U.S. Indian Reservations are relevant.
Documents about ballot initiatives are relevant, whether or not passed by the voters. Documents about negotiations with companies that provide casino operations are irrelevant. Negotiations between states and the tribes are relevant. Documents about riverboat casinos are irrelevant.

116   Archaeology discoveries

Description Find current documents on new archaeological discoveries in the world.

Narrative Documents interpreting former discoveries should be excluded.

117   Organ transplants in the UK

Description Research reports on organ transplantation in the United Kingdom.

Narrative Reports on actual organ transplant cases are relevant. Also relevant are research programs in the UK and elsewhere that are developing drugs to enhance the transplant acceptance rate.

118   Progress in treatment of schizophrenia

Description Provide documents reflecting any progress in medical research for the treatment of schizophrenia.

Narrative Documents providing the names of drugs used as treatment of schizophrenia are relevant. Documents that described drugs that showed an improvement in the severity of schizophrenia are relevant.

119   U.S. gas prices

Description Find documents discussing possible reasons for the wide fluctuation in U.S. automobile gasoline prices.

Narrative Documents that provide reasons why U.S. gasoline prices fluctuate are relevant. Documents concerning gas prices in other countries are not relevant.
120  Deaths mining accidents

Description  Identify any documents mentioning deaths in mining accidents.

Narrative  Documents listing statistics on number of mining deaths are relevant. Documents about ethnic clashes, and resultant deaths of mine workers near a mine are not relevant.

121  China Pakistan nuclear missile

Description  Search for evidence of whether or not China is aiding Pakistan in developing military nuclear or missile capabilities.

Narrative  Documents which contain information confirming or denying China’s aiding Pakistan in developing military nuclear and missile capabilities are relevant. General references to the subject with no details are irrelevant.

122  Symptoms Parkinson’s disease

Description  Find early symptoms of diagnosing Parkinson’s disease. What changes take place indicating that one has the early stages of the disease?

Narrative  Documents discussing people with Parkinsons without giving the symptoms are irrelevant. If a document gave known symptoms, but does not identify them as Parkinsons it is irrelevant.

123  Newspaper circulation decline

Description  Collect documents which address the decline of newspaper circulation and the reasons for its occurrence.

Narrative  Documents which cite both circulation decline and reasons for the decline are relevant. Documents showing circulation decline without attribution for cause are irrelevant.

124  Aborigine health
Description Research reports on the health of aborigine peoples.

Narrative Relevant documents will address current attempts to improve the health of the aborigine peoples in Australia.

125 Scottish Independence

Description The Scottish people have been pushing for independence from Great Britain. What is being reported on their progress?

Narrative Documents that only discuss creation of a Scottish Parliament without full independence are not relevant. Documents reporting support for an independent Scottish Parliament are relevant.

126 Nuclear plants U.S.

Description Find the location and status of United States nuclear power plants.

Narrative Documents giving a specific location of a nuclear power plant regardless of status are relevant.

127 U.S. automobile seat belt

Description Find documents concerning the use of automobile seat belts by the U.S. population.

Narrative Relevant documents show the use of seat belts by the U.S. population. Documents encouraging the use of seat belts and/or describe the proper use of seat belts, especially for children, are relevant.

128 Child labor laws

Description Research documents covering the current state of child labor laws.

Narrative Relevant documents discuss the creation of laws to establish the base age for children to work, the hours they can work and the conditions under which they may work.
129 Problems illegal aliens U.S.

**Description** Find documents referencing problems resulting from illegal aliens residing in the U.S..

**Narrative** Documents that mention illegal alien activity without citing difficulties caused by illegal aliens are irrelevant, as are references to illegal alien problems in countries other than the U.S..

130 College tuition planning

**Description** Find documents discussing the spiraling cost of college tuition and what families are doing to prepare for it.

**Narrative** Documents that describe a plan where parents can contribute to offset expensive tuition costs are relevant. Documents that discuss saving money by pre-planning for college expenses are relevant.

131 Television U.S. children

**Description** Produce documents reflecting actions taken to improve the quality of children’s television in the U.S..

**Narrative** Documents discussing actions taken in the U.S. to provide better quality television programs for children are relevant.

132 Friendly fire deaths

**Description** Identify any instances where death has resulted due to ”Friendly Fire” or military training accidents.

**Narrative** Relevant documents describe death occurring during performance of official duty. Civilian deaths occurring as a result of official military duty are relevant. Suspected ”Friendly Fire” are irrelevant.

133 Anti-rejection transplant drugs
**Description** Identify immune-suppressing drugs that are used or being studied to prevent rejection of organ transplants in humans or animals.

