HAL open science

Accès à de l'information de type patrimoine culturel

Kian Lam Tan

To cite this version:

Kian Lam Tan. Accès à de l'information de type patrimoine culturel. Recherche d'information [cs.IR]. Université de Grenoble; Universiti Sains Malaysia (Malaisie), 2014. Français. NNT : 2014GRENM018 . tel-01548603

HAL Id: tel-01548603

https://theses.hal.science/tel-01548603v1

Submitted on 27 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE GRENOBLE

UNIVERSITI SAINS MALAYSIA

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

préparée dans le cadre d'une cotutelle entre l'*Université de Grenoble* et Universiti Sains Malaysia

Spécialité : **Informatique**

Arrêté ministériel : 7 août 2006

Présentée par

Kian Lam TAN

Thèse dirigée par **Jean-Pierre CHEVALLET** et
codirigée par **Abdullah Zawawi TALIB**

préparée au sein du **Laboratoire d'Informatique de Grenoble**
dans **l'École Doctorale Mathematiques, Sciences et Technologies de l'Information, Informatique (MSTII)**

Accès à de l'informations de type patrimoine culturel

Thèse soutenue publiquement le **« 30 avril 2014 »**,
devant le jury composé de :

M, Ali Selamat
Professor, Universiti Teknologi Malaysia, Malaysia, Président
Mme, Sylvie Calabretto
Professeur, LIRIS-INSA, Lyon, Rapporteur
Mme, Josiane, Mothe
Professeur, IRIT, Toulouse, Rapporteur
M, Cheah Yu-N
Associate Professor, Universiti Sains Malaysia, Malaysia, Examinateur
M, Jean-Pierre Chevallet
MCF-HDR, UPMF, Grenoble, Directeur de these
M, Abdullah Zawawi Talib
Professeur, University Sains Malaysia, Malaysia, Co-directeur de these

Acknowledgements

The completion of my dissertation has been a long journey. It is true that "Life is what happens" when you are completing your dissertation. Life does not stand still, nor wait until you are finished and have time to manage it. Much as happened and changed in the time I have been involved with this study.

In preparing for this research paper, I am truly indebted to my supervisors Dr. Jean-Pierre Chevallet and Dr. Abdullah Zawawi Talib for their guidance and inspiration without which I would not have been able to complete this task. In addition, I would like to express my deepest appreciation and thanks to Dr. Mathieu Guerin from the French Embassy, Malaysia who gave me this opportunity to work together with the researchers from Laboratoire d'Informatique de Grenoble (LIG). In addition, partially of the works are supported by Research University Grant (U0829) and Graduate Assistant Scheme.

Dr. Jean-Pierre Chevallet and Dr. Abdullah Zawawi Talib who read the whole draft and provided me with a host of valuable suggestions and critical comments. I would also like to thank Dr. Philippe Mulhem, Dr. Catherine Berrut and the researchers from the team of Multimedia Information Modeling And Retrieval (MRIM) who have been very kind to give me many valuable suggestions for the improvements of my study.

Furthermore, I would like thanks to Dr. Vivien Petras for providing me the collection on cultural heritage and valuable advice.

I owe a debt of gratitude to my wife (Kim) for her constructive suggestions and invaluable assistance and all those who have assisted me in one way or the other.

Last but not least, I would like to thank to my parents for being so patient and understanding in giving me the time and space to complete this course. My sister (GC) and my youngest brother (Julius) who constantly asked me "are you done yet?". You both kept things light and me smiling. May God bless and keep them.

Abstract

With the explosive growth of digitization, many cultural heritage institutions have been converting physical objects of cultural heritage into digital representation or descriptive representation and to make it publicly available for everyone. The main goal is to construct a nation identity with the use of cultural heritage. However, the conversion resulted in several issues such as: 1) the documents are descriptive in nature, 2) the documents are ambiguity and brevity, 3) the documents used dedicated vocabulary, and 4) the documents contain a lot of term variations. Most of the time, the issues are caused by the aggregated fault in annotating the documents while the short query problem is caused by naive user who has little prior background knowledge and experience in cultural heritage domain. In this research, the main aim is to model information access system to overcome partially the issues arising from the documentation process (conversion from physical to descriptive representation) and the background knowledge of the users of digital cultural heritage. Therefore, three types of information access tool are introduced and established namely information retrieval system, context search (contextual information retrieval system), and mobile game on cultural heritage that allow the user to access, learn, and explore the information on cultural heritage. Basically, the main idea for information retrieval system and context search is to incorporate the link relationship between terms into the Language Model by extending of Dirichlet Smoothing to solve the problems arising from both the documentation process and background of the users. In addition, a Preference Model is introduced based on the Theory of Charging a Capacitor to quantify the cognitive context based on time and integrate into the extended Dirichlet Smoothing. Besides, a mobile game is introduced and integrated the elements of the games of monopoly and treasure hunt to mitigate the problems arising from the background of the users especially their casual behaviour. The first and second approaches were tested on the Cultural Heritage in CLEF (CHiC) collection that consists of short queries and documents. The results show that the approach is effective and yields better accuracy. Finally, a survey was carried out to investigate the third approach, and the result suggests that the game is able to help the participants to explore and learn the information on cultural heritage. In addition, the participants also felt that an information seeking tool that is integrated with the game can provide more information to the user in a more convenient manner while playing the game and visiting the heritage sites in the game. In conclusion, the results show that the proposed solutions are able to solve the problems arising from the documentation process and the background knowledge of the users of digital cultural heritage.

Keywords. Information retrieval, Language model, Term similarity, Cognitive context, Mobile game

Contents

Contents iii

List of Figures vii

List of Tables xi

I INTRODUCTION 1

1 Introduction 2
1.1 Motivation 5
1.2 Problem Statements 5
1.2.1 Collection 5
1.2.2 Users 8
1.3 Objectives 9
1.4 Scope and Limitation 10
1.5 Contributions 11
1.6 Research Questions 11
1.7 Thesis Outline 11

II ANALYSIS, BACKGROUND AND PROCEDURES 13

2 Analysis, Background and Procedures 14
2.1 Overview of the Procedures 14
2.2 Cultural Heritage 15
2.3 Process in Cultural Heritage 15
2.4 Information Access 19
2.4.1 Information Access based on Free-text Query 20
2.4.1.1 Basic Concepts in Information Retrieval 20
2.4.1.2 Information Retrieval in Cultural Heritage 23
2.4.1.3 Information Retrieval with Context 25
2.4.1.4 Discussion 27
2.4.2 Information Access based on Application 28
2.4.2.1 Discussion 30

2.5 Evaluation . . . 31
2.5.1 Laboratory-based Evaluation . . . 31
2.5.2 Questionnaires . . . 33
2.5.2.1 Discussion . . . 34
2.6 Discussion . . . 34
2.7 Summary . . . 35

III STATE OF THE ART 36

3 State of the Art 37
3.1 Information Access - Information Retrieval System . . . 37
3.1.1 Relevance Feedback . . . 37
3.1.2 Query Expansion . . . 38
3.1.3 Query Reformulation . . . 38
3.1.4 Dimension Reduction . . . 38
3.1.5 Statistical Translation Model . . . 39
3.1.6 Exploiting the Term Similarity Matrix in Vector Space Model . . . 40
3.1.7 Discussion . . . 41
3.2 Information Access - Context Search . . . 42
3.2.1 Gathering and Representing Interest . . . 42
3.2.2 Gathering User Profile through Explicit Relevance Feedback . . . 43
3.2.3 Gathering User Profile through Implicit Relevance Feedback . . . 43
3.2.4 User Profile . . . 44
3.2.5 Discussion . . . 46
3.3 Information Access - Mobile Game . . . 47
3.3.1 Mobile Learning . . . 47
3.3.2 Game as Learning . . . 49
3.3.3 Theoretical Foundations . . . 52
3.3.4 Inquiry Learning . . . 53
3.3.5 Experiential Learning . . . 54
3.3.6 Situated Learning . . . 55
3.3.7 Discussion . . . 56
3.4 Summary . . . 56

IV PROPOSED SOLUTIONS 57

4 Proposed Solutions 58
4.1 Information Access - Information Retrieval System . . . 58
4.1.1 Extracting Term Links . . . 61
4.1.2 Extended Dirichlet Smoothing . . . 65
4.1.3 Extended Dirichlet Smoothing with Increase of Document Size (EDS+) 68
4.1.4 Extended Dirichlet Smoothing with Decrease of Document Size (EDS-) 71
4.1.5 Discussion . . . 72

4.2 Information Access - Context Search 72
4.2.1 Preference Model 72
4.2.2 Theory of Charging a Capacitor 79
4.2.3 Charging a Preference Model (CP) 80
4.2.4 A Statistical Box Plot 83
4.2.5 Discussion 84
4.3 Information Access - Mobile Game 85
4.3.1 Overall Framework of M-Heritage Hunt 86
4.3.2 Discussion 92
4.4 Summary 94

V EVALUATIONS, RESULTS AND DISCUSSION 95

5 Evaluations, Results and Discussion 96
5.1 Data Collection 96
5.2 Environment for the Mobile Game 97
5.3 Evaluation of Information Access - Information Retrieval System 99
5.3.1 Experiment Setup 99
5.3.2 Results based on EDS+ 103
5.3.3 Results based on EDS- 103
5.3.4 Comparison between EDS+ with EDS- 106
5.3.5 Comparison between Dice Coefficient with Cosine Similarity and Tanimoto Similarity with Baseline based on ESD+ 110
5.3.6 Discussion based on the result from EDS+ 111
5.3.7 Comparison between Pseudo-Relevance Feedback with EDS+ 113
5.3.8 Comparison between Query Expansion with EDS+ 114
5.4 Evaluation of Information Access - Context Search 116
5.4.1 Experiment Setup 117
5.4.2 Result and Discussion 118
5.5 Evaluation of Information Access - Mobile Game 119
5.6 Summary 121

VI CONCLUSION AND FUTURE WORDS 122

6 Conclusion and Future Work 123
6.1 Conclusion 123
6.2 Future Work 125
6.2.1 Information Retrieval System 125
6.2.2 Context Search 126
6.2.3 Mobile Game 126

VII APPENDIXES 128

A A Statistical Box Plot 129

B Analysis in CHiC 2012 133
B.1 Analysis for Query 7 “schlesien maps” 133
B.2 Analysis for Query 25 “syriac” 133

C Three Distinct Phases that Exploit the User Profile 137

D Questionnaire 139

E Simulation Run 142

F Resume etendu en francais 152

G List of Publications 154

Bibliography 156

List of Figures

1.1 Example of the Current State-of-the-art system: The Website of Europeana . . 3
1.2 Example of the Current State-of-the-art system: The Website of Digital Public Library of America 4
1.3 Types of Document in Cultural Heritage Institutions 6
1.4 Physical Representation - Manuscript (Deposition of Henry Jones) from 1641 Depositions 7
1.5 Digital Representation - Manuscript (Deposition of Henry Jones) from 1641 Depositions 7
1.6 Descriptive Representation - Manuscript (Deposition of Henry Jones) from 1641 Depositions 8
1.7 Information Gap between the user and the Collection 9
1.8 Heritage Zone of George Town, Penang in Malaysia 10

2.1 Overall Procedures for Research on Information Access in Cultural Heritage . . 15
2.2 Overview of Cultural Heritage 16
2.3 Digital Representation - Steam Traction Engine 19
2.4 Evolution of Information Access in Cultural Heritage 21
2.5 The Overall Framework for Pseudo Relevance Feedback with Query Expansion 24
2.6 Search Result for query (all persons mentioned together with "phelim o'neil") from EoS [Carmel *et al.*, 2012] 26
2.7 Overview over an Example $path$ on the topic of "Railways" and the Ability to Structure the Information by Creating Multiple Branches [Clough *et al.*, 2011] . 27
2.8 AmbieSense: Screenshots of Content in Mobile Device which Ayuntamiento (City Council, Seville) is the Main Tourist Attraction near Plaza de San Francisco [Göker & Myrhaug, 2008] 28
2.9 iCity: Screenshot of Content in PC which Lists Out the Cultural Events [Carmagnola *et al.*, 2008] 29
2.10 A Venn Diagram to Illustrate Recall and Precision 31
2.11 The Flow of this Research 35

3.1 A Weighted Keywords User Profile 45
3.2 A User Profile based on Semantic Networks 46
3.3 Screenshot of *Can You See Me Now?*: Display in Mobile Device [Benford *et al.*, 2006] 48

3.4 Screenshot of *Can You See Me Now?*: The Location of all the Players [Benford *et al.*, 2006] . . . 49
3.5 Screenshot of *Can You See Me Now?*: Emily (player) inside the Virtual World [Benford *et al.*, 2006] . . . 50
3.6 Screenshots of *MyArtSpace*: (a) Initial Screen, (b) Main Menu, (c and d) Interfaces of Collecting an Exhibit [Vavoula *et al.*, 2009] . . . 50
3.7 Screenshot of *MyArtSpace*: An Example of Student Gallery [Vavoula *et al.*, 2009] 51
3.8 Children Taking Moisture Measurements around the Wood. The Physical Position of the Child Taking Measurements Indicates to the Team-mates that the Activity is Taking Place [Cole & Stanton, 2003] . . . 51
3.9 Children are Exploring the Physical Environment by Searching for the Objects through Mobile Device [Cole & Stanton, 2003] . . . 52
3.10 Screenshots of *Quest Atlantis*: Showing a Scene from a Village on the Left, the Homepage for a Student on the Right, and the Synchronous Chat Space at the Bottom [Barab *et al.*, 2005] . . . 53
3.11 Screenshot of *MadCityMystery*: Map Interface, Environmental Detectives [Squire & Jan, 2007] . . . 54
3.12 Screenshots of *MadCityMystery*: Map and Information [Squire & Jan, 2007] . . . 55

4.1 System Architecture to Exploit Non-matching Terms during Retrieval . . . 60
4.2 The terms "Last Supper" in Library of Congress Subject Headings . . . 63
4.3 The terms "Last Supper" in WordNet . . . 64
4.4 The terms "Last Supper" in Wikipedia . . . 64
4.5 The terms "Last Supper" in ConceptNet5 . . . 65
4.6 Graphical Interpretation of Extended Dirichlet Smoothing . . . 66
4.7 Graphical Interpretation of the Document Size: (a) Increase and (b) Decrease . 66
4.8 Evolution from Maximum Likelihood Estimator to Extended Dirichlet Smoothing 67
4.9 Query Expansion . . . 67
4.10 Document Expansion . . . 68
4.11 Proposed Framework . . . 68
4.12 Global Change . . . 69
4.13 Local Change . . . 69
4.14 A function f with domain denoted as DI and codomain denoted as $Duration$. 74
4.15 A function *Pref* with domain denoted as $Duration$ and codomain denoted as *Preference* . . . 75
4.16 Exponential Function . . . 76
4.17 Different μ in the Exponential Function . . . 77
4.18 Bar Chart: With Different Preferences . . . 79
4.19 Scatter Plot: With Different Preferences . . . 79
4.20 Capacitor in the Circuit . . . 80
4.21 Capacitor Charging Graph . . . 81
4.22 Mapping between Capacitor with Preference Model . . . 82
4.23 Multiple Preferences by Using Analogy from Capacitor . . . 82
4.24 Box Plot Diagram according to Table 4.2, Table 4.3 and Table 4.4 . . . 84
4.25 Overview of M-Heritage . . . 86

4.26 Gallery Module (Situated Learning) 87
4.27 Game Play Interface (Experiential Learning) 88
4.28 Sample of Treasure Hunt Question (Inquiry Learning) 89
4.29 Graphical Interpretation of the Movable Cells 89
4.30 Interface for Monopoly 90
4.31 The Value of the Dice 90
4.32 Treasure Hunt 92
4.33 Panoramic View: Goddess of Mercy Temple without Scrolling 93
4.34 Panoramic View: City Hall without Scrolling 93
4.35 A 360-degree Panoramic View with Highlighted Display on Figure 4.34 94
4.36 Building Information in Table Views 94

5.1 Example of Data Collection - One of the Documents 98
5.2 Example of Data Collection - One of the Queries 98
5.3 Example of the Environment - Fort Cornwallis 99
5.4 Example of the Environment - Goddess of Mercy Temple 99
5.5 Steps of Pre-processing 100
5.6 Optimal Value for Dirichlet Prior Smoothing is 200 101
5.7 Optimal Value for Extended Dirichlet Smoothing is 350 101
5.8 Comparison between the result of baseline and Dice Coefficient based on EDS+ 104
5.9 Comparison of Baseline with LMED-D by using Averaged Eleven-point Precision/Recall Curve based on EDS+ 105
5.10 Comparison between the results of the baseline and Dice Coefficient based on EDS- 106
5.11 Comparison of Baseline with LMED-D by using Averaged Eleven-point Precision/Recall Curve based on EDS- 107
5.12 Comparison between EDS+ and EDS- based on the approach of LMED-D . . . 109
5.13 Comparison between EDS+ and EDS- based on the Approach of LMED-D by using Average Eleven-point Precision/Recall Curve 110
5.14 Top 10 Similar Values between CC, CS, DC, OC and TS of the Term “Syriac”. 111
5.15 Further Highlights of the Top 10 Similar Values between TS and DC of the Term “Syriac”. 112
5.16 Comparison of LMED-D based on EDS+ with QE-WordNet, QE-Wikipedia, and Pseudo-relevance Feedback by using Average Eleven-point Precision/Recall Curve 115
5.17 Comparison between the results of LMED-D based on EDS+ with QE-WordNet, QE-Wikipedia, and Pseudo-relevance Feedback 116

6.1 Integration between Information Retrieval System, Context Search, and Mobile Game 124
6.2 Multiple Users from different Domains to Access the Information on Cultural Heritage 125

A.1 A Statistical Box Plot 129
A.2 Landscape of A Statistical Box Plot 131

A.3 Outliers and Extremes of a Statistical Box Plot 132

C.1 Context Search: User Profile Occurs during the Process of Retrieval 137
C.2 Context Search: Distinct Re-ranking Activity 138
C.3 Context Search: Pre-processing of the User Query 138

E.1 Screenshots of "How to Start" 143
E.2 Further Highlight of Figure E.1(f) 144
E.3 Information for Each Player 144
E.4 Screenshots of Starting a "New Game" 146
E.5 Screenshots of "Resume Game" 147
E.6 A Sample Screenshot of the "Gallery" 148
E.7 Screenshots of "Help" 149
E.8 Continued from Figure E.7 150
E.9 Screenshots of "Upgrade Property" 151

List of Tables

2.1 Summary of Features of Several Applications 30

4.1 Mean Average Precision by different Information Retrieval Models in CHiC Collection 59
4.2 Pref for item "Series1" 78
4.3 Pref for item "Series2" 78
4.4 Pref for item "Series3" 78
4.5 The Sample Output based on *Pref* from Table 4.2, Table 4.3 and Table 4.4 . . . 85

5.1 Number of Documents in CHIC Collection by Languages 97
5.2 Sample Term Using Dice Coefficient and Cosine Similarity. Note that Words are Stemmed and q Refers to Term from Query. 102
5.3 Performance of Generating the Term Similarity Matrix 102
5.4 The Specification of the Server to Generate the Term Similarity Matrix 102
5.5 Performance with Various Types of Statistics from the First Paragraph of the Articles from Wikipedia based on EDS+ model (* = Statistical Significance at $p < 0.06$ Using the Student's Paired T-Test). 104
5.6 Results on Average Precision at Precison-10 based on EDS+. (*=Statistical Significance at p<0.01 and ** at p<0.03 Using the Student's Paired T-Test) . . 105
5.7 Performance with Various Types of Statistic from the First Paragraph of the Articles from Wikipedia based on EDS- model (* = Statistical Significance at $p < 0.05$ Using the Student's Paired T-Test). 106
5.8 Results on Average Precision at Precison-10 based on EDS-. (*=Statistical Significance at p<0.01 and ** at p<0.03 Using the Student's Paired T-Test) 107
5.9 Comparison between EDS+ and EDS- with Various Types of Statistics from the First Paragraph of the Articles from Wikipedia 108
5.10 Comparison between EDS+ with EDS- and Various Type of Statistic at Precision-10 109
5.11 Comparison between Baseline, EDS+ with EDS- in terms of MAP and Precision-10 in CLEFMed2011 109
5.12 Average Precision (AP) and Total Number of Relevant Documents Retrieved (Rel) for BL, CC, CS, TS, DC and OC of the Query (CHIC-032 where the Title is "syriac"). The total number of the Relevant Documents is 80 for this Query. . 111
5.13 Different Terms Provided by Using CC, CS, TS, DC and OC by Term "syriac". Note that Words are Stemmed. 113

5.14 Correlation among CC, CS, DC, OC and TS. 113
5.15 Mean Average Precision by using Pseudo-Relevance Feedback($*$ represents the highest value while ˆ represents the lowest value) 114
5.16 Results on Mean Average Precision Compared to Query Expansion method . . 115
5.17 Number of Documents according to Category 117
5.18 Number of Documents according to CHIC-032 (syriac) 117
5.19 Results on Mean Average Precision and R-Precision Compared to Various Types of Methods for Context Search 119
5.20 Results on R-Precision compared to various types of Approach for Context Search 119
5.21 Result of the Evaluation of Mobile Game 121

B.1 Query 7 - “schlesien maps” 133

B.2 One of the Relevant Document for Query 7 134
B.3 One of the Relevant Document for Query 7 which Contains the Term of “silesia” from Term Similarity Matrix 135
B.4 Query 25 - “syriac” 135
B.5 41 of the Total Relevant Documents are Using the Same Terms inside the Title and Description for Query 25 (“syriac”) 136

E.1 Description of Games Icon for Figure E.3 145
E.2 Description from Figure E.3 145

Part I

INTRODUCTION

Chapter 1

Introduction

Over the last 30 years, United Nations Educational, Scientific and Cultural Organization (UNESCO) have been encouraging all the nations around the world to identify, protect and preserve their own cultural and natural heritage because they are considered to have outstanding value to humanity and can provide a concrete basis for ideas from the past. In addition, UNESCO also compiled a World Heritage List which includes 981 properties that are considered as outstanding universal value in the world. Furthermore, several projects have been successfully restored by UNESCO such as Angkor Watt in Cambodia and the Old City of Dubrovnik in Croatia and they have successfully removed the sites from the List of World Heritage in Danger[1]. In general, cultural heritage is a significant treasure that holds the nation and the humanity of the past together. It gives many clues about the history and also clears up the future. Sometimes, this treasure can be damaged due to natural disasters such as earthquake, fire, flood and sometimes it can be locked away in the storerooms.

Digitization is a good opportunity to overcome these problems because digitization converts a physical document into machine-readable form and it can help to preserve and store these information. With the encouragement by UNESCO and the rise of new communication technologies, institutions and museums around the world have put up a lot of efforts to translate and describe the cultural heritage objects in order to preserve, store and organize these information digitally. Through this process, cultural heritage objects can be made available and be used easily by certain group of users such as historians and curators.

In recent years, institutions and museums has started to convert all the information on cultural heritage into digital format and push these information to online in order to attract more users. So, it becomes easier for people around the world such as students, educators, genealogists, and some users who are interested in the local history and heritage to access these types of information. Several examples of online information access in cultural heritage include the Louvre from France[2], the British Museum from the United Kingdom[3], Rijksmuseum from the Netherlands[4], and Museo del Prado from Spain[5]. In order to ease the job of looking for informa-

[1] http://whc.unesco.org/en/107/
[2] http://www.louvre.fr
[3] http://www.britishmuseum.org
[4] https://www.rijksmuseum.nl
[5] http://www.museodelprado.es

tion on cultural heritage from one website to another website, Europeana[1] as shown in Figure 1.1 has taken the initiative by collecting the information which is in digital representation to represent the physical objects across many European Union (EU) member states. Besides, Digital Public Library of America[2] as shown in Figure 1.2 also serves the same objective like Europeana and aims to bring together all the collections from the institutions and museums across the United States of America into a large-scale public digital library.

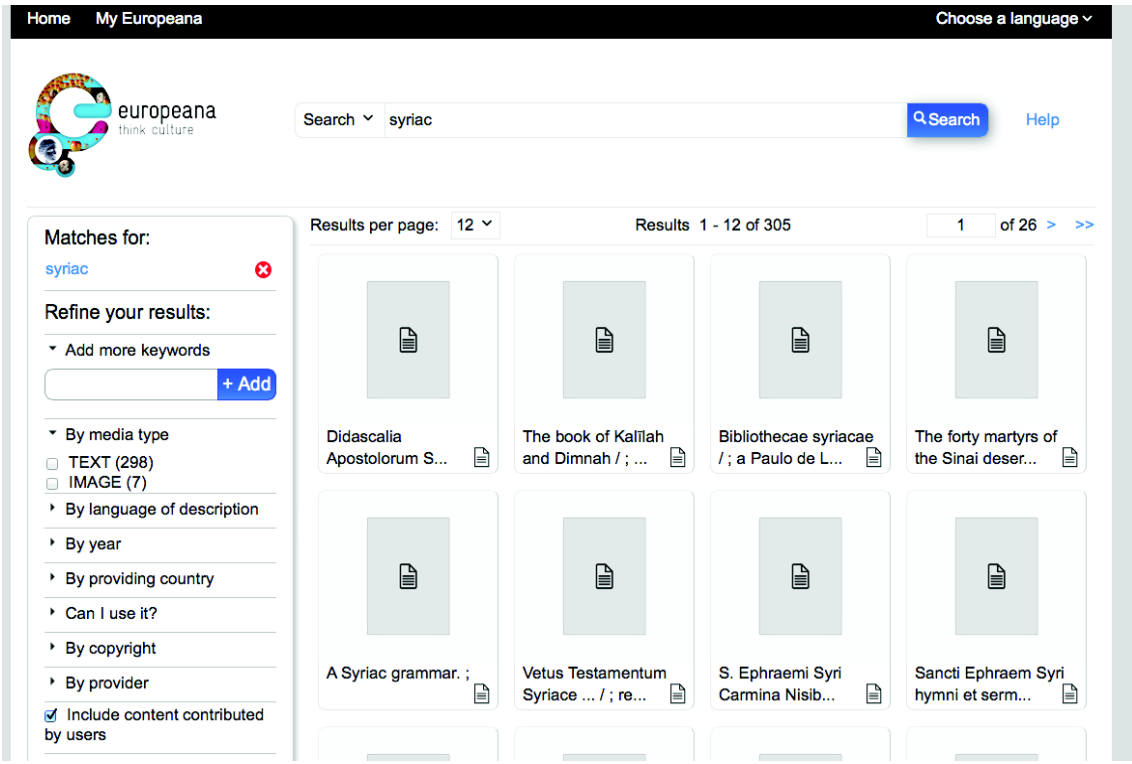


Figure 1.1: Example of the Current State-of-the-art system: The Website of Europeana

With mobile devices becoming the new Personal Computer (PC), people are increasingly using smart phones and tablets instead of desktop computers to access products and services. In the last quarter of 2010, smart phones sold more than PCs for the first time and the growth for tablets, such as Apple iPad, Motorola Xoom and Samsung is expected to expand from 17 million in 2010 to 270 million in 2014 [3]. Furthermore, smart phones are expected to climb from 1,700,000 million in 2012 to 1,900,000 million in 2014 with an increase of 11%. In addition, International Data Corporation (IDC) forecasted that the number of downloaded mobile applications is to grow to 76.9 billion in 2014 which represent $35 billion in revenue. The main reason that mobile devices are getting popular is because the users can use the mobile device to access any information anywhere and anytime. However, the experience in accessing information by using mobile devices is different from the experience in accessing the web from a PC or laptop and the main factor is the size of the screen. Several institutions and museums

[1] http://www.europeana.eu
[2] http://www.dp.la
[3] http://www.gartner.com/newsroom/id/2525515

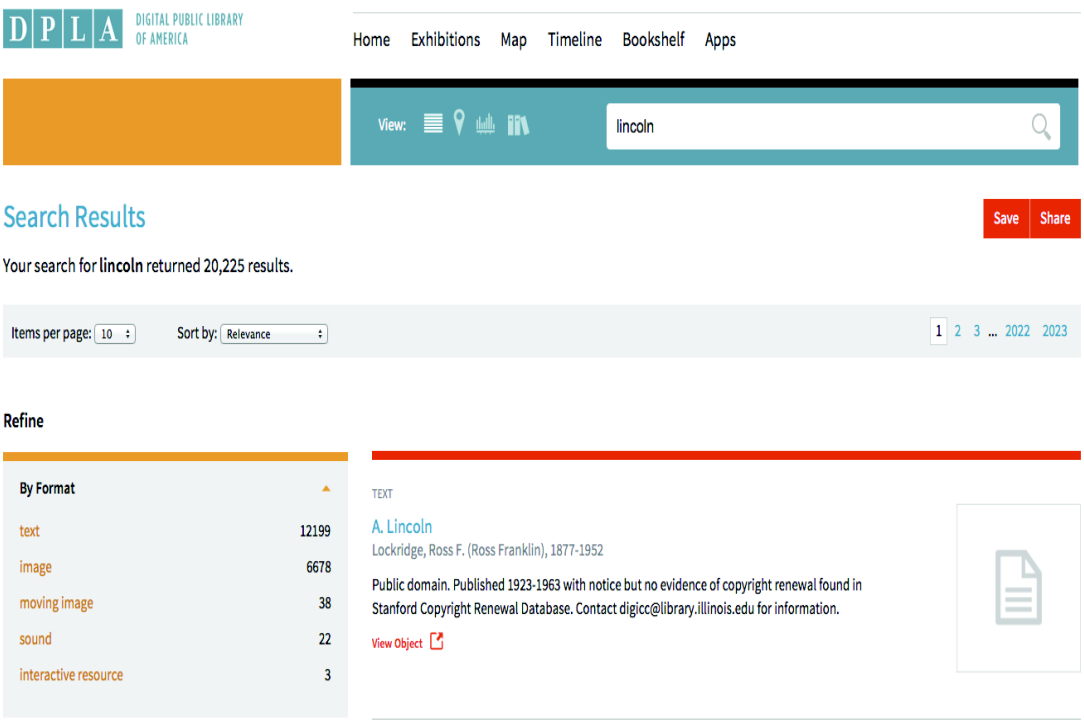


Figure 1.2: Example of the Current State-of-the-art system: The Website of Digital Public Library of America

also provide "lite" interface for users with mobile devices based on this issue. Moreover, Cui & Roto [2008] have conducted a case study that concluded that three types of information-seeking behaviours on mobile devices can be found which are: 1) fact finding (seeking specific information, 2) casual browsing (accessing general information, and 3) gathering information from multiple sources to achieve the goal (making a decision or to collect knowledge around a topic).

There are a variety of studies focusing on information access. However in seeking to cater for the accessing needs, game-based information access have been popular to engage accessing, learning, and exploring while still enjoying the satisfaction as a learner. There are some game-based mobile accessing that integrates the elements of playing and learning as found in Savill-Smith [2005], Shiratuddin & Zaibon [2010]. In these mobile games, the problems of fundamental science is embedded into the game, while the mobile technology makes it possible to simulate the problem in the natural context without losing the motivation of the games as proposed by Schwabe [2005]. In addition, game-based mobile accessing can promote the skills of decision-making and help the young adults in lifelong learning [Carr, 2010], [Song *et al.*, 2004], [Prayaga, 2005], [Blanco *et al.*, 2010], [Berland *et al.*, 2010], and [Harpstead *et al.*, 2013]

In a nutshell, there is a need to provide an information access tool to help the user to access the information on cultural heritage by using mobile devices. Even though the free-text retrieval has received an enormous success, information access in the domain of cultural heritage is still very challenging and requires a lot of research efforts in this domain. The scope of this research intends to provide different types of information access tool to the user which are able to help

them to access, learn, and explore the information on cultural heritage.

1.1 Motivation

During the last decade, there has been a growing interest in building large digital libraries with the large projects receiving national (such as Gallica) or international support (The European Library, or Europeana). The main motivation behind such projects is the preservation of the cultural heritage and allow a worldwide users to access to this valuable material. With the growing capabilities and the failing prices of current hardware system, there are ever increasing possibilities to access the information on cultural heritage. Also with ever increasing broadband capabilities, it is now possible to view the information on cultural heritage online.

The main motivation of this research comes from a case study involving a mobile game on cultural heritage. Basically, the intention of the mobile game is to evaluate the effectiveness of delivering information on cultural heritage to users (uncertain on their information needs) through mobile game. However, the majority of the respondents highlighted that the information and knowledge acquired in the mobile game is just enough for them to learn and use to explore the heritage sites. Therefore, these is a need for better and effective tool such as Information Retrieval System that can be integrated with the mobile game to provide instant access when the user is playing the game (please bear in mind that the questionnaire for this case study was brief and it was not intended for empirical study). So, this is the main motivation of this research since the Information Retrieval System would allow the user to access, learn, and explore the information on cultural heritage especially in mobile game environment.

1.2 Problem Statements

Basically, the problem statements can be split into two parts. The first part is related to the collection of cultural heritage and the second part is related to the information seeking behaviours by the users in particular in the domain of cultural heritage.

1.2.1 Collection

The cultural heritage collection is a special collection because the physical objects of cultural heritage can contain different resources such as books, manuscript, pamphlets, paintings, sculptures, clothing, jewellery, and coins. Each of these objects has their own characteristics such as the books contain the authors and the number of pages, but coin and jewellery do not contain these type of information. Therefore, there is no universal set of characteristics to capture all the different aspects of different objects.

In addition, each physical object can be represented by three types of representation namely: 1) physical representation, 2) digital representation and 3) descriptive representation. Each representation can be considered as a document because all of the representations fulfil the features of a document as proposed by Pédauque [2003] and each physical representation can be associated with digital representation or descriptive representation. However, these representations

can become the main obstacle in blocking the users to access the information on cultural heritage because the Information Retrieval System indexes the descriptive representation rather than the physical representation while the user is looking for the physical representation. In other words, it creates a gap between the physical representation and the descriptive representation.

Lets use manuscript as an example to describe the physical representation, digital representation and descriptive representation as shown in Figure 1.3. Physical representation refers to the objects from cultural heritage such as antiquities, applied arts (sculpture, porcelain, glass, armour, weapon, ceramic), coins, manuscripts, and painting. So, the physical representation of a manuscript is itself a manuscript which includes the whole contents inside the manuscript and Figure 1.4 shows the example of the manuscript from 1641 Depositions [1]. In addition, Portier *et al.* [2012] considered two main structures on the manuscript that may overlap namely 1) book structure (a sequence of columns and lines), and 2) syntactic structure (a sequence of phrase and words) in ubiquitous case.

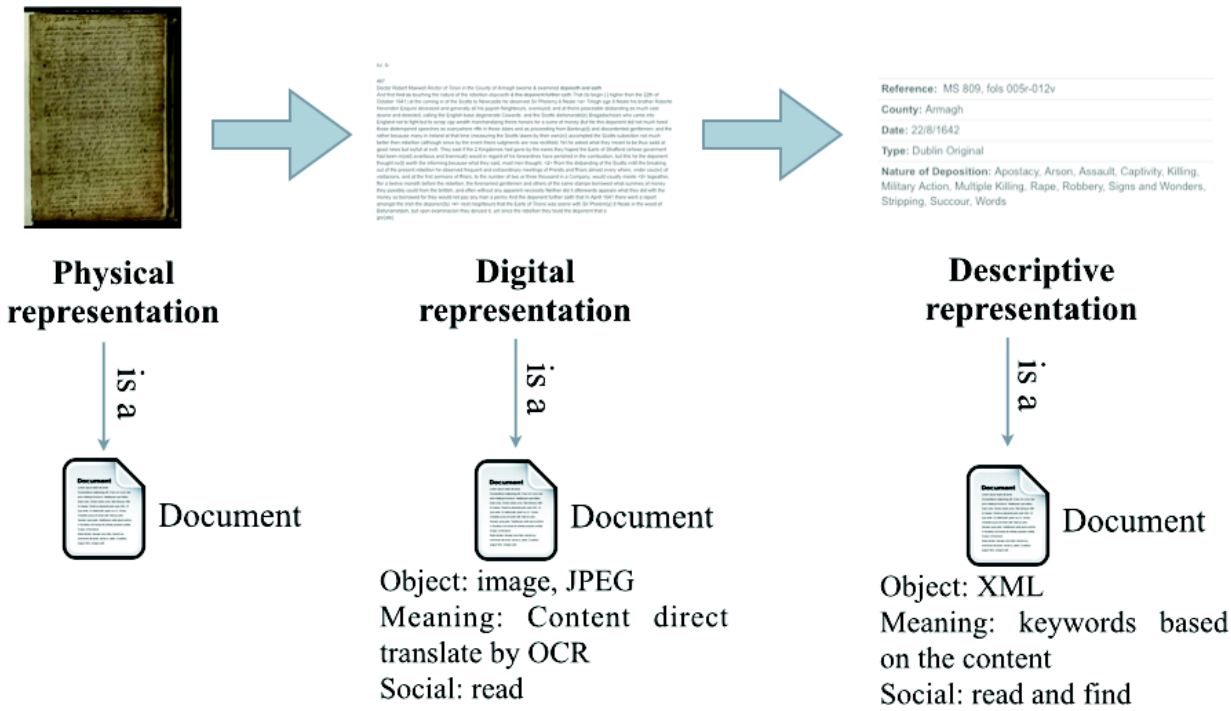


Figure 1.3: Types of Document in Cultural Heritage Institutions

Digital representation refers to the digital object that can be represented by the physical object. The digital representation can be scans, photographs, videos, audio recordings or any types of digital file that represents the physical objects. So, the digital representation of a manuscript is the digital file created by scanning the manuscript using an Object Recognition Tool or Optical Character Recognition (OCR) as shown in Figure 1.5. Furthermore, the digital

[1] http://1641.tcd.ie/deposition.php?depID=809001r001

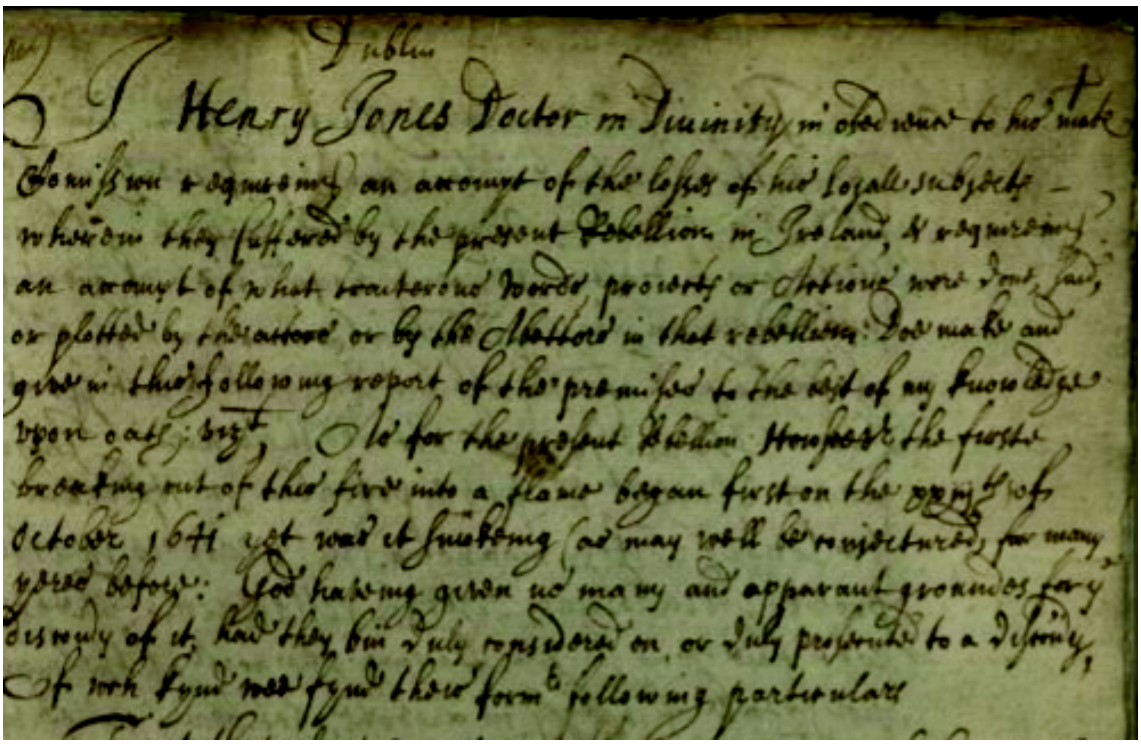


Figure 1.4: Physical Representation - Manuscript (Deposition of Henry Jones) from 1641 Depositions

representation has a deep impact on the preservation and several factors can become a hindrance in optimizing the management of the resources as discussed in [Ogier, 2006]. The content inside the digital representation may contain some errors due to spelling, old language [Hampson *et al.*, 2012] and imperfection of the Object Recognition Tool [Naji & Savoy, 2011].