**Narrative** Research using human stem cell cultures are irrelevant. Documents referring to transplant medicine without identifying it as anti-rejection are irrelevant.

**134 Crime Statistics Great Britain**

**Description** Find all documents relating to the increase or decrease of crime in Great Britain.

**Narrative** Parliamentary debate, political speeches, calls by citizen groups and clergy for government action against crime are considered relevant only if statistics are included. Reports on individual crimes and war crime tribunals are not relevant.

**135 WTO trade debates**

**Description** The WTO has had an impact upon world trade. What are the current trade issues being debated by the WTO?

**Narrative** Relevant documents will contain information pertaining to an issue between two or more members of the WTO such as tariff rates imposed by one entity against others for a specific commodity.

**136 Substance abuse crime**

**Description** Find documents linking substance abuse to other criminal behavior.

**Narrative** Relevant documents directly associated substance abuse (e.g. drugs, alcohol) with criminal activity. Crime committed while under the influence of a drug is relevant. Documents referring to drugs and/or crime without providing a direct relationship between the two are not relevant. Drug crimes such as smuggling and trafficking are not relevant.
137 Sea turtle deaths

**Description** Identify any information relevant to the deaths of sea turtles.

**Narrative** Relevant documents will provide any information with information on the deaths of sea turtles including where and reasons for their death.

138 Creutzfeldt-Jakob, mad cow disease

**Description** Find documents which contain information on cases of Creutzfeldt-Jakob disease (CJD) in humans attributable to contact with or the consumption of beef products from cattle infected with Bovine Spongiform Encephalopathy (BSE) also known as mad cow disease.

**Narrative** Relevant documents cite specific cases or the current tally of cases of CJD believed to have been caused by contact with or ingestion of BSE-infected cattle, beef, or related products. General discussion of the possibility or likelihood of its occurrence is irrelevant.

139 Pig organ transplants

**Description** Research reports on the use of pigs for organ transplants in humans.

**Narrative** Relevant documents show the development of pigs for organ transplants and the actual use of pig organs for transplants. Development of drugs to assist organ transplants are not relevant.

140 Computer simulation

**Description** Reports on how computer simulation and modelling techniques are being used by business and government.

**Narrative** Documents reporting the use of simulation and modelling techniques to improve business and to understand and predict happenings in the real world (such as weather predictions) are relevant.

141 Environment National Park
Description Find documents relating to environmental problems in U.S. National Parks and any Congressional actions which address these problems.

Narrative Documents addressing National Forests problems, individual hiker accidents not caused by an environmental problem and park improvement documents not directly related to the environment are not relevant. Congressional debate over monies to support acquisition and clean-up are relevant if deemed a result of an environmental issue impacting the park.

142 Illiteracy Arab Africa

Description Research reports on the illiteracy rates in African and Arab countries.

Narrative Relevant documents discuss illiteracy in Africa and the Arab world, or indicate the percentage of African and Arab people that are illiterate.

143 Improving aircraft safety

Description What is being done by U.S. airplane manufacturers to improve the safety of their passenger aircraft?

Narrative Relevant documents reflect independent actions taken by airlines, under their own initiative, to improve the safety of their passenger aircraft. Documents citing actions taken by the manufacturers as a result of safety mandates imposed by Federal regulations are not relevant.

144 Mountain climbing deaths

Description Identify any information where mountain climbing has resulted in death of an individual.

Narrative Relevant documents identify any instance of death due to mountain climbing. Documents that provide information on where the accident occurred, the cause such as avalanche, falling or victim freezing to death are relevant.
145 Airline passenger disruptions

**Description** Identify any disruptions brought about by unruly airline passengers.

**Narrative** Documents that identify any instance where a disruption to normal operation of an aircraft has been brought about by the unruly behavior of a passenger are relevant. Hijacking, or attempts to hijack a plane by a passenger are not relevant.

146 Germ warfare

**Description** Research reports on germ warfare. Including development of germ warfare weapons and the use of germ warfare tactics.

**Narrative** Reports on the use or development of germ or biological weapons are relevant. Reports on the use or development of chemical warfare weapons (i.e. gases) are not relevant. Delivery systems for mass destruction weapons are not relevant.

147 Natural gas vehicles

**Description** What are the pros and cons regarding the use of natural gas vehicles.

**Narrative** Documents that are indicative of the pro’s and con’s of natural gas vehicles are relevant. Only the use of natural gas as applied to vehicles should be considered relevant.

148 North American Free Trade Agreement

**Description** The NAFTA was created in the 90s. What are the current issues?

**Narrative** Documents containing information about current issues that are being considered by NAFTA such as: raising tariffs without approval of the
NAFTA members; breaking agreements on production quotas; and production of items when specifically prohibited by NAFTA agreement would all be relevant. Considerations for new members are not relevant.

149 Aid to handicapped people

**Description** Find documents reflecting actions to aid handicapped people.

**Narrative** Relevant documents clearly demonstrate efforts undertaken to improve conditions for handicapped people. Documents mentioning handicapped people and problems associated with their handicap without actions taken to improve or correct these problems are not relevant.

150 Drive-by shootings

**Description** Research documents on drive-by shootings.

**Narrative** Documents indicating shots fired from a passing car are relevant. Documents about shots fired at a passing car are not relevant.
Appendix B

Subjects in the Semi-automatic User Profile Acquiring Model

The list outlines the subjects identified in the experiments conducted for the Semi-automatic User Profiles Acquiring Model, as discussed in Chapter 4 for the preliminary study and in Chapters 7 and 8 for evaluation experiments. Note that the “-” symbols indicate that the information related to the indicated concepts is discarded. For example, the “crime -convicts -repeat offends” means any information about “convicts” or “repeat offends” is discarded when searching information for “crime”.

101 Economic espionage

Positive Economic espionage; Commercial espionage; Technical espionage; Industrial espionage.

Negative Military espionage; Political espionage.

102 Convicts, repeat offenders

Positive Repeat offenders crime; Ex-convicts crime.
Appendix B. Subjects in the Semi-automatic User Profile Acquiring Model

Negative crime -convicts -repeat offends.

103 Ferry Boat sinkings

Positive Ferry boat sink.

Negative Boat -Ferry -sink.

104 Rescue of kidnapped children

Positive children Rescue kidnapped; children Rescue abducted; children Rescue murder; children Rescue not found.

Negative children abuse -kidnapped -abducted.

105 Sport Utility Vehicles U.S.

Positive Sport Utility Vehicles ownership in the United States; Sport Utility Vehicles consumer groups in the United States.

Negative light trucks -Sport -Utility -United States.

106 Government supported school vouchers

Positive private schools; religious schools; Government voucher; Government funds; Government support; public money.

Negative schools -Government -voucher -funds -public money.

107 Tourism Great Britain

Positive Great Britain Tourism; Scotland Tourism; Wales Tourism; Northern Ireland Tourism.

Negative Tourism -Great Britain -Scotland -Wales -Northern Ireland.

108 Harmful weight-loss drugs

Positive Drugs harmful obesity; drugs harmful weight-loss.
Negative Drugs -obesity -weight-loss -harmful.

109 Child custody cases

Positive Blood relative child custody; parent child custody; grandparent child custody.

Negative Child -custody -blood relative -parent -grandparent.

110 Terrorism Middle East tourism

Positive Middle East tourism and terrorism;

Negative Tourism -Middle East -Terrorism .

111 Telemarketing practices U.S.

Positive The United States telemarketing abusive practice; the United States telemarketing intrusive practice; the United States telemarketing evasive practice; the United States telemarketing deceptive practice; the United States telemarketing fraudulent practice; the United States telemarketing regulation.


112 School bus accidents

Positive School bus accident; location; death.

Negative Accident -bus -school -death -student.

113 Ford foreign ventures

Positive Ford motor foreign venture; Ford motor foreign partnership; Ford motor foreign cooperative alliance.

Negative Ford motor -foreign -venture -partnership -cooperative -alliance.
114 Effects of global warming

**Positive** Global warming effect on climate; greenhouse effect on environment.

**Negative** Climate -global -warming -greenhouse; environment -global -warming -greenhouse.

115 Indian casino laws

**Positive** Casino gamble law in the United States; Indian reservation.

**Negative** Company; riverboat -casino -gamble.

116 Archaeology discoveries

**Positive** New archaeology discovery.

**Negative** Archaeology discovery -new.

117 Organ transplants in the UK

**Positive** Organ transplant; Great Britain; UK; England; United Kingdom.

**Negative** Drug -Organ -transplant -United Kingdom.

118 Progress in treatment of schizophrenia

**Positive** Schizophrenia treatment.

**Negative** Medical treatment -schizophrenia.

119 U.S. gas prices

**Positive** Gas; gasoline; price fluctuation; United States

**Negative** Gas; gasoline -price -fluctuation -United States.

120 Deaths mining accidents

**Positive** Mining accident death.
Negative Mining Death -accident.

121 China Pakistan nuclear missile

Positive China; Pakistan; nuclear missile; military.

Negative Nuclear -missile -military -China -Pakistan.

122 Symptoms Parkinson’s disease

Positive Parkinson disease Symptom.

Negative Disease -symptom -Parkinson.