Dublin
<Exw
137> I Henry Jones Doctor in Diuinity in obedience to his maiesties Comission requireing an accompt of the losses of his lojall subjects wherein they suffered by the present Rebellion in Ireland, & requireing an accompt of what traiterous words, proiects or Actions were done, said, or plotted by the actors, or by the Abettors in that rebellion: Doe make and give in this following report of the premises to the best of my knowledge vpon oath: vizt, As for the present Rebellion Howsoever the firste breaking out of this fire into a flame began first on the xxiijth of October 1641 yet was it smokeing (as may well be conjectured, for many yeres before: God haveing given us many and apparant groundes for the discouery of it, had they bin duly considered on, or duly prosecuted to a discouery, Of which kynd wee fynd theis former following particulars

Figure 1.5: Digital Representation - Manuscript (Deposition of Henry Jones) from 1641 Depositions

Descriptive representation refers to the descriptive information in the form of digital representation or physical representation. So, the descriptive representation of a manuscript is the description of the manuscript and not the whole content of the manuscript (also known as meta data). Basically, the description of the manuscript is indexed manually by the annotator. The main challenge in descriptive representation is that the description contains a lot of term variations such as name aliases, multiple languages, multiple terms and time dimension (refer [Accary-Barbier & Calabretto, 2008] for further details) as described below:

Reference: MS 809, fols 001r-004v
County: Unknown
Date: 3/3/1642
Type: Waring Copy
Nature of Deposition: Military Action, Words

Figure 1.6: Descriptive Representation - Manuscript (Deposition of Henry Jones) from 1641 Depositions

- *Name aliases* - The terms used to describe the object may not always specify the name correctly. For example, the name of the artist may have different name aliases such as ”Raffaello” and ”Raphael” which refer to the same artist who was an Italian painter and architect of the High Renaissance.
- *Multiple languages* - Objects may have multilingual annotations. The objects may come from all the parts of the world and it can be annotated in vernacular terms or other languages such as French, English, German and Dutch. For example, the annotator from United Kingdom will use English to describe the term ”silesia” while the annotator from Germany will use German to describe the term ”schlesien” and both terms refer to the same object.
- *Multiple terms* - There are many potential terms that can be used to annotate objects. For example, different terms such ”Islam”, ”Islamic”, ”Moslem” and ”Muslim” can be used to describe the Islamic collection.
- *Time dimension* - The terms used to describe the object can contain simple dates and interval dates. Simple dates refer to exact year such as 1945 while interval dates refers to an event such as from 1900 until 1905. In addition, it is usual to have dates like beginning 1800 or mid 19th century in the documents of cultural heritage.

In a nutshell, the problem statements in the collection can be summarized as follows:

- The documents are the descriptive representation of cultural heritage.
- The information on the documents are ambiguous and short.
- The vocabulary used in the documents are usually dedicated to cultural heritage.
- Term variation is caused by name aliases, multiple languages, multiple terms and time dimension.

1.2.2 Users

Several log studies have been investigated in Information Retrieval System and queries [Nicholas *et al.*, 2013]; [Church *et al.*, 2008]; [Church & Smyth, 2009]; [Yi *et al.*, 2008]; [Kamvar *et al.*, 2009]. The most relevant findings from the log studies (cultural heritage) is that the mean term count for the queries is between 2.1 and 2.7, and Koolen & Kamps [2010] found out that non-experts have more complex information needs. Therefore, the problems from the perspective of users can be summarized as follows:

- Short query.
- The information gap between the user and the collection makes it difficult for the users to form a query as shown in Figure 1.7.
- Casual behaviour

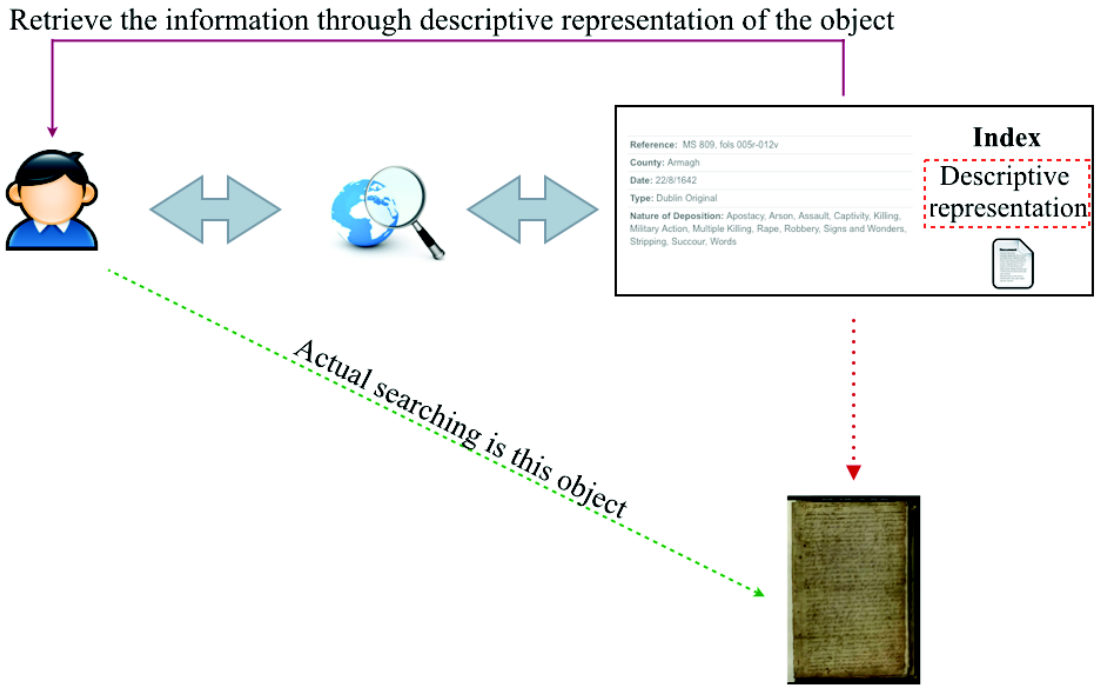


Figure 1.7: Information Gap between the user and the Collection

1.3 Objectives

The main goal of this research is to model information access system to overcome partially the issues arising from the documentation process (conversion from physical to descriptive representation) and the background knowledge of the users of digital cultural heritage. Therefore, the objectives of this research are as follows:

- To model an information access system or more specifically an Information Retrieval System to allow the user to access, learn, and explore the information on cultural heritage.
- To model and integrate the user's cognitive context into the Information Retrieval System.
- To design and model a mobile game that is able to attract users to access, learn, and explore the information of cultural heritage and investigate the need for an information access tool in such a game.

1.4 Scope and Limitation

The scope and limitation of this research are as follows:

- Two types of data collection that have been used in this research. The first type of the data is from Cultural Heritage in Clef (CHiC) which will be used in the first and second objectives, and the second is from George Town, Penang, Malaysia which will be used in the third objective.
- The main reason to use the data from CHiC is because the collection of CHiC overlaps 80% with the current data in Europeana (one of the largest digital library in the domain of cultural heritage in Europe).
- A mobile game is developed to study and investigate the need for instant access of information on cultural heritage. The background of the mobile game was customized for the core heritage zone (green colour) of George Town, Penang in Malaysia as shown in Figure 1.8[1].

Figure 1.8: Heritage Zone of George Town, Penang in Malaysia

- The context search uses the analogy from Physics to quantify the level of preference (cognitive context) of a user. However, only the integration between level of preference of a user with the proposed framework is evaluated since the level of preference is qualitative.

[1] http://www.visitpenang.gov.my/UnescoHeritageBrochure.pdf

1.5 Contributions

The contributions of this research can be summarized as follows:

- **An Information Retrieval System by using external knowledge to improve the effectiveness of the system:** A framework based on Dirichlet Smoothing to exploit the non-matching terms during the retrieval time by using external knowledge. Two models will be introduced based on the previous framework which are Extended Dirichlet Smoothing with Increase of Document Size (EDS+), and Extended Dirichlet Smoothing with Decrease of Document Size (EDS-).
- **A method to quantify the cognitive context and integrate into the Information Retrieval System:** A Preference Model based on the Theory of Charging a Capacitor that is able to quantify the cognitive context and integrate into the model of EDS+.
- **An environment that integrates education and entertainment that allows the user to access, learn, and explore the information on cultural heritage:** A Mobile Game based on the Theory of Social Constructivism that is able to support inquiry learning, experiential learning, and situated learning.

1.6 Research Questions

The research questions of this research can be summarized as follows:

- How can knowledge or context help to improve the retrieval effectiveness in Information Retrieval System?
- How can matching adaptation (local change) perform better than query modification (global change)?
- Is instant access to information in a mobile game able to help the user to access, learn, and explore the information on cultural heritage?

1.7 Thesis Outline

The remainder of this thesis is structured as follows:

- Chapter 2 analyses the process in the domain of cultural heritage and the overall procedures that involved in this research. In addition, Chapter 2 also presents the basic concepts of Information Retrieval.
- Chapter 3 surveys the literature review in the fields of Information Retrieval System, context search, and mobile game.
- Chapter 4 presents the proposed solutions for Information Retrieval System, and context search, and the design of the mobile game.
- Chapter 5 evaluates the proposed solution for Information Retrieval System and context search. In addition, a case study have been investigated to show that mobile games is able to support learning activities on cultural heritage and there is a need for an information access tool for such a game.

- Chapter 6 summarizes the outcomes of this research, discusses limitations, and suggests future directions.

Part II

ANALYSIS, BACKGROUND AND PROCEDURES

Chapter 2

Analysis, Background and Procedures

This chapter discusses the analysis in the domain of cultural heritage and procedures involved in this research. Section 2.1 presents the overview of the procedures of this research while Section 2.2, Section 2.3, Section 2.4, and Section 2.5 describe in more details on the stages of the procedures involved in this research. Basically, Section 2.4.1 and Section 2.4.2 presents an overview of analysis of information access in cultural heritage while basic concepts are also given in Section 2.4.1.1.

2.1 Overview of the Procedures

As shown in Figure 2.1, the overall procedures of this research starts with a study on Cultural Heritage before the other steps. In this stage, a study on the concept of cultural heritage which include *tangible* and *intangible* forms of cultural heritage is carried out. Next, the second stage of the overall procedures examines the processes in cultural heritage which covers the basic guidelines used by the Cultural Heritage Institution (CHI) to document the objects of cultural heritage. This includes the process of registering before the process of documenting (also known as cataloguing) by referring to the process from Museums Australia Victoria[1], Europeana Professional[2], and the World Museum Community[3]. The third stage is the most interesting and important stage. In this stage, a comparative analysis on the tool for information access especially in the Information Retrieval System is carried out. In addition, several applications of cultural heritage are also examined because these applications represent another type of Information Access. Several methods and solutions for Information Access are proposed in this stage. Finally, the models of evaluation are established in order to determine the effectiveness and quality of the proposed methods in this research.

[1] http://www.mavic.asn.au/services/small-museums-cataloguing-manual

[2] http://www.pro.europeana.eu/web/guest/publications

[3] http://icom.museum/programmes/art-and-cultural-heritage-mediation/mediation-workshop-in-art-and-cultural-heritage/

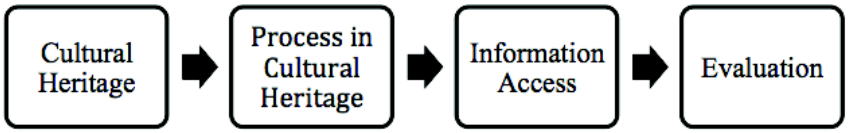


Figure 2.1: Overall Procedures for Research on Information Access in Cultural Heritage

2.2 Cultural Heritage

In general, Cultural Heritage is defined as the legacy of physical artefacts and intangible attributes of a group or society that are inherited from past generations, maintained in the present and bestowed for the benefit of future generations. So, the concept of Cultural Heritage is not only manifested through the *tangible forms* [1] such as artefacts, buildings, historic places, monuments, work of art and natural environment which are considered worthy of preservation for the future, but also through *intangible forms* [2] which represent the arts, social practices, festive events, storytelling, cuisine, knowledge and traditional craftsmanship. Therefore, the concept of Cultural Heritage can be expressed in two types: 1) tangible or 2) intangible as shown in Figure 2.2.

Physical objects are important to the study of human history because they provide a concrete basis for the ideas and the physical objects can only be validated through the process of discovery and analysis. The preservation of the physical objects allows the recognition of the necessity of the past, because the physical objects are able to tell the story about the past. Although the concept of Cultural Heritage is a broad area, information access tools that cover certain parts of tangible and intangible can be provided as long as both forms are properly documented digitally. Based on the concept of Cultural Heritage, it is clear that Cultural Heritage encompasses a vast range of different phenomena ranging from fine arts, artefacts, building, natural environment, traditional event to archival records. In other words, this research focuses more on the digital content provided by the Cultural Heritage Institution (CHI) which are properly documented.

2.3 Process in Cultural Heritage

This stage is to understand the process of documenting (also known as cataloguing) the collections by CHIs in order to propose and design a model of information access tool in the domain of cultural heritage. Although this stage serves to provide the basic guidelines to document the physical objects of cultural heritage, it covers mostly the important part of the process. The main reason to document the physical objects is to enrich the value of the physical objects before the CHIs display the objects to public. The physical objects will be useless if the CHIs do not document it properly. So, documentation is a vital activity in all the CHIs. In addition, documenting the physical objects properly will enhance the effectiveness in accessing the cultural heritage objects by the administrators in CHIs.

Basically, the process of registering physical object occurs before the process of cataloguing the physical objects. The process of registering an object is naturally linked to the cataloguing

[1] http://www.unesco.org/new/en/cairo/culture/tangible-cultural-heritage/
[2] http://www.unesco.org/new/en/cairo/culture/intangible-cultural-heritage/

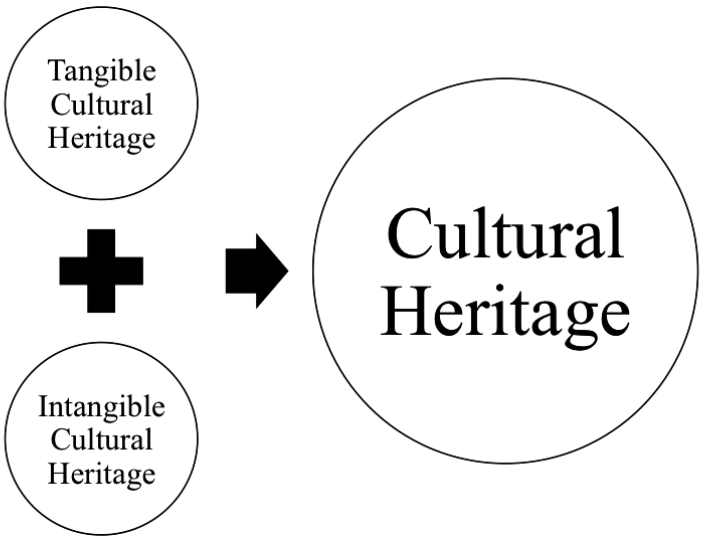


Figure 2.2: Overview of Cultural Heritage

process, because an object must be formally registered before it can be documented in the catalogue system. During the registration, the object will be assigned with a unique and permanent identification number in order to distinguish it from other physical objects in the CHIs.

Each object acquired by the CHIs must be entered into the registration book or also known as the "acquisition book". It is used to enter it into the permanent collection and a separate book is used to record the physical objects that are on loan from other CHIs. Basically, an ideal way to register an object is to include a) registration date, b) registration number, c) object's name and description, d) acquisition method, e) acquisition date, f) source and address, and g) comments. Their details are as follows:

- **Registration Date:** This refers to the date when the physical object was entered into the book and the date is often different from the date it was acquired. For example, the physical object may have been acquired in 2005 but the object has not been registered for another five years or more. Therefore, the registration date must indicate the date at which the process of registration is carried so that it is more meaningful.
- **Registration Number:** This is a unique identifying number that distinguishes one physical object from other physical objects. The registration number is usually assigned sequentially and according to the date that the object was acquired. The simplest method is to start at number 1, but this method is not imperative. Some CHIs use a compound numbering system that start with the year of acquisition followed by a sequential number. The compound numbering system can quickly indicate when the physical object was received and may avoid very large numbers.
- **Object's name and description:** The name of the physical object is specified in a few words such as two or three words. The description of the physical object is a brief description such as "silver-coloured metal, with two-tone blue enamel inset". A more comprehensive description is given in the catalogue.
- **Acquisition method:** It is useful to know whether the physical object was acquired by the

CHI through donation, purchase or by other methods. The method of acquisition should be specified in this field.

- **Acquisition date:** This indicates the arrival of the physical objects at the CHI, and not the date it was registered. Dates should always be written in full and in the same day-month-year form such as 16-04-1994.
- **Source and address:** This is to indicate whether the physical object was donated, purchased or collected, and the details of the source such as personal or company must be recorded.
- **Comments:** Important additional information is given in this field. It may highlights the significance of the physical object or relate it to other physical objects in the collection. Subsequently, the information on the physical object can also be recorded here, such as if it is missing or stolen, or has been transferred.

After reviewing the process of registering physical objects, then comes the most important part of the process which is documenting the physical objects in the system of catalogue. A catalogue is a powerful system to centralize the knowledge about the collection. The catalogue has been used for a long time in the CHIs. Nowadays, most of the CHIs are using computer-based catalogue such as Adlib [1], Maxus [2], and Vernon [3], but still a small number of CHIs are using paper-based catalogues. The catalogue can assist the curators and experts in their daily activities if the catalogue is created, managed and used properly. Normally, the computer-based catalogues are equipped with some tools such as authority lists and thesauri.

Authority lists and thesauri are useful tools to help the administrator or expert to document and retrieve the object of cultural heritage easily. Basically, the authority list is self-generated by the curators in CHI and it may cause ambiguity and confusion for the other curators because some curators prefer to use their own terms to document the object. In order to solve the problem of ambiguity, thesauri are introduced into the system as a means to control the terminology. Essentially, thesaurus can be viewed as a classification system that is able to control the preferred terms for the curators or experts to describe the objects. For example, the term “plane” and “aircraft” can be used to describe an airplane. Although, the term “plane” and “aircraft” are correct, it will cause difficulty when the curators or experts perform a search in the collection. So, thesauri can help to avoid such confusion by classifying a term according to a preferred term, and hierarchically group any related terms. Then, they can select a general or specific term by scrolling through the given hierarchy. In addition, thesauri are also used to place an object of cultural heritage within a group of similar objects. For example, the CHIs may wish to group all the objects that relate to the container of food and drink under the same name. So, the terms “food and drink containers” can become the terms for all similar objects associated with the kitchen. On the other hand, CHIs may also want to include narrower or broader terms such as “ceramic cup” or “domestic technology” to document the objects.

The main objective of cataloguing is to create a useful resource for knowing, accessing and managing the collection. Some careful consideration is required before building the catalogue. Basically, the CHIs need to consider the strengths of the collection and how the CHIs can

[1] http://www.adlibsoft.com
[2] http://www.maxus.net.au
[3] http://www.vernonsystems.com

gain most from having a particular object before the process of cataloguing. When the CHI is cataloguing a visual material, the CHI needs to be as accurate as possible about the medium such as photograph, lantern slide, transparency, postcard, lithograph, and water colour.

Fundamentally, the physical objects are not always easy to identify and indirectly it causes difficulty for the experts to document the physical objects of cultural heritage. Basically, there are many excellent reference books available to help the experts in identifying the objects, but the experts still face some challenges in cataloguing the objects. Ideally, most of the CHIs should build a small reference library to aid the experts in their work. Therefore, a basic and ideal case of a catalogue system should contain several fields such as a) registration number, b) object's name, c) title, d) description, e) keywords, f) inscriptions and markings, g) size, h) details of the maker, i) where made, j) when made, k) where used, l) when used, m) acquisition details, n) condition, o) storage location, p) current location, q) supplementary file, r) restriction and s) notes. Only selected fields such as title, description, and keywords or subjects are described and they are elaborated as follows:

- **Title:** This field is used for the books, manuscripts, arts, titles of the documents and photographs. It is to record the title given by the author and not by the creator in the CHI. Usually, the titles should be recorded exactly. In addition, a series of images can have a single title (a series title), which makes this field very useful for someone to perform the search for all images in a series. For example, the word "Caught" might be the title of a photographing that consists of a series of 15 photographs of local fishing along the Borneo River in 1964. Therefore, for certain physical representation which are not books, manuscripts, arts or photographs, this field will be empty.
- **Description:** The purpose of this field is to provide a description for the physical object of cultural heritage, but not its history or interpretation. Basically, a lot of people make the wrong assumption that this field is to show the history of the object, but in reality this field is the place for a formal statement on its significance. For example, if the physical object has some obvious significance such as the Bible that belonged to the first settler of the local community, then this detail should be included in this field. The aim of this field is to allow the reader to visualize the physical object. Ideally, the opening sentence should insert the name of the physical object and include some details about the physical object. The rest of the entry should clearly describe its physical attributes, including any damage or missing parts and it will be no more than about five sentences. Furthermore, it is useful to start describing the physical object from one point such as from top to bottom or vice versa. Besides, it also requires focusing on the most prominent features of the object such as shape, colour, materials, texture, ornamentation, method of production, and relative size. The following example describes a showman's steam traction engine held in one of the collections on history and the digital representation is shown in Figure 2.3.

  This model of a showman's steam traction engine was hand-fashioned around 1980. About half a meter in length, it is made of found materials from bits of metal, cardboard, tin boxes, screws, and small light globes. It is painted fire engine red, with yellow and black detailing. A narrow black cylinder is attached to its long, curved roof, with signage on the roof's two long edges reading "CHARGES", "SMITHS-ENTERPRISE SMITHS",

"CHARGES". Its four spoke-wheel rotate, the rear pair being larger than the front pair.[1]

Figure 2.3: Digital Representation - Steam Traction Engine

- **Keywords or Subjects:** This is an important search field when performing the search function because the keywords will become the indexer for all the physical objects. So, CHIs should consider very carefully the terms to be used as the keywords. Subject areas as well as names pertinent to the physical object should be listed, including names of individuals, organizations, buildings, places, and events. The ideal case is four or five keywords.

In a nutshell, it can be concluded that in most of the CHIs collecting the information about the physical objects are still done manually, and storing and organizing them in the management software like library catalogues are also done manually. The information such as description of the physical object is provided manually by the curators and experts. Even with the help of strict rules and software tools such as authority lists and thesauri, the problem of inconsistency in describing the objects is still prevalent in the CHIs. For example, the curators or experts use "syriac" to describe an object and another curators use "language of ancient syria" to describe the same physical object, then these two information are inconsistent with regard to the content of the information. Therefore, there is a need to consider the problems of 1) inconsistency, 2) documents of the physical representation is the descriptive representation of cultural heritage, 3) ambiguous and short information (for example the field of "title" as mentioned), 4) vocabulary used in the documents are dedicated to cultural heritage (the description is based on curators and experts) and 5) term variations (authority lists and thesauri), when the solutions are proposed.

2.4 Information Access

Over the past 20 years, the research on information access in cultural heritage have been experimented with different settings and proposed with various solutions. All of these solutions

[1] http://www.mavic.asn.au/services/small-museums-cataloguing-manual

were developed during the evolution of Information and Communications Technologies (ICT), specifically concerning 1) communication networks (from wired to broadband wireless connections), 2) end-user devices (from desktop computers to mobile smart phones) and 3) user interfaces (from textual to virtual reality). Information access can be represented in two methods namely 1) push and 2) pull. The *push* method of information access is to provide the direct information to the user while the *pull* method is to let the user decides on the information. In this research, the focus is purely in the pull method of information access.

The most common pull method is based on free-text query (refer Section 2.4.1) where the user understands the needs of the information by himself. Such information needs several combinations of ideas such as what the target information might look like, where it might be found, or how can the user tracks the information. In other words, the information needs can be summarized with the words of *look*, *where*, and *how* as proposed by Campbell [2000]. With the increase in casual search behaviour as mentioned by Villa *et al.* [2013], and Elsweiler *et al.* [2010] and overloading of information in cultural heritage, some users get lost and do not know how to navigate and explore information on cultural heritage. So, this type of user prefers to have a tool that is able to help them to navigate the information on cultural heritage rather than just inserting the query. This type of tools is also called as application-based information access tool. Therefore, several applications have been proposed to help this type of user to access, learn and explore the information on cultural heritage and this is discussed further in Section 2.4.2.

2.4.1 Information Access based on Free-text Query

Figure 2.4 shows the evaluation of *Exact Match* to *Best Match* in the domain of information access of cultural heritage. Exact match refers to the database solution which is mainly used by the curators in the past. For example, if the user searches the term “Mona Lisa”, the system returns only the results that are directly related to the term. The limited results retrieved by exact match approach are not sufficient to the users. In addition, other problem such as the data is fragmented over different databases results in a difficult access and retrieval by the users. Besides, the external users also cannot search the information due to unfamiliarity with the schema of the database and the data collections are fragmented and heterogeneous. The term “Best Match” refers to the information retrieval solution. This solution has unified access to all the data in the museum that provides score to each document based on the free-text query by the user. Koolen *et al.* [2009] and Koolen & Kamps [2010] showed that the approach from “Best Match” outperformed the “Exact Match” in the domain of information access of cultural heritage. The main reason to shift from “Exact Match” to “Best Match” is that the external users would still be able to search the information on cultural heritage although they do not know the schema of the database by using the free-text query.

2.4.1.1 Basic Concepts in Information Retrieval

In general, Information Retrieval is the underlying science of search engines. As a research field, it is primarily concerned with developing theories, algorithms, principles, and systems to help a user to find relevant information from a collection (text documents) to satisfy the needs of the user (query). However, Information Retrieval is also concerned with many other tasks

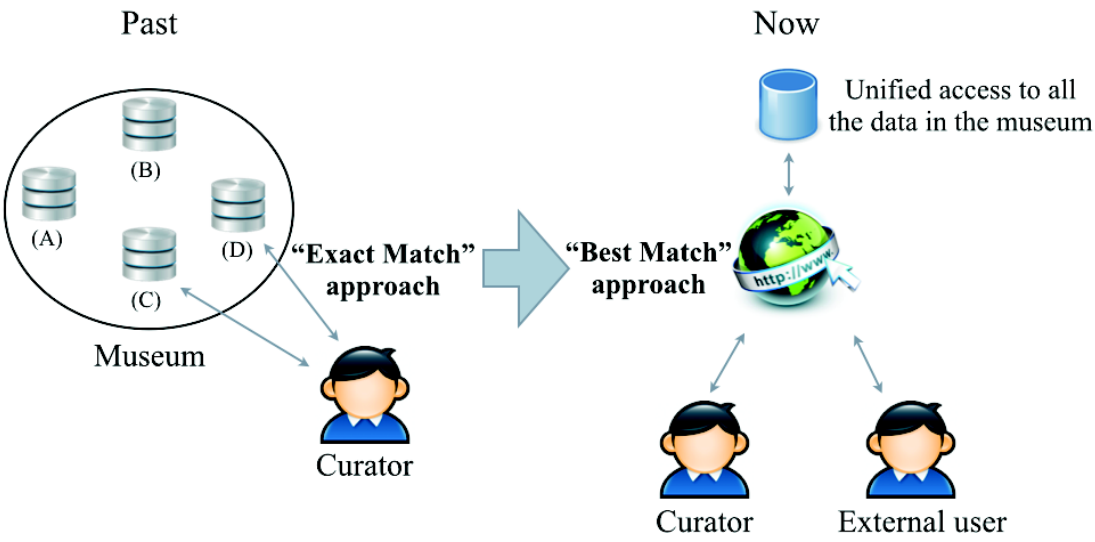


Figure 2.4: Evolution of Information Access in Cultural Heritage

which help to manage and exploit the information such as text categorization, text clustering, text summarization, and information filtering.

A main goal of any Information Retrieval System is to rank the documents optimally by given a query where the relevant documents should be ranked above than less relevant and non-relevant documents. In order to achieve this goal, the Information Retrieval System must be able to score the documents so that the highly relevant documents will tend to have a higher score than a non-relevant documents. Therefore, the retrieval accuracy of Information Retrieval System is directly impacted by the quality of the scoring function (also known as retrieval model).

Over the decades, the researchers from Information Retrieval have developed several types of retrieval models namely 1) Vector Space Model (VSM), 2) Probabilistic Models (such as Okapi, Divergence from Randomness) and 3) Language Model and their details are as follows:

- **Vector Space Model:** Vector Space Model (proposed by Salton [1971]) is a model where the terms are represented as vectors in a multi-dimensional (linear) space. The main aim of this model is to match the documents with the queries which both are represented as vectors of terms. Documents are retrieved by computing the distance between both vectors (the closer the distance between the vectors, then the more relevant is the document to the search query). Various techniques have been studied to compute the distance between vectors as given by Salton [1991].
- **Probabilistic Models:** Okapi and Divergence from Randomness are the examples which are based on the probability ranking principle (proposed by Robertson [1997]) to estimate the probability of a document being relevant to a given query. With this model, the documents in a collection are considered to be either relevant or non-relevant to an information needs expressed in the query. Then, the documents should be ranked based on the

probability of being relevant as follows:

$$\frac{P(\overrightarrow{d}|r)}{P(\overrightarrow{d}|n)} \tag{2.1}$$

where $P(\overrightarrow{d})$ is a document, r indicates relevance of the document, n indicated non-relevant documents of the document and $P(\overrightarrow{d}|r)$ is the probability of document $P(\overrightarrow{d})$ being relevant. A survey on probabilistic ranking models are given by Robertson [1997].

- **Language Model:** The Language Model was proposed by Ponte & Croft [1998]. The basic idea of Language Model is to assume that a query q is generated by using the Probabilistic Model based on a document d. Given a query $q = q_1, q_2, ...q_m$, and a document d, the idea is to estimate $P(d|q)$ which is the probability that document d has been used to generate query q. Therefore, by applying the formula of Bayes, the formula can be written as follows:

$$P(d|q) \propto P(q|d).P(d) \tag{2.2}$$

Based on Equation (2.2), the $P(d)$ is the prior belief that document d is relevant to any query. Then, $P(q|d)$ is the query likelihood for the given document d that can match with the query q. $P(d)$ is often assumed to be uniform and does not affect the document ranking. So, in this case, it can be ignored. In addition, Ponte & Croft [1998] assumed that each query word is generated independently, then we can rewrite Equation (2.2) into the form of log likelihood as follows:

$$logP(d|q) = \sum_{w \in V} c(w;q).logP(w|d) \tag{2.3}$$

where $c(w;q)$ is the count of words w in query q and V is a set of vocabulary. Based on multinomial distribution, the simplest way to estimate $P(w|d)$ is through the *maximum likelihood estimator* as given below :

$$P_{ml}(w|d) = \frac{c(w;d)}{|d|} \tag{2.4}$$

where $|d|$ is the length of the document d. Due to the problem of data spareness, the maximum likelihood estimator directly assigns the value of $null$ to the unseen words in a document. Therefore, smoothing is a technique to assign extra probability mass to the unseen words in order to solve this problem.

Basically, Dirichlet Smoothing is one of the smoothing techniques which is based on the principle of adding an extra pseudo term frequency which is $\mu P(w|C)$ according to Zhai & Lafferty [2004]. The Dirichlet Smoothing is obtained by taking into account the extra pseudo term frequency distribution:

$$P_{\mu}(w|d) = \frac{c(w;d) + \mu P(w|C)}{|d| + \mu} \tag{2.5}$$

Besides the retrieval models, query expansion, stemming and pseudo-relevance feedback or blind feedback are the basic methods which are often used in the Information Retrieval to improve the retrieval results as follows:

- **Query Expansion:** Query expansion is to expand the terms in the original query in order to match with a large number of the relevant documents. Basically, query expansion relies on techniques such as the synonyms of the terms through thesaurus [Jing & Croft, 1994] either through manual or automatic way that is able to fix the spelling error [Xu & Croft, 1996]. Most of the manual thesauri are compiled by linguists and domain experts and a well known thesauri is WordNet [Miller *et al.*, 1990]. WordNet is a large lexical database of English which groups the nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets) and each synsets can express a distinct concept. In addition, WordNet also resembles a thesaurus that groups all the terms together based on the meaning. In general, the thesaurus consists of a list of synonyms of the terms and identified by the human experts. Moreover, the thesaurus is consulted during the process of the query, so the synonyms of the term from the query can be added to the original query. As a result, manual thesaurus is costly to build, because it costs a lot of human labour. In order to overcome this issue, automatic thesaurus was introduced. Most of the automatic thesauri are based on the association hypothesis which stated that the terms that are related in a corpus tend to co-occur in the documents of that corpus [Rijsbergen, 1979].
- **Stemming:** Stemming is routinely used to reduce different form of terms to common roots. The purpose of stemming is to retrieve documents which contain morphological variants of the query words. English stemming is traditionally carried out by stripping off the common suffixes. Two widely used stemmers are the Porter stemmer (proposed by Porter [1997]) and Lovins stemmer (proposed by Lovins [1968]). The Lovins stemmer simply removes the longest suffix of a term while the Porter stemmer is more complex and it removes the endings from a term according to a set of rules iteratively until no more can be removed.
- **Pseudo-relevance feedback:** Pseudo-relevance feedback assumed that the top-ranked documents of an initial retrieval based on the original query are likely to be relevant [Croft & Harper, 1988]; [Xu & Croft, 1996]; [Mitra *et al.*, 1998]; [Lavrenko & Croft, 2001]. Figure 2.5 shows the overall framework for pseudo relevance feedback and the retrieval model will return the initial ranking documents before expanding the query. Basically, pseudo relevance feedback will select the top terms based on initial ranking documents (select top 20-30 terms from the initial ranking documents using for instance $tf.idf$ weights) and expand the query accordingly. Then, the retrieval model will match the new query with the documents and provide the final ranking documents.

After explaining the basic concept in Information Retrieval, the next section will show the several works which used various types of retrieval models and methods in the corpus of cultural heritage (CHiC).

2.4.1.2 Information Retrieval in Cultural Heritage

The Cultural Heritage in CLEF (CHiC) is an evaluation challenge and has been around since 2012 [Petras *et al.*, 2012]. It is organized by PROMISE Network of Excellence from the 7th

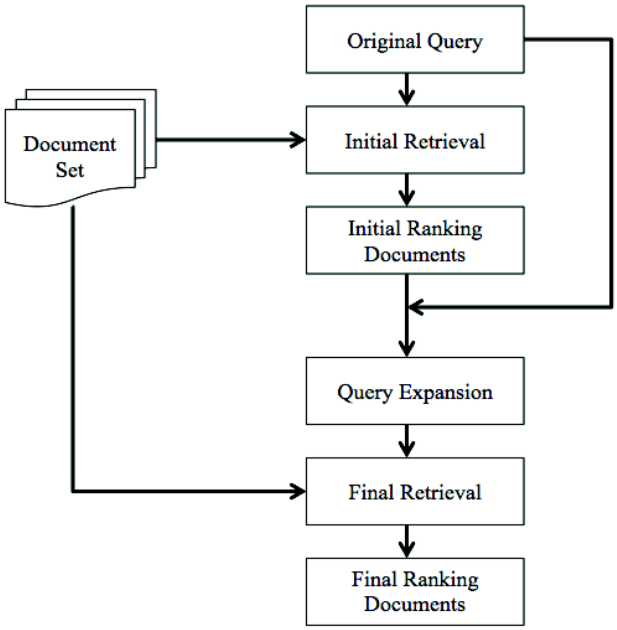


Figure 2.5: The Overall Framework for Pseudo Relevance Feedback with Query Expansion

Framework Programme of the European Commission to help the researchers from the domain of Information Retrieval to evaluate the effectiveness of the information access based on free-text query to the cultural heritage materials.

The test collection used by CHiC is from Europeana which consists of the complete Europeana metadata index as downloaded from the production system in March 2012. It contains 23,300,932 documents. With the move by Europeana to an open data license in the summer of 2012 and the subsequent changes in the content, the test collection represents a snapshot of the Europeana data from a particular time. However, the overlapping of the test collection with the current content is about 80%.

In 2012, six research groups namely 1) Chemnitz University of Technology, Department of Computer Science [Kürsten *et al.*, 2012], 2) GESIS, Leibniz Institute for the Social Sciences [Schaer *et al.*, 2012], 3) Unit for Natural Language Processing, Digital Enterprise Research Institute, National University of Ireland [Aggarwal & Buitelaar, 2012], 4) University of the Basque Country UPV/EHU and the University of Sheffield [Agirre *et al.*, 2012], 5) School of Information, University of California, Berkeley [Petras *et al.*, 2012], and 6) Computer Science Department, University of Neuchatel [Akasereh *et al.*, 2012] participated in the evaluation challenge in order to evaluate the retrieval models and other retrieval methods.

The research group from Basque Country and the University of Sheffield achieved the best result in the evaluation challenge of that year [Agirre *et al.*, 2012]. The proposed model used by this research group is the Language Model. Based on Agirre *et al.* [2012], the researchers argued that the query expansion approach together with external knowledge would give a little improvement which is less than 0.3% or even result in a slightly degrading performance.

The research group from Computer Science Department, University of Neuchatel [Akasereh *et al.*, 2012] used different models of Information Retrieval namely Vector Space Model (pro-

posed by Salton [1991]) and Probabilistic Models (proposed by Robertson [1997], Amati & van Rijsbergen [2002]) to compare and define the most effective model. The best performing model suggested by this research group is the Probabilistic Models. Besides, the research group [Akasereh *et al.*, 2012] also used different methods namely pseudo-relevance feedback (proposed by Rocchio [1971]) and light stemmer (only removes the plural "-s"). In the result, the pseudo relevance feedback method does not help to improve the retrieval performance while the light stemmer method manages to enhance the retrieval performance.

The research group from Chemnitz University of Technology, Department of Computer Science [Kürsten *et al.*, 2012] and GESIS [Schaer *et al.*, 2012] used the Apache Lucene [1] and Solr [2] which is based on Vector Space Model in the evaluation challenge. Both Apache Lucene and Solr are the open source Information Retrieval System. In addition, both research groups also applied the query expansion approach together with external knowledge. Unfortunately, the retrieval performance reported by Chemnitz is even worse than the one without the external knowledge while GESIS did not show the result of their method without the query expansion approach. So, it is not possible to conclude whether the the approach suggested by GESIS will help to improve the result of the retrieval performance.

Apart from free-text query, Carmel *et al.* [2012] and Yogev *et al.* [2012] proposed an entity oriented search (EoS) which the user can express the initial information need by using a wide range of queries (from free-text to more structured queries). The output of EoS is a ranked list of entities that match the query from the user. Figure 2.6 shows the search result by searching all the persons that are mentioned together with "Phelim O' Neill" (a person) in one of the deposition. In addition, Clough *et al.* [2011] proposed a method that models path or trails in order to provide a way for the users to access and to utilize the contents of digital libraries by enriching the experiences of these resources. Figure 2.7 shows the user can arrange the items into a path structure to support the final sense-making step in the information seeking journey. The path is an ordered structure of items that supports branching and the paths can be treated as a personal overview over a topic that can evolve as the user further explores the collection.

2.4.1.3 Information Retrieval with Context

Besides the Cultural Heritage in Clef, several researchers Tamine-Lechani *et al.* [2010], Kim & Chan [2003], and Sugiyama *et al.* [2004] also agreed that the free-text query should be integrated with other contexts, and this may help to improve the retrieval result. Besides, Tamine-Lechani *et al.* [2010] defined the context as:

"any knowledge or elementary information characterizing the surrounding application (user, objects, interactions) and having an important relationship with the application itself".

Basically, context may refer back to the whole data, metadata, applications, and cognitive structures which are embedded in situation of information seeking. Several studies have been proposed by Goker & Myrhaug [2008] and Vieira *et al.* [2007] and the context can be categorized into 5 types namely: 1) device, 2) spatio-temporal context, 3) user context, 4) task/problem, and 5) document context. The detail of the categories are as follows:

[1] http://lucene.apache.org/core/
[2] http://lucene.apache.org/solr/

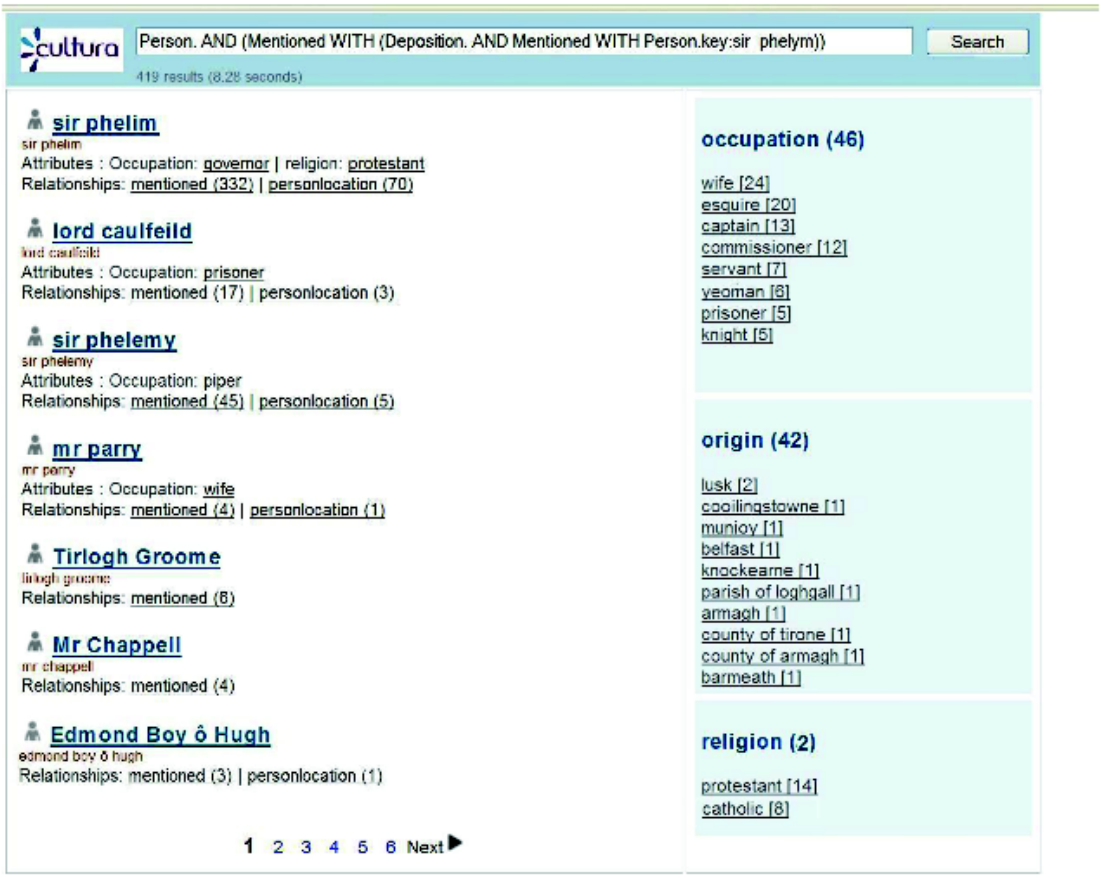


Figure 2.6: Search Result for query (all persons mentioned together with "phelim o'neil") from EoS [Carmel *et al.*, 2012]

- **Device:** Device refers to a physical tool that gives information to the user through computer, ubiquitous wireless computer, mobile phone, smart phone, or tablet. This context refers to the characteristics of the device. For example, working with small devices implies that high level of interaction is difficult, because the resources of the memory are limited and the queries are very short [Goker & Myrhaug, 2002]. Therefore, the retrieval process should avoid sophisticated outputs and be able to deal with ambiguous queries.
- **Spatio-temporal context:** Spatio-temporal context refers to the geographical location and time of the location. This context helps to deliver the information that adapt to the situation of a user where the data or query change and will not be the same all the time if the user changes the location [Goker & Myrhaug, 2008].
- **User context:** User context refers to the information about the user context and the user context can be collected through explicit approach [Hancock-Beaulieu & Walker, 1992]; [Belkin *et al.*, 2001] or implicit approach [Nichols, 1997] [Agichtein *et al.*, 2006]. The explicit approach requires the user to do explicit information action such as filling in the survey while the implicit approach requires an agent to collect the information on

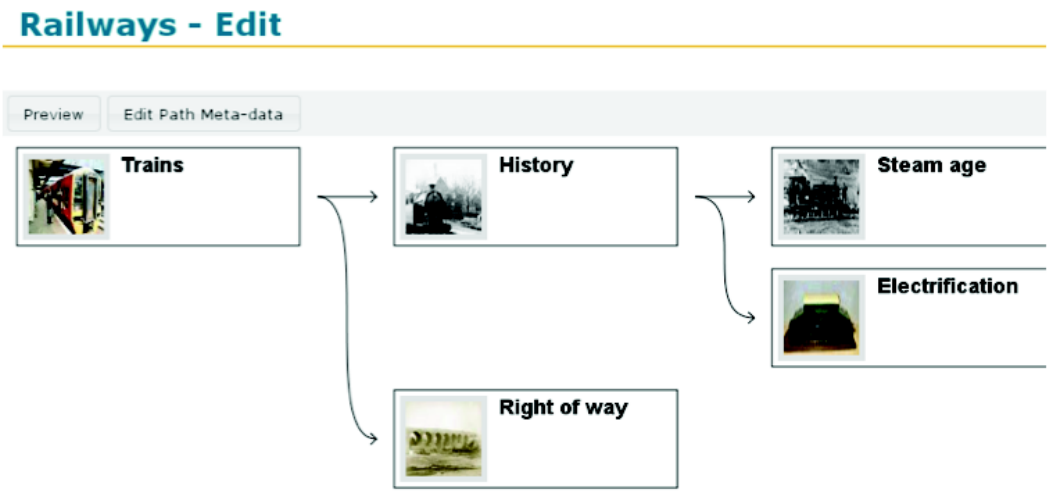


Figure 2.7: Overview over an Example *path* on the topic of "Railways" and the Ability to Structure the Information by Creating Multiple Branches [Clough *et al.*, 2011]

the activity of the user. In this dimension, there are two sub-dimensions namely *Personal Context* and *Social Context* under the User Context. In the Personal Context, the information of a user can be collected through: 1) *demographic context*, 2) *psychological context*, and 3) *cognitive context*. The demographic context refers to the personal attributes such as *language* and *gender* of the user [Frias-Martinez *et al.*, 2007] while psychological context refers to the *anxiety* and *frustration* of a user which may influence the information-seeking behavior [Kim, 2008]. Cognitive context refers to the *level of expertise* [Sugiyama *et al.*, 2004]; [Kim & Chan, 2003]; [Takano *et al.*, 2012] and *preference* of a user either in short-term or long-term.

- **Task/problem:** This context refers to the basic goal or intention behind the search activity such as explore, navigate, and fact finding [Jansen *et al.*, 2007]. In addition, this context also refers to the main application that requires information like in location-aware IR in mobile contexts such as tourism guide.
- **Document context:** Document context refers to the principle of poly representation and it was proposed by Ingwersen [1994]. The concept of poly representation is based on the hypothesis that the documents are classified into a multi-variable space such as document structure, document genre, document layout, and document hyperlink structure that uses multiple evidences from these variables to represent the documents.

2.4.1.4 Discussion

In a nutshell, it can be concluded that most of the research groups agreed that by integrating the external knowledge, the retrieval result can be improved. In addition, Adrian [2013] showed that explicit semantic analysis (refer [Znaidia *et al.*, 2012] for further information) clearly outperforms the baseline result. Unfortunately, the improvement is less than 1% or even worse. Besides, the research group from Computer Science Department, University of Neucha-

tel [Akasereh *et al.*, 2012] showed that the light stemmer which removes only the plural "-s" can help to improve the retrieval result in the range of 3.8% to 6.3%. In addition, various types of context such as user context can be integrated into the information retrieval system in order to improve the retrieval result based on the user because each user may contribute different context.

2.4.2 Information Access based on Application

A number of applications have been developed to help the user to access, learn and explore the information on cultural heritage such as GUIDE (proposed by Cheverst *et al.* [2000]), AmbieSense (proposed by Göker & Myrhaug [2008], and shown in Figure 2.8), LISTEN (mentioned by Ardissono *et al.* [2012]), COMPASS (proposed by Setten *et al.* [2004]), iCITY (proposed by Carmagnola *et al.* [2008], and refer to Figure 2.9), AgentSalon (proposed by Sumi & Mase [2001]) and Smartmuseum (mentioned by Ardissono *et al.* [2012]). A summary of the features of this type of approach is given in Table 2.1 based on three types of dimension namely 1) Setting, 2) Device, and 3) Presentation Style. The detail of the dimensions is as follows:

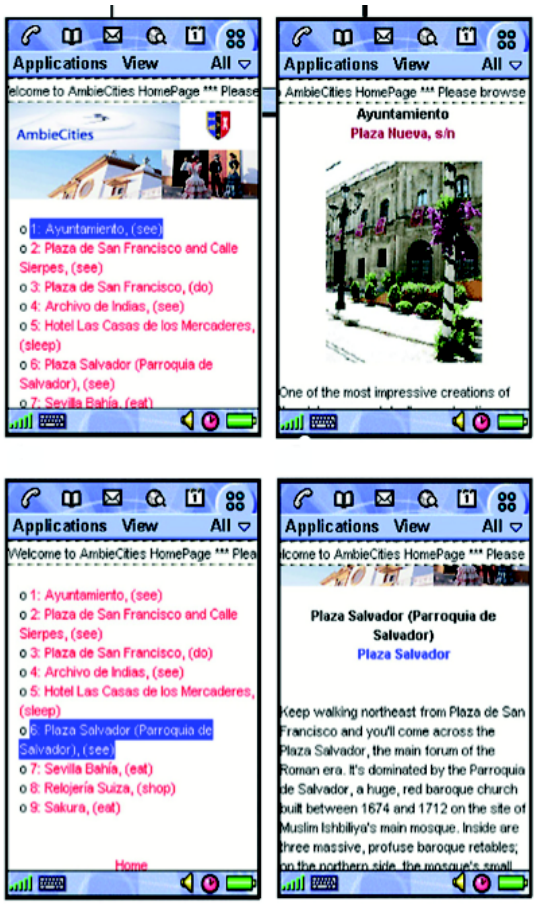


Figure 2.8: AmbieSense: Screenshots of Content in Mobile Device which Ayuntamiento (City Council, Seville) is the Main Tourist Attraction near Plaza de San Francisco [Göker & Myrhaug, 2008]

Figure 2.9: iCity: Screenshot of Content in PC which Lists Out the Cultural Events [Carmagnola *et al.*, 2008]

- **Setting:** This dimension represents the situation where the interaction occurs. With this dimension, three types of setting can be represented such as: 1) Indoor (In), 2) Outdoor (Out), and 3) Virtual (Vir). Indoor and outdoor indicate the physical presence at the site, while virtual indicates that the user is using virtual reality to view the gallery or collection.
- **Device:** This dimension represents the different devices that are able to afford different interactions. Therefore, three types of device can be distinguished that can be used to access the information on cultural heritage namely 1) Desktop (PC), 2) Mobile (Mobile), and 3) Wearable (Wear) devices.
- **Presentation Style:** This dimension represents the way to deliver the information on cultural heritage to the user according to the dimensions of Setting and Device. In this dimension, there are four types of Presentation Style namely 1) Map-based (Map), 2) Web Page (Web), 3) Character (Char), and 4) Audio-Visual (A-V). Map-based is used to show the point of interest through a map such as Google Map or Bing Map. Web Page refers to traditional Web layout, while Character uses a virtual character or agent to deliver the content. Finally, Audio-Video refers to the animation clips.

Table 2.1: Summary of Features of Several Applications

Application	Setting			Device			Presentation Style			
	In	Out	Vir	PC	Mobile	Wear	Map	Web	Char	A-V
GUIDE		*			*		*	*		
AmbieSense		*		*	*			*		
LISTEN	*					*				*
COMPASS		*			*			*		
iCITY		*		*	*		*	*		
AgentSalon		*			*				*	
Smartmuseum	*	*		*	*		*	*		*

Some early applications such as AVANTI (proposed by Fink *et al.* [1999]), and INTRIGUE (proposed by Oberlander *et al.* [1998]) stored the information on cultural heritage into structure knowledge base. Unfortunately, this approach is hard to develop and maintain. In order to overcome such limitations, some applications employed natural language generation techniques to create the presentation directly from the archival descriptions of items such as Tiddler (proposed by Paris *et al.* [2001]). Conversely in other projects, the lightweight representations and techniques were explored to minimize the amount of information on the items used by the applications. Basically, an application to explore the objects should ideally include:

- A simple list of objects which can represent the exhibition as a "visit paths" [Clough *et al.*, 2011] [Hall *et al.*, 2013].
- Display the textual descriptions of the objects.
- Classification of the concept of the object through natural language processing techniques [Fernando *et al.*, 2012].
- A specific ontology that is able to do multi-classification of artworks, such as location, culture, and multi-faceted search based on the collection [Yogev *et al.*, 2012] [Carmel *et al.*, 2012].

2.4.2.1 Discussion

In a nutshell, mobile technologies have become the common sight in the sector of cultural heritage while social web technology is also spreading fast in this sector. In addition, the focus has shifted from individuals engaged in a single visit to: 1) supporting a more realistic scenario which complies with the fact that most people visit physical museums in small groups such as families and classes of schoolmates, 2) managing a long lasting interaction between the visitors and the cultural heritage site by allowing online and on site visits, 3) enabling group collaboration in both physical and online cultural heritage sites in order to support the formation of virtual communities. Therefore, there is a need to consider these information before proposing an application on mobile platform.

2.5 Evaluation

Finally, the final stage is to determine the acceptability and the quality of the techniques and methods derived in this research. The main objective is to gather the results and comments of the proposed solutions in this research. In general, two types of evaluation methods will be introduced namely 1) Laboratory-based evaluation, and 2) Questionnaires.

2.5.1 Laboratory-based Evaluation

The laboratory-based evaluation in Information Retrieval dates back to the Cranfield project that involves experiments in a batch mode as proposed by Cleverdon [1997]. It is the evaluation paradigm of system-oriented Information Retrieval research that aims to measure the performance of Information Retrieval System [Kekäläinen & Järvelin, 2002]. More precisely, Cleverdon [1997] listed several criteria namely 1) coverage, 2) time lag, 3) recall, 4) precision, 5) presentation and 6) user effort, that could be used to evaluate the performance of an Information Retrieval System. Among these criteria, *recall* and *precision* are the most popular measurement used to estimate the effectiveness of the Information Retrieval System by returning the relevant documents. In Figure 2.10, a Venn diagram is given showing a set of relevant documents denoted as (R) for an information need, and a set of retrieved documents which denoted as (R') by Information Retrieval System. Recall is the portion of relevant documents retrieved, and precision is the portion of the retrieved documents that are relevant. In more specific, recall is the ratio of the number of relevant documents retrieved among the relevant documents in the collection as shown in the Equation (2.6) while precision is the ratio of the number of relevant documents retrieved among all the documents in the collection as shown in the Equation (2.7).

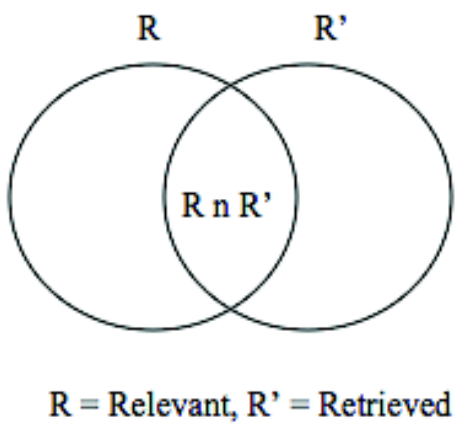


Figure 2.10: A Venn Diagram to Illustrate Recall and Precision

$$recall = \frac{|R \cap R'|}{|R|} \tag{2.6}$$

$$precision = \frac{|R \cap R'|}{|R'|} \tag{2.7}$$

Unfortunately, the criteria of presentation and user effort have not been studied in depth and there is no consensus agreement on related formal measures mentioned by Gwizdka & Chignell [1999] as compared to the criteria of precision and recall. Basically, the laboratory-based evaluation is based on 1) test collection, and 2) evaluation measures.

The test collection usually contains a set of documents, a set of well defined *topics* and a set of *relevance* assessments to identify those documents that are relevant to each topic. The early attempts to build the test collections is an exhaustive method which is based on the judgement of every document based on the topic. Basically, a group of people will review the documents and identify whether the document is relevant or non-relevant according to the topic as proposed by Cleverdon [1997]. In order to achieve scalable assessment in the case of large collections, the most widely method is the pool assessment. Pool assessment uses the top-ranked documents returned by several Information Retrieval System and re-judge the top-ranked documents manually as proposed by Voorhees [1998].

An evaluation measure estimates the ability and the effectiveness of a system in order to return the relevant documents. Basically, the value of relevance is binary where "1" is relevant while "0" is irrelevant. The measure is mainly based on precision and recall. In addition, Text Retrieval Conference (TREC) proposed by Harman [1995] also introduced several measurements namely 1) interpolated recall-precision average, 2) non-interpolated average precision over all relevant documents, 3) precision@X and 4) R-precision as described below:

- **The interpolated recall-precision average:** Computes precision at recalls of 11 equal intervals from 0.00 to 1.00. The values are averaged over all the queries for each of the recall levels.
- **The non-interpolated average precision over all relevant documents:** Computes precision after each relevant document is retrieved. All precision values are then averaged to get a single value of the query performance, called Average Precision (AP). Mean Average Precision (MAP) is the averaged value across all the queries. In other words, average precision is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved, and his value is then average over information needs. If the set of relevant documents for an information need $q_j \in Q$ is $d_1, ...d_{mj}$ and R_{jk} is the set of ranked retrieval results form the top result up to document d_k, then:

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk}) \tag{2.8}$$

- **Precision@X (P@X):** Computes the precision after X retrieved documents.
- **R-precision**: Computes the precision after R retrieved documents where R is the number of relevant documents for the query.

In recent years, Mean Average Precision (MAP) is the most standard measurement among the TREC community which provides a single-figure measure of quality across recall levels. Among the measurements, MAP has been shown to have especially good discrimination and stability.

The laboratory-based evaluation model has been undertaken in the Information Retrieval research for several years and contributed to several advances in this area. In a nutshell, it can be summarized that the merits and limitations of such evaluation design are as follows:

Merits

- The ability to repeat an experiment is a key feature of empirical research according to the principle of TREC.
- The iterative evaluation identifies the problems of the proposed model and help the researchers to understand the problems and the expected the solutions.
- A critical requirement of a retrieval test collection is the ability to help in comparing with other models by using the same collection.

Limitations

- The test collection is not realistic because it neither involves real user's tasks nor situations during the experiments. In addition, the queries do not always reflect the user's tasks or situations.
- The laboratory-based evaluation is based solely on algorithms which ignore the user's context (situation and knowledge).
- Binary relevance assumption which may have a negative impact because it lead to identify all relevant documents rather than medium relevant.

2.5.2 Questionnaires

Bulmer [2004] defined questionnaires as a "structured research instrument which is used to collect social research data in a face-to-face interview, self-completion survey, telephone interview or Web survey". It consists of a series of questions set out in a schedule, which may be a form, on an interview schedule on paper, or on a Web page. Both the interview and self-completion survey (electronic or via pen-and-paper) questionnaire modes are commonly used to gather user opinion about the proposed solutions. Various types of questions which are commonly used are as follows:

- **Open questions:** Open questions are useful to find out more about the reasons, why users behave the way they do and provide the chance to give free comments on aspects of the systems. In the experiment, open questions are used to gather the opinion from the user about a specific feature of the system. Furthermore, open questions are used to identify positive and negative features from the user point of view.
- **Closed questions:** Users can respond to a given set of responses. The closed questions can be in the form of a statement such as "I was satisfied with the application". The *Likert Scale* technique is used for quantifying the expression of agreement or disagreement of a user. It presents a set of attitudes. For measuring the level of agreement, a numerical value from one to seven is used. The value can be measured in calculating the average of all received responses.

In general, a combination of both question types is commonly used because each type of question has the advantages and disadvantages. Kelly *et al.* [2008] concluded that the pen-and-paper method is the most efficient mode to gather information. In addition, Kelly *et al.* [2008] also showed that the user's qualitative evaluation was significantly higher using an electronic questionnaire than face-to-face interview.

2.5.2.1 Discussion

In a nutshell, the laboratory model is a good evaluation model to evaluate the proposed methods and able to repeat the experiments. In addition, the laboratory model also provides the flexibility that allows the researchers to evaluate and compare the other models or approaches easily. Unfortunately, the laboratory model lacks of considering the number of search session, other contexts and tasks. As a result, laboratory model is more towards to system-centred, while questionnaire is more towards to user-centred evaluation.

A short discussion and summary which cover information access based on free-text query and information access based on application will be given in the next section.

2.6 Discussion

In general, most users will turn to the Information Retrieval System as the quickest way to access, learn and explore the information. So, an information access tool is needed in the domain of cultural heritage. In the context of Information Retrieval, there are two types of users namely: 1) the user who understands the needs, and 2) the user who is not sure about the needs. In order to provide an information access in cultural heritage, two types of solution are proposed on this research namely: 1) Information Retrieval System for the user who understands the needs, and 2) an application for the user who is not sure about the needs and helps them to access the information in cultural heritage.

In the nutshell, it can be summarized that the problems mentioned namely 1) inconsistency, 2) documents of the physical representation is the descriptive representation of cultural heritage, 3) ambiguous and short information, 4) vocabulary used in the documents are dedicated to cultural heritage and 5) term variations fall into the problem of term mismatch. The term "term mismatch" is defined as a mismatch between the term in the query and the term in the document. For example, if a document is about "penguin", a query is about "animal", and without any knowledge, then the retrieval model will not be able to return this document. If we have the knowledge that "penguin is an animal", then the retrieval model is able to retrieve the document. In addition, several researchers namely Radhouani *et al.* [2008], Vallet *et al.* [2010], Radhouani & Falquet [2006], Chevallet *et al.* [2007], Adrian [2013], Petras *et al.* [2013], and Lacoste *et al.* [2007] have shown that using domain knowledge would improve the retrieval result.

As a result, this research proposes to integrate knowledge into existing retrieval model to solve partially the problem of term mismatch in order to improve the precision of the Information Retrieval System for the user who understand the needs. In addition, this research provides and investigates an application on mobile device which integrates image, panoramic view, virtual environment, audio, and text to help the user who are not sure about the needs. Both

proposed solutions are able to help the user to access, learn and explore the information on cultural heritage.

The overall flow of this research is shown in Figure 2.11. As mentioned before, the main aim to model an indirect information access system for the user to access, learn, and explore the information on cultural heritage. So, the first stage of this research comes from a case study involving a mobile game on cultural heritage. Then, the second stage of this research is on Information Retrieval System, and the third stage is on Context Search. In Context Search, the user profile will be exploited after the initial result from Information Retrieval System and this is also known as re-ranking. Basically, the first stage is to tackle the users who are uncertain on their information needs, while second and third stages tackle the users who understand their needs (seeking information). As stated in Section 1.1, the main motivation of this research comes from a case study involving a mobile game on cultural heritage and the aim of the mobile game in this research is to investigate the importance of instant access when the user is playing the game. Therefore, the first, second, and third stages would allow the user to access, learn, and explore the information on cultural heritage. Since the first stage of this research is only a preliminary case study (mobile game), therefore it will be discussed after the chapters on Information Retrieval System (second stage) and Context Search (third stage).

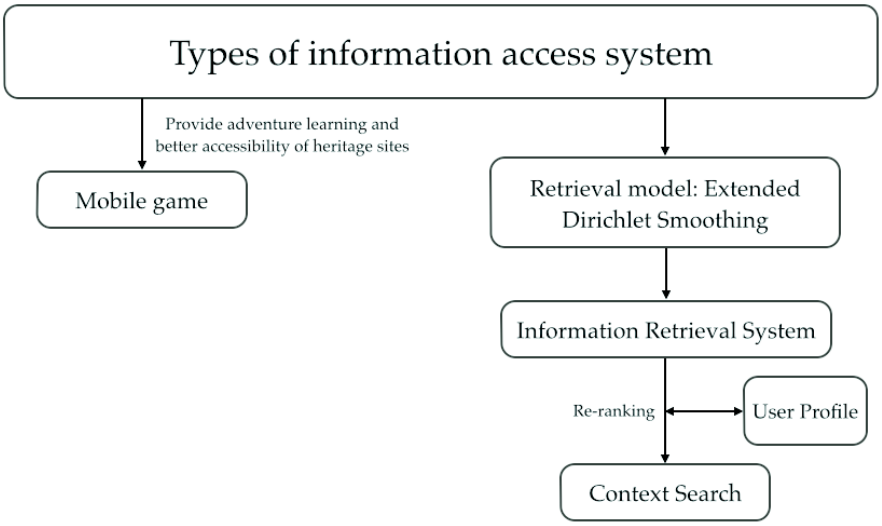


Figure 2.11: The Flow of this Research

2.7 Summary

Firstly, a study on the concept of cultural heritage is carried out in Section 2.2. Secondly, an examination on the processes in the cultural heritage which covers the basic guidelines used by the CHIs to document the object of cultural heritage in Section 2.3. Thirdly, a comparative analysis on the tool for Information Retrieval System is shown in Section 2.4.1 and several application of cultural heritage are also examined in Section 2.4.2. Finally, the model of evaluation is established in the last section which is Section 2.5 of this chapter.

Part III

STATE OF THE ART

Chapter 3

State of the Art

This chapter discusses the state of the art of this research. Section 3.1 presents the state of the art on solving the problem of term mismatch while Section 3.2 presents the state of the art an gathering and representing the interest of a user. Finally, section 3.3 presents the state of the art on mobile game as a learning platform.

3.1 Information Access - Information Retrieval System

Several techniques have been proposed to tackle the problem of term mismatch. Some of these important techniques include *query expansion* (proposed by Xu & Croft [1996], Zhao & Callan [2012]), *relevance feedback* (proposed by Salton [1971], Lavrenko & Croft [2001]), *dimension reduction* (proposed by Porter [1997], Krovetz [1993], Deerwester *et al.* [1990], Bendersky & Croft [2008], Jing & Croft [1994]), and *statistical translation model* (proposed by Berger & Lafferty [1999], Karimzadehgan & Zhai [2010], Karimzadehgan *et al.* [2013]).

3.1.1 Relevance Feedback

The idea of relevance feedback is to involve the user into the process of information retrieval in order to improve the final result. In particular, the user gives feedback on the relevance of documents in an initial set of results. Basically, the basic procedure is as follows:

- The user submits a query.
- The retrieval model returns an initial set of retrieval results.
- The user marks some returned documents as relevant or non relevant.
- The retrieval model computes a better representation of the information need based on the feedback from the user.
- The retrieval model displays a revised set of retrieval results.

Usually, relevance feedback consists of three types: 1) Explicit feedback (will be explained later, refer to Section 3.2.2) 2) Implicit feedback (will be explained later, refer to Section 3.2.3) and 3) Pseudo or Blind feedback (refer to Section 2.4.1.1 for more information).

3.1.2 Query Expansion

Query expansion is an approach to provide a more direct way to solve the term mismatch by including additional terms into the original query (Refer to Section 2.4.1.1 for more information on query expansion).

A classical query expansion is the combination of the relevance feedback. Lease [2008] has shown that simple pseudo-relevance feedback can bring a decent gain in the overall retrieval performance. However, pseudo-relevance feedback is usually applicable to all the cases, but Akasereh *et al.* [2012] showed that pseudo-relevance feedback return bad result in the CHiC collection. In recent years, many query expansion techniques and algorithms were developed and derived from Rocchio relevance feedback. A popular and successful automatic query expansion was proposed by Mitra *et al.* [1998] for the Okapi system, and query expansion in Divergence from Randomness (DFR) was proposed by Carpineto *et al.* [2001]. The recent works by Agirre *et al.* [2012], Kürsten *et al.* [2012], Schaer *et al.* [2012], and Yin *et al.* [2009] have shown that by integrating query expansion with external knowledge such as WordNet, DBpedia, and Wikipedia the retrieval result would be improved.

3.1.3 Query Reformulation

Query expansion aims to solve the particular kind of term mismatch problems caused by synonyms of query terms. More generally, query reformulation techniques provide the framework to solve different types of mismatch problem. These reformulations include 1) removing terms from the query [Kumaran & Carvalho, 2009]; [Dang & Croft, 2010], 2) adding new terms, and 3) replacing certain terms with others. Basically, query reformulation consists of two types: 1) query term reduction for long queries, and 2) query term substitution as described below:

- **Query Term Reduction for Long Queries:** Query term reduction for long queries is a particular form of reformulation that improves the result of retrieval. The main assumption for this approach is to assume that the long queries may contain several noise (the terms inside the query that are unnecessary), and it can help to achieve a better retrieval performance if those terms are removed according to Kumaran & Carvalho [2009], and Bendersky & Croft [2008]. The main reason is that these query terms may represent the unnecessary concept.
- **Query Term Substitution:** Query term substitution aims to substitute the terms inside the query which are the terms from the query that are not effective in the process of matching. Wang & Zhai [2008], Dang & Croft [2010], Cao *et al.* [2008], and Jones *et al.* [2006] performed automatic per term query reformulation and showed that it is able to improve the retrieval result.

3.1.4 Dimension Reduction

Dimension reduction is the process of reducing the number of random variables in cases where the query and the document refer to the same concept but using different terms. This can be achieved by using *concept-based approach* (proposed by Abdulahhad *et al.* [2011], Bendersky & Croft [2008]), *stemming* (proposed by Porter [1997], Krovetz [1993]), and *latent semantic*

indexing (proposed by Deerwester *et al.* [1990]). All these techniques use different strategies to reduce the chances that the query and document refer to the same concept but using a different term.

- **Concept-based approach:** Yang & Chute [1993] introduced a technique that uses a set of training queries with relevance judgements to learn a term and mapped the term with a concept and this is also known as concept mapping. The mapping is applied from the terms of the query to infer the related concepts. Then, the query concepts are used to represent the query and to rank the documents according to similarity in the space of concept. This method can be unsupervised or supervised by utilizing the relevance judgements in a training set in order to improve the retrieval result.
- **Stemming:** The major problem with Lovins (proposed by Lovins [1968]) and Porter (proposed by Porter [1997]) is that both do not pay any attention to the meaning of the terms. In later development, Peng *et al.* [2007] performed stemming according to the context of the query which helps to improve the accuracy and the performance of retrieval compared to the query independent stemmers such as Porter [Porter, 1997] and Krovetz [Krovetz, 1993]. Recent works from Akasereh *et al.* [2012], and Petras *et al.* [2013] have shown that a light stemmer is able to improve the retrieval result in the domain of cultural heritage rather than the standard stemmer. (Refer to Section 2.4.1.1 for more information on stemming)
- **Latent Semantic Indexing:** Latent Semantic Indexing (LSI) (proposed by Deerwester *et al.* [1990]) is a technique that was motivated by one of the earliest quantitative research on the term mismatch problem by Furnas *et al.* [1987]. The main aim is to address the term mismatch by representing the terms and the documents in a latent semantic space and compute the similarity between the query and document in the semantic space. Basically, the terms that are similar in the latent semantic space tend to be the terms that not only co-occur in the collection, but also appear in a similar contexts. LSI is typically only applied to a sampled set of documents instead of the whole collection. Schütze *et al.* [1995] extended the use of LSI by applying LSI on the top ranked documents from an initial retrieval. This localizes the semantic space to the context of the current query, and improved the accuracy. However, this local LSI method still does not show robust improvements over a keyword based method.

3.1.5 Statistical Translation Model

Statistical Translation Model solves the problem of term mismatch between the queries and documents through a learning model. In other words, Statistical Translation Model is a model where all the translations are generated on the basis of statistical models according to Berger & Lafferty [1999]. The idea is based on information theory where a document is translated according to the probability distribution $P(u|v)$, which gives the probability that term v can be semantically translated to term u in order to address the problem of term mismatch. Basically, training these translation models require a large enough query to cover enough terms in order to gather enough accurate training data. Gao *et al.* [2010] obtained large amounts of fairly accurate training data from user clicks based on query logs, and used the data to train a translation-based retrieval model to improve retrieval. Recent works from Karimzadehgan *et al.* [2013],

and Karimzadehgan & Zhai [2010] have shown that Statistical Translation Model is able to bridge the vocabulary gap and able to improve the retrieval result. Karimzadehgan & Zhai [2010] proposed to integrate the similarities (translated into probabilities) for translating the usual maximum likelihood estimations of terms in documents according to the query terms. In addition, Karimzadehgan & Zhai [2010] relied on the data from the corpus and not from other sources like synthetic queries which was proposed by Berger & Lafferty [1999]. Furthermore, Karimzadehgan & Zhai [2010] mentioned that if the similarities expressed as probabilities will lead to unwanted behaviour, like the fact that "it is possible that $p(w|u) > p(w|w)$" for a term w. To overcome this problem, Karimzadehgan & Zhai [2010] defined a regularization of the self-translation probability.

3.1.6 Exploiting the Term Similarity Matrix in Vector Space Model

Crestani [2000] proposed a general framework to exploit the term similarity matrix into the matching process based on a variation of inner product. At first, Crestani [2000] assumed the matching process is a variant of a vector inner product as shown below;

$$RSV(d,q) = \sum_{t \in q} w_d(t) w_q(t) \tag{3.1}$$

Furthermore, Crestani [2000] noted that $q \triangleright d$ exploits the non-matching terms in the documents by utilizing the similarity Sim function between all the terms. $Sim(t,t') = 1$ *iff* $t = t'$, means that t and t' refer to the same term. If $Sim(t,t')$ is close to 1, it means that t and t' are semantically closed and t and t' can be used to express the same concepts and otherwise it is 0.

In general, the proposed idea by Crestani [2000] is an extension of the Retrieval Status Value (RSV) formula as shown in Equation (3.1). The idea is to extend the matching process that includes a new intermediate term t^* into the relation of the term t which appears in the query. Essentially, this term t^* does not necessarily appears in the query although the term t^* appears in the document d. Given a term t from the query[1] means to consider all the terms in the document. The main idea of this approach is to take into account of the term t^* which maximizes the term similarity function Sim for a given t in the query as shown below;

$$t^* = \underset{t' \in d}{argmax}(Sim(t,t')) \tag{3.2}$$

This means that t^* is chosen among the terms that belong to d, i.e. among the terms where $w_d(t) \neq 0$, by maximizing the Sim value. If the term t appears in the document d, then the best term t^* is t itself. If t does not appear in document d, then the weight of t is substituted by a term t^* in the document with the lower score if there is a similarity value between t and t^*. The similarity value should be lower than the the value of exact term and it changes the formula in the following way:

$$RSV_{max(q \triangleright d)}(d,q) = \sum_{t \in q} Sim(t,t^*) w_d(t^*) w_q(t) \tag{3.3}$$

[1] When we say "t from the query" or "from document" it simply means the weight of t is not null in the query, or in the document i.e. $w_q(t) > 0$ or $w_d(t) > 0$

In other words, t^* is the most similar term to t in document d. If t^* exists in document d, then the similarity score should be 1 and the formula is the same as Equation (3.1). For this reason, we can conclude that the formula is an extension of the inner product of the vector d where the term does not appear in the document, but through the Sim function.

Crestani [2000] proposed another solution that selected the term t^* with the maximum similarity. In addition, Crestani [2000] proposed to add all contributions of non matching terms in the evaluation of the RSV. This new value $RSV_{tot(q \triangleright d)}$ is defined by:

$$RSV_{tot(q \triangleright d)}(d, q) = \sum_{t \in q} \left(\sum_{t^* \in d} Sim(t, t^*) w_d(t^*) \right) w_q(t) \tag{3.4}$$

The difference is that instead of using only the term t^* that is the most similar, all possible similar terms are used to compute the new weight $w_d^{tot}(t)$ as follow:

$$w_d^{tot}(t) = \sum_{t^* \in d} Sim(t, t^*) w_d(t^*) \tag{3.5}$$

From a matrix point of view, this computation is equivalent to a matrix product between the similarity matrix Sim and the document vector d. This produces a new extended document vector d_1:

$$d_1 = Sim \times d \tag{3.6}$$

From a graph point of view, if we consider the Sim matrix as a weighted graph, then this is equivalent to move one step into this graph, and sum all the weights for the term. Hence, it is equivalent to *extent the document* d by using the similarity graph (matrix) and the formula can then be rewritten as :

$$RSV_{tot(q \triangleright d)}(d, q) = \left(d^\top \times Sim\right) \times q = d_1^\top \times q \tag{3.7}$$

3.1.7 Discussion

All the approaches such as Relevance Feedback, Query Expansion, Query Reformulation, Dimension Reduction and Statistical Translation Model have been shown to be able to solve the problem of term mismatch and tend to improve the retrieval result. In a nutshell, it is necessary to highlight that the Query Expansion with feedback approach requires the initial set of retrieval result before expanding the terms into the query. In other words, Query Expansion requires two processes in order to get the final set of retrieval result. Although the recent works by Agirre *et al.* [2012] have integrated the Query Expansion with WordNet while Otegi *et al.* [2011] have integrated the Query Expansion with knowledge bases, both works still require two processes in order to get the final ranking. In order to formalize this scenario, Q is used to represent the query, and Q' to represent the query after expanding the terms based on the initial set of retrieval result. Then, Q' will be matched with the documents again in order to get the final ranking. This process is considered as global adaptation because Q' is the final query which matches with all the documents. The main idea is to transform Q to Q' after integrating the knowledge. There are a number of approaches to solve the term mismatch problem by using Language Model. The recent works from Berger & Lafferty [1999], Karimzadehgan & Zhai

[2010], and Karimzadehgan *et al.* [2013] proposed to use Statistical Translation Model to solve the term mismatch problem. The main difference between these works is that Berger & Lafferty [1999] use synthetic queries while Karimzadehgan & Zhai [2010] use mutual information to generate the relationship between the two terms. In addition, Karimzadehgan *et al.* [2013] use the Conditional Context estimation from Karimzadehgan & Zhai [2012] to generate the relationship between the two terms. In some way, the proposed idea by Karimzadehgan & Zhai [2010] is related to the second proposition of [Crestani, 2000] where the idea is to consider the similarities between each query term and all the document terms.

3.2 Information Access - Context Search

It is a necessity to integrate other contexts (besides the query) into the Information Retrieval System in order to satisfy the needs of the user. So, it is a necessary to interpret the needs of a user correctly. Spink *et al.* [1998] indicated that a user is often not sure about the information needs by themselves and the most fundamental problem is that the user is not familiar with the data inside the collection. As a result, the user does not know what type of information that they can retrieve from the collection as shown in Salton & Buckley [1997]. Furthermore, Nicholas *et al.* [2013], and Church & Smyth [2009] have shown that search queries in cultural heritage are rather short, usually consisting of approximately three terms. Considering these observation, it is hence challenging to satisfy the information needs of a user. Without further knowledge, it is a demanding task to understand the intention of a user. Contextual Information Retrieval System (also known as context search) aims to improve the classical information retrieval model such as Vector Space Model, Probabilistic Model, and Language Model where the classical models focus more on the topic of the search and ignore the context from the user.

The next section will explain several ways to gather the interest (implicit and explicit), and representation (weighted keywords and semantic networks) for the user profile.

3.2.1 Gathering and Representing Interest

Most of the approaches for Contextual Information Retrieval System are based on user profile. In general, user profile is a repository of a user which contains all the information of a user. In other words, user profile is a process to learn the information of a user such as language, gender, and preference through a period of time. User profile can be expressed through relevance feedback [Salton & Buckley, 1997]. Ruthven & Lalmas [2003] have shown an overview to exploit the context of a user through relevance feedback. The most fundamental principle to gather the context of a user is through Rocchio algorithm (proposed by Rocchio [1971]). The formula and variable definitions for Rocchio algorithm are as follows:

$$\overrightarrow{Q_m} = (a.\overrightarrow{Q_0}) + (b \cdot \frac{1}{|D_r|} \cdot \sum_{\overrightarrow{D_j} \in D_r} \overrightarrow{D_j}) - (c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\overrightarrow{D_k} \in D_{nr}} \overrightarrow{D_k}) \tag{3.8}$$

where $\overrightarrow{Q_m}$ is the modified query vector; $\overrightarrow{Q_0}$ is the original query vector; $\overrightarrow{D_j}$ is the related document vector; $\overrightarrow{D_k}$ is the non-related document vector; a is the weight for original query; b is the

weight for related documents; c is the weight for non-related documents; D_r is a set of related documents and D_{nr} is a set of non-related documents.

Basically, there are three types of relevance feedback in the Contextual Information Retrieval System namely: 1) explicit, 2) implicit, and 3) pseudo relevance feedback.

3.2.2 Gathering User Profile through Explicit Relevance Feedback

A simple approach to identify the interest of a user is to explicitly inform the user about their opinion. In other words, the user can express the opinion by providing explicit relevance feedback to the Information Retrieval System. Basically, the user can inform the Information Retrieval System on which documents are relevant in the list of the result and the relevance judgement can be given as binary ("0" to indicate non-relevant, "1" to indicate relevant) or graded relevance scale. Unfortunately, binary relevance scale is considered as a strong judgement since it only considers two choices which are relevant or non-relevant. As a result, graded relevance scale was introduced and allows the user to define the different grades such as "highly relevant", "relevant", and "maybe relevant" or "somewhat relevant". In a nutshell, explicit relevance feedback is very reliable and it also has shown that the explicit relevance feedback can affect the search process as shown in Belkin *et al.* [2001].

3.2.3 Gathering User Profile through Implicit Relevance Feedback

Implicit feedback techniques helps to learn the interest of a user unobtrusively rather than asking or "to ask" the user to rate the documents. The main advantage of this approach is that the user is not required to provide any explicit feedback. As a result, a large quantity of implicit data can be gathered without disturbing the work flow of a user. Therefore, implicit relevance feedback is an attractive alternative approach. Nichols [1997] in his research concluded that implicit approach is less accurate than information based on explicit feedback. Then, Agichtein *et al.* [2006] evaluated the effect of user feedback based on web retrieval by using 3000 queries and the data of the user interactions. They concluded that implicit relevance feedback can improve the retrieval performance as much as 31% compared to the systems that do not incorporate any feedback. Furthermore, both implicit and explicit measures can be combined to provide an accurate representation of the interests of a user.

From the psychological point of view, a promising indicator of interest is viewing time. Basically, people will spend more time to view objects that they like than objects that they are not interested in. For instance, Faw & Nunnally [1967] showed that a positive correlation between "pleasant ratings" with the "viewing time". In addition, Oostendorp & Berlyne [1978] found that most people look longer at the images that they prefer, and this action indirectly evokes the emotions of pleasure by the user. In the context of Information Retrieval, users are expected to spend more time to view the relevant documents than the non-relevant documents. So, Claypool *et al.* [2001] introduced a set of indicators in the web which includes explicit and implicit indicators. As a result, they concluded that three features namely 1) "the time spend on a page", 2) "the amount of scrolling on a page", and 3) combination of the first and second features are the valid implicit indicators to show the interest of a user. Furthermore, they found that the action of "scrolling measures" and the action of "number of mouse clicks" are ineffective indicators. Furthermore, Morita & Shinoda [1994] investigated that the user behaviour such as "coping",

”saving or following up of an entry”, and ”the time spend for reading” could be used as implicit indicators for interest when reading the articles from the newsgroup. As a result, they revealed that the action of ”the time spend for reading” is an effective indicator because user tends to spend more time to read interesting documents than uninteresting documents. Besides, they highlighted that the relation between interest with ”saving or following-up of an entry”, and the relation between interest and ”copying” cannot be considered to be an effective indicator. White *et al.* [2002] considered the reading time as a technique to re-rank sentence-based summaries, but the results were inconclusive. Kelly & Teevan [2003] assumed that information-seeking behaviour is not influenced by contextual factors such as topic, task, and collection. Therefore, Kelly & Teevan [2003] performed a study to investigate the relationship between information-seeking task and the display time. Unfortunately, the results cast doubt on the straightforward interpretation of dwell time as an indicator of interest. Seo & Zhang [2000] introduced a method to learn the preference of a user from observing the web-browsing behaviour unobtrusively, and concluded that the approach can improve retrieval performance. However, the adaptation of the interest of a user over a longer period of time has not been taken into account as their search sessions were set up for a short period only.