123 Newspaper circulation decline

Positive Newspaper; circulation; reason; cause.

Negative newspaper circulation -reason -cause.

124 Aborigine health

Positive Aborigine health; Australia.

Negative Aborigine -health -Australia.

125 Scottish Independence

Positive Scottish independence; independent Scottish Parliament.

Negative Scottish -independence.

126 Nuclear plants U.S.

Positive Nuclear plants; United States.

Negative Nuclear plants -United States.

127 U.S. automobile seat belt

Positive Seat belt; automobile; United States; children.
Negative  Automobile -seat belt -United States.

128  Child labor laws

Positive  Child labor; law; work hour; condition.

Negative  Law -labor -child.

129  Problems illegal aliens U.S.

Positive  Illegal aliens; difficulty; problem; United States.

Negative  Aliens -illegal -problem -difficulty -United States.

130  College tuition planning

Positive  College expense; tuition expense; plan; prepare; family; parent.

Negative  College -tuition -expense.

131  Television U.S. children

Positive  Television TV program; children; United States.

Negative  Television TV program -children -United States.

132  Friendly fire deaths

Positive  Death; die; friendly fire; military.

Negative  Death -friendly -fire -military -training; die -friendly -fire -military -training.

133  Anti-rejection transplant drugs

Positive  Anti-rejection -human stem cell; prevent rejection -human stem cell; organ transplant.

Negative  Organ transplant -rejection.
134 Crime Statistics Great Britain

Positive Crime statistics; crime figures; Great Britain; United Kingdom; England.

Negative Crime -figures -Great Britain -United Kingdom.

135 WTO trade debates

Positive WTO; World Trade Organization; trade debates; tariff rates.

Negative WTO -trade -debates -tariff rates.

136 Substance abuse crime

Positive Substance abuse Crime; drugs; alcohol.

Negative crime trafficking -substance -abuse -drugs -alcohol; crime smuggling -substance -abuse -drugs -alcohol.

137 Sea turtle deaths

Positive Sea turtles deaths.

Negative Sea turtles -deaths.

138 Creutzfeldt-Jakob, mad cow disease

Positive disease; Creutzfeldt-Jakob; mad cow; Bovine Spongiform Encephalopathy; humans attributable beef.

Negative disease -Creutzfeldt-Jakob -mad cow -Bovine Spongiform Encephalopathy.

139 Pig organ transplants

Positive Pig; organ transplants; human.

Negative drug -organ -transplants -pig -human.
140 Computer simulation

**Positive** Computer simulation; computer modelling; business; government.

**Negative** Computer business government -simulation -modelling.

141 Environment National Park

**Positive** Environmental problems; United States national Parks; United States national forests.

**Negative** National parks; national forests; accidents -environmental -problems -United States.

142 Illiteracy Arab Africa

**Positive** Illiteracy; rates; Arab; Africa.

**Negative** Arab -illiteracy; Africa -illiteracy.

143 Improving aircraft safety

**Positive** Safety; aircraft manufacturers; airlines; United States.

**Negative** Aircraft -safety -United States.

144 Mountain climbing deaths

**Positive** Deaths; mountain climbing.

**Negative** Deaths -mountain -climbing.

145 Airline passenger disruptions

**Positive** Airline passenger; disruptions; unruly behavior.

**Negative** Airline Hijack -disruptions -unruly behavior.

146 Germ warfare
Positive  Germ weapons; tactics; warfare; biological weapons.

Negative  Chemical weapons -germ -warfare -biological; mass destruction -germ -warfare -biological; gas -germ -warfare -biological. weapons.

147  Natural gas vehicles

Positive  Natural gas; vehicles; pros; cons.


148  North American Free Trade Agreement

Positive  NAFTA; raising tariffs; without approval; breaking agreements; production of items; new members.

Negative  NAFTA -raising tariffs -without approval -breaking agreements -production of items.

149  Aid to handicapped people

Positive  Handicapped people aid; improve conditions.

Negative  Handicapped people -aid -improve conditions. conditions

150  Drive-by shootings

Positive  Drive-by shootings; shots fired from passing; car.

Negative  Shootings; car -drive-by -fired from passing.
Appendix B. Subjects in the Semi-automatic User Profile Acquiring Model
Bibliography


[47] B. Espinasse, S. Fournier, and F. Freitas. Agent and ontology based information gathering on restricted web domains with AGATHE. In *Proceed-


conference on Information and knowledge management, pages 1023–1032, New York, NY, USA, 2008. ACM.


BIBLIOGRAPHY


on Conference on information and knowledge management, pages 525–534, New York, NY, USA, 2007. ACM.