In addition, Sun *et al.* [2005], Jiang *et al.* [2013] and Dou *et al.* [2007] proposed to mine the query log based on the clickthrough information by the user and integrate this information into context search. So, the Information Retrieval System is able to return the documents that were clicked by the user previously. Basically, clickthrough data is limited to the query that the user has executed into the Information Retrieval System. Then the user opens and clicks the document returned by the Information Retrieval System. Furthermore, Sun *et al.* [2005] applied a dimensional reduction technique which is a pre-processing step on the clickthrough data in order to find the latent semantic links between users, queries, and documents. Moreover, Dou *et al.* [2007], Huang *et al.* [2013] and Gao *et al.* [2011] also added these latent relationships with user-topic and document-topic similarity measures. In addition, Craswell & Szummer [2007] used a bipartite graph to represent the clickthrough data of an image retrieval system where queries and documents are the nodes and links are the ones directly captured in clickthrough data. A random walk is then applied in order to recommend images based on the last query of the user. Furthermore, Aktolga & Allan [2011] proposed a method to boost rarely-clicked queries in the Information Retrieval System by using probabilistic approach to carry out the estimation and then using this result to re-rank the document before Information Retrieval System shows the final ranking to the user.

3.2.4 User Profile

Considering the large amount of personal data that can be captured, most context search rely on user profile. User profiling is the process of learning the interest of a user through a period of time. User profile can be expressed in explicitly or implicitly as mentioned in Sections 3.2.2 and 3.2.3. This section, highlights the basic principles of the user profile, and the state-of-the-art survey is given by Gauch *et al.* [2007], who distinguished between two types of user profiles namely 1) short-term, and 2) long-term profiles. Short-term user profile is used within one session such as any feedback that the user provides during the current information seeking task. Long-term user profile aims to keep track of long-term interests of a user. In general, there are two types of user profile representations namely: 1) weighted keywords, and 2) semantic

networks.

Weighted Keywords: The most popular representation of user interests is the weighted keyword approach as shown by Nanas *et al.* [2010]. The interests are represented as a vector of weighted terms that have either been extracted from those documents that users have shown interest in or that have been provided manually by the users. The weighting indicates the importance of the corresponding term in the user profile. Basically, these can be automatically extracted from Web documents or directly provided by the user. So, the weights which are usually associated with keywords are numerical representation of the user's interest. Then, each keyword can be represented as a topic of interest or keywords can be grouped in categories to reflect a more standard representation of user's interest as shown in Figure 3.1 which was taken from Gauch *et al.* [2007].

Art

0.55	0.67	0.31
Portrait	Sculpture	Watercolour

Sports

0.89	0.25	0.31
Soccer	Bat	Score

Music

0.12	0.88	0.31
Rock	Orchestra	Symphony

Figure 3.1: A Weighted Keywords User Profile

The main disadvantage of this approach is the problem of polysemy, that is the multiple meanings that each word can have. In an early example, Moukas & Maes [1998] proposed *Amalthaea* which is an agent to extract the keywords from the websites and assigned the weight for the keywords based on $tf \cdot idf$. Furthermore, Micarelli *et al.* [2007] also used the same approach to generate the weight based on $tf \cdot idf$. Sakagami & Kamba [1997] proposed *Anatagonomy* which is a personalized online newspapers, Lieberman [1995] proposed *Letizia* which is a browsing assistant, and Pazzani *et al.* [1996] proposed *Syskill & Webert* which is a recommender system that exploits weighted keyword profiles.

Semantic Networks: In the semantic network approach, keywords are replaced with concepts. User interests are represented as weighted nodes of a graph where each node is a concept that is of interest to the user as shown by Parkhomenko *et al.* [2003]. The main difference between Semantic Network and Keywords is that Semantic Network considers the abstract of the topic while Keywords approach considers words to represent the interest of the user. Gentili *et al.* [2003] proposed a filtering system which is *InfoWeb* which user profile is represented as a semantic network. In general, each semantic network contains a collection of unlinked nodes

in which each node represents a concept (also known as "planets"). As more information on the user is gathered, the profile is enriched to include additional weighted keywords associated with the concepts. These keywords are stored in subsidiary nodes (also known as "satellites") which are linked to their associated concept nodes (planets). Figure 3.2 shows an example of a user profile based on semantic networks. Daoud *et al.* [2008] analysed the documents based on the implicit relevance feedback, and then mapped the concepts of these documents with the Open Directory Project (ODP) ontology and stored them in the user profile at the end of each search session. Other contextual search techniques based on ODP includes Sieg *et al.* [2007]. Chirita *et al.* [2005] have shown that incorporating this taxonomy can significantly outperform the technique without context search.

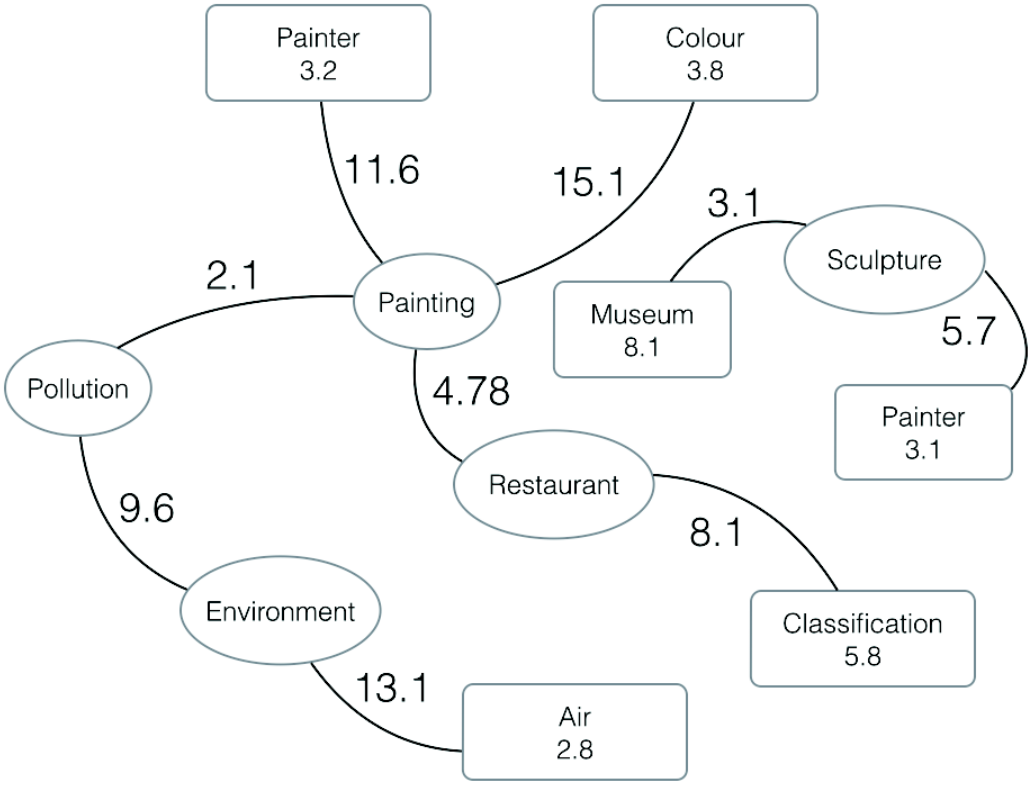


Figure 3.2: A User Profile based on Semantic Networks

3.2.5 Discussion

In the past, implicit feedback approach have been successfully applied to retrieval systems. Furthermore, White *et al.* [2002], and Joachims *et al.* [2005] defined and evaluated several implicit feedback based on text-based retrieval system. Both have shown that implicit feedback is able to obtain a better performance compared to explicit feedback.

In a nutshell, various state-of-the-art user profiling approaches have been surveyed in this section. As mentioned, the most recent research approaches exploit the clickthrough data.

Hence, implicit relevance feedback is employed to determine the user's interests. Kelly & Teevan [2003] also agreed that time is an important factor and carried out several experiments based on implicit feedback. It is agreed that time is the main factor that distinguishes the interest of a user, but most of the approaches are based on the interactions with the system such as reading time, saving, printing, selecting, and clickthrough without quantifying the level of preferences based on the time. Most of the approaches are based on the time spent on a page (longest time), and then the terms are extracted based on the page, and $tf.idf$ is used to calculate the weight of the interest. Although Nichols [1997] also argued that the time spent on a page cannot purely become an indicator to judge the interest of a user, but it is believed that it is good to consider other activities rather than the actions of reading time, saving, printing and clickthrough. Therefore, there is a need to quantify the preference by including other activities to judge the interest of the user rather than purely based on the actions of reading time, saving, printing and clickthrough.

3.3 Information Access - Mobile Game

A range of projects have successfully used mobile technologies to link with games such as *Can You See Me Now?* [1] proposed by Benford *et al.* [2006] who demonstrated that the technologies are able to bridge the real world together with virtual world as shown in Figure 3.3, Figure 3.4 and Figure 3.5. In addition, some researchers also exploited mobile technologies to enhance the experience of teaching and learning such as *MyArtSpace* by [Vavoula *et al.*, 2009] who have shown that even simple mobile technologies can enhance the experience of learning as shown in Figure 3.6 and Figure 3.7. Furthermore, Wilensky & Stroup [2000] also proposed to use mobile technologies to support outdoor learning based on the simulations where the user can play an active role in the simulation of a physical or social system.

In general, mobile technologies are also able to facilitate the inquiry based activities where the user can be engaged in active inquiry by using realistic tools in an authentic environment. Recent work has shown enthusiasm in the education sector for the increased use of Information and Communications Technologies (ICT) to enhance science learning such as McFarlane & Sakellariou [2002], Hennessy *et al.*, Squire & Jan [2007], and Anastopoulou *et al.* [2008]. In addition, Linn [2003] has shown that mobile technologies can help the user to improve the skills of learning and investigating outside the classroom. Therefore, games have been used to engage learners with the system and several projects proposed by Squire & Jan [2007], and Huizenga *et al.* [2009]) have shown that games have the power to motivate learners to take part in learning activity and may provide a suitable support for their activities in the field.

3.3.1 Mobile Learning

For several years, mobile learning has been offering modern ways to support learning process through mobile devices, such as hand-held and desktop computers, MP3 players, smart phones and mobile phones. Mobile learning presents unique attributes compared to conventional e-learning such as personal, portable, collaborative, interactive, contextual and situated.

[1] http://www.blasttheory.co.uk

Figure 3.3: Screenshot of *Can You See Me Now?*: Display in Mobile Device [Benford *et al.*, 2006]

In general, mobile learning emphasizes ”just-in-time-learning” as instruction can be delivered anywhere and anytime. Moreover, it is an aid to hold enormous potential in transforming the delivery of education and training.

Mobile technologies allow contents to be provided to the learner in whatever location, and also allow the location itself to be used as part of interactive learning activities. These new technological capabilities offer the promise of new forms of educational experience situated away from the classroom [Roschelle & Pea, 2002]. In addition, the technologies allow more learners to interact simultaneously with the physical world or digital world through a mobile device. Several researchers namely Rogers Yvonne [2002], Sharples *et al.* [2002], and Cole & Stanton [2003] have shown that the combination of physical activity with digital activity (being able to simultaneously view digital resources is able to promote reflection and new ways of assisting children in learning) as shown in Figure 3.8 and Figure 3.9. Using mobile technologies away from the classroom like this, where the physical environment itself has a meaningful role to play in the learning activity, has been termed as the “physical context” by Frohberg *et al.* [2009] who offered a recent review of state-of-the-art in mobile learning. In addition, Frohberg *et al.* [2009] re-used the framework developed by Taylor *et al.* [2006] and carried a survey on the project of mobile learning and categorized the finding according to a set of meaningful dimensions.

In recent years, researchers and educators such as Savill-Smith [2005], Shiratuddin & Zaibon [2010], and Carr [2010] have shown that game can create a structured activities to support mobile learning. In addition, Shaffer *et al.* [2005] agreed that mobile game has already changed

Figure 3.4: Screenshot of *Can You See Me Now?*: The Location of all the Players [Benford *et al.*, 2006]

the notion of learning from "untapped educational resource" to "give a glimpse of how we might create new and more powerful ways to learn in schools".

3.3.2 Game as Learning

The growing popularity of games and digital technologies has led to exploitation of technologies and modern games in providing a better environment for learning. Facer *et al.* [2004] also highlighted that the school should facilitate the digital technologies in order to help the students to learn outside the classroom.

In general terms, the purpose of using games in education is to help the learners who are in the situation of: 1) disengaged from the learning process, 2) lack of interest or confidence, and 3) lack of self-esteem. Schwabe [2005] identified that digital games are able to encourage the player/learner to experience different ways of learning and thinking. Leutner [1993] highlighted that games which involve some kind of manipulation of the objects can stimulate the learning experience of a user, and Wilson *et al.* [2009] also mentioned that visualization, experimentation, and creative activities also help to enhance the learning experience. Furthermore, Doolittle [1995] has shown that game is one of the effective way to address a particular problem especially in the field of Mathematics and Science, because some concepts can be hard to visualize by the students. Therefore, game can help the student to visualize the concept in digital world. In addition, Barab *et al.* [2005], Prayaga [2005], Blanco *et al.* [2010], Berland *et al.* [2010], Harpstead *et al.* [2013] and Shaffer *et al.* [2005] proposed an innovative strategy game

Figure 3.5: Screenshot of *Can You See Me Now?*: Emily (player) inside the Virtual World [Benford *et al.*, 2006]

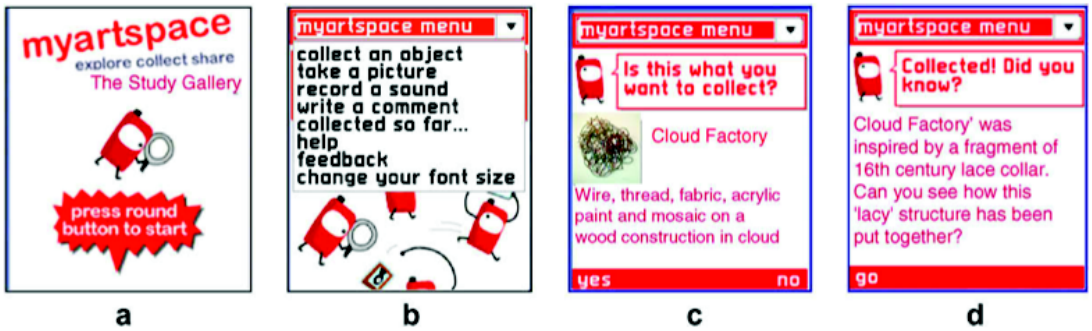


Figure 3.6: Screenshots of *MyArtSpace*: (a) Initial Screen, (b) Main Menu, (c and d) Interfaces of Collecting an Exhibit [Vavoula *et al.*, 2009]

that requires the learners to think, do, and provide a set of tools to guide the learners to achieve the goal of the game in order to help the children to improve their brain development as shown in Figure 3.10 (as proposed by Barab *et al.* [2005]). In addition, Squire & Jan [2007] mentioned that games should consist of several core features that are relevant to learning namely 1) the cycles of making choices, 2) the experience of consequences after the choice, and 3) interpreting the state of the game as shown in specific examples such as the strategy game *Civilization* as shown in Figure 3.11 and Figure 3.12.

Recent advances in technologies, especially the smart phones and tablets have led to more

Figure 3.7: Screenshot of *MyArtSpace*: An Example of Student Gallery [Vavoula *et al.*, 2009]

Figure 3.8: Children Taking Moisture Measurements around the Wood. The Physical Position of the Child Taking Measurements Indicates to the Team-mates that the Activity is Taking Place [Cole & Stanton, 2003]

recent works by using off-the-shelf components with only software required. For example, *Mad*

Figure 3.9: Children are Exploring the Physical Environment by Searching for the Objects through Mobile Device [Cole & Stanton, 2003]

City Mystery proposed by Squire & Jan [2007] who used smart phones to present the place by requiring the students to investigate through a mystery-based game. The goals of the simulation are to help students to develop the skills of investigating and observing by relating them to scientific processes. Students are presented with an open-ended problem (a death) and are able to gather location-based evidence by exploring a physical area with a Global Positioning System (GPS) technology from the smart phone. In addition, Huizenga *et al.* [2009] developed a system *Frequency 1550* which uses standard components to provide an environment to explore local history by using a mystery based game platform similar to *Mad City Mystery* as proposed by Squire & Jan [2007].

3.3.3 Theoretical Foundations

After reviewing the recent works on mobile technologies to support the learning activities, it can be argued that most of the projects are based on the theory of *social constructivism* (proposed by Vygotsky [1978]). The definition of social constructivism according to Vygotsky [1978]) is as follows:

Every function in the child's cultural development appears twice: first, on the social level, and, later on, on the individual level; first, between people (inter psychological) and then inside the child (intra psychological). This applies equally to voluntary attention, to logical memory, and to the formation of concepts. All the higher functions originate as actual relationship between individuals

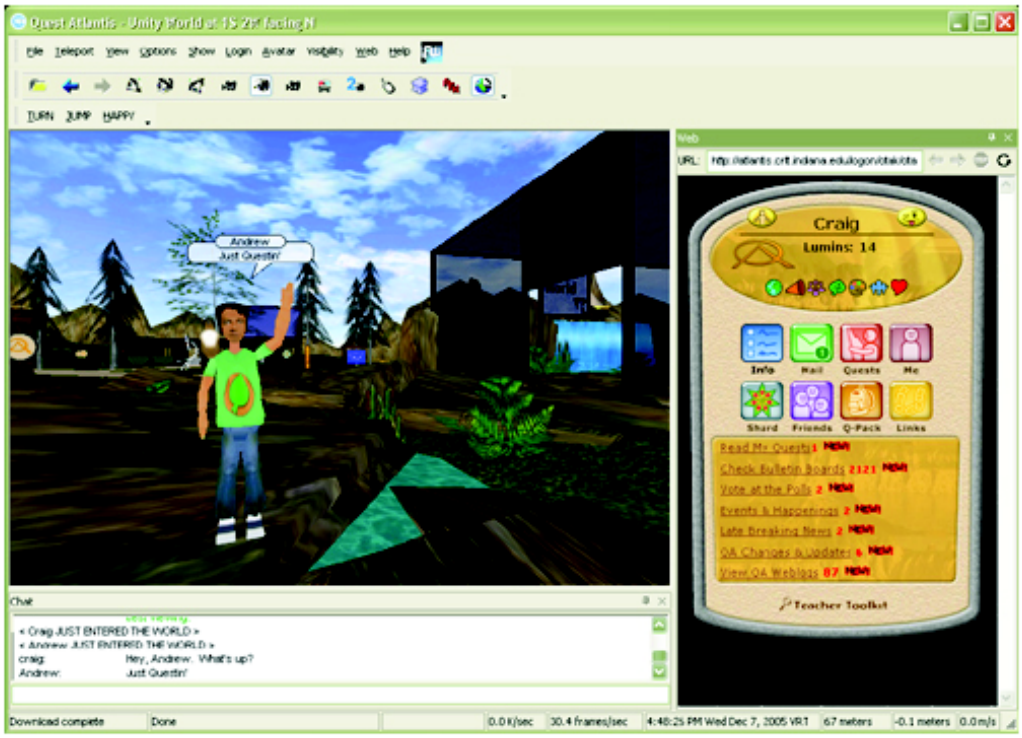


Figure 3.10: Screenshots of *Quest Atlantis*: Showing a Scene from a Village on the Left, the Homepage for a Student on the Right, and the Synchronous Chat Space at the Bottom [Barab *et al.*, 2005]

Actually, this theory was the extension of the theory of constructivism from Lovell [1959] who provided a solid framework about the nature and development of human intelligence. In other words, it deals with the nature of knowledge itself and how humans come gradually to acquire, construct, and use it. In addition, Lovell [1959] described two mechanisms namely 1) accommodation, and 2) assimilation. They form the processes for a learner to build up new knowledge from the experience. Assimilation is the process which the new knowledge is incorporated into existing knowledge structures without modifying the structures while accommodation is the process of re-framing the internal representation in order to fit with new experiences that do not fit into existing knowledge. Several projects proposed by Taylor *et al.* [2006], and Anastopoulou *et al.* [2008] are based on social constructivism. Basically, social constructivism consists of three specific learning approaches namely:

- Inquiry Learning
- Experiential Learning
- Situated Learning

3.3.4 Inquiry Learning

Inquiry learning is an instructional strategy that is used as the basis to design active learning where the students are engaged in some kind of investigation that involves questions and require

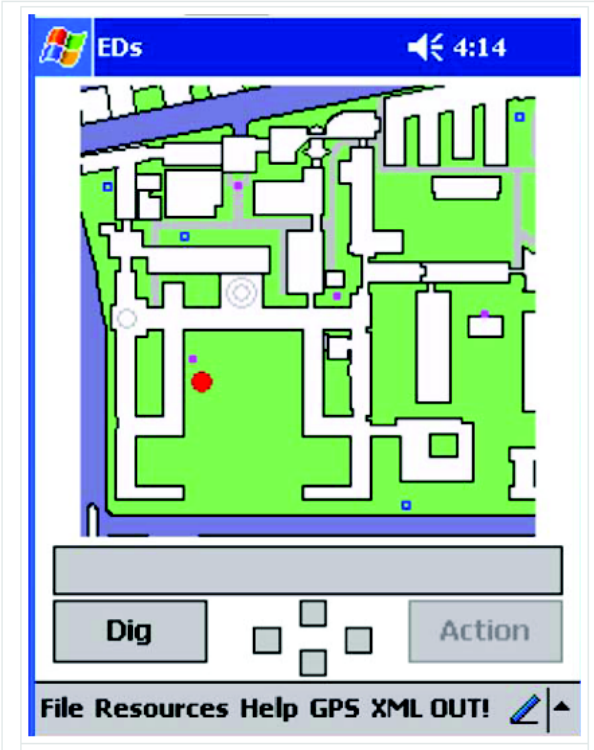


Figure 3.11: Screenshot of *MadCityMystery*: Map Interface, Environmental Detectives [Squire & Jan, 2007]

the student to find the answers based on the questions. In general, Keselman [2003] described that inquiry learning as "an educational activity in which students are placed in the position of scientists gathering knowledge about the world" and several researches such as Rogers *et al.* [2002], and Stamatina Anastopoulou [2008] used this theory to develop the learning content.

3.3.5 Experiential Learning

Experiential learning is viewed as a process to create a new knowledge through transformation of the experience. In other words, learners are able to link with new concepts based on past experience. In addition, experiential learning also emphasizes the use of tangible learning concepts that learners encounter and able to engage it directly. As mentioned before, Lovell [1959] demonstrated that effective learning requires an environment where learners can have appropriate *experience*. Experiential learning further emphasizes the role of environments and experiences in the process of learning such as providing an environment to let the learner encounters the problem rather than learns the solution based on others' experience. This means that the learner is able to understand and discover the solution based on their previous experience.

In general, several projects have been developed by using experiential learning in mobile devices such as *Mad City Mystery* (proposed by Squire & Jan [2007]) and *Frequency 1550* (proposed by Huizenga *et al.* [2009]). The projects developed a learning activity that allows learners to experience directly a simulation of a physical system and creates a direct link between the

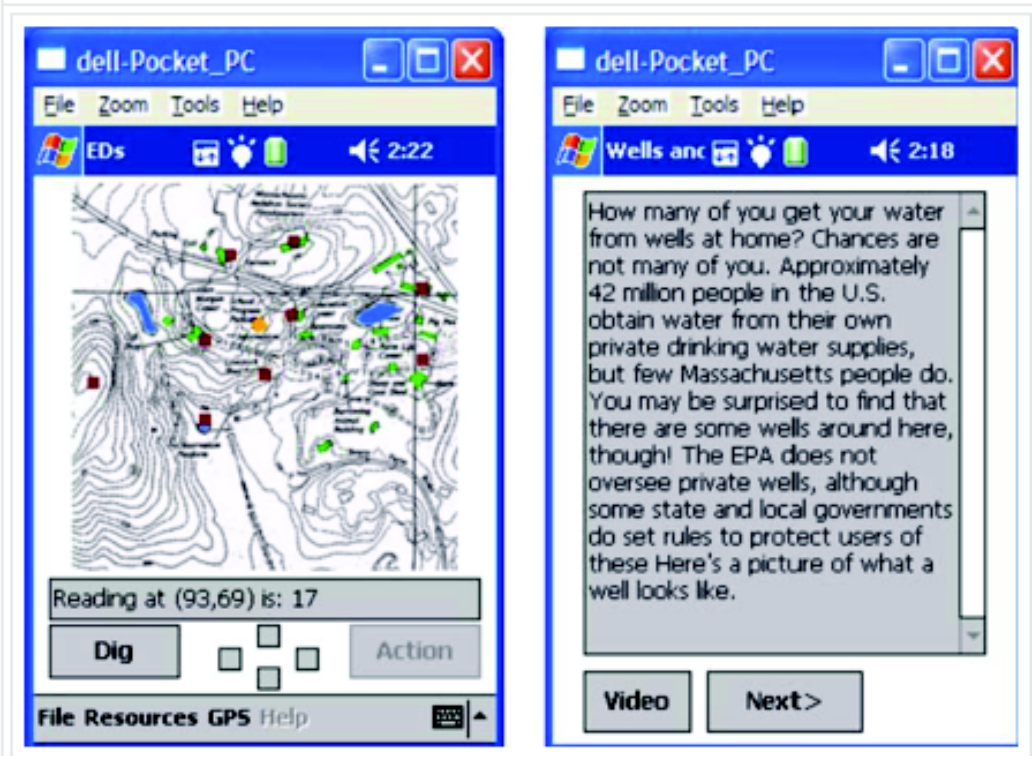


Figure 3.12: Screenshots of *MadCityMystery*: Map and Information [Squire & Jan, 2007]

personal experience in physical space with the underlying rules that govern the underlying simulation.

3.3.6 Situated Learning

Situated learning can be viewed as a complement to the experiential learning. Situated learning emphasizes the role of exploratory spaces and practical activities to enhance the learning process. In other words, situated learning focuses more on the implementation of authentic activities rather than innovative learning practice.

Mobile devices are seen as highly effective in bridging the real world and learning activities. They can be used to enhance the learning experience by carrying knowledge from the classroom into the real word as shown by Falk & Dierking [2000]. In addition, Vavoula *et al.* [2009] proposed a system *Myartspace* to bridge the gap between the classroom and real life learning situation while Sharples [2006] has shown that several projects such as Mobilearn [1] which is a European research centre under the European Commission to explore situated learning by using key advances in mobile technologies.

[1] http://www.mobilearn.org

3.3.7 Discussion

In a nutshell, the field of ”mobile learning” and ”game as learning” have been reviewed and both fields are co-related. In addition, it has also been shown that several projects combine the elements from mobile technologies and game in order to support and provide a better learning environment for the learner. In addition, there is a need to transform the way of teaching (abstract knowledge delivered in the classroom) into something that relies more on ”doing” and ”experiencing”, and the best way to achieve this goal is by using the combination of mobile technologies and games to encourage the learners to gain the new knowledge. This research uses the theory of social constructivism to model a mobile game which consists of inquiry learning, experiential learning, and situated learning that will form a powerful learning model. Therefore, there is a need to investigate whether the mobile game can help the learners to learn the information on cultural heritage.

3.4 Summary

Several methods have been discussed in Section 3.1 to solve the problem of term mismatch namely relevance feedback, query expansion, query reformulation, dimension reduction, and statistical translation model. An important perquisite for context search is to identify the user’s preference and several methods (how to gather and represent the preference) have been discussed in Section 3.2. Finally, Section 3.3 discusses the mobile game as a learning platform.

Part IV

PROPOSED SOLUTIONS

Chapter 4

Proposed Solutions

This chapter presents the details of the proposed solutions of this research. Section 4.1 describes the proposed solution for information access in Information Retrieval System which includes the proposed framework and proposed models which are Extended Dirichlet Smoothing with Increase of Document Size (EDS+) and Extended Dirichlet Smoothing with Decrease of Document Size (EDS-). Section 4.2 describes the proposed method to quantify the preferences of a user by using the analogy of charging a capacitor, while Section 4.3 describes the overall idea and the framework of the mobile game (M-Heritage Hunt). As mentioned before, Section 4.2 will use the result generated from Section 4.1 to do the context search while Section 4.3 is to carry out a case study involving a mobile game on cultural heritage.

4.1 Information Access - Information Retrieval System

In the past, a number of Information Retrieval models such as Vector Space Model (VSM) [Salton, 1991], [Salton & Buckley, 1988], Probabilistic Model (Okapi, BM25) as mentioned in [Robertson, 1997], [Amati & van Rijsbergen, 2002] and Language Model [Ponte & Croft, 1998]; [Zhai & Lafferty, 2004] were proposed which are based on term intersection approach. Term intersection is an approach in which both document and query should share the same terms. Although this approach provides a good result in terms of speed and accuracy, it may cause a problem - *term mismatch*. The term "term mismatch" is defined as a mismatch between the term in the query and the term in the document.

For example, a user is searching for the information about a "schlesien [1] map" and submits the query:

$$q = (schlesien, map)$$

and an Information Retrieval System considers the following documents:

$d_1 = (map, germany)$
$d_2 = (china, map)$
$d_3 = (germany, silesia)$

[1] Silesia (English) or Schlesien (German) is a region of Central Europe located mostly in Poland, with smaller parts in the Czech Republic and Germany.

Table 4.1: Mean Average Precision by different Information Retrieval Models in CHiC Collection

Types of Retrieval Model	**Mean Average Precision(MAP)**
Vector Space Model	0.2933
Probabilistic Model (Okapi)	0.3902
Language Model based on Dirichlet Smoothing	**0.4929**
Language Model based on Jelinek-Mercer Smoothing	0.4918

Basically, the Information Retrieval System assigns a very similar Retrieval Status Value (RSV) to d_1 and d_2 which is highly dependent on the indexing weights because these documents contain the similar terms as the query which is the "map". However, we know that d_2 is surely not relevant since d_2 contains the information on "china map" and not the information of "schlesien map". In addition, we can attest that d_3 is more relevant than d_2 based on the needs of the user.

Several important techniques have been proposed to tackle this problem such as 1) pseudo relevance feedback, 2) exploiting the term similarity matrix in Vector Space Model, and 3) statistical translation models (Refer the techniques in Section 3.1 for further information).

As mentioned earlier, our main goal in this research is to incorporate the term links into the Language Model based on Dirichlet Smoothing in order to solve the problem of term mismatch. The main reason to choose the Language Model with Dirichlet Smoothing is because Language Model outperforms the rest of the retrieval models such as Vector Space Model and Probabilistic Models with the same setting in the evaluation challenge (Cultural Heritage in CLEF 2012) as shown by Petras *et al.* [2012]. In addition, Table 4.1 shows that Language Model with Dirichlet Smoothing outperforms the rest of the retrieval models in the experiment (default setting) and this is the main motivation to choose Language Model based on Dirichlet Smoothing in this research.

After the reviewing Crestani [2000], Karimzadehgan & Zhai [2010], Zhai & Lafferty [2004] and considering the problems, the following approaches are proposed:

- Use the maximum or highest value instead of the total value from the term similarity between the terms from query with the terms from document. In addition, only consider the closest semantic terms from the documents if a term in the document from the query cannot be found.
- Use statistical approach to generate the knowledge rather than the probability approach in order to avoid the value of $P(w|u)$ which is higher than $P(w|w)$ for a term w as obtained by Karimzadehgan & Zhai [2010].
- Use w to represent the term because most of the Language Models use w to represent the term rather than t as proposed by Crestani [2000].

To do so, an overview of the proposed framework (EDS framework) of this research to solve the problem of term mismatch is presented in Figure 4.1. The process of pre-processing is a standard process in Information Retrieval which covers remove all the stopwords, remove all non-character, convert all upper case to lower case, and apply the Porter stemming. In general,

the framework from Language Model based on Dirichlet Smoothing is extended (Extended Dirichlet Smoothing (EDS)) to exploit the non-matching terms during retrieval. The proposed framework (EDS) is able to accept any terms such as w_1 and w_2 from the knowledge base during retrieval. In this research, a matrix is used to represent the knowledge by storing a list of pair terms from any kind of external resources which is computed by various types of statistical approach. The main difference between the proposed framework (Figure 2.5, query expansion) and Figure 4.1 is that in Figure 4.1 the terms from external resources are exploited into the matching function before the process of ranking while in Figure 2.5 the terms are exploited from external resources after the process of ranking and another round of matching process is computed before the final ranking of documents. In other words, pseudo relevance feedback and query expansion with external evidence (proposed by Yin *et al.* [2009]) need two retrieval processes per query. The first is to expand the query, and the second is to retrieve the results for the expanded version. In addition, the framework as shown in Figure 2.5 is considered as global adaptation while the proposed framework is considered as local adaptation. The advantages of the EDS framework are as follows:

- It enhances the classical models where the classical models do not take into account of non-matching terms.
- The EDS framework exploits the terms during the process of matching.

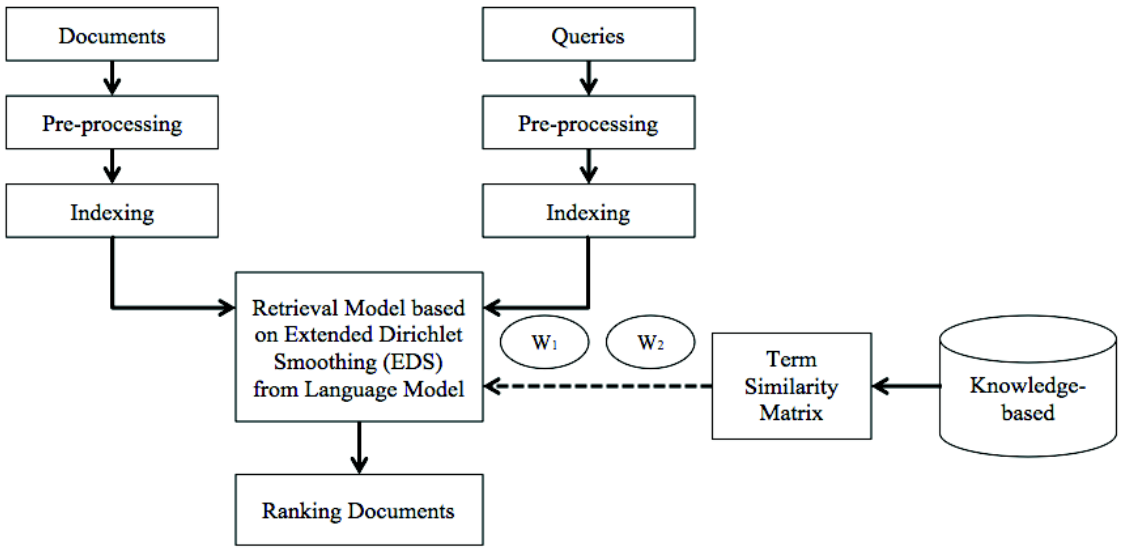


Figure 4.1: System Architecture to Exploit Non-matching Terms during Retrieval

Based on the EDS framework, two models are derived which are Extended Dirichlet Smoothing with Increase of Document Size (EDS+) and Extended Dirichlet Smoothing with Decrease of Document Size (EDS-). Before these two models are explained, it is necessary to explain the ways to find the links between the two terms or in other words to extract the term links in the next section.

4.1.1 Extracting Term Links

A similarity measure between pairs of terms is necessary in order to let EDS framework to exploit the non-matching terms during the retrieval time as shown in Figure 4.1. In general, several studies have been carried out to measure the similarity such as clustering, thesauri construction, and hierarchical structuring based on semantic similarity in the domain of cultural heritage as proposed by Wang *et al.* [2013]. Semantic similarity is a metric over a set of documents or terms within the term lists, where the idea of distance between terms is based on the likeness of their meaning or semantic content. The two main approaches to determine the semantic similarity between a pair of text are 1) knowledge-based, and 2) corpus-based. Knowledge-based refers to the use of algorithms to measure similarity on structured knowledge resources such as thesauri, dictionaries, and semantic networks. Corpus-based refers to the use of information about occurrences of terms in collection of documents or corpus. The terms that occur in a document are used as features to represent it and approaches used to compare pairs of documents.

Knowledge-based approach is a techniques to compute the similarity between pairs of terms. The aim is to assign high values to pairs of words with similar meaning, such as "graveyard" and "cemetery", and low scores to dissimilar pairs such as "wizard" and "coffee". A wide range of methods for determining the similarity between terms has been proposed by Aletras *et al.* [2013]. Basically, these algorithms have often made use of WordNet [Bentivogli *et al.*, 2004] and Wikipedia [Fernando *et al.*, 2012].

Corpus-based approach is based on the assumption that the similarity between documents can be measured by observing the number of terms they have in common. A widely used approach is to represent each document as a vector consisting of the terms and to estimate the similarity between documents by computing the cosine of the angle between their vectors. Thus, if two documents are identical, the cosine value of their vectors is 1. If they share no common terms the cosine value is 0. This approach can be improved by taking into account of the frequency of terms, both in each document and in the entire corpus, and using these values to weight the vector.

Basically, the EDS framework is able to integrate with knowledge-based or corpus-based approaches. In this research, corpus-based approach is used to generate the term links where each pair of terms in the term space contains a real value which estimates how semantically close between the terms. Therefore, the links between all the terms in the collection named as V which is a set of vocabulary needs to be found, w refers to the term from query, and w' refers to the term from document.

$$w, w' \in V, 0 \leq Sim(w, w') \leq 1 \tag{4.1}$$

1. $Sim(w, w') = 0$, there is no link between the terms w and w'
2. $Sim(w, w') < 1$, there is a link between the terms w and w'
3. $Sim(w, w') = 1$, there is an equal match between the terms w and w'

Based on Equation (4.1), $Sim(w, w') = 0$ means the term w does not contain any semantic link with the term w', while $Sim(w, w') < 1$ shows that w contains semantic link with the w'. If the value of $Sim(w, w') = 1$ it means that there is an exact match between the term w and w'.

In this section, an easier and lightweight (linear fashion) way to generate the knowledge compared to Expected Mutual Information Measure (EMIM) proposed by Crestani [2000]. Although mutual information is the best to be extracted for semantic similarity information, mutual information is the most expensive in terms of computation (quadratic complexity, refer to Equation 4.2) according to Markines *et al.* [2009]. In other words, mutual information computes all the possible combinations of attributes pairs. In information theory, $EMIM(w_i, w_j)$ is often interpreted as a measure of the statistical information contained in w_i about w_j (or vice versa). The EMIM measure is defined as follows:

$$EMIM(w_i, w_j) = \sum_{w_i, w_j} P(w_i \in d, w_j \in d) log \frac{P(w_i \in d, w_j \in d)}{P(w_i \in d)P(w_j \in d)} \tag{4.2}$$

where w_i and w_j are any two terms of the term space T, d represents the document, and P represents the probability.

In the corpus-based approach, similarity between terms can be computed in a variety of ways. In the proposed approach, Confidence Coefficient (CC), Tanimoto Similarity (TS), Dice Coefficient (DC), Cosine Similarity (CS) and Overlap Coefficient (OC) are used to generate the statistical information as stated by Rijsbergen [1979]. The CC between terms w_i and w_j are calculated as follows:

$$Sim_{conf}(w_i, w_j) = \frac{c(w_i \cap w_j)}{c(w_i)} \tag{4.3}$$

where $c(w_i)$ is the number of terms (w_i) in the corpus, and $c(w_i \cap w_j)$ is the number of terms that term w_i co-occurs together with w_j in the corpus.

The TS between terms w_i and w_j is calculated as follows:

$$Sim_{tani}(w_i, w_j) = \frac{c(w_i \cap w_j)}{c(w_i) + c(w_j) - c(w_i \cap w_j)} \tag{4.4}$$

The DC between terms w_i and w_j is calculated as follows:

$$Sim_{dice}(w_i, w_j) = \frac{2c|w_i \cap w_j|}{c(w_i) + c(w_j)} \tag{4.5}$$

The CS between terms w_i and w_j is presented using a dot product and magnitude as follows:

$$Sim_{cosine}(w_i, w_j) = \sqrt{\frac{c(w_i \cap w_j)}{c(w_i).n(w_j)}} \tag{4.6}$$

The OC between terms w_i and w_j is calculated as follows:

$$Sim_{over}(w_i, w_j) = \frac{c(w_i \cap w_j)}{min(c(w_i), c(w_j))} \tag{4.7}$$

The reason to choose CC, TS, DC, CS and OC is because all of the approaches are a linear approach through overlapping attributes of the two terms and based on the experiment by Srinivas *et al.* [2010] who showed that Dice Coefficient and Cosine Similarity are the best corpus

based measurement. In a nutshell, this research decided to use CC, TS, DC, CS, and OC to generate the matrix (also known as Term Similarity Matrix) instead of EMIM.

Several types of external knowledge based can be used to extract the term links such as 1) Library of Congress Subject Headings [1] (LCSH), 2) WordNet [2], 3) Wikipedia [3], 4) DBpedia [4], 5) Wikipedia Miner [5], and 6) ConceptNet5 [6]. In general, DBpedia and Wikipedia Miner are using the same source from Wikipedia, but with the different purpose. Short description are given for each knowledge based are shown below;

- **LCSH:** LCSH provides a controlled vocabulary of keywords (or subject headings) which are widely used in libraries to document the materials and facilitate the information access.

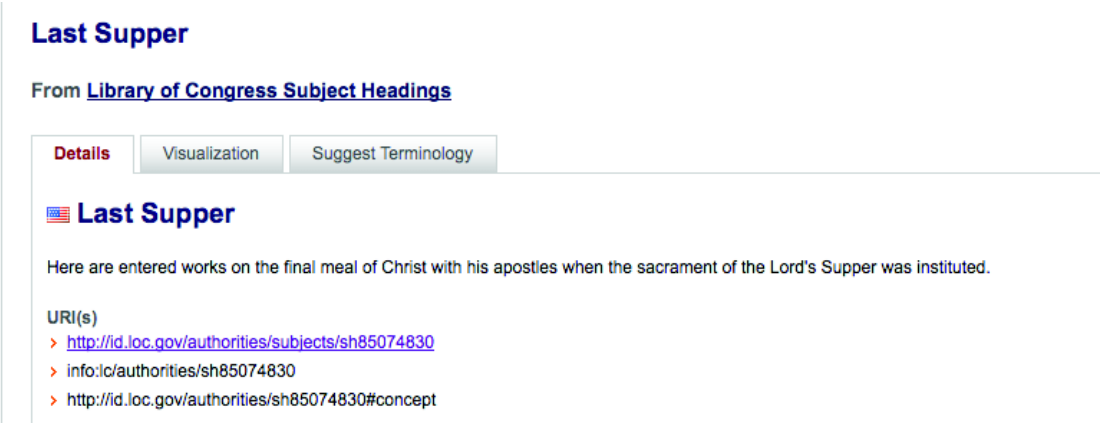


Figure 4.2: The terms “Last Supper” in Library of Congress Subject Headings

- **WordNet:** WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. In addition, synsets are interlinked by means of conceptual-semantic and lexical relations.
- **Wikipedia:** Wikipedia is a collaboratively edited, multilingual, and free Internet encyclopaedia which contains 30 million articles in 287 languages. In addition, Wikipedia is the largest and most popular general reference work on the Internet.
- **DBpedia:** DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia. Furthermore, DBpedia allows the use of sophisticated queries against Wikipedia, and is able to link the different datasets on the Web to Wikipedia data.
- **Wikipedia Miner:** Wikipedia Miner is a toolkit for tapping the rich semantics encoded within Wikipedia. In addition, Wikipedia Miner is a Wikification tool which adds inline links to Wikipedia articles into free text.

[1] http://www.loc.gov/aba/cataloging/subject
[2] http://wordnet.princeton.edu
[3] http://www.wikipedia.org
[4] http://wiki.dbpedia.org/
[5] http://wikipedia-miner.cms.waikato.ac.nz
[6] http://conceptnet5.media.mit.edu

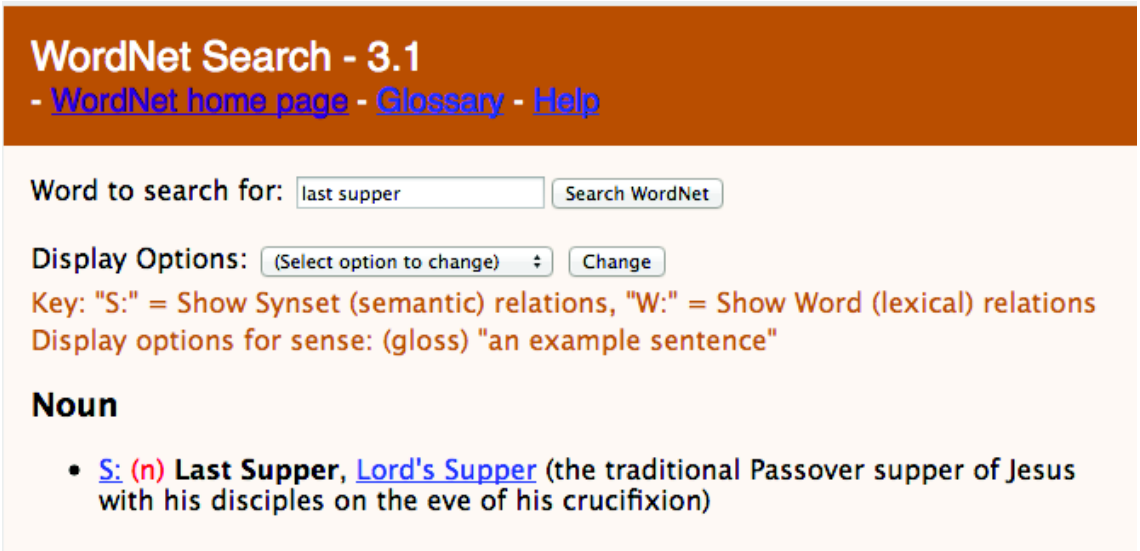


Figure 4.3: The terms “Last Supper” in WordNet

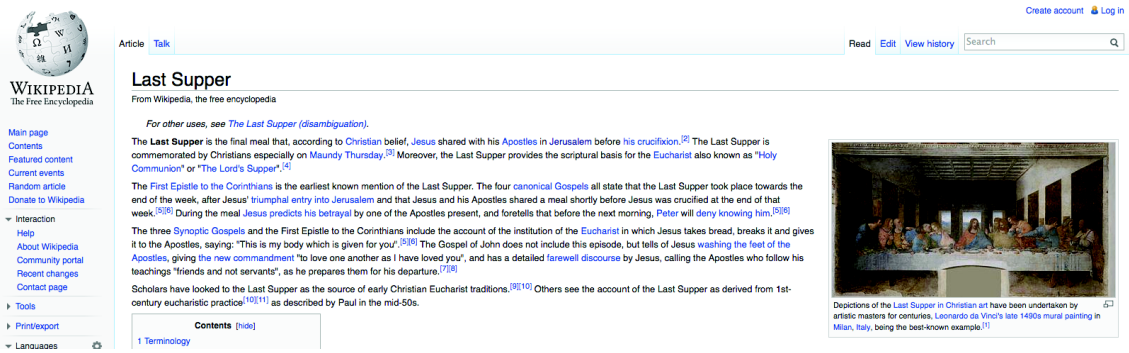


Figure 4.4: The terms “Last Supper” in Wikipedia

- **ConceptNet5:** ConceptNet5 is a semantic network which built from nodes that represent the concepts in the form of words or short phrases of natural language, and labelled relationship between them.

After considering the several types of external knowledge bases, it is decided that Wikipedia is used as the source to extract the term links since DBpedia and Wikipedia Miner are using the same source from Wikipedia. As mentioned before, the goal is to find the links of the term and the term can act as widely as the term can. Figure 4.2, Figure 4.3, Figure 4.4 and Figure 4.5 show the terms “Last Supper” in various types of external knowledge-based, and Wikipedia is one of the choices to provide the most critical idea for the terms “Last Supper”. Therefore, it is decided that CC, TS, DC, CS, and OC are to be used to generate the matrix (also known as Term Similarity Matrix) from Wikipedia as the external knowledge-based. Although the effectiveness of the proposed framework presented in this research depends very much on the quality of the similarity measure, the proposed framework could make use of any available similarity information of the term space.

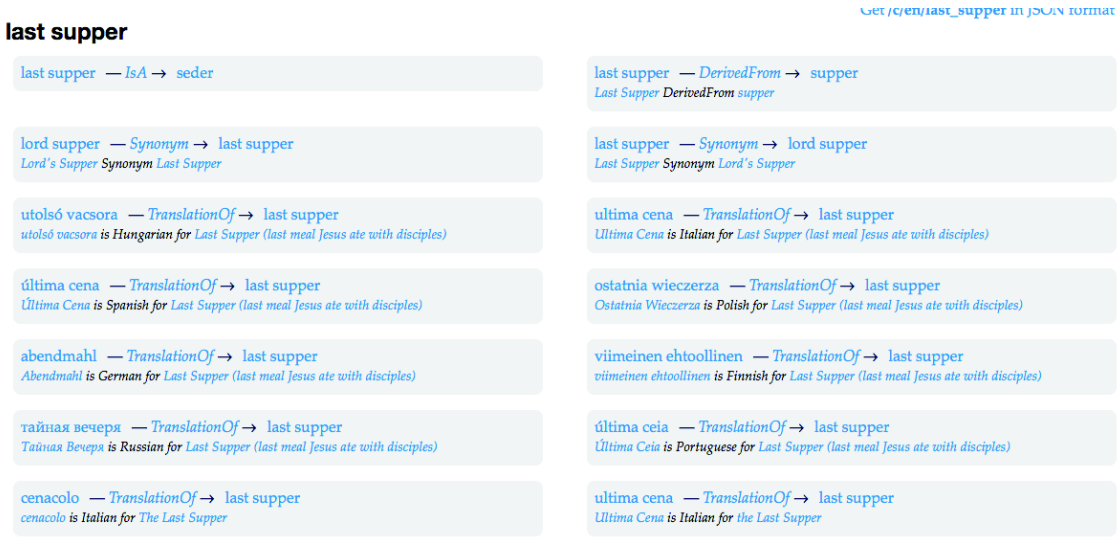


Figure 4.5: The terms "Last Supper" in ConceptNet5

4.1.2 Extended Dirichlet Smoothing

The main idea of the EDS framework is to integrate the Term Similarity Matrix which represents the knowledge into the current Dirichlet Smoothing formula in order to solve the problem of term mismatch. Let's use a document which contains several terms such as "A", "B", and "C" as an example to describe the overall idea of Extended Dirichlet Smoothing as shown in Figure 4.6(a). Basically, the document is dipping inside a basket with full knowledge which contains several terms such as "D", "E", "F", "G", "H", and "I". If the query is "F" and "D", then this document will not be returned by the retrieval model by default. If there is a link between "C" with "F" and "B" with "D" as shown in Figure 4.6(b), then EDS will change the length of the document (red-dashed line) by considering that "F" and "D" can be inside as shown in Figure 4.7(a) or replacing the terms in the document as shown in Figure 4.7(b). With this notion, two types of behaviour will occur in the size of the document namely 1) Increase and 2) Decrease. Increase indicates that extra terms ("D" and "F") occur inside the document, while Decrease indicates that the extra terms ("D" and "F") can be replaced by the terms "B" and "C". Therefore, the length of the document can be increased (Figure 4.7(a)) or decreased (Figure 4.7(b)).

Figure 4.8(a) shows the normal maximum likelihood estimator while Figure 4.8(b) shows the Dirichlet Smoothing. Basically, the Dirichlet Smoothing is derived from maximum likelihood estimator by adding an extra pseudo term frequency (refer to (1) in the Figure 4.8) and normalizing it by μ (refer to (2) in the Figure 4.8). In other words, EDS acts as another smoothing (α as an example, refer to (3) in Figure 4.8) and normalized by α (refer to (4) in the Figure 4.8) in the Dirichlet Smoothing as shown in Figure 4.8(c). The value of α refers to the value from Term Similarity Matrix. Since it can integrate the term links inside the Document, then it needs to normalize the length of the Document accordingly.

Basically, query expansion and document expansion are able to solve the problem of term mismatch in which query expansion requires integration of the knowledge from Query to Query* before the matching process with other Documents as shown in Figure 4.9. This scenario is the

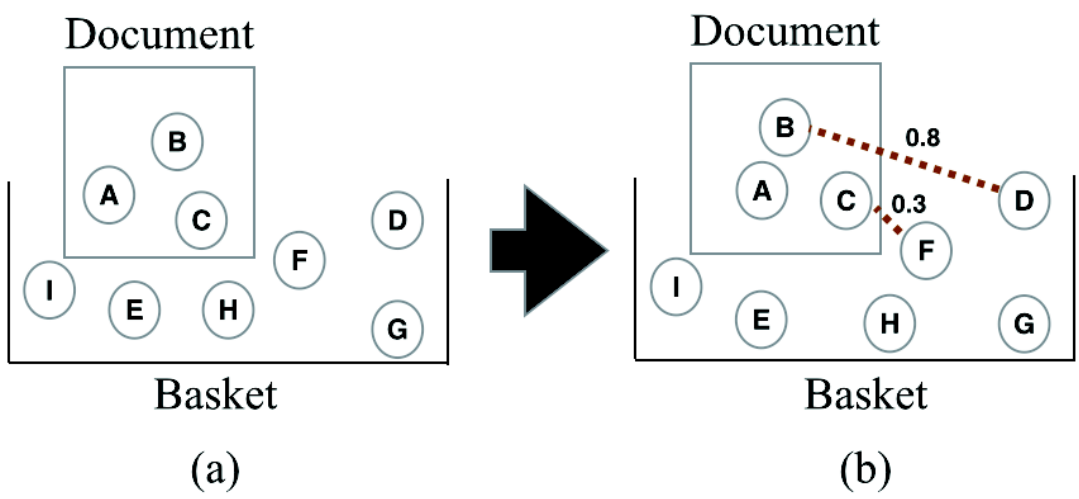


Figure 4.6: Graphical Interpretation of Extended Dirichlet Smoothing

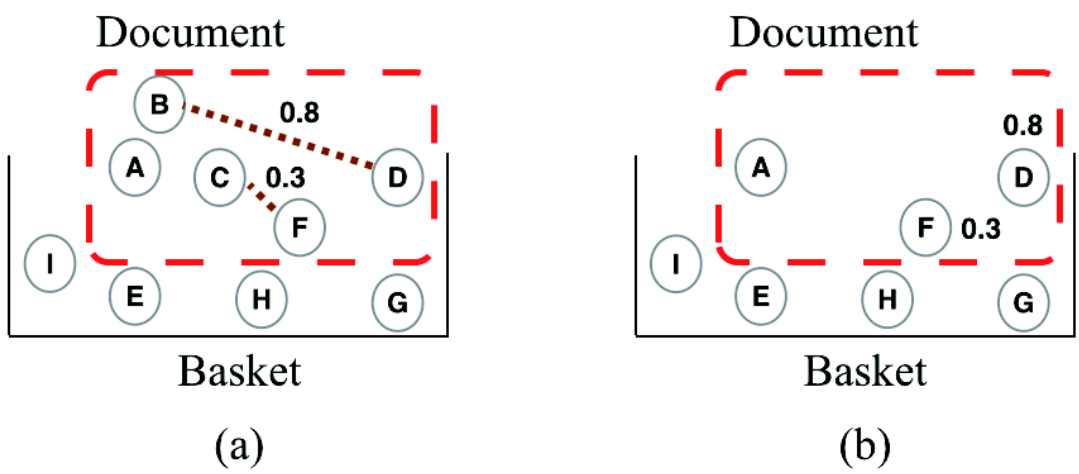


Figure 4.7: Graphical Interpretation of the Document Size: (a) Increase and (b) Decrease

same with Document Expansion because it also requires integration of the knowledge from Documents to Documents* before the matching process with other Query as shown in Figure 4.10. Basically, the proposed framework directly integrates the knowledge during the matching process as shown in Figure 4.11 while Query Expansion and Document Expansion requires the integration of knowledge in the second steps before the matching. In addition, Query Expansion and Document Expansion are considered as global change as shown in Figure 4.12, while the proposed framework is considered as local change as shown in Figure 4.13. As mentioned in the example of Section 4.1 (a user is searching for the information about a "schlesien map"), if there is a knowledge which contains the semantic links between the term "silesia" and "schlesien", then the proposed framework is able to exploit these links during the retrieval and retrieve d_3 based on the needs of the user.

Based on the EDS framework, two models are proposed namely Extended Dirichlet Smooth-

$$P_{ml}(w|d) = \frac{c(w;d)}{|d|}$$

Maximum Likelihood Estimator (a)

$$P_{\mu}(w|d) = \frac{c(w;d) + \mu P(w|C)}{|d| + \mu} \quad (1) \quad (2)$$

Dirichlet Smoothing (b)

$$P_{\mu}(w|d) = \frac{c(w;d).\alpha + \mu P(w|C)}{|d| + \mu + \alpha} \quad (3) \quad (4)$$

Extended Dirichlet Smoothing (c)

Figure 4.8: Evolution from Maximum Likelihood Estimator to Extended Dirichlet Smoothing

ing with Increase of Document Size (EDS+) and Extended Dirichlet Smoothing with Decrease of Document Size (EDS-). Both models are able to exploit the non-matching terms during retrieval before the process of ranking.

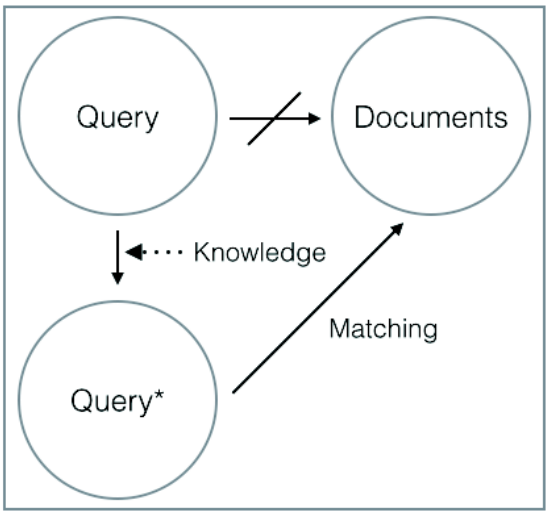


Figure 4.9: Query Expansion

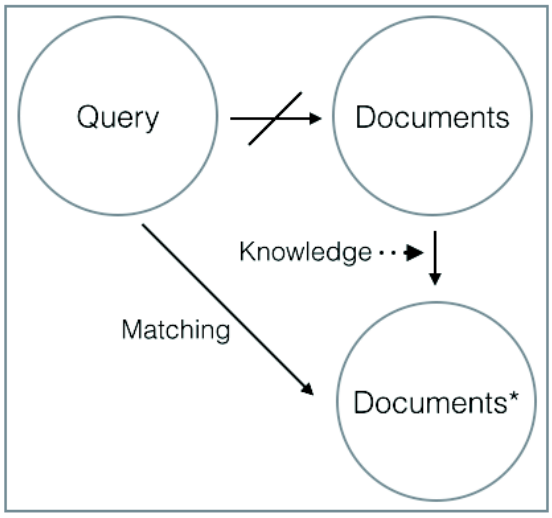


Figure 4.10: Document Expansion

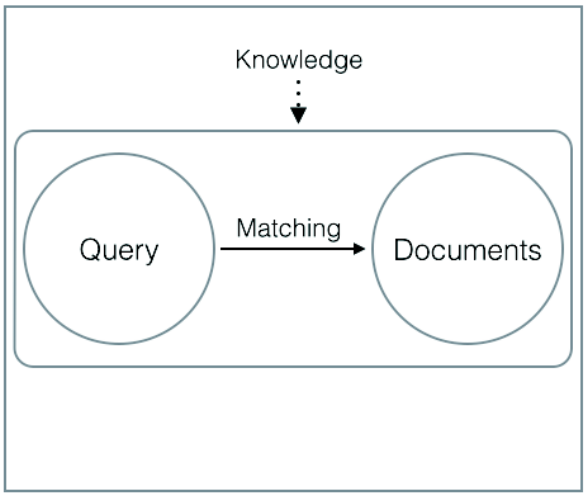


Figure 4.11: Proposed Framework

4.1.3 Extended Dirichlet Smoothing with Increase of Document Size (EDS+)

Firstly, we need to assume that a term $w' \in d$ can play the role of w where $w \in q$ during the matching process. More specifically, it is considered that if w does not occur in the initial document d but occurs in the *document* $d_{increase}$, which is the result of increasing d according to the query and some knowledge [1], then the probability of the term w' is defined according to the size of increasing the document $d_{increse}$.

Basically, the knowledge can be represented either in symmetrical or non-symmetrical similarity function. In the example, the knowledge is assumed to form a symmetrical similarity function which is $Sim : V \times V \rightarrow [0, 1]$, that denotes the strength of the similarity between

[1] The knowledge refers to the Term Similarity Matrix

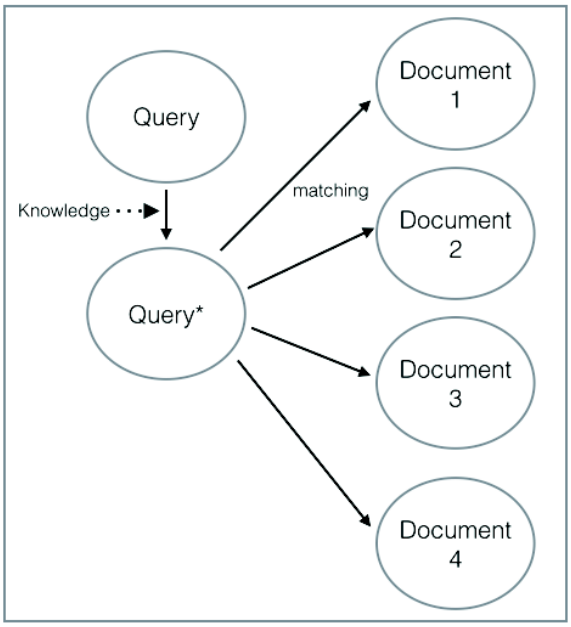


Figure 4.12: Global Change

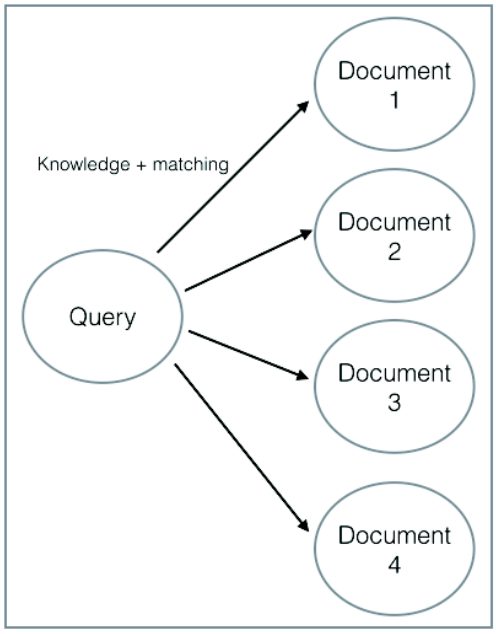


Figure 4.13: Local Change

the two terms from the vocabulary (the larger the value, the higher the strength). It is proposed that: $\forall w, w' \in V, Sim(w, w') = 1$ if there is exact match between w with w', and $\forall w, w' \in V, Sim(w, w') = 0$ if w does not contain any link with w'.

In order to avoid any complex extensions (see Section 4.1.1), the following constraints are defined :

- One query term w, must only impact occurrences in one document term w'.
- Only the maximum value of the term that can be found in the knowledge is to be used to exploit the non-matching terms during the retrieval.

To achieve this, some simple and sensible heuristics are used:

1. If a query term w, occurs in a document d, then the term will not change the length of the document.
2. If a query term w, does not occur in a document d, but the term w contains a link with w' (term from document), then w^* (please take note that $*$ is a function) is defined as:

$$w^* = \underset{w' \in d}{argmax}\ (Sim(w, w')) \tag{4.8}$$

The term from the document will serve as the basic count of the pseudo occurrences of w in d as:

$$c(w^*; d).Sim(w^*, w) \tag{4.9}$$

This pseudo occurrences of the term w^* are then included and increase the size of the document.

Eventually, using the usual set of notations for the terms that occur in the document and the query, then the new length of the document ($|d_{increase}|$) is: $|d_{increase}| = \sum_{w \in d \cap q} c(w; d) + \sum_{w \in q} c(w^*; d).Sim(w^*, w) + \sum_{w \in d \setminus q} c(w; d)$

Using the fact above, the expression of $|d_{increase}|$ can be easily simplified into:

$$|d_{increase}| = |d| + \sum_{w \in q} c(w^*; d).Sim(w^*, w) \tag{4.10}$$

Note that the idea is to increase the size of the document according to the query. With all the elements described above, the Extended Dirichlet Smoothing with Increase of Document Size (EDS+) model leads to the following probability for the term w, of the vocabulary V in the document $d_{increase}$ according to a query q. Note that $P_\mu(w|d_{increase})$[1] is defined as:

1. if $w \in d \cap q$:

$$P_\mu(w|d_{increase}) = \frac{c(w; d) + \mu P(w|C)}{|d_{increase}| + \mu} \tag{4.11}$$

2. $w \notin d \cap q$ and if $\exists w^* \in d \setminus q; Sim(w, w^*) \neq 0$:

$$P_\mu(w|d_{increase}) = \frac{c(w^*; d).Sim(w, w^*) + \mu P(w^*|C)}{|d_{increase}| + \mu} \tag{4.12}$$

with $w^* = argmax_{w' \in d} Sim(w, w')$.

[1] P refers to probability

3. $w \notin d \cap q$ and if $\nexists w^* \in d \setminus q; Sim(w, w^*) \neq 0$

$$P_\mu(w|d_{increase}) = \frac{c(w;d) + \mu P(w|C)}{|d_{increase}| + \mu} \tag{4.13}$$

with $w^* = argmax_{w' \in d} Sim(w, w')$.

In a specific case when all the query terms from q occur in the document d, the first case from the above is used where $|d_{increase}| = |d|$ leads to $P_\mu(w|d) = P_\mu(w|d_{increase})$.

4.1.4 Extended Dirichlet Smoothing with Decrease of Document Size (EDS-)

Firstly, it is assumed that a term, $w' \in d$ can play the role of w where $w \in q$ during the process of matching. More specifically, it is only considered that if w does not occur in the initial document d, but w have the links with w' which occurs in the Term Similarity Matrix, then it is considered as a match with the value lower than 1.

The main difference between EDS+ and EDS- is that EDS+ is assumed to add additional terms into the original documents in order to have exact match between query with document. So, the current length of the document will be more than the original length of the document. EDS- assumes that there is a term from the knowledge that is able to replace the term in the original document. So, the weight for the replacement term should be lower than "1" since it is not an exact match. Therefore, the current length of the document will become less than the original length of the document.

To achieve this goal, some simple and sensible heuristics are used as follows:

1. If a query term w occurs in a document d, then the term will not change the length of the document. It means that the document will maintain the same length.
2. If a query term w does not occur in a document d but the term w contains a link with w' (term from document), then w^* is defined as the same in EDS+ (refer to Equation (4.8)).

   The term from the document will serve as the basic count of the pseudo occurrences of w in d as the same in EDS+ (refer to Equation (4.9)).

   This pseudo occurrences of the term w^* is then included in the length of the document. In other words, instead of the size of the document increasing, the length of the document is decreased based on the value of $c(w^*; d).Sim(w^*, w)$.

Then, the new length of the document ($|d_{decrease}|$) is defined as:

$$|d_{decrease}| = |d| - \sum_{w \in q} c(w^*; d).(1 - Sim(w^*, w)) \tag{4.14}$$

Note that the proposed model which is EDS- is to decrease the size of the document according to the term of the same query. With all the elements described above, the EDS- leads to the following probability for the term w of the vocabulary V in the document $d_{decrease}$ according to a query q. Therefore, the only difference in term of the equation is to replace the $|d_{increase}|$ to $|d_{decrease}|$ based on Equation 4.11, Equation 4.12, and Equation 4.13 .

4.1.5 Discussion

In a nutshell, the proposed framework is able to integrate the knowledge into the retrieval model and exploit the non-matching terms during retrieval. In addition, two models are proposed namely 1) EDS+, and 2) EDS-. Both proposed models are based on the proposed framework and the main difference between these two models is summarized as follows:

- EDS+ increases the length of the document as shown in Figure 4.7(a). EDS+ is assumed to increase the document if the term from the query has a link with the term in Term Similarity Matrix based on the knowledge.
- EDS- decreases the length of the document as shown in Figure 4.7(b). EDS- is assumed to decrease the document if the term from the query has a link with the term in Term Similarity Matrix based on the knowledge. The maximum value of a term is "1" and if the term from query can find a link with the term from Term Similary Matrix, then the value of that term should be lower than "1" according to the similarity value.

4.2 Information Access - Context Search

Most existing Information Retrieval Systems assume that the query is the sole knowledge of the information needed by a user and then return the same results regardless of who submitted the query. This fact entails a crucial problem nowadays as the information is authored and exploited by millions of different people with different backgrounds, knowledge, and preferences. In order to tackle this problem, a key challenge in Information Retrieval is: 1) how to capture and 2) how to integrate contextual information in the retrieval process in order to increase the search performance.

Context search is defined by Tamine-Lechani *et al.* [2010] as *"combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for information needs of a user"*. Thus, context search aims to optimize the retrieval accuracy by including two related steps: 1) define the context (information needs) of a user (also known as search context), and 2) adapt the context into the Information Retrieval System. Basically, context can be gathered through 1) device, 2) spatio-temporal context, 3) user context, 4) task/problem, and 5) document context. In this section, a method is proposed only to quantify the level of preference by a user and this context is integrated into the Information Retrieval System by utilizing the proposed framework (EDS).

4.2.1 Preference Model

In general, the scope of the preference is defined as how much a user likes the interest from the range of 0 to 100. The interest can be referred to various types such as sports or hobbies. This research proposes to use the *duration* (total duration spent in a particular interest) to quantify the preference within the *time frame* because the duration that a user spends can represent the preference. As an example for the notion of duration can be 1 second, 1 minute, or 1 hour, while as an example for the time frame may refer to any period such as a day, a week, or a year. In order to avoid any confusion, we would like to fix the time frame as a day and duration as an

hour (Please bear in mind that time frame can be any period of time, and duration can be any unit). Before go to details about Preference Model, we would like to discuss the definition of the term *preference* given by Oxford Dictionary is:

Definition of "preference": A greater liking for one alternative over another
but it is believed that a *preference* should associate with duration and the definition of preference is refined as:

Definition of "preference": A greater liking for one alternative over another with continuing practice within the duration.

Based on the above definition, it helps us to define two types of preferences namely 1) long-term-based preference, and 2) short-term preference. So, the hypothesis for long-term preference is the stable preference of a user where the user keeps to practise the preference and increase the *duration* accordingly. Then, the hypothesis for short-term preference is referring to a kind of state that the user is interested in something for a short period of duration. (Will explain later how to define long-term-based preference and short-term preference by using Statistical Box Plot)

After knew the definition, we would like to fix the time frame as a day, and the duration as hour. Lets use *Day* as a set to represents the time frame (as mentioned before), *Interest* is a set to represent the interests (Interest can be considered as a set of preferred elements) provided by the user such as basketball, jogging, or football. Basically, the set of *Interest* comes from the entry of the ontology. In general, an ontology represents the knowledge as a set of concepts within a domain by using a shared vocabulary to denote the types, properties, and interrelationships of those concepts. The ontology that we referred in this section is the Sport Ontology from BBC[1] and the concept that we referred is "Sport:Discipline". Therefore, the attributes of "Sport:Discipline" are "basketball", "jogging", and "football". *Duration* is a set to represent duration in hour (as mentioned before), and *preference* is a value to represent the level of preference from the range of 0 to 100. Day and Duration are independent set.

$$Day = \{1, 2, 3, ..., d\}$$
$$Interest = \{football, basketball, ..., i\}$$
$$Duration = \{0, 1, 2, ..., t\}$$

Basically, the Cartesian Product is used to return a set which is called as DI from Day and $Interest$. So, the Cartesian Product $Day \times Interest$ is the set of all ordered pairs (d, i) where $d \in Day$ and $i \in Interest$ as shown below:

$$DI = Day \times Interest = \{d, i\} \mid d \in Day \wedge i \in Interest \tag{4.15}$$

For an example, the Cartesian Product for DI is $DI = \{\{1, football\}, \{1, basketball\}, ... di\}$. The main reason to use the Cartesian Product is to differentiate the Day and Interest. For an example, the value of $1, football$ is to represent the user practise the football in first day. Then, a function is introduced to gather the total duration spend in one of the interests in a day. The

[1] http://www.bbc.co.uk/ontologies/sport/2011-02-17.shtmlclass$_{sportsDiscipline}$

input is from the Cartesian Product $(Day \times Interest)$ and the output is the total duration in hour.

The Cartesian Product of two sets DI and $Duration$ are the set of all ordered pairs written in (di, t), where di is an element of DI and t is an element of $Duration$. The di and the t are the components of the ordered pair. The Cartesian Product of DI and $Duration$ is denoted by $DI \times Duration$. A function F from DI to $Duration$ is a subset of the Cartesian Product $(DI \times Duration)$ to the following condition: every element of DI is the first component of one and only one ordered pair in the subset. In other words, for every di from DI there is exactly one element t such that the ordered pair (di, t) is contained in the subset of defining the function F where DI is a set of natural numbers and $Duration$ is a set of real number.

$$F : DI \rightarrow Duration \tag{4.16}$$

Since we fix the time frame as a day and duration in hour. Therefore, all the duration that spend in all the interest in one day should be less than 24 (the constant value) in this example.

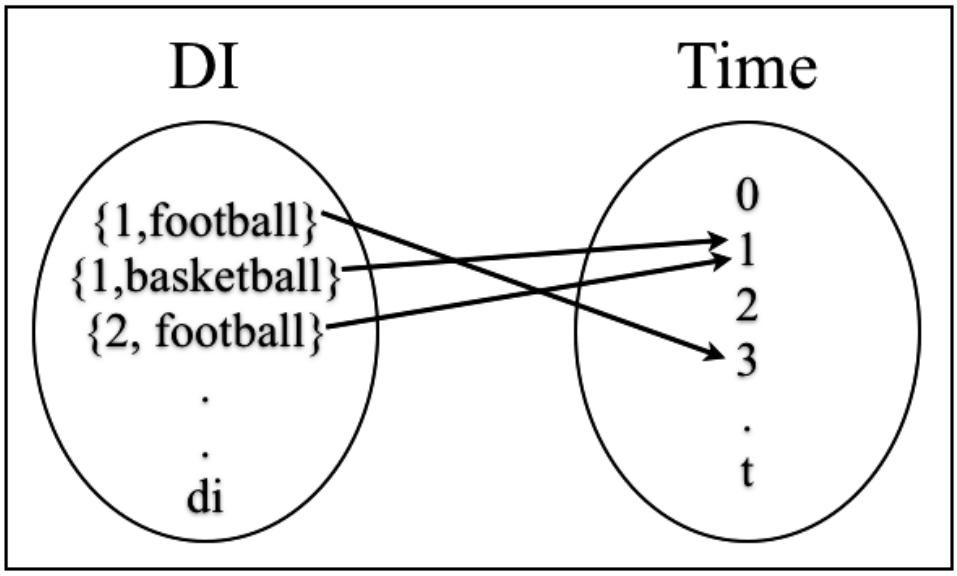


Figure 4.14: A function f with domain denoted as DI and codomain denoted as $Duration$

Based on Figure 4.14, there is a function F with domain $\{\{1, football\}, \{1, basketball\}, ..., di\}$, codomain $\{0, 1, 2, ..., t\}$, and there is also a set of ordered pairs which are $(\{1, football\}, 3)$, $(\{1, basketball\}, 1)$, and $(\{2, football\}, 1)$. With the function F, it returns 3 as the duration spent on football, and 1 as basketball in the first day.

Since we are able to collect the duration spent on the interest, then it is proposed that the $Duration$ in every day is used to quantify the level of preference for the particular interest. Therefore, a function *Pref* is introduced to quantify the level of preference for the interest from the range 0 to 100 as shown below;

$$Pref : Duration \rightarrow Preference \tag{4.17}$$

Both functions, F and *Pref* are composition of functions. Therefore, $F : DI \rightarrow Duration$ and $Pref : Duration \rightarrow Preference$ can rewrite as shown below;

$$h : DI \rightarrow Preference, h(x) = Pref(F(x)) \tag{4.18}$$

is the composition of F and *Pref*, and is also denoted by Pref ° F.

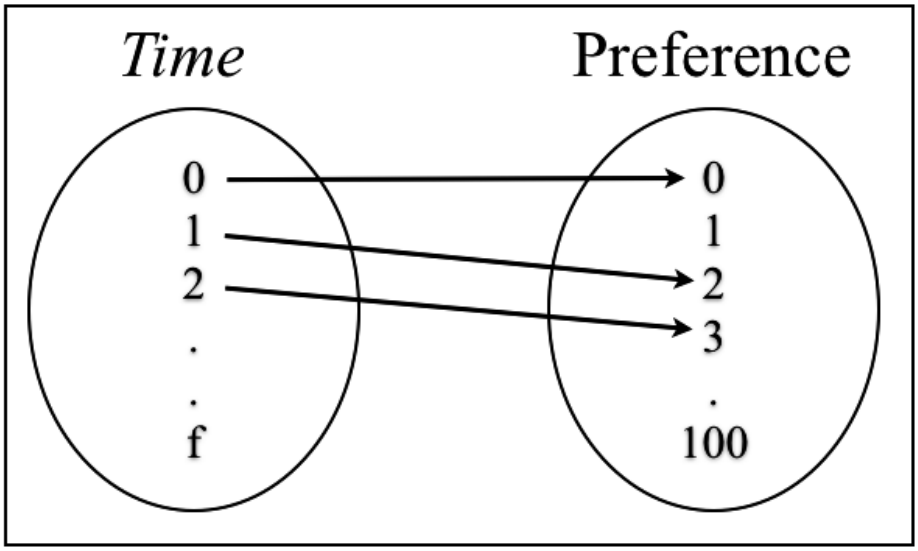


Figure 4.15: A function *Pref* with domain denoted as $Duration$ and codomain denoted as *Preference*

Based on Figure 4.15, a function *Pref* with domain is $\{0, 1, 2, .., f\}$, and codomain is $\{0, 1, 2, .., 100\}$. With the function *Pref*, it returns the level of preference from range 0 to 100. The characteristic of the function *Pref* are as follows;

- The value of *Pref* behave like a monotonic function where the value of *Pref* is either increasing or decreasing.
- The maximum value of *Pref* is approximate to horizontal asymptotes.

After considering the characteristics of the function *Pref*, it has been decided to choose the exponential as the function to represent the level of preference of a user as shown below;

$$Pref(duration) = (1 - e^{\frac{-duration}{\mu}}) * 100 \tag{4.19}$$

Based on the Equation (4.19), the value of Pref is direct relation with the duration. So, if the value of $duration$ is increase, then the value of *Pref* will keep increasing too as shown in Figure 4.16 where the x-axis represents the value of $duration$, and y-axis represents the value of *Pref* with the input of $duration$. Basically, μ is a constant value in all the cases for a user. So, we need a training data to generate the value of μ. Therefore, we proposed to use the total average duration of all the interests within a timeslot to represent the μ. The notion of timeslot

is defined by the user. The user can choose any value (timeslot) to train the μ. In this example, we used a week (7 days) to represent the timeslot is shown below:

$$\mu = \frac{\sum duration\ for\ all\ the\ interets\ in\ a\ week}{\sum interests} \tag{4.20}$$

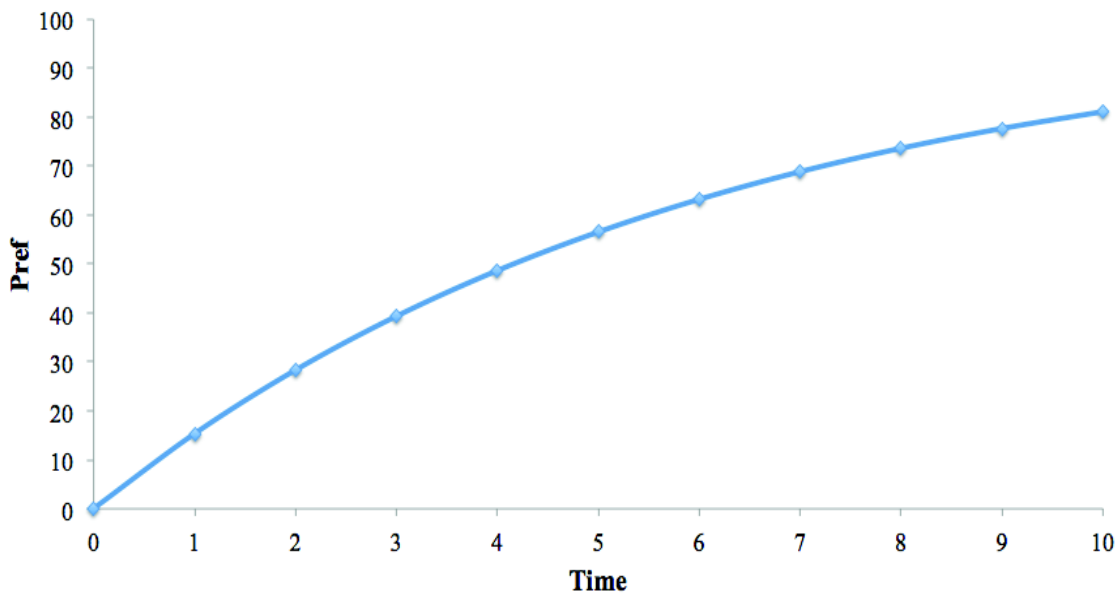


Figure 4.16: Exponential Function

In other words, each user will have their own value of μ. Basically, different value can be used such as the Maximum or Minimum value to represent the μ. The main reason for using the average time is because average time helps to reduce the risk of extreme values. The extreme value refers to the inconsistency of a user to spend in a particular interest. For example, a user dislike football and only gets influenced by his friends to play football in a day by 8 hours (highest value in all the interest per day). So, if we take this value as the μ, then the value of μ will be directly affected. Therefore, we proposed to use average time to represent the μ, because it helps to reduce the amount of random errors present in practice. In addition, it allows the most probable future of the duration spend in the interest of a user to be predicted since the value of μ is only represented by the particular user.

As mentioned before, the total duration for all the interest in a day should less than 24. So, assume that the total duration for all the interests in a day is maximum which is 24, then we can represent this value as constant as shown below;

$$\mu = \frac{constant}{\sum interests} \tag{4.21}$$

Based on the Equation 4.21, if the value of μ is big, then the value of $interests$ should be small and it means that the user needs to spend longer duration in the interest.

The main reason to use the exponential function (in the context of mathematics) to represent the level of preference is because when the value of $duration$ starts from 0 to infinity, the value

of *Pref* is still able to provide the value. For example, if the value of $duration$ is infinity, then the value of *Pref* is approximated to 100. This is true in this case because if the user does not spend any *duration* in the interest, then the value of *Pref* should be 0, and if the user spends infinity *Duration* in one of the Interest, then the exponential function is able to provide the value approximated to 100. The second reason is that exponential function has the line of asymptote. The line of asymptote is important in this case because it controls the level of preference in the range from 0 to 100. So, this is valid in this case again because the maximum value for the level of preference is less than 100. The third reason to use exponential function is that there are no intersection points with different μ as shown in Figure 4.17 because each user spends the duration and interest differently. Based on Figure 4.17, more curve means the value of μ is small.

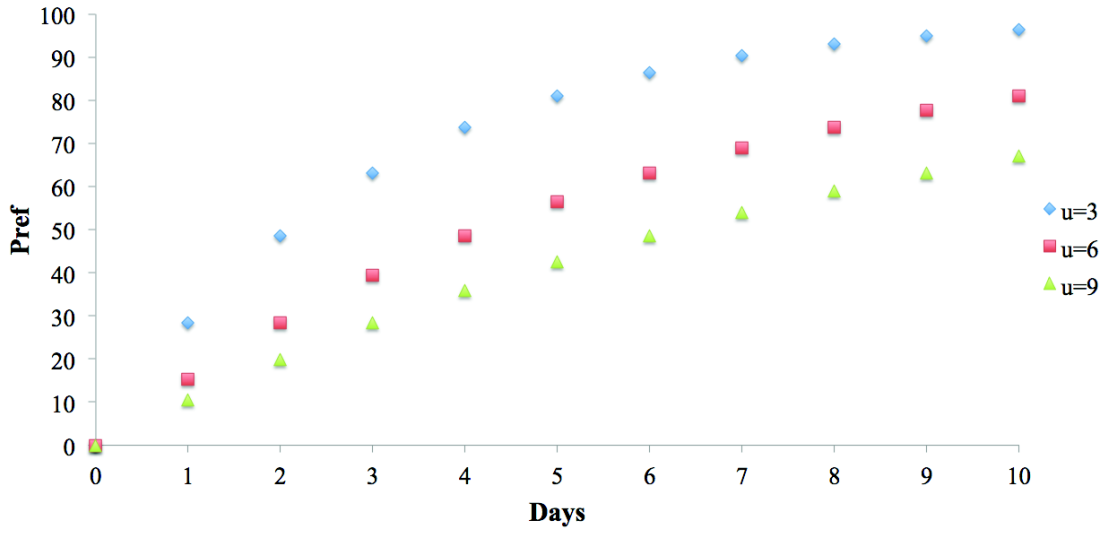


Figure 4.17: Different μ in the Exponential Function

As mentioned before, we fix the time frame as a day, and the examples below are the user's history which contains 5 days. So, the day starts from *first* day until the *fifth* day. Let's use the Day to represent $\{1, 2, 3, 4, 5\}$, then the Interest for a user is $\{Series1, ..., Series3\}$, DI represents $\{\{1, Series1\}, \{1, Series2\}, \{1, Series3\}, ..., \{5, Series3\}\}$, and Duration represents $\{1, 2, 3, 4, 5, 6, 7\}$. Each DI will use the function F as shown in Equation (4.16) to get the elements in the set of Duration. Then, Equation (4.19) is used to calculate the level of preference for each series as shown in Table 4.2 (value of *Pref* for Series1), Table 4.3 (value of Pref for Series2), and Table 4.4 (value of Pref for Series3) in the time frame of 5 days. As mentioned before, we need a training data to represent the μ. In this case, we defined the time slot is 1 day. So, we take all the duration for all the preferences (duration spend for a day in day 1) from the Table 4.2, Table 4.3 and Table 4.4 which are $(2+3+4)/3 = 3$. Therefore, the value of μ is 3 in this example. Then, Figure 4.18 shows the bar chart and Figure 4.19 shows the scatter plot for all the series which are Series1, Series2, and Series3 from the first day to the fifth days based on the history of the user, and the x-axis represents the days (time frame) while y-axis represents the value of Pref.

In Figure 4.18, it can be seen that Series2 is increasing steadily because the user increase the

Table 4.2: Pref for item "Series1"

Day	**Total duration spend for a day**	**Pref(duration) with $\mu = 3$**
1	2	48.65
2	5	81.11
3	3	63.21
4	6	86.47
5	2	48.66

Table 4.3: Pref for item "Series2"

Day	**Total duration spend for a day**	**Pref(duration) with $\mu = 3$**
1	3	63.21
2	4	73.64
3	5	81.11
4	6	86.47
5	7	90.30

Table 4.4: Pref for item "Series3"

Day	**Total duration spend for a day**	**Pref(duration) with $\mu = 3$**
1	4	73.64
2	4	73.64
3	4	73.64
4	4	73.64
5	4	73.64

$duration$ from the *first* day until the *fifth* day. Then, it can also be seen that Series1 fluctuates because the user does not increase the duration for the interest consistently. Finally, Series3 shows that the user spends the exact duration for the interest from *first* day until *fifth* day.

In a nutshell, we can use the exponential function to represent the value of Preference of a user, then we may refer to the theory of charging a capacitor (next section) to charge the Preference Model. It is because the theory of charging a capacitor is using the exponential function to calculate the value of voltage (from 0 to 100 in %) for the capacitor in the circuit. Since both are using the exponential function, then we can use the analogy from the theory of charging a capacitor to charge the Preference Model. Section 4.2.2 will explain the theory of charging a capacitor and Section 4.2.3 will explain how the analogy is used from the theory of charging a capacitor to charge the Preference Model.

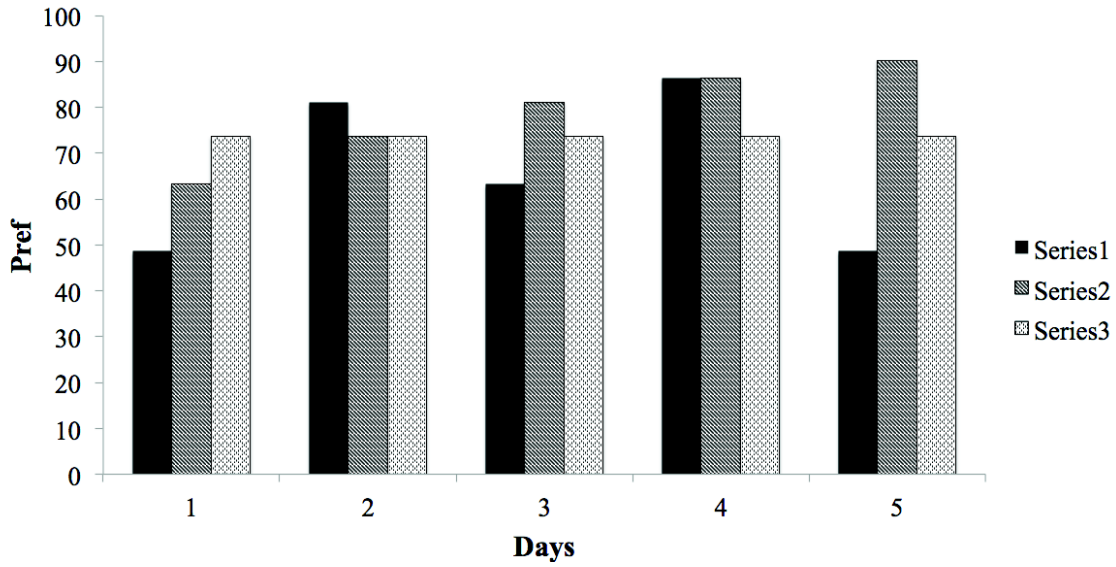


Figure 4.18: Bar Chart: With Different Preferences

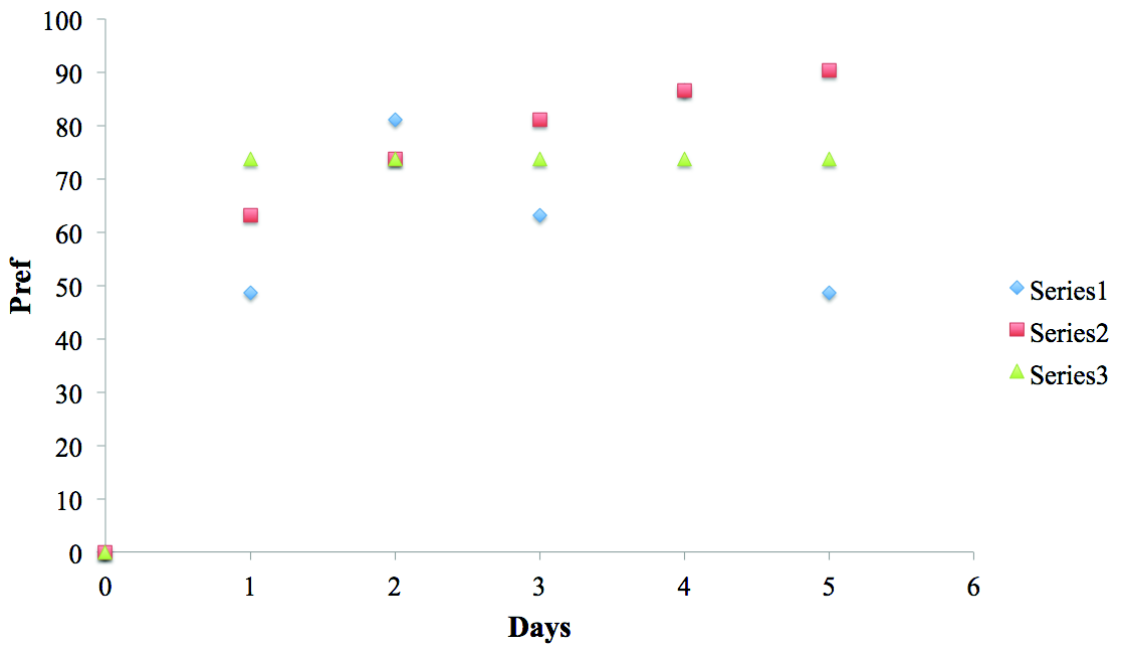


Figure 4.19: Scatter Plot: With Different Preferences

4.2.2 Theory of Charging a Capacitor

A capacitor is a passive device that stores energy in its electric field and returns the energy to the circuit whenever it is required. Basically, a capacitor consists of two conducting plates separated by an insulating material or dielectric. So, when a capacitor is connected to a circuit with Direct Current (DC) source, two processes will occur which are “charging” and “discharging” under the specific condition. (Please refer to the book by Ulaby [1997] for further information about the charging of a capacitor).

In Figure 4.20, the capacitor is connected by the Direct Current (DC) power supply and the current can only flow through the circuit. The capacitor will get equal charges at both plates, and potential difference named (V_c) is created while the capacitor is charging. Once the Voltage at the terminals of the capacitor V_c is equal to the power supply Voltage, ($V_c = V$), then the capacitor is fully charged and the current will move slowly through the circuit as shown in Figure 4.21 where the increment of the percentage of the voltage is not as high as at the beginning. When the capacitor is disconnected from the power supply Voltage and the circuit is maintained closed, the capacitor discharges through the resistor R_D and the Voltage between the plate will drop gradually to zero.

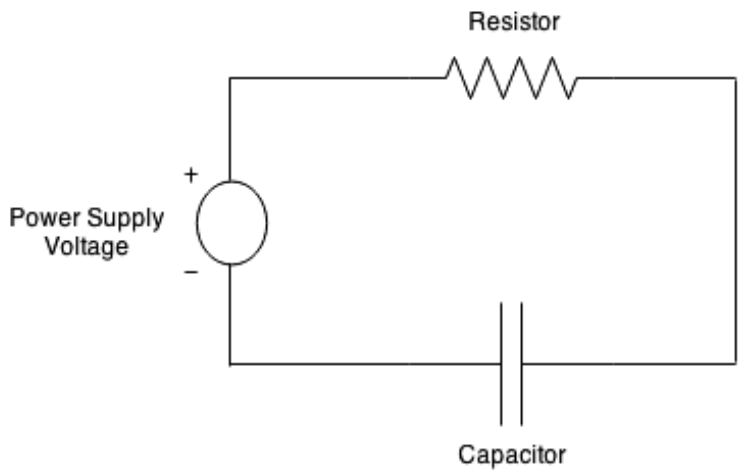


Figure 4.20: Capacitor in the Circuit

The rate of charging is typically described in terms of a time constant which is τ. The charge value of a capacitor depends on the charge voltage from the power supply voltage. Therefore, the value of charge can be calculated as follows:

$$Q = CV_{max}[1 - e^{-t/\tau}] \tag{4.22}$$

where Q refers to the charge (in coulomb), C refers to the capacitance (in farad), V refers to voltage (in volt) from battery, and t refers to the time (in second). The unit of τ is called as the Time Constant which characterizes the rate of charging of a capacitor. The smaller value of τ will impact the time required to charge or discharge the capacitor. Basically, τ is the value of time constant and it will not change during the process of charging the capacitor. After understanding the theory of charging a capacitor, we will explain how we can use this analogy to charge the Preference Model in next section.

4.2.3 Charging a Preference Model (CP)

We proposed that the analogy from ”charging a capacitor” can be used to represent the level of preference of a user. Each capacitor is represented by a Preference to store the charge for one interest and each circuit contains only one capacitor or Preference as shown in Figure 4.22.

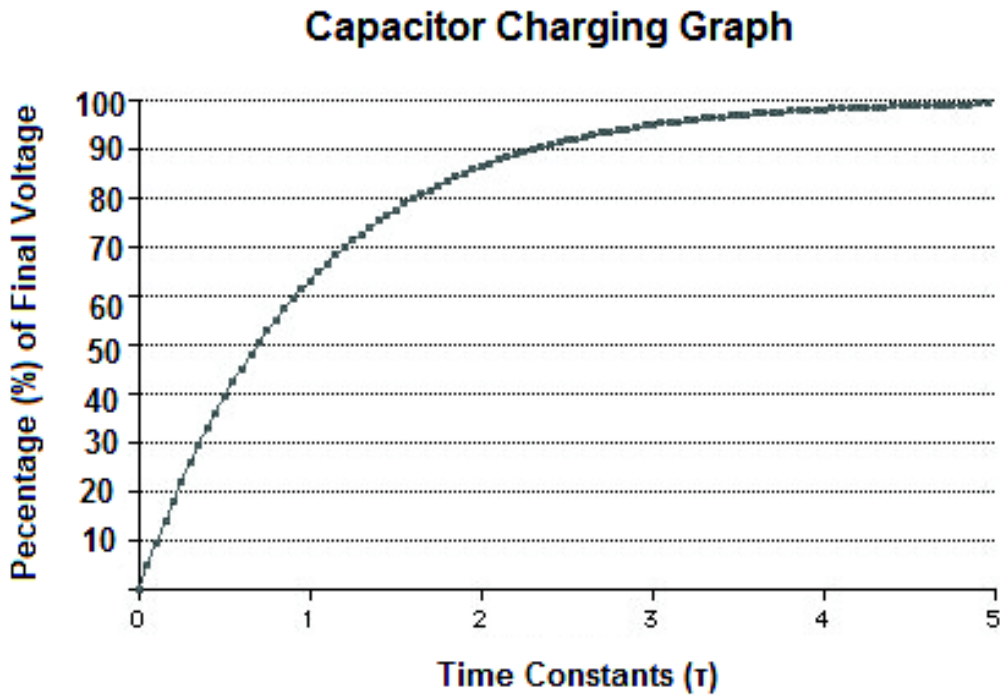


Figure 4.21: Capacitor Charging Graph

Based on Figure 4.22, the mapping between the Capacitor with Preference Model is summarized as follows:

- Each circuit represents only one storage to keep the charge for one preference.
- Inside this circuit, the power supply voltage replaces by Human Energy from the user. Essentially, the user will provide the energy to the preference by using the time. Therefore, the power supply voltage is the function F as shown in Equation (4.16).
- Inside this circuit, the capacitor replaces by preference where the preference is the function *Pref* as shown in Equation (4.17).

As mentioned before, each circuit only can contain one preference. If the user has two preferences, then the user will have two circuits as shown in Figure 4.23 and each circuit acts independently. In general, the user will provide the energy to one of the preferences by using the duration. So, the *duration* (t) that a user spends on a preference is exactly the duration used by power supply voltage to charge the capacitor. In other words, the amount of charge received by a capacitor represents the level of preference of a user in percentage (%). The higher the charge of a capacitor, the higher the percentage of level of preference will be. The main reason for visiting the theory from "charging a capacitor" is because the charging of a capacitor is dependent on duration t, while a preference of a user is also dependent on the duration t. Based on psychological point of view [Oostendorp & Berlyne, 1978], and several researchers [Claypool *et al.*, 2001]; [Morita & Shinoda, 1994] concluded that a user will spend more duration to view the relevant document than non-relevant document. Therefore, it is believed that the analogy from "charging a capacitor" can be used to quantify the level of preference of a user.

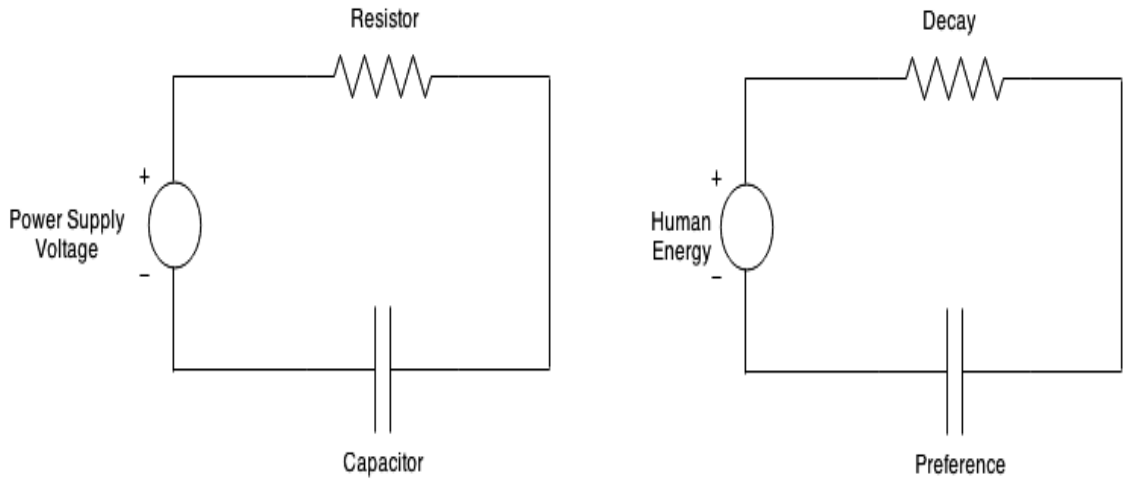


Figure 4.22: Mapping between Capacitor with Preference Model

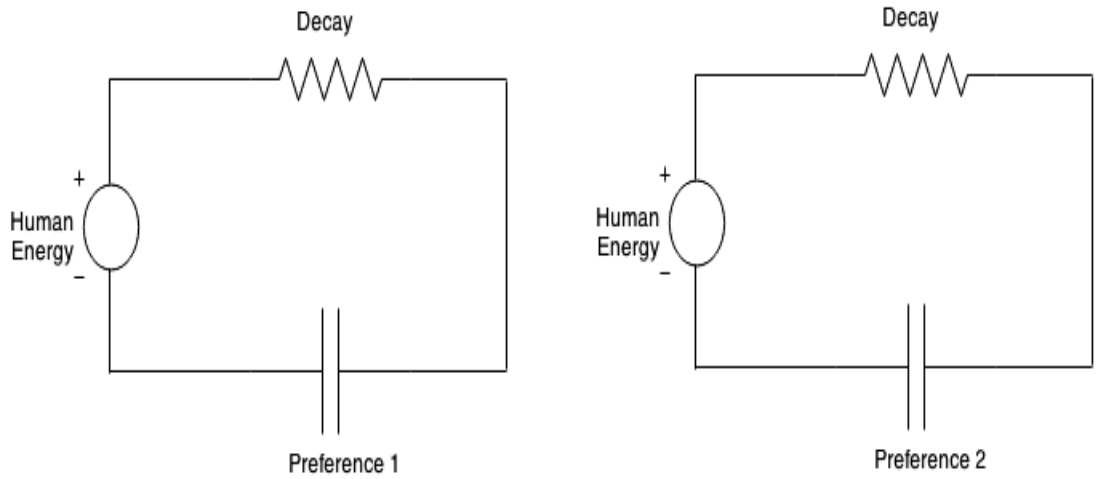


Figure 4.23: Multiple Preferences by Using Analogy from Capacitor

Basically, we replaced the CV_{max} to Q_{max} based on the formula from Capacitance C where $C = Q/V$. Then, the formula can be rewritten from $C = Q/V$ to $Q = CV$ where Q refers to Q_{max} while V in the circuit refers to the maximum Voltage of the power supply Voltage supplied to the circuit. Therefore, according to the formula $Q = CV$ where C is a constant and when V is maximum, then Q is also maximum (V is directly proportional to Q). Then, Q_{max} can be used to represent the voltage of the power supply Voltage in the circuit that always provides the maximum voltage and μ represents the value of time constant. Then, the level of preference can be calculated which is Q for a user as shown below;

$$Q = Q_{max}[1 - e^{-t/\mu}] \tag{4.23}$$

Based on the Equation (4.23), Q_{max} refers to the maximum value of a preference while t represents the duration, and μ represents the time constant. The total average duration of all the preferences is chosen as the time constant because a ratio is needed to define the level of preference in all the preferences belonging to the same user. Therefore, μ can be represented by using the Equation (4.20)

The value of t is different in all the preferences, because each user will spend the duration differently according to the preferences. Therefore, there is a need to use the ratio to represent the level of preference by each preference (Refer to Equation (4.20) for more information).

In order to make the level of the preference to become more meaningful, it is suggested to use Q_{max} to represent the percentage. In other words, each preference will have a value from the range of "0" (minimum) to "100" (maximum) to represent the level of preference.

Based on the $Q = Q_{max}[1 - e^{-t/\mu}]$ graph pattern, it can be shown that from the capacitor charging analogy can be used to represent the level of preference of a user if the user continues to spend the duration t from time to time.

The capacitor charging analogy clarifies a few aspects of the level of preference:

- The duration "t" alters the level of the preference as the current is charging the capacitor. More specifically, the effect of the level of the preference and the amount of the charge inside the capacitor depends on the value of "t".
- The more level of preference is charged, then the longer the Preference can hold the charge.

Based on the capacitor charging analogy, the following hypothesis is established:

- Duration taken within the time frame reflects the preference of a user.

As mentioned in the beginning of Section 4.2.1, two types of preferences namely 1) long-term preference, and 2) short-term preference can be defined by using Statistical Box Plot. Several types of statistical methods can be used to define the long-term preference and short-term preference. The main reason for proposing to use the statistical box plot is because it allows to display the differences between the populations without making any assumptions of the underlying statistical distribution. The mechanism to use statistical box plot is to define the long-term and short-term preference and it will be shown in the next section.

4.2.4 A Statistical Box Plot

Basically, statistical box plot consists of Maximum, Minimum, Median, Quartiles, First Quartile (Q_1), Third Quartile (Q_3), Mean, Interquartile Range, Outliers, and Extreme (Refer to Appendix A for further information about Statistical Box Plot). By applying the statistical box plot, the following constrains are defined:

- Considered as **Long-term based preference** if the value of $Pref > Q3$
- Considered as **Short-term based preference** if the value of *Pref* falls into the range of $Q1 < Pref < Q3$

Then, we re-used back the data (only the Pref value) from the Table 4.2, Table 4.3 and Table 4.4 to define the long-term and short-term preferences. Based on Table 4.5, we can conclude after the *fifth* day (Table 4.2, Table 4.3 and Table 4.4), only 2 preferences are considered long-term preference which are Series 1 at *fouth* day, and Series 2 at *fourth* and *fifth* day because the value of $Pref$ is larger than 81.11 (Table 4.24). In addition, only 2 preferences are considered short-term preference which are Series 2 at *second* day and Series 3 at *first* day until *fifth* day. Figure 4.24 shows the box plot diagram based on Table 4.2, Table 4.3 and Table 4.4.

Basically, the quantity of long-term and short-term preference allows us to understand that the user prefer Series 2 other than Series 1 and Series 3. So, we can propose Series 2 to the user and this can be part of the user context. If there is a tie between the three of them, then all the series will be used to build the user profile.

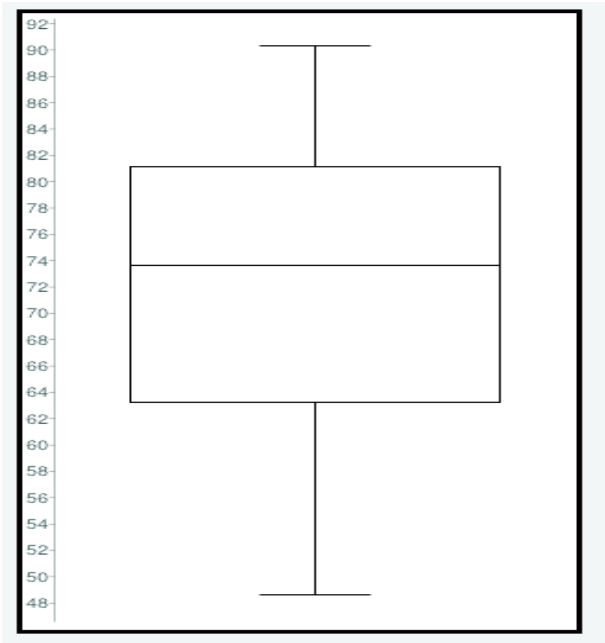


Figure 4.24: Box Plot Diagram according to Table 4.2, Table 4.3 and Table 4.4

4.2.5 Discussion

In a nutshell, this section proposed a Charging Preference (CP) Model based on Equation (4.19) to quantify the level of preference according to the time spent by a user. In addition, the statistical box plot is used to define the long-term and short-term based preferences. Finally, the output from CP model is used to integrate the user context into the Information Retrieval System.

As mentioned in Section 1.3, the last sub-objectives of this research is to design and model a mobile game that is able to attract the users to access, learn, and explore the information of cultural heritage in mobile device. So, we will explain how we design and model the mobile game that able to help the user to learn the information of cultural heritage in the next section.

Table 4.5: The Sample Output based on *Pref* from Table 4.2, Table 4.3 and Table 4.4

Population size	15
Median	73.64
Minimum	48.65
Maximum	90.30
First quartile	63.21
Third quartile	81.11
Interquartile range	17.9
Outlier	-
Extreme	-

4.3 Information Access - Mobile Game

The main idea in designing a mobile game is to assist the user who has little knowledge of the content to search on cultural heritage and at the same time to help the user to learn the information on cultural heritage. In addition, the mobile game is also suitable for the users to access, learn and explore the content of cultural heritage and this type of users exhibits casual search behaviour. Basically, the concept of the mobile game is to provide an information on cultural heritage through the game based on the theory of social constructivism as suggested by Lovell [1959]. So, the proposed mobile game should be able to carry out three types of learning which are 1) inquiry learning, 2) experiential learning, and 3) situated learning. Inquiry learning is an instructional strategy that is used as the basis to design the active learning where the students are engaged in some kind of investigation that involves questions and require the student to find the answers based on the questions. Experiential learning is viewed as a process to create a new knowledge through transformation of the experience. Situated learning can be viewed as complementary to the experiential learning and emphasizes the role of exploratory spaces and practical activities to enhance the learning process. Therefore, the proposed mobile game should provide an indirect way to access the information through text, visual and audio effect. In general, the mobile game can be used by various users in the following different situations.

- **Inquiry Learning:** In a history class, students are working in groups of three or four. The students can use the mobile game to find out more information on certain heritage site inside the classroom. Through the mobile game, the students can get more meaningful information through the panoramic views and acoustical heritage simulation. In addition, the students are required to answer a question in order to proceed with the game.
- **Experiential Learning:** Peter is relating to his friends on his experience in visiting George Town, Penang. All his friends are so excited with the way in which Peter presents it, that is through his mobile device and by using the game. They get to learn and know the heritage sites through the mobile game.
- **Situated Learning 1:** A family is looking for a city to visit and they plan to get to know

the heritage sites before the real visit. The family can launch the mobile game to get first hand image and audio experience before their physical visit.

- **Situated Learning 2:** A tourist group arrives very early in the morning at a heritage site. When they arrive in one of the heritage buildings, they could not find proper information on the heritage building. The tourists can launch the mobile game to get the information of the heritage building.

The proposed mobile game, M-Heritage Hunt is designed based on the classical game called Monopoly together with the game of Treasure Hunt. In this board game, the player can control the character (an icon) inside the game by rolling a dice in order to move around the path. The objectives of the game are to earn properties, to manage own resources and to visit historical buildings in a heritage area. The winner of the game is the player who have successfully monopolised the other players in the game. M-Heritage Hunt is a multi player game for two to four players. Moreover, before starting to play the game, the player is able to visit the various historic sites in M-Heritage Hunt.

4.3.1 Overall Framework of M-Heritage Hunt

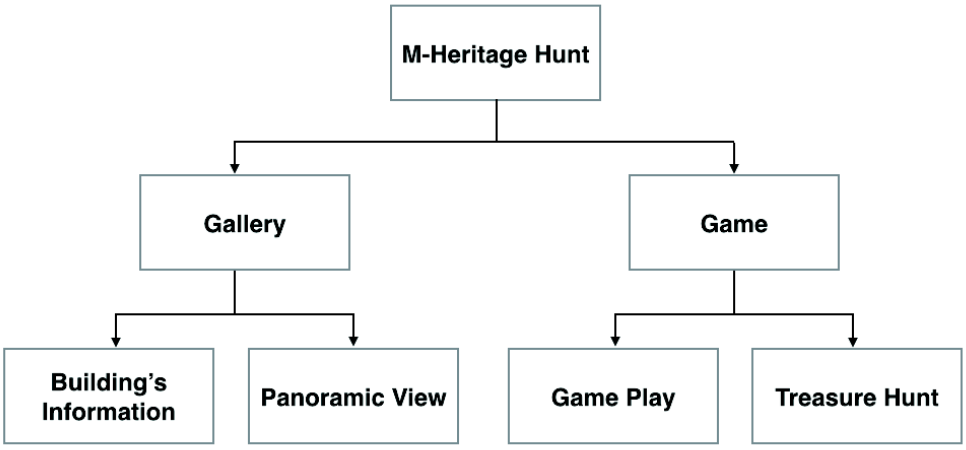


Figure 4.25: Overview of M-Heritage

M-Heritage Hunt consists of two components namely Gallery and Game as shown in Figure 4.25. The Gallery component consists of Information and Panoramic View of the building and the surrounding area. The Information module contains information on a building such as pictures, descriptions and stories of the building. The Panoramic View module allows the users to monitor and control the view by themselves while viewing a heritage building or site. The Game component consists of Game Play and Treasure Hunt modules. The Game Play (Monopoly) module is where the players can play the Heritage Hunt and at the same time are able to visit the heritage sits. The Treasure Hunt module is where users are required to answer some questions based on pre-defined's questions or panoramic view in the Gallery. The Gallery

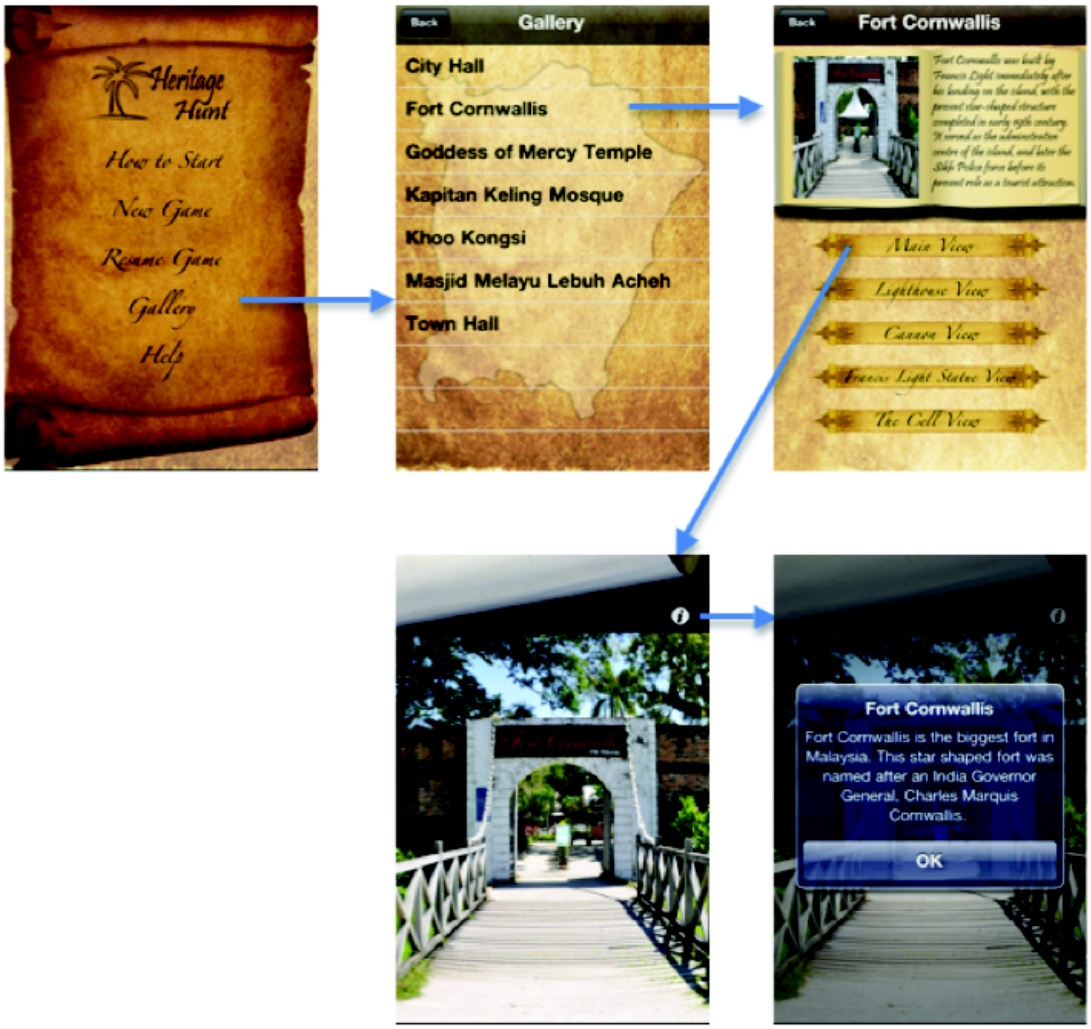


Figure 4.26: Gallery Module (Situated Learning)

module stores the views and information of a heritage building or site. The information is presented in a table format and each row in the table represents one building or site. Each of these buildings contains information such as pictures, descriptions and panoramic views as shown in Figure 4.26 (situated learning).

The Game component is the section in which the Game Play module as shown in Figure 4.27 (experiential learning) and the Treasure Hunt module as shown in Figure 4.28 (inquiry learning) are executed. The Game Play module consists of 40 movable cells as shown in Figure 4.29. Four main cells are distributed at each corner of the board as in the game of monopoly, the other six cells are for the fate corner, another two cells are for the service tax and 26 cells are for the building objects, while the last two cells are for the public transport service.

The four main cells represent the start point, the hospital, the free car park and the ambulance. The main reason to include the cells of ambulance and hospital is to make the player misses the turn. For example, if the icon of the player steps in the cell of ambulance, then the icon will direct move to the cell of hospital and will miss the turn according to the value of the

Figure 4.27: Game Play Interface (Experiential Learning)

dice. The function for the free car park is to collect all the money from the service tax.

In the cells of fate, there are three main types of fates namely the prize, the payment and the movement. The fate card consists of three possible actions based on the information of the fate card. The fate icon of the player is executed by the system automatically. The service tax card requires player to pay for the tax in the range of 100 to 200 depending on which type of icon the player steps on. The user is allowed to purchase the public transport service if the particular cell is not taken off by any other players. The layout for monopoly is as shown in Figure 4.30. It is designed in horizontally so that the users can play the game easily with wider screen.

The general steps of the game are as follows:

- **Step 1:** The player needs to roll the Dice icon in order to roll the dice. The player will move according to the dice number as shown in Figure 4.31 where the number is 11.
- **Step 2:** The next player needs to tap the Next Player icon in order to proceed with his/her turn.

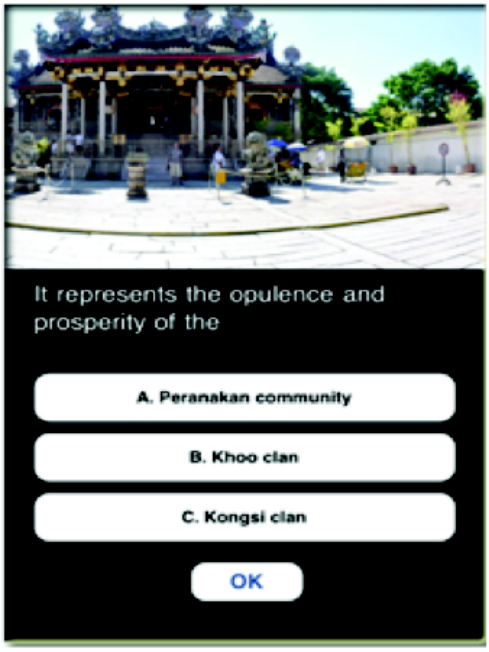


Figure 4.28: Sample of Treasure Hunt Question (Inquiry Learning)

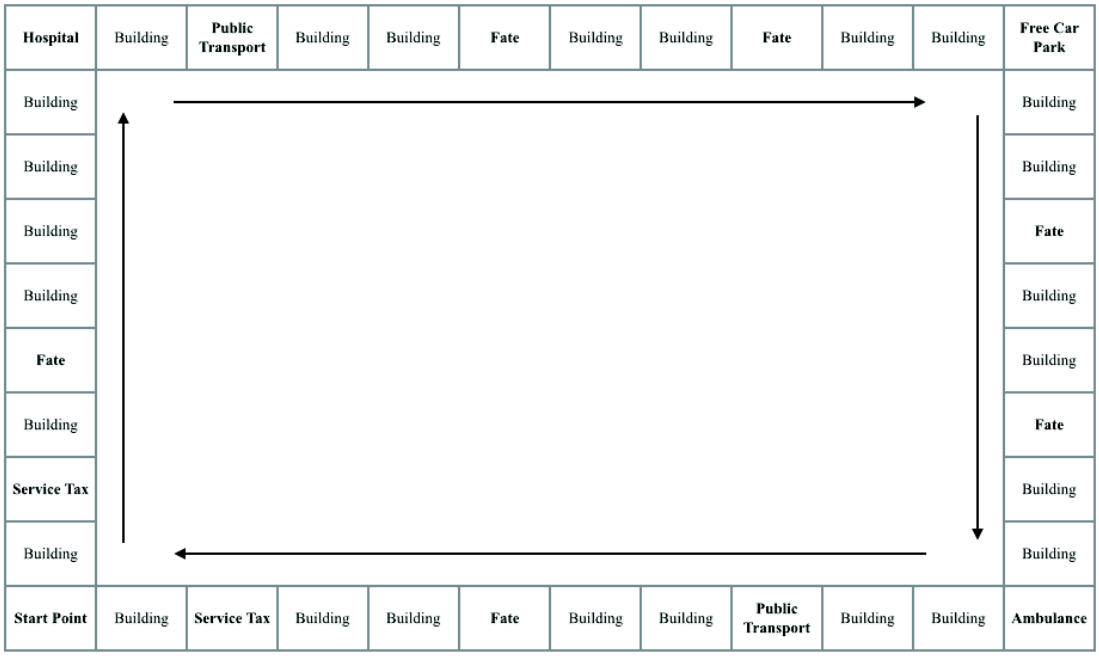


Figure 4.29: Graphical Interpretation of the Movable Cells

- **Step 3:** The player needs to repeat Step 1 and Step 2.
- **Step 4:** The player needs to purchase more properties in order to win the game. The player needs to pay certain amount of money when they step into the property which is

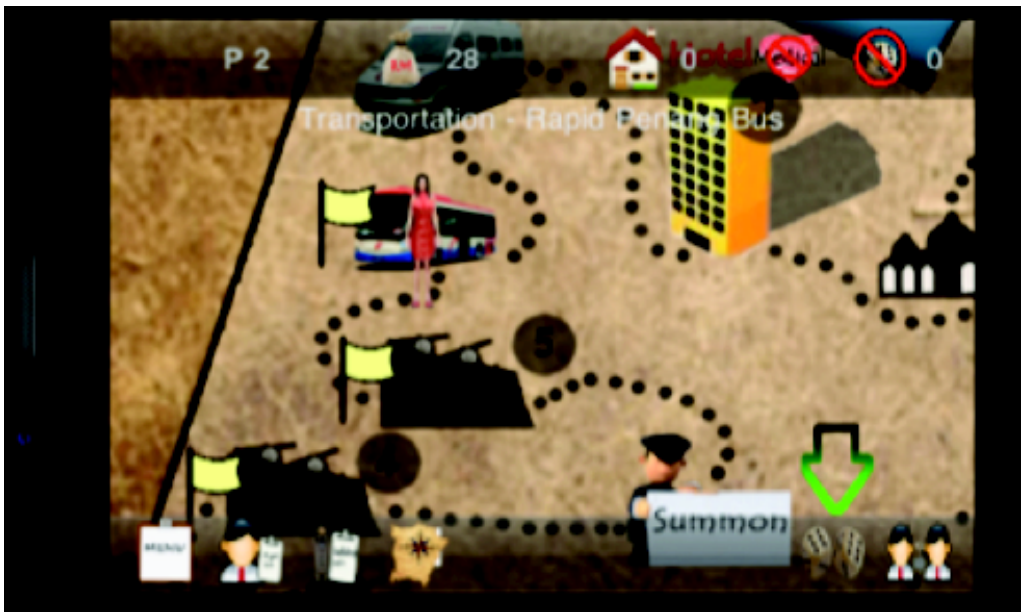


Figure 4.30: Interface for Monopoly

owned by other players. Finally the remaming player is declared as the winner of the game.

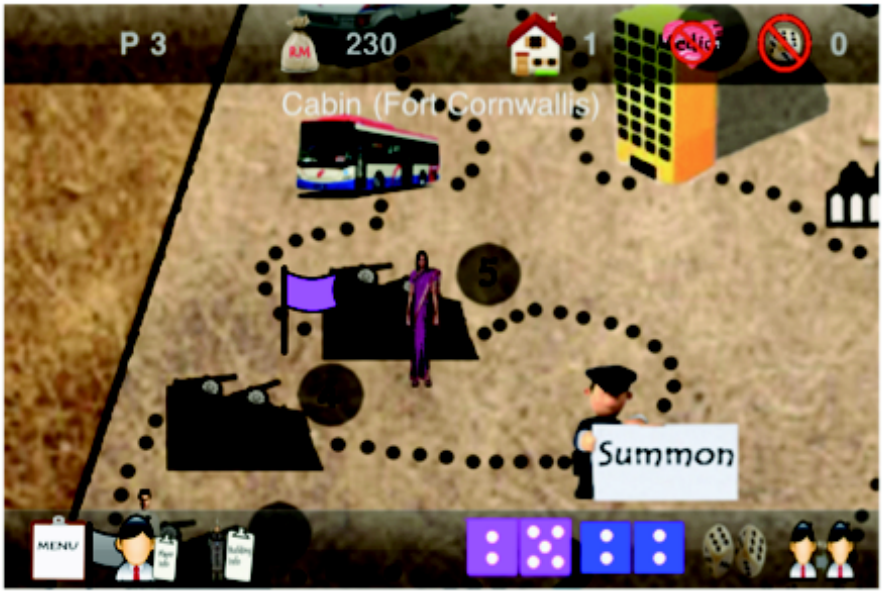


Figure 4.31: The Value of the Dice

The rules of M-Heritage Hunt are as follows;

1. Get double (both numbers on the dice are the same):
    - If the player rolls the dice and get doubles, the player can roll the dice again.

- If the player rolls the dice and get doubles for the second time, the player will be sent to hospital automatically.

2. Not enough cash:

   - If the player does not have enough cash to make payment such as rental or penalty, his/her properties will be mortgaged automatically in order to get extra cash.

3. Send to hospital:

   - Scenario 1: The player gets double two times in a row.
   - Scenario 2: The player lands on the Ambulance point.
   - Scenario 3: The player lands on the Fate point and get a fate card that requires the player to be sent to hospital.

4. Bankrupt:

   - If the player does not have enough cash and no more properties that can be mortgaged, then the player will be declared a bankrupt.

5. Winner:

   - The last player that remains in the game is declared the winner of the game.

Each type of the cell can perform different event once the icon steps and stops on the cell. The explanations for each event are as follows:

- Three possible events may happen when the icon stops in a particular cell. The starting point cell will reward the player with the money when a player passes by or stops on the cell. The ambulance cell will direct the player to the hospital cell when the player steps on the ambulance cell and the player will miss three turns automatically. Lastly, nothing will happen when the player stops on the hospital and free car park cells.
- For the fate cell, the player will receive random fates which are prize, move or penalty. The prize fate means that the player can receive money from the banker or the opponents. The move fate means that the player can make their move accordingly. Finally, the penalty fate means that the player needs to pay money to the banker.
- For the service tax cell, the player is required to pay tax when the player stops on the cell. The amount of tax to be paid depends on where the player stops.
- The building object and the public transport cells have the same function. The only different between them is that the public transport cell requires the player to roll the dice again to determine the payment before moving to the next step. The player can buy, mortgage, or upgrade the building except for the public transport cells.

As for the Treasure Hunt module, there are mini games for users to play. Figure 4.32 shows the title for the Alert Dialogue which is “Upgrade Property” where the message is “Do you want to upgrade the building by treasure hunting”. If the user agrees to “Hunt it!”, then the user will be linked to another interface as shown in Figure 4.28 to answer a series of questions in order to upgrade their property’s value. If the user fails to answer the question, the hunt is considered

to be unsuccessful and the game goes on to the next player. Through the panoramic views in the Gallery module, the players would be able to understand better the buildings, background, culture and architecture, and thus would be able to answer correctly. Furthermore, some hints are also given inside the questions to help the players to complete the Treasure Hunt as well as in learning the historical aspects of the heritage building or site.

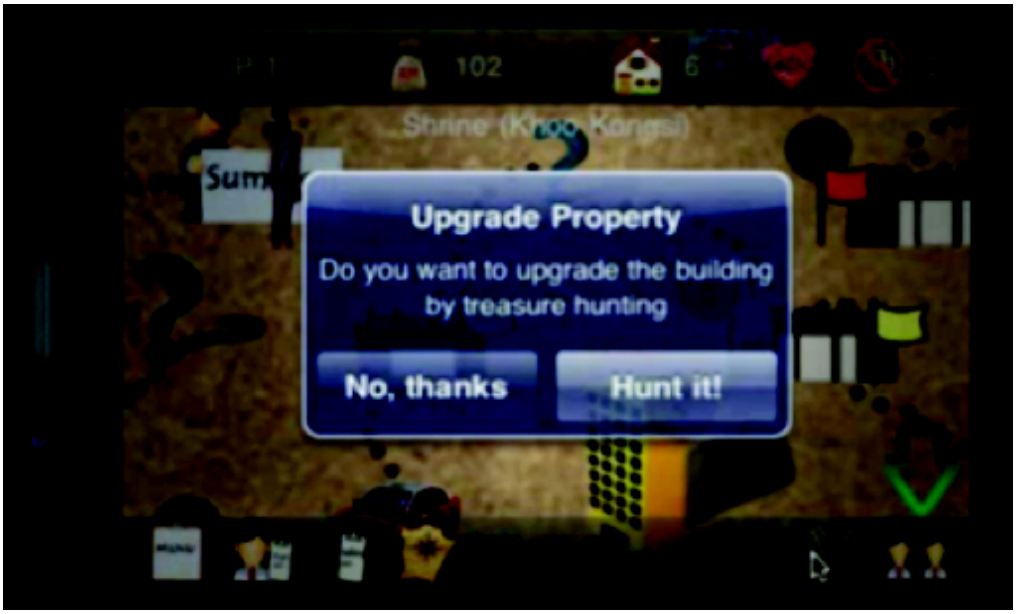


Figure 4.32: Treasure Hunt

Panoramic View is an image made with an ultra wide angle for fish eye lens which provides a 360-degree view. A user is able to view both inner view and outer view of a heritage site using panoramic view. Panoramic view allows the user to view the building and control the view by themselves while viewing a heritage site as shown in Figure 4.33 and Figure 4.34 while Figure 4.35 shows a 360-degree cylindrical panoramic view.

Building Information module is vital in giving constructive information to the user on the historical aspect of a place. It is designed in table views in order to communicate long lists of information to the user effectively and efficiently. This module allows the user to obtain the information and provides the flexibility of viewing it anytime and anywhere as shown in Figure 4.36.

4.3.2 Discussion

The mobile game - M-Heritage Hunt has great potential to be used as one of the tool for information access. The game which is a hybrid of the traditional monopoly game and treasure hunt game would be able to impart better understanding and help the user to learn and explore the information on cultural heritage. The game is designed in such a way that the panoramic view of the heritage buildings and their information are integrated with the game environment in order to help the user who normally exhibits the casual behaviour in information access to navigate the information on cultural heritage rather than requiring them to insert a query to

Figure 4.33: Panoramic View: Goddess of Mercy Temple without Scrolling

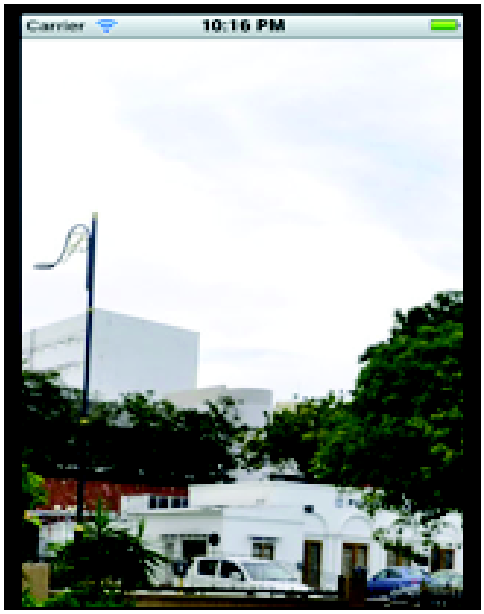


Figure 4.34: Panoramic View: City Hall without Scrolling

information retrieval system. In addition, the M-Heritage Hunt is able to support three types of learning which are inquiry learning, experiential learning and situated learning based on the theory of social constructivism as suggested by Lovell [1959]. The inquiry learning is based on the Questions and Answers from the Treasure Hunt. The user needs to answer correctly in order to upgrade their building on the cell. The experiential learning is based on the environments simulated by the Panoramic Views, and the experiences is based on the Game. The situated learning is based on exploratory and practical activities through the Building's Informa-

Figure 4.35: A 360-degree Panoramic View with Highlighted Display on Figure 4.34

Figure 4.36: Building Information in Table Views

tion where Building's Information provides pictures, descriptions, and stories on the building to the user.

4.4 Summary

In general, this chapter has focused more on the proposed solutions. Firstly, a proposed framework which is Extended Dirichlet Smoothing is introduced to enhance the shortcoming of the term intersection approach. Then, two proposed models namely EDS+ and EDS- have been discussed in Section 4.1 to solve partially the problems in the problem statements of this research. Secondly, a Preference Model based on the Theory of Charging a Capacitor have been discussed in Section 4.2. Finally, the overall framework for the mobile game on cultural heritage that is based on the social constructivism have been shown in Section 4.3.

Part V

EVALUATIONS, RESULTS AND DISCUSSION

Chapter 5

Evaluations, Results and Discussion

This chapter presents several experiments that have been carried out for benchmarking and continued improvements of the entire research. Section 5.1 describes a short summary about the collection of Cultural Heritage in Clef (CHiC) and Section 5.2 describes a short summary about the history of George Town, Penang, Malaysia where it was chosen for the development of a demonstration system for the mobile game. Section 5.3 describes the steps to setup the experiments and presents the results of the proposed model of Information Retrieval System which is Extended Dirichlet Smoothing with Increase of Document Size (EDS+) model and Extended Dirichlet Smoothing with Decrease of Document Size (EDS-) model. In addition, this section also shows the results of comparing the results of the proposed models with other approaches namely pseudo-relevance feedback and query expansion. Section 5.4 describes the steps to setup the experiments and presents the results of the proposed context search with the aim of showing that User Profile can be integrated into the framework of Extended Dirichlet Smoothing. Finally, Section 5.5 presents the results of the mobile game through questionnaire. Basically, Section 5.3 and Section 5.4 are based on system-centred evaluation while Section 5.5 is based on user-centred evaluation.

5.1 Data Collection

The text collection of Cultural Heritage in Clef (CHiC) has been collected by Europeana since 2011 to help the researchers from the domain of Information Retrieval to evaluate the effectiveness of the information access to the materials on cultural heritage. In addition, it also helps to provide feedback into the community of cultural heritage in order to provide a better document representation. So, these materials can be accessed by all the people in the world.

Basically, the CHiC collection contains the data from cultural heritage institutions in Europe. The CHiC collection contains almost 23.3 millions of documents recorded in thirteen different languages (see Table 5.1).

For the experimental purpose, the metadata contains several types of information on an object such as identifier, subject, title, type, country, dataProvider, isShownAt, language, provider, type and url (see Figure 5.1 for an example) . Basically, the collection can be very sparse because the collection relies on the aggregators. Relying on the aggregators indirectly causes the problem of inconsistency and incompleteness in the metadata. For example, if one annotator

Table 5.1: Number of Documents in CHIC Collection by Languages

Sub-Collection Language	Number of Documents in Collection
German	3,865,680
French	3,635,388
Swedish	2,360,050
Italian	2,120,059
Spanish	1,953,124
Norwegian	1,557,820
Dutch	1,251,027
English	1,107,176
Polish	1,093,705
Finnish	800,302
Slovenian	246,952
Greek	197,371
Hungarian	121,771
Others	2,990,507
Total	**23,300,932**

uses "syriac" to describe an object and another annotator uses "language of ancient syria" to describe the same object, then two annotations are inconsistent with regard to the content of the annotation. Inconsistency may also refer to the structure of the annotation themselves. For instance, some annotators might insert all the information into one description field and others may split it into multiple metadata fields such as field of description and title. In such cases, the information of an object may differ depending on the human annotators. As a result, it is hard for a user to search an object since the characteristics of the cultural heritage are not formatted in the same way. In addition, the uniqueness of the CHiC collection is in the average mean of the length of the query which is 2.84. Basically, the topic set provided by CHiC consists of 50 topics sampled from Europeana logs (see Figure 5.2 for an example). The topics from the query logs comprise queries for people, places, work titles, events or subjects. In general, only 36 queries return the relevant document while 14 queries do not contain any relevant documents.

5.2 Environment for the Mobile Game

George Town and Melaka, the historic cities of the Straits of Malacca in Malaysia have been honoured with UNESCO prestigious World Heritage Sites. In the last 500 years, George Town and Melaka have been well known for their multi-cultural trading activities. As a result, these cities as well as other major port cities in East and Southeast Asia were very much influenced by traders from all over the world and the conquests of European powers. With a combination

```
<ims:metadata
        ims:identifier="http://www.europeana.eu/resolve/record/10105/662DC5085397837
        C8C8891836EA6431C4A477CB2"ims:namespace="http://www.europeana.eu/"
        ims:language="eng">
    <ims:fields>
        <dc:identifier>Orn.0446</dc:identifier>
        <dc:subject>Australian Pelican</dc:subject>
        <dc:title>Australian Pelican (Orn.0446)</dc:title>
        <dc:type>mounted specimen</dc:type>
        <europeana:country>malta</europeana:country>
        <europeana:dataProvider>Heritage Malta</europeana:dataProvider>
        <europeana:isShownAt>http://www.heritagemalta.org/sterna/orn.php?id=0446
        </europeana:isShownAt>
        <europeana:language>en</europeana:language>
        <europeana:provider>STERNA</europeana:provider>
        <europeana:type>IMAGE</europeana:type>
        <europeana:uri>
        http://www.europeana.eu/resolve/record/10105/662DC5085397837C8C8891
        836EA6431C4A477CB2</europeana:uri>
    </ims:fields>
</ims:metadata>
```


Figure 5.1: Example of Data Collection - One of the Documents


```
<?xml version="1.0" encoding="UTF-8"?>
<topics>
<topic lang="en">
<identifier>CHIC-001</identifier>
<title>hiroshima</title>
<description>Town in Japan</description>
</topic>
```


Figure 5.2: Example of Data Collection - One of the Queries

of Malay, Chinese, Indian and other cultures, in particular under the British colonial influences, different architectures, religions, dialects, costumes, cuisines, festivities and lifestyles have been established in these historic cities. The heritage site of George Town which has been included in UNESCO World Heritage Listing since July 2008 covers a 109.38 hectare site on the north-east coast of Penang Island consisting of the core zone (the inner boundary) and the buffer zone (the outer boundary) which surrounds the core zone which is as large as 150.04 hectares. In this work, the UNESCO World Heritage Sites of George Town, Penang in Malaysia are chosen for the development of a demonstration system for the mobile game. In addition, several heritage sites such as City Hall, Fort Cornwallis (refer to Figure 5.3), Goddess of Mercy Temple (refer to Figure 5.4), Kapitan Keling Mosque, Khoo Kongsi, Masjid Melayu Lebuh Acheh, and Town Hall are used to provide panoramic views inside the mobile game so that the users are able to

view both the inner view and the outer view of the heritage buildings.

Figure 5.3: Example of the Environment - Fort Cornwallis

Figure 5.4: Example of the Environment - Goddess of Mercy Temple

5.3 Evaluation of Information Access - Information Retrieval System

In order to evaluate the effectiveness of the proposed models, several experiments with different settings were carried out. The main objective of the evaluation is to show that the EDS+ model and EDS- model can impact the performance of retrieval by exploiting the non-matching terms during the retrieval.

5.3.1 Experiment Setup

Before applying the process of tokenization, there is a need to pre-process all the queries and documents as shown in Figure 5.5. The main reason to do the pre-process is to simplify the process of carrying out the experiments. It is acknowledged that some information may be lost such as the meaning of name entities (upper case). Firstly, there is a need to remove all the

stopwords from the documents and queries by using a list of stopwords which contains 571 words. Basically, the list of stopwords contains common words, conjunction and preposition such as “a”, “about”, “the” and “her”. The purpose is to remove all the stopwords because they cannot represent the needs of the user. Secondly, all non-characters are removed such as “#” and “$” since the symbols cannot represent any meaning inside the documents and queries. Thirdly, all the upper case letters are converted to lower case inside the documents and queries because the retrieval model is case sensitive. Finally, there is a need to apply Porter (stemmer) inside the documents and queries. The usage of Porter is to stem all the words to become a root word such as from “runs” to “run” in order to reduce the number of random variables that the query and the document refer to but using different tenses.

After the pre-processing, the unigram language model or often called as bag-of-words model was used to index the title without any description from the queries, and only the metadata from documents such as title, subject and description were used during the experiments. Therefore, the exact ordering of the terms in the document is ignored.

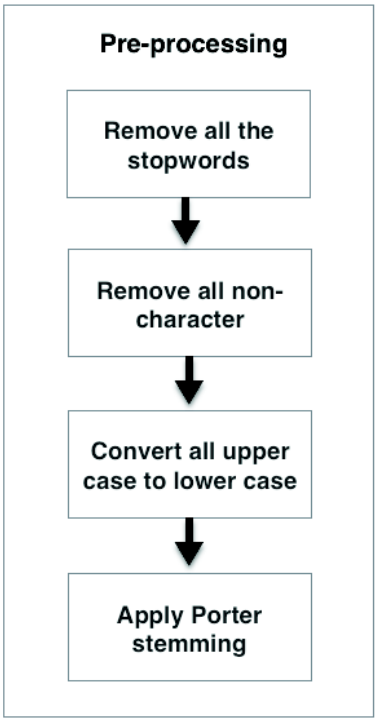


Figure 5.5: Steps of Pre-processing

All the experiments were carried out by using the XIOTA engine proposed by Chevallet [2005]. The performance is measured by Mean Average Precision (MAP). The optimal value for Dirichlet prior smoothing for baseline is 200 and 350 for all the EDS+ model. Basically, a range from 100 to 1000 with the interval is 100 have been tested and the optimal value for Dirichlet prior smoothing is 200 (refer to Figure 5.6) and 350 (refer to Figure 5.7) for all the EDS+ and EDS- model. The reason that the value for Dirichlet prior smoothing is small because most of the size of the documents inside the CHiC collection is small (the total number of terms in dc:subject and dc:title are in the range of 1 to 100. Refer to Figure 5.1 as an example). The

purpose is to increase the optimal value for Dirichlet prior smoothing from 200 to 350 because the EDS+ model increases the length of the document (during the computation) according to the term that appears in the Term Similarity Matrix. Then, it is assumed that 350 will also provide the optimal value for EDS- model as well since the optimal value for Dirichlet prior smoothing is in the range of 300 to 400. Besides, student's paired t-test (at $p < 0.01$ and $p < 0.06$) have been applied to assess the significance of different measurements between several types of statistical approach.

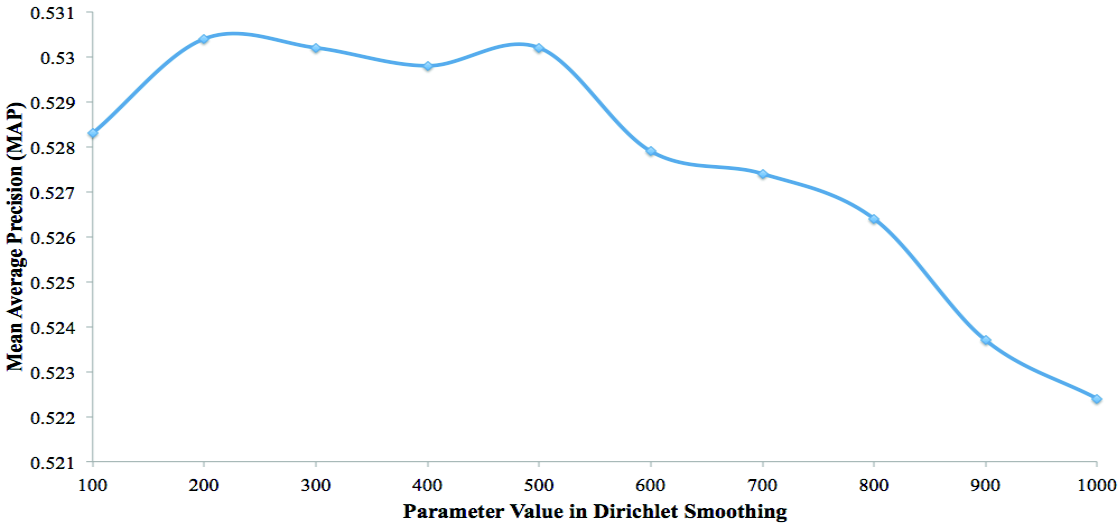


Figure 5.6: Optimal Value for Dirichlet Prior Smoothing is 200

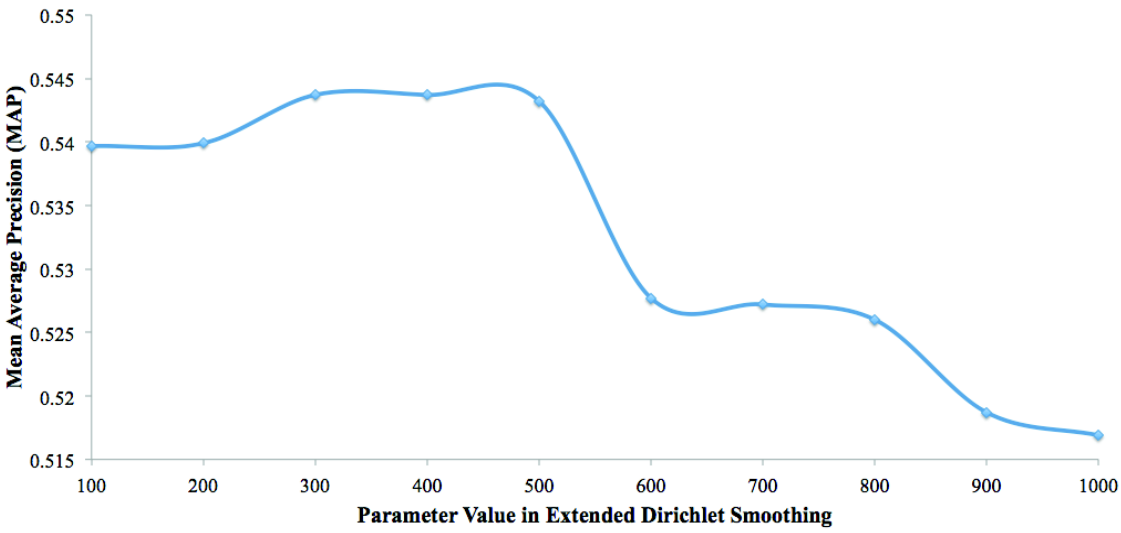


Figure 5.7: Optimal Value for Extended Dirichlet Smoothing is 350

Furthermore, external resources such as Wikipedia have been used to generate the Term Similarity Matrix. First and foremost, the English Wikipedia (version 2012-01-01) which contains 3.835 million articles in the corpus were used. In this experiment, only the first paragraph

Table 5.2: Sample Term Using Dice Coefficient and Cosine Similarity. Note that Words are Stemmed and q Refers to Term from Query.

$Sim(syriac, wj)$,	**DC**	**CS**
$Sim(q, wj = assyrian)$	0.1797	0.1839
$Sim(q, wj = chaldean)$	0.1291	0.1470
$Sim(q, wj = ephrem)$	0.0586	0.1150
$Sim(q, wj = nestorian)$	0.0833	0.1079
$Sim(q, wj = syrian)$	0.0852	0.1011
$Sim(q, wj = antioch)$	0.0971	0.0988
$Sim(q, wj = edessa)$	0.0772	0.0959
$Sim(q, wj = patriarch)$	0.0672	0.0853
$Sim(q, wj = maronit)$	0.0726	0.0803
$Sim(q, wj = coptic)$	0.0715	0.0715

Table 5.3: Performance of Generating the Term Similarity Matrix

Duration	$\approx$ 15 min
Documents	$\approx$ 4 million
Vocabulary	$\approx$ 1.2 million
Co-occurrences	$\approx$ 1.4 billion

Table 5.4: The Specification of the Server to Generate the Term Similarity Matrix

CPU	Quad Core, 2.66 GHZ
Memory	48 GB
Swap Memory	32 GB
Bits	64

of each article from the Wikipedia was used to generate the Term Similarity Matrix because the first paragraph of each article in the Wikipedia pertains to the most critical idea of an article and it can stand on it owns as a concise version of this article according to the guideline from Wikipedia. Basically, around 1.4 billion pairs of terms have been generated based on the first paragraph of each article from Wikipedia and Table 5.2 shows a sample term and values generated by using Dice Coefficient and Cosine Similarity. The proposed model is a generic solution to all application domains. Table 5.3 shows the duration required to generate the Term Similarity Matrix. Based on Figure 5.3, it shows that the server required around 15 minutes to generate around 1.4 billion pair of terms and is consider fast if compare to EMIM. Table 5.4 shows the specification of the server to generate the Term Similarity Matrix. In addition, the complexity of the algorithm to generate the Term Similarity Matrix is as follows:

$$O(n \times m) \approx O(n^2) \tag{5.1}$$

The following approaches were used in the experiments of the next section:

- BL (baseline): LM with Dirichlet smoothing
- LMED-Conf: LM with Extended Dirichlet and Confidence Coefficient (CC)
- LMED-T: LM with Extended Dirichlet and Tanimoto Similarity (TS)
- LMED-D: LM with Extended Dirichlet and Dice Coefficient (DC)
- LMED-Cos: LM with Extended Dirichlet and Cosine Similarity (CS)
- LMED-O: LM with Extended Dirichlet and Overlap Coefficient (OC)

Please aware that the statistical approaches such as CC, TS, DC, CS, and OC are one of the way to generate the Term Similarity Matrix which represent the knowledge.

5.3.2 Results based on EDS+

Table 5.5 shows clearly that the proposed approach based on EDS+ outperforms the baseline result. The most statistically significant improvement ($p<0.06$) is with the LMED-D from 0.5273 to 0.5451 while the most depreciation is with the LMED-O. The reason for these bad results (LMED-O) is that most of the non-null values inside the Term Similarity Matrix is equal "1" which is abnormal because the value of "1" should represent the exact match. Please bear in mind that only the top 10 highest similarity value are used to build the Term Similarity Matrix in order to reduce the noise.

Overall, 16 queries show increments, 9 queries show fluctuations and 11 queries remain the same by using LMED-D as shown in Figure 5.8. The most increment is in Query 25 (query is "syriac") which increases to around 2088% from 0.0025 to 0.0547 in terms of Average Precision (This huge increment will be explained in Section 5.3.5). The most decrement is in Query 28 (query is "fishermen people") which decreases to around -15.38% from 0.0015(AP) to 0.0013(AP). The reason is that the Term Similarity Matrix generated several noise and pull down the result. It can be noticed that these extreme variations occur at rather low values of AP.

In addition, it is necessary to highlight that both LMED-D and LMED-Cos achieve the most statistically significant improvement which is $p<0.01$ based on Precision-10 and the Average Precision (AP) with the values of 0.5333 and 0.5250 respectively as shown in Table 5.6. Besides, LMED-D generates the best AP among other approaches and LMED-D achieves almost 9.7% of increment if compared to the baseline.

Figure 5.9 shows the Precision-Recall curves for BL, and LMED-D only (LMED-D achieves the most gain, and the difference value between LMED-D with LMED-T and LMED-Cos is less than 1%). This figure indeed indicates that the precision of LMED-D method at different recall points is higher than BL. This is an interesting conclusion that LMED-D brings in co-occurrence term knowledge that helps to improve the retrieval result.

5.3.3 Results based on EDS-

Table 5.7 shows clearly again that the proposed approach based on EDS- outperforms the baseline result. The most statistically significant improvement ($p<0.05$) is with the LMED-D from 0.5273 to 0.5447 while the most depreciation is with the LMED-O. As mention before in Section 5.3.2, the reason for these bad results (LMED-O) is that most of the non-null values inside

Table 5.5: Performance with Various Types of Statistics from the First Paragraph of the Articles from Wikipedia based on EDS+ model (* = Statistical Significance at $p < 0.06$ Using the Student's Paired T-Test).

Types of Approach	MAP	MAP Gain/Lost
BL	0.5273	
LMED-Conf	0.5128	-2.82%
LMED-T	0.5395	+2.31%
LMED-D	**0.5451***	**+3.38%**
LMED-Cos	0.5435	+3.07%
LMED-O	0.4890	-7.81%

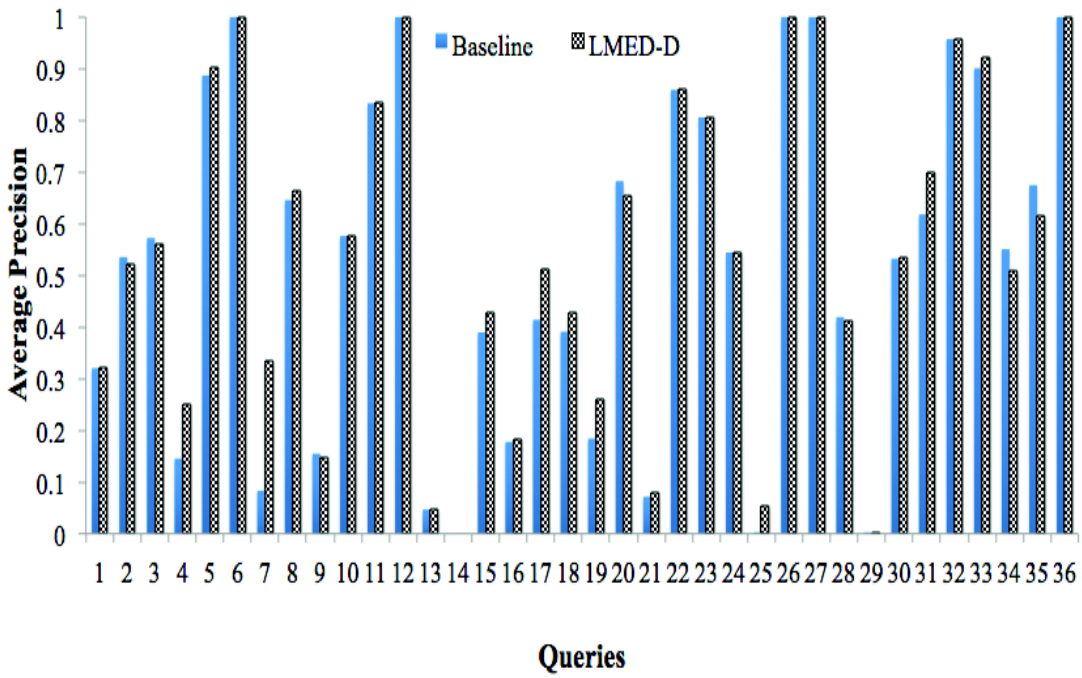


Figure 5.8: Comparison between the result of baseline and Dice Coefficient based on EDS+

the Term Similarity Matrix is equal to "1" which is abnormal because the value of "1" should represent the exact match.

Overall, 17 queries show increments, 8 queries show fluctuations and 11 queries remain the same by using LMED-D as shown in Figure 5.10. The most increment is in Query 25 (the query is "syriac") which increases to around 2088% from 0.0025 to 0.0547 in terms of AP. The main reason with this huge increment is because EDS- manage to retrieve the total of relevant documents from "1"(baseline) to "12"(proposed model). The most decrements is in Query 29 (the query is "1809 combat") which decreases to around -25% from 0.0015(AP) to 0.0012(AP). The main reason with this huge decrement (the poorest result among all the queries) is because both approaches which are baseline and proposed model are able to retrieve the only "1" relevant

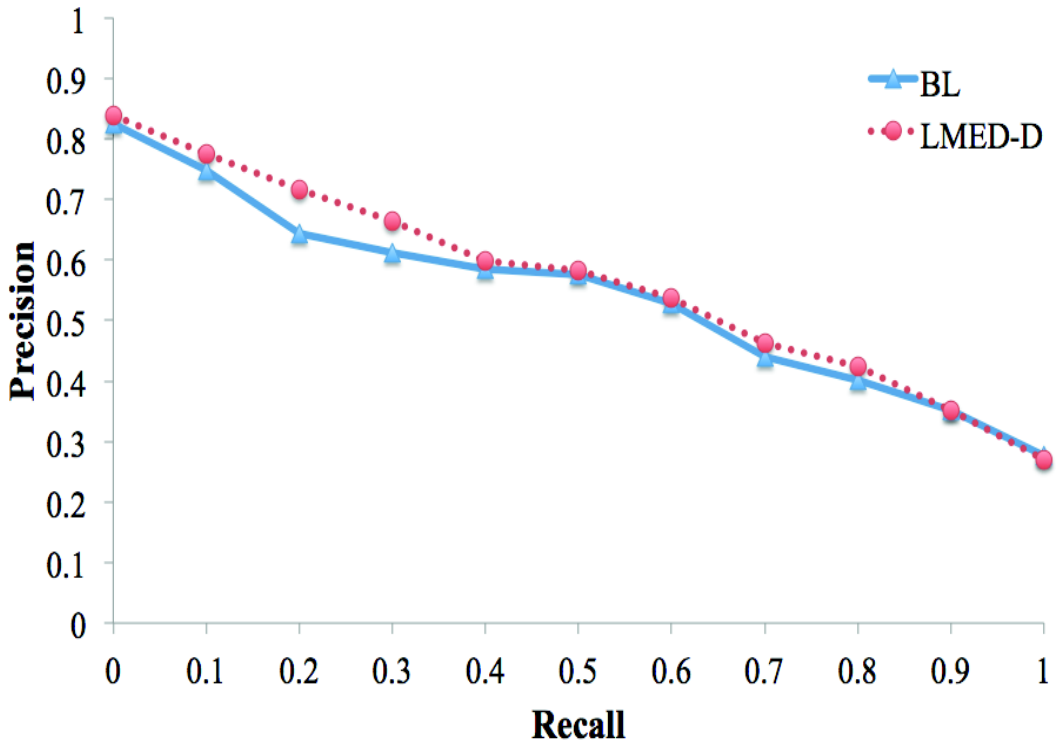


Figure 5.9: Comparison of Baseline with LMED-D by using Averaged Eleven-point Precision/Recall Curve based on EDS+

Table 5.6: Results on Average Precision at Precison-10 based on EDS+. (*=Statistical Significance at p<0.01 and ** at p<0.03 Using the Student's Paired T-Test)

Types of Approaches	P@10	P@10 Gain/Lost
BL	0.4861	
LMED-Conf	0.5083	+4.59%
LMED-T	**0.5250****	+8.00%
LMED-D	**0.5333***	+9.70%
LMED-Cos	**0.5250***	+8.00%
LMED-O	0.4861	-

document, but with different ranking. It can be noticed that these extreme variations occur at rather low values of AP.

In addition, it is necessary to highlight that both LMED-D and LMED-Cos achieve the most statistically significant improvement which is p<0.01 based on Precision-10 and the Average Precision (AP) is 0.5333 and 0.5250 respectively as shown in Table 5.8. Besides, LMED-D generates the best AP among other approaches and LMED-D achieves almost 9.7% of increment compared to the baseline.

Figure 5.11 shows the Precision-Recall curves for BL, and LMED-D. Based on this figure,

Table 5.7: Performance with Various Types of Statistic from the First Paragraph of the Articles from Wikipedia based on EDS- model (* = Statistical Significance at $p < 0.05$ Using the Student's Paired T-Test).

Types of Approach	MAP	MAP Gain/Lost
BL	0.5273	
LMED-Conf	0.5168	-2.03%
LMED-T	0.5390	+2.21%
LMED-D	**0.5447***	**+3.22%**
LMED-Cos	0.5432	+3.01%
LMED-O	0.4921	-7.15%

the precision of LMED-D method once again is higher than BL at different recall points and it shows that EDS- helps to improve the retrieval result.

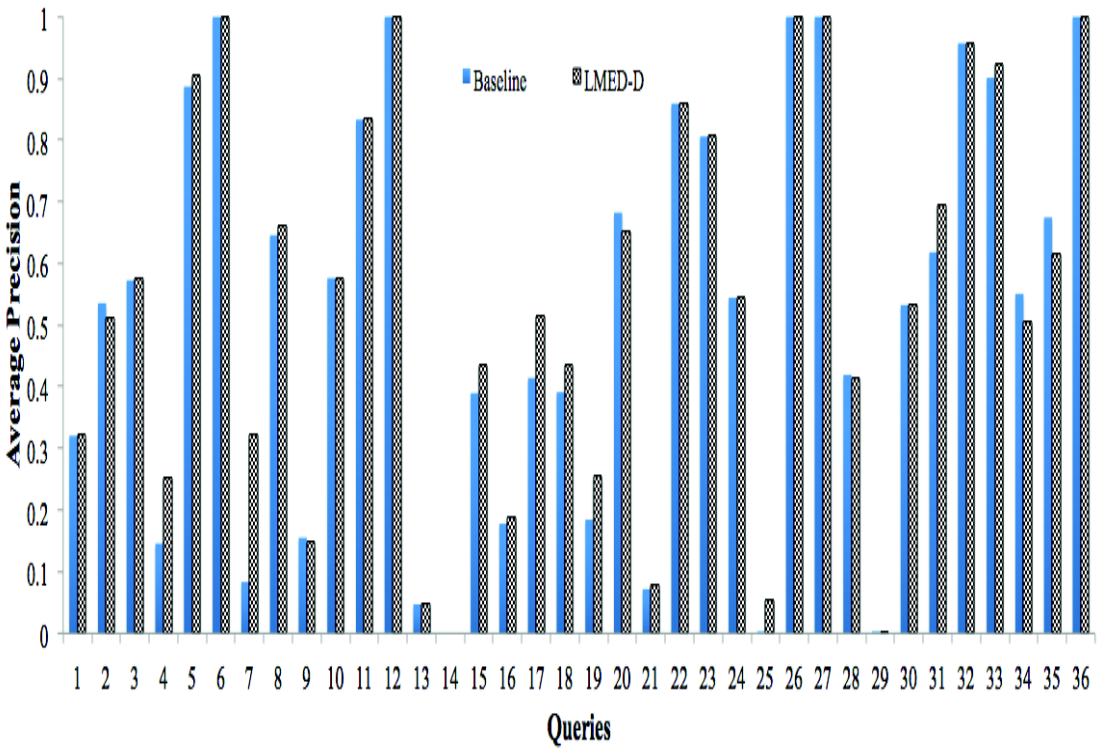


Figure 5.10: Comparison between the results of the baseline and Dice Coefficient based on EDS-

5.3.4 Comparison between EDS+ with EDS-

Table 5.9 shows the comparison between EDS+ with EDS- with various types of statistical approach. In Table 5.9, it is important to highlight that EDS+ performs better than EDS- because

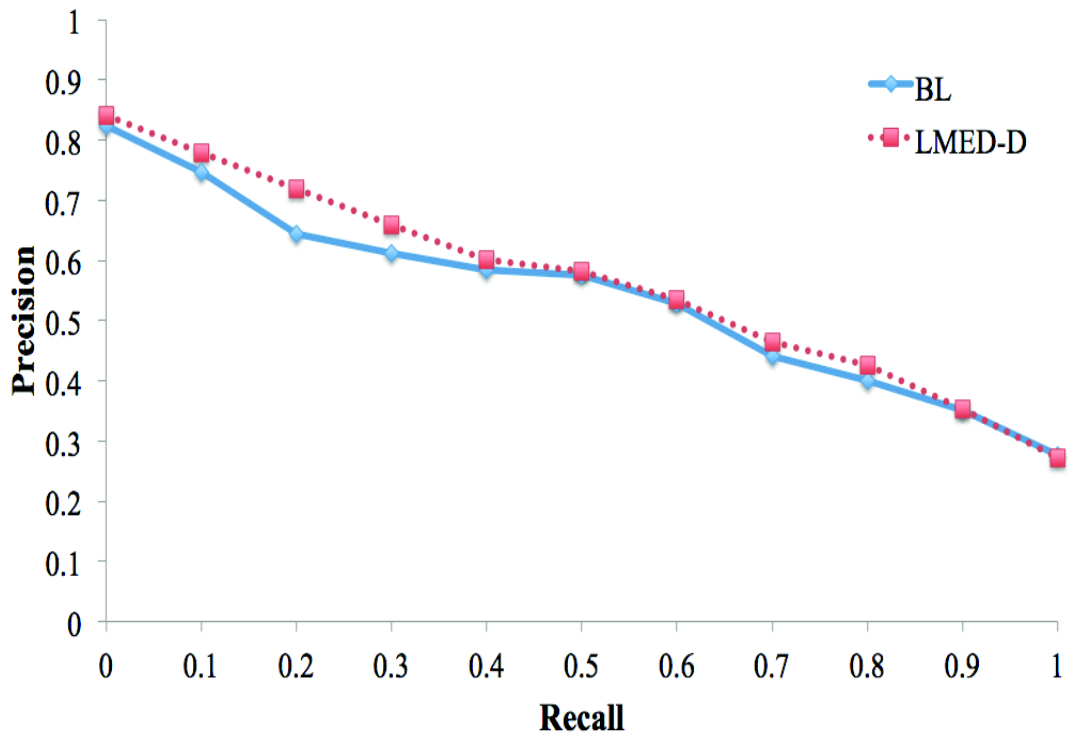


Figure 5.11: Comparison of Baseline with LMED-D by using Averaged Eleven-point Precision/Recall Curve based on EDS-

Table 5.8: Results on Average Precision at Precison-10 based on EDS-. (*=Statistical Significance at p<0.01 and ** at p<0.03 Using the Student's Paired T-Test)

Types of Approach	P@10	P@10 Gain/Lost
BL	0.4861	
LMED-Conf	0.5111	+5.14%
LMED-T	**0.5222****	+7.42%
LMED-D	**0.5333***	+9.70%
LMED-Cos	**0.5278***	+8.57%
LMED-O	0.4917	+1.15

the highest MAP (0.5451) is from the approach of LMED-D based on EDS+. In addition, three types of approach (LMED-T, LMED-D and LMED-Cos) based on EDS+ outperform the corresponding approaches based on EDS-. In general, the difference between these two models are less than 0.0006 except for the LMED-Conf (less than 0.0029) and LMED-O (less than 0.0032).

Overall, 10 queries from EDS+ are much better than EDS-, 10 queries from EDS- show better result than EDS+ and 16 queries remain the same as shown in Figure 5.12. The best performance based on EDS+ is in Query 7 (the query is "schlesien maps") where the difference

Table 5.9: Comparison between EDS+ and EDS- with Various Types of Statistics from the First Paragraph of the Articles from Wikipedia

Types of Approach	MAP based on EDS+	MAP based on EDS-
LMED-Conf	0.5128	**0.5168**
LMED-T	**0.5395**	0.5390
LMED-D	**0.5451**	0.5447
LMED-Cos	**0.5435**	0.5432
LMED-O	0.4890	**0.4921**

between EDS+ and EDS- is 3.96%. Basically, both proposed models (EDS+ and EDS-) are able to retrieve the total number of relevant documents from "1" (baseline) to "2" (proposed models), but both proposed models give different score for the document and this affects the final ranking (please bear in mind that EDS+ increases while EDS- decreases the size of the document). The best performance based on EDS- is in Query 29 (the query is "1809 combat") where the difference between EDS- with EDS+ is 8.3%. Both models are able to retrieve "1" relevant document, but in different ranking due to the score. In addition, it is interesting to highlight that EDS+ is able to show improvement from the range of 0.07% (minimum) to 3.96% (maximum), while EDS- is able to show the improvement from the range of 0.04% (minimum) to 8.3% (maximum).

However, it can be pointed out that EDS- performs better than EDS+ based on Precision-10. The main reason is because the value of the normalization in Language Model. Please be aware than EDS+ is to increase the size of the document while EDS- is to decrease the size of the document. So, assuming that the same query and the same document are used, the only difference between EDS+ and EDS- is the value of the normalization. Therefore, the score for EDS- will be lower than EDS+ if we use the same query and same document are used.

In general, three approaches (LMED-Conf, LMED-Cos, and LMED-O) which are based on EDS- outperform the approaches (LMED-Conf, LMED-Cos, and LMED-O) based on EDS+. In general, the value of the difference between these two models are less than 0.0029 except the approaches from LMED-D (remain the same), and LMED-O (less than 0.0057). The main reason that EDS- performs better than EDS+ in terms of Precision-10 is because EDS- decreases the size of the document

In general, Figure 5.13 shows the Precision-Recall curves for BL, LMED-D based on EDS+, and LMED-D based on EDS-, and both proposed models outperform BL at different recall points. As a conclusion, both models behave almost the same based on the curve as shown in Figure 5.13.

In order to validate the proposed models (EDS+ works better in recall and EDS- works better in $P@10$), the same setting (μ is 200 for baseline, and 350 for EDS+ and EDS-) was tried in the collection from CLEFMed2011. Basically, the case-based collection was used from CLEFMed2011 where the topics are from Casimage. Then, the hierarchical link was extracted from the domain knowledge which is Unified Medical Language System *UMLS* as the Term Similarity Matrix. Table 5.11 shows that both proposed models which are EDS+ and EDS- outperform the baseline in terms of MAP and Precision-10.

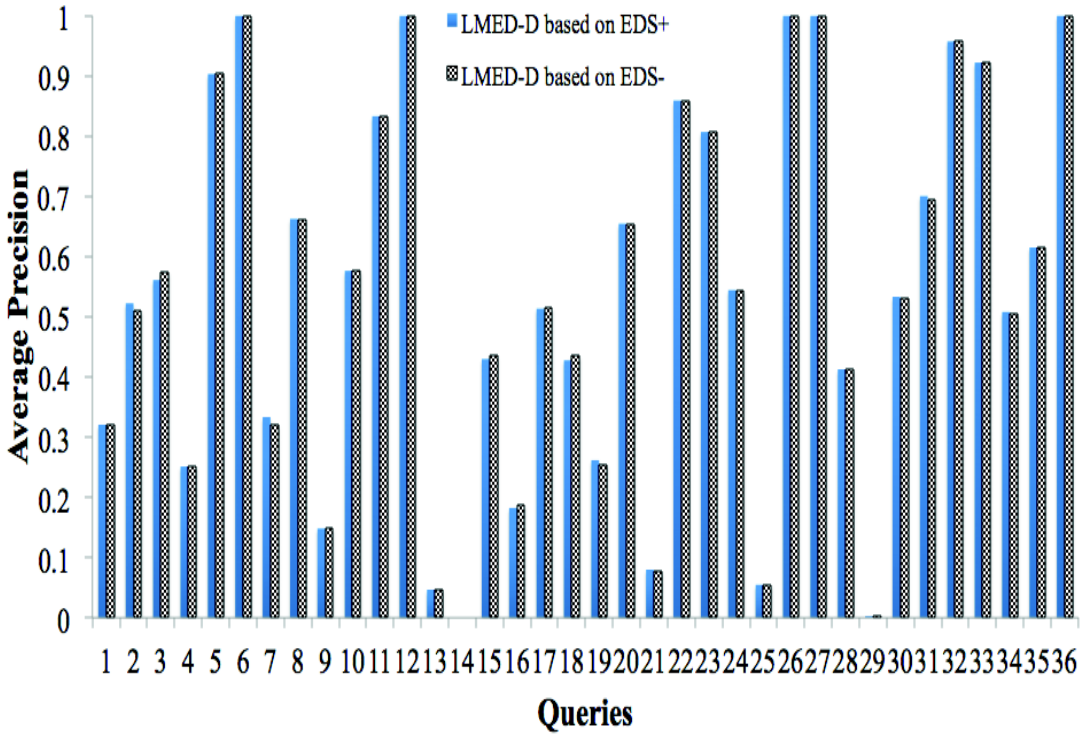


Figure 5.12: Comparison between EDS+ and EDS- based on the approach of LMED-D

Table 5.10: Comparison between EDS+ with EDS- and Various Type of Statistic at Precision-10

Types of Approach	P@10 based on EDS+	P@10 based on EDS-
LMED-Conf	0.5083	**0.5111**
LMED-T	**0.5250**	0.5222
LMED-D	0.5333	0.5333
LMED-Cos	0.5250	**0.5278**
LMED-O	0.4861	**0.4917**

Table 5.11: Comparison between Baseline, EDS+ with EDS- in terms of MAP and Precision-10 in CLEFMed2011

Types of Approach	MAP	MAP Gain	P@10	P@10 Gain
Basline	0.1044	-	0.16	-
EDS+	0.1134	+8.62%	0.17	+6.25%
EDS-	0.1136	**+8.81%**	0.17	+6.25%

In a nutshell, it can be concluded that EDS+ works better in recall (EDS+ is able to retrieve more relevant document), while EDS- works better in $P@10$ (EDS- is able to retrieve more precise document) based on Table 5.9 and Table 5.10, but it never shows that EDS+ works

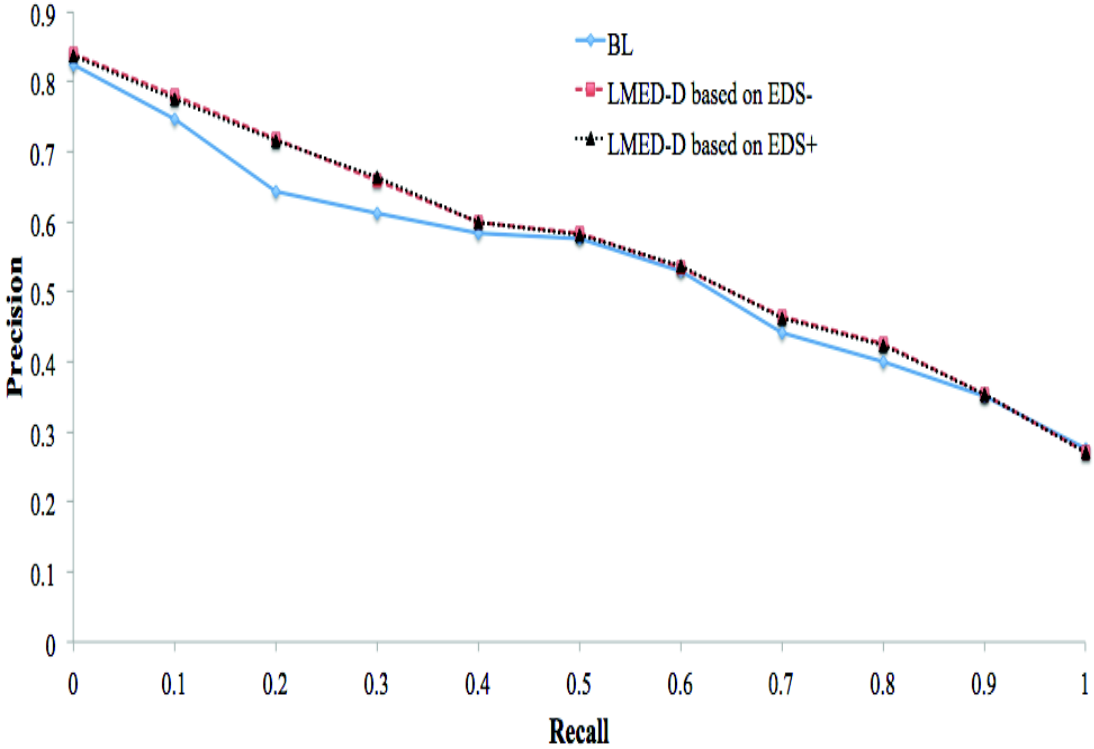


Figure 5.13: Comparison between EDS+ and EDS- based on the Approach of LMED-D by using Average Eleven-point Precision/Recall Curve

better in MAP and EDS- works better in $P@10$ when they were tried in CLEFMed2011 as shown in Table 5.11.

5.3.5 Comparison between Dice Coefficient with Cosine Similarity and Tanimoto Similarity with Baseline based on ESD+

In the experiment, only the top 10 similar values [1] were selected for each term from the queries to evaluate the EDS+ model. Figure 5.14 shows the top 10 similar values between CS, TS, DC, CC and OC by the term "syriac' (please note that the similarity value for DC and TS are in the between of 0.00009 until 0.0001) while Figure 5.15 shows further highlights between TS and DC by the same term. Basically, the similar values generated by DC is the lowest among the various statistical approaches. However, DC generates better results compared to CC, TS and OC in terms of average precision as shown in Table 5.12. Table 5.12 shows the Average Precision (AP) and the total number of relevant documents retrieved by CS, TS, DC, CC and OC. In addition, fifteen and twelve relevant documents were successfully retrieved by using CS and DC approaches respectively. One of the main reason for CS to perform the best is because CS can generate more useful terms which is related to the term "syriac" such as "assyrian" and "antioch" as shown in Table 5.13. Besides, OC generated a lot of unrelated terms such as "overscor" and "recepti" which do not have any relation to the term "syriac" and indirectly it

[1] Similar values refer to the output generated by statistical approaches

influences the average precision for OC. (Please refer to Appendix B for further remarks about this query (Query 25 - “syriac”)).

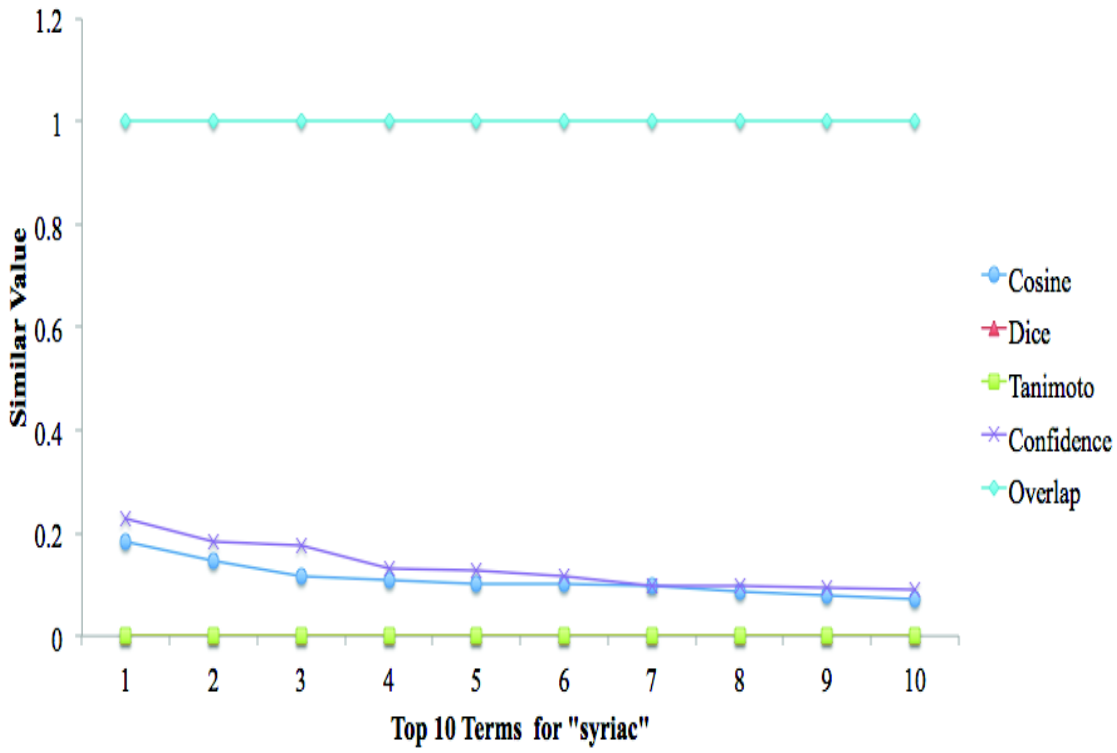


Figure 5.14: Top 10 Similar Values between CC, CS, DC, OC and TS of the Term “Syriac”.

In a nutshell, LMED-D, LMED-Cos and LMED-T outperform the baseline as shown in Table 5.5 where the baseline method is purely based on the term intersection approach. The performance will drop if the exact matching is not possible. This indicates that the proposed model especially DC, CS and TS work better than the baseline when there is a problems of inconsistency, incompleteness and short query between the queries and documents. Besides, the proposed model is able to bridge the gap of vocabulary between the queries and documents.

Table 5.12: Average Precision (AP) and Total Number of Relevant Documents Retrieved (Rel) for BL, CC, CS, TS, DC and OC of the Query (CHIC-032 where the Title is “syriac”). The total number of the Relevant Documents is 80 for this Query.

	BL	**CC**	**CS**	**TS**	**DC**	**OC**
AP	0.0025	0.0002	0.0495	0.0025	0.0210	0.0003
Rel	1	2	15	1	12	1

5.3.6 Discussion based on the result from EDS+

In order to explain the findings, a table was generated as shown in Table 5.14 which is the correlation among all the statistical approaches presented namely Confidence Coefficient, Tanimoto Similarity, Dice Coefficient, Cosine Similarity and Overlap Coefficient. Basically, *high*

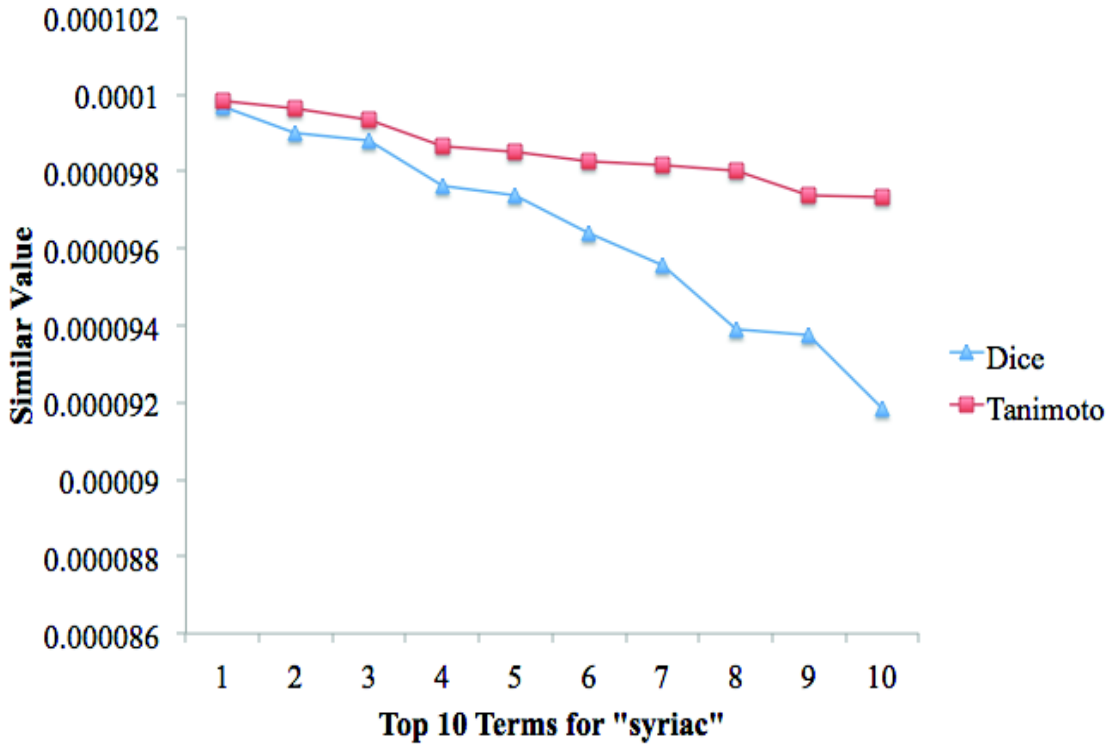


Figure 5.15: Further Highlights of the Top 10 Similar Values between TS and DC of the Term "Syriac".

correlation means that two sets of data are strongly linked together. So, correlation is *positive* when both the values increase, and correlation is *negative* when one value decreases as the other increases. Please bear in mind that the approach of correlation is used to evaluate the correlation between all the statistical approaches without the integration of EDS+ and EDS-. The formula to calculate the correlation is as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{5.2}$$

where:

- Two sets of data namely x and y
- $(x_i - \overline{x})$ is each x-value minus the mean of x
- $(y_i - \overline{y})$ is each y-value minus the mean of y

The findings are as follows:

1. OC has low positive correlation (<0.4) with CC, CS, TS and DC based on Table 5.14.
2. DC has a high positive correlation (>0.8) with CC, CS and TS based on Table 5.14.

Table 5.13: Different Terms Provided by Using CC, CS, TS, DC and OC by Term "syriac". Note that Words are Stemmed.

CC	CS	TS	DC	OC
assyrian	assyrian	younger	ic	truxali
syrian	chaldean	boat	antioch	bass
patriarch	ephrem	zealand	savariana	survei
manuscript	nestorian	im	hill	recepti
antioch	syrian	bush	photograph	racham
syria	antioch	crash	philippin	quival
testament	edessa	shell	assyrian	paolu
st	patriarch	dj	guitarist	overscor
iraq	maronit	tune	sydnei	nezem
monasteri	coptic	syrian	syrian	koftgari

Table 5.14: Correlation among CC, CS, DC, OC and TS.

	CC	CS	DC	OC	TS
CC	1	0.7752	0.8203	0.1454	0.8003
CS	0.7752	1	0.9526	0.3465	0.9141
DC	0.8203	0.9526	1	0.1889	0.9746
OC	0.1454	0.3465	0.1889	1	0.1782
TS	0.8003	0.9141	0.9746	0.1782	1

As a conclusion, CS, DC and TS are the best among all the statistical approaches. In addition, the MAP for LMED-Cos, LMED-D and LMED-T are the best where the increment starts from 2.31% until 3.38% compared to the baseline. Besides, the result of MAP for LMED-Cos, LMED-D and LMED-T are close to each other because DC (LMED-D) has a high correlation with CS (LMED-Cos) and TS (LMED-T).

5.3.7 Comparison between Pseudo-Relevance Feedback with EDS+

In order to confirm the validity of the proposed methods, the results were compared with Pseudo-relevance feedback based on Lavrenko's relevance models [Lavrenko & Croft, 2001] and the settings are exactly the same as stated in Section 5.3.1.

Table 5.15 shows that different setting such as the number of documents (docs), and the number of terms (ts) to be used as feedback which can bring different results. The highest value of MAP that can be achieved by using Pseudo-relevance feedback based on Lavrenko's relevance models (proposed by Lavrenko & Croft [2001]) is 0.5376 by using 10 documents and 30 terms while the lowest value is 0.4962 by using 15 documents and 15 terms. In a nutshell, the proposed model (LMED-D) with a value of 0.5451 as shown in Table 5.5 still outperforms the highest value in Table 5.15 which is 0.5376. It can be seen that in all the cases, the proposed model (LMED-D) still outperforms and achieves better result than the results produced by using different types of parameter in pseudo-relevance feedback.

Table 5.15: Mean Average Precision by using Pseudo-Relevance Feedback(∗ represents the highest value while ˆ represents the lowest value)

	5 ts	10 ts	15 ts	20 ts	25 ts	30 ts
5 docs	0.5146	0.5218	0.5073	0.5072	0.5147	0.5209
10 docs	0.5144	0.5143	0.5084	0.5223	0.5331	**0.5376***
15 docs	0.5087	0.5111	0.4962	0.5103	0.5236	0.5300
20 docs	0.5065	0.5044	0.4918	0.5047	0.5156	0.5209
25 docs	0.5026	0.5021	0.4904	0.5008	0.5096	0.5195
30 docs	0.5037	0.5014	**0.4896ˆ**	0.5016	0.5096	0.5204

5.3.8 Comparison between Query Expansion with EDS+

Furthermore, the proposed model were also compared with other techniques such as query expansion (suggested by Agirre *et al.* [2012]). The author used the method of query expansion based on random walks over the graph representation of concepts and relations in a knowledge base to obtain concepts related to queries. In order to confirm the validity of the proposed methods, results of the proposed methods were compared with their result which can be found in Agirre *et al.* [2012]. The settings for this experiment is exactly the same as stated in Section 5.3.1.

The following approaches were used in the experiments of the next section:

- LMED-D: LM with Extended Dirichlet and DC (the proposed model)
- QE-WordNet : Query expansion with Personalized Page Rank over graphs constructed from WordNet
- QE-Wikipedia : Query expansion with Personalized Page Rank over graphs constructed from Wikipedia

Table 5.16 summarizes the results obtained from the CHiC collection with the three methods namely LMED-D, QE-WordNet and QE-Wikipedia. It can be seen that in all cases, the proposed model (LMED-D) outperforms the other two methods. More precisely, the proposed model (LMED-D) outperforms the method of query expansion. Moreover, LMED-D achieves +7.00% more than the method of QE-Wikipedia. In a nutshell, this has proved that the proposed model performs better than the method of query expansion. In addition, the proposed model (LMED-D) achieves the increment of about 5.62% compared to the value from QE-WordNet which is highest ranked during the evaluation challenge of 2012.

In a nutshell, Figure 5.16 and Figure 5.17 shows the comparison between the LMED-D based on EDS+ with QE-WordNet, QE-Wikipedia, and the highest value in Pseudo-relevance feedback (10 documents, 30 terms). Basically, LMED-D based on EDS+ has a clear advantage over QE-WordNet, QE-Wikipedia, and Pseudo-relevance feedback. However, pseudo-relevance feedback performs better than LMED-D based on EDS+ at recall levels (X-axis) of 0.4, 0.5 and 0.6 (the gap is almost the same). In addition, Figure 5.17 shows the comparison of each query in different methods, and 8 queries from LMED-D based on EDS+ outperforms the other methods. The biggest gap between LMED-D based on EDS+ with other methods is Query 7 (schlesien map). Basically, Query 7 contains 12 relevant documents and only one document which contains the term of “schlesien” and “map” (Refer to Appendix B for more information on the real

Table 5.16: Results on Mean Average Precision Compared to Query Expansion method

Types of Approach	MAP	MAP Gain/Lost
LMED-D	**0.5451***	
QE-WordNet	0.5161	+5.62%
QE-Wikipedia	0.5094	+7.00%

documents in CHiC for this query). With this result, it shows the level of importance of the proposed models which is able to retrieve more relevant documents by exploiting the non-matching terms. It can be noticed that 12 queries from Pseudo-relevance feedback outperforms the other methods including LMED-D based on EDS+. However, the MAP for Pseudo-relevance feedback is non-statistically significant result ($p<0.62$).

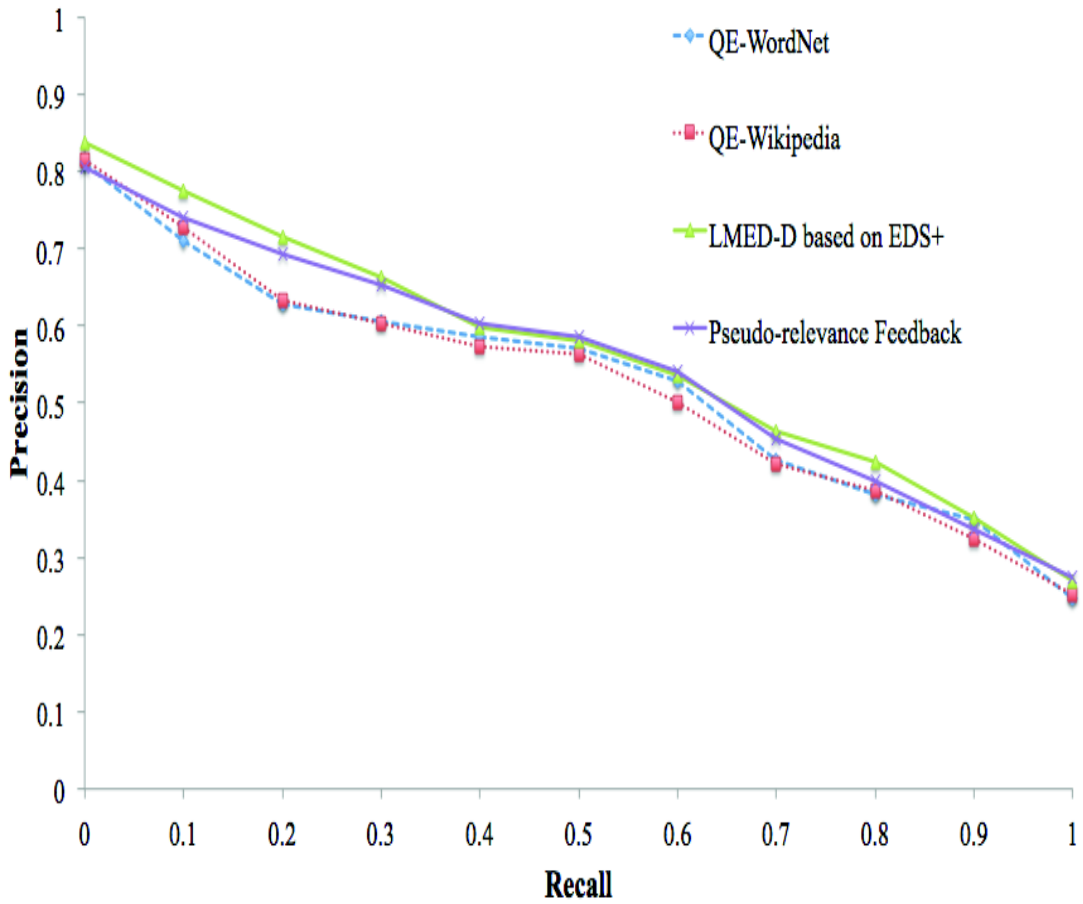


Figure 5.16: Comparison of LMED-D based on EDS+ with QE-WordNet, QE-Wikipedia, and Pseudo-relevance Feedback by using Average Eleven-point Precision/Recall Curve

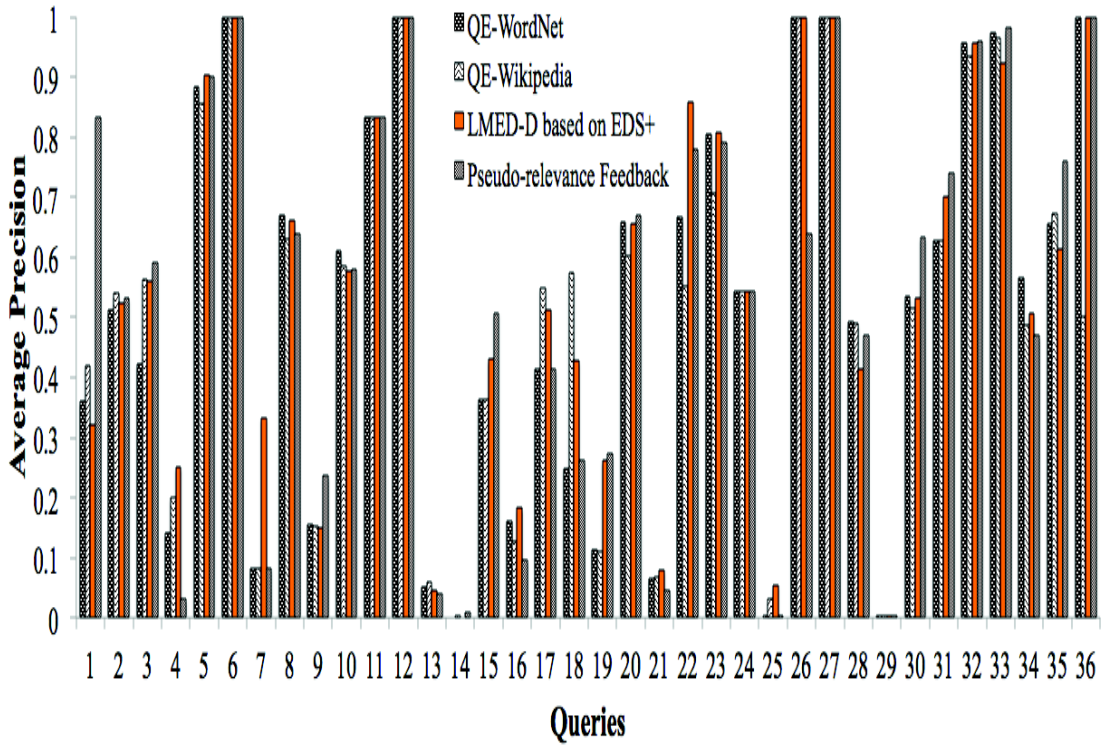


Figure 5.17: Comparison between the results of LMED-D based on EDS+ with QE-WordNet, QE-Wikipedia, and Pseudo-relevance Feedback

5.4 Evaluation of Information Access - Context Search

Several experiments were conducted in order to verify whether the user profile can be integrated together with the EDS framework. The same collection which is the CHiC collection were reused for this purpose. As mentioned before, the CHiC collection is a special collection which consists of books, bibles, paintings, and physical objects. Therefore, the documents in the collection are categorized based on media such as image, text, video and sound, and the number according to the category as given in Table 5.17. Firstly, an assumption is made that the user preference is based on the types of media. For example, when the user submits a query such as “mona lisa”, the initiative of the user is to retrieve the document which is of the type “text” such as the script of the opera written by Max von Schilings and not the painting by Leonardo da Vinci. So, if the user’s preference is “text”, then the Information Retrieval System manages to retrieve the documents based on the query of “mona lisa” and filters these documents based on the user’s preference.

In the experiments, the R-Precision was used to evaluate the result since R-Precision is the precision after R documents have been retrieved, where the value of R is the number of the relevant documents for the topic. So, it is a good measurement to de-emphasize the exact ranking of the retrieved relevant documents.

As explained before in the context search system, the component of the user’s profile can affect the retrieval in three distinct phases (refer to Appendix C for graphical interpretation) namely:

Table 5.17: Number of Documents according to Category

Media Types	Quantity	Percentage (%)
Image	1049622	94.81
Text	45821	4.13
Video	6564	0.59
Sound	5169	0.47
Total	**1107112**	**100**

Table 5.18: Number of Documents according to CHIC-032 (syriac)

Media Types	Quantity	Percentage (%)
Image	33	41.25
Text	47	58.75
Total Relevant Documents	**80**	**100**

- **Part of Retrieval Process**: the ranking is a unified process wherein user's profile is employed to retrieve the contents of the documents.
- **Re-ranking Activity**: user's profile takes part in a second step after evaluating the documents ranked via the approach of term intersection .
- **Query Modification**: user's profile affects and expands or modifies the query based on the needs of the user.

5.4.1 Experiment Setup

Only the CHIC-032 from the CHiC collection was used as the query in this experiment. Basically, this query contains 80 of the relevant documents as shown in Table 5.18 where the number of relevant documents of the media types *image* and *text* is 33 and 47 respectively.

In the experiments, three types of methods that were carried out namely: 1) Part of Retrieval Process, 2) Re-ranking Activity and 3) Query Modification. Firstly, both query and user profile were combined together in order to do the process of matching in the first method. Secondly, the term from query was used to retrieve the first one thousand documents, and then the documents were filtered based on the terms from the user's profile in the second method. Thirdly, the query was modified based on the preference of a user and several terms were added into the query since the size of the query is short for the third method.

The following approaches were used for experiments in the next section based on EDS+:

- LMDS-Q-UP: LM with Dirichlet Smoothing, query and user's profile
- LMDS-UP: LM with Dirichlet Smoothing and user's profile
- LMED-D-Q-UP: LM with Extended Dirichlet Smoothing based on EDS+ model, Dice Coefficient, query and user's profile

- LMED-D-UP: LM with Extended Dirichlet Smoothing based on EDS+ model, Dice Coefficient and user's profile
- LMDS-QM: LM with Dirichlet Smoothing and query modification

After the pre-processing as shown in Figure 5.5, the unigram language model or often called as bag-of-words model was used to index the title without any description from the queries, and the metadata from documents such as title, subject, description and types of media were used in the experiment (please be aware that the setting for this experiment is different from the setting in Section 5.3.1. The main difference over here is with additional metadata which is "types of media"). So, the exact ordering of the terms in the document is ignored. The purpose is to combine the content and the metadata together in order to evaluate the effectiveness of the ESD framework.

5.4.2 Result and Discussion

Table 5.19 shows the result of MAP and R-Precision with different types of method namely 1) Part of Retrieval Process, 2) Re-ranking Activity and 3) Query Modification. Firstly, LMDS-Q-UP achieves only 0.0303 (R-Precision) in the first method. The main reason for the poor result is because of the fact that it only returned one relevant document. As mentioned in Table 5.18, the total number of relevant documents is thirty three and the IRS only managed to retrieve one relevant document. So, this explains the poor result.

Four types of approach were carried out namely LMDS-Q-UP, LMDS-UP, LMED-D-Q-UP and LMED-D-UP in the methods of Re-ranking activity. Please note that the output from LMED-D from Table 5.5 was used which means that only the first one thousand documents returned by LMED-D were used before the process of re-ranking. The total number of relevant documents in the first one thousand containing twelve out of thirty three by using the output from LMED-D. In other words, the output contains the quantity of twelve documents which are relevant before the process of re-ranking. Based on the re-ranking activity, LMED-D-Q-UP achieves the best result which is 0.2424 (R-Precision) and 0.1428 (AP) compared to LMDS-Q-UP, LMDS-Q-UP and LMED-D-UP. The main reason that LMED-D-Q-UP outperforms others is because of the fact that the proposed model (EDS-DS) which is LMED-D was used in the process of re-ranking again. In other words, the non-matching terms are exploited again during the process of re-ranking compared to LMDS-Q-UP which does not exploit the non-matching terms. Therefore, LMDS-Q-UP only achieves 0.0303 in terms of R-Precision which means that only one relevant document appears in the ranking of the first thirty three documents. So, that is the reason that LMDS-Q-UP achieves the lowest score for R-Precision in the method of re-ranking activity. It is necessary to explain the reason for LMDS-UP and LMED-D-UP to achieve the same score in terms of AP and R-Precision. This is because LMED-D-UP does not contain any terms such as "image" in the Term Similarity Matrix. Since there is no similar terms found in the Term Similarity Matrix, then the EDS+ model will retrieve the documents based on the approach of term intersection only. Although, LMDS-Q-UP achieves better score for AP when compared to LMDS-UP, LMDS-UP achieves better score for R-Precision. This is due to the fact that LMDS-UP retrieved two relevant documents in the ranking of the first thirty three documents.

Table 5.19: Results on Mean Average Precision and R-Precision Compared to Various Types of Methods for Context Search

Types of Method	Types of Approach	MAP	R-Precision
1) Part of Retrieval Process	LMDS-Q-UP	0.0061	0.0303
2) Re-ranking Activity	LMDS-Q-UP	0.0103	0.0303
	LMDS-UP	0.0089	0.0606
	LMED-D-Q-UP	0.1428	**0.2424**
	LMED-D-UP	0.0089	0.0606
3) Query Modification	LMDS-QM	0.0563	0.1515

Table 5.20: Results on R-Precision compared to various types of Approach for Context Search

Types of Approach	R-Precision	Gain
LMED-D-Q (baseline)	0.2121	-
LMED-D-Q-UP	0.2424	+14.70%

Finally, LMDS-QM achieves 0.1515 (R-Precision) in the method of query modification. Although LMDS-QM managed to retrieve twelve relevant documents, only five relevant documents appeared in the ranking of the first thirty three documents.

Based on Table 5.20, LMED-D-Q-UP achieves 0.2424 while the baseline only achieves 0.2121 which is almost on increment of 14.70% compared to the baseline. Please take note that Language Model with Extended Dirichlet and Dice Coefficient (LMED-D) were used as the baseline and the approach of LMED-D is without specifying the preference of a user. In other words, the baseline is not integrated with any user profile. In a nutshell, LMED-D-Q-UP in the method of re-ranking activity outperforms the other methods. So, it is showed that EDS framework can be integrated with user profile.

As mentioned in Section 1.1, the main motivation of this research comes from a case study involving a mobile game on cultural heritage. Therefore, a case study was carried out in the next section to evaluate the effectiveness of delivering information on cultural heritage to the user.

5.5 Evaluation of Information Access - Mobile Game

A questionnaire has been designed to obtain feedback from various parties in order to evaluate the effectiveness of delivering information on cultural heritage to the user through mobile game. User-centred evaluation is used rather than system-centred evaluation because knowledge of a user is qualitative. Therefore, questionnaire was used to evaluate the mobile game. The main objective of the questionnaire is to gather the feedback and comments on the mobile game based on the perspective of the user. The questionnaire consists of four questions, namely 1) Is the game interesting?, 2) Have you learnt any basic information of the heritage sites?, 3) Have you

gained any new knowledge from the games?, and 4) Would the information be sufficient in guiding you when a visit is made to George Town?. The respondents were required to rate their satisfaction based on a scale of 1 to 7 where the value of 1 is the lowest and the value of 7 is the highest degree of satisfaction in all of the questions. Question (1) is a general question while the objective of Question (2) is to evaluate the inquiry learning of the participants where the active learning and the investigative nature through questions and answers are expected to impart some basic information of the heritage sites to the participants. Question (3) aims to evaluate the experiential learning of the participants in order to gauge whether any new knowledge has been gained through the transformation of experience as provided by the game. Questions (4) is the most important question with regard to this research which focuses on information access and for the mobile game the focus is also on the user of information access that exhibits a casual behaviour. It aims to evaluate the situated learning of the game by gauging the needs for better information access through exploration and practical activities as provided by the mobile game. The questionnaire is given in Appendix D.

In the evaluation of the mobile game, forty participants were invited to evaluate the game contents. The respondents comprise non-experts in the field of Information Technology and most of the respondents are Master or PhD students from overseas who arrived in Penang, Malaysia for the last 3 months (little or without any cultural heritage's knowledge in Penang). The group consisted of 34 males and 6 females with an age 32.4 and a proficiency with English. The result of the evaluation is summarized in Table 5.21. 65% of the respondents have given an excellent rating (6 or 7) on whether the game is an interesting game. For the second question on whether the respondents have learnt any basic information of the heritage sites in George Town, only a minority which is three respondents (15%) have given an average rating of 5 while the rest have given an excellent rating (6 or 7). All respondents gave an excellent rating (6 or 7) (with 65% of them gave the highest rating) on whether they have gained any new knowledge after playing the game in the Question (2). For the last question, 75% of the respondents (with rating of 4 or 5) highlighted that the information provided by the mobile game is just enough for them to explore George Town when a visit is made to the city. Therefore, there is a need for better and effective tool such as information seeking tool that can be integrated with the mobile game to provide more information to the user while playing the game.

Also from the evaluation, some useful feedbacks on the effectiveness of the mobile game were obtained. Many respondents thought that the content is interesting with regard to the heritage information on the heritage area and its environment and culture. The respondents also commented that more heritage sites should be included and should not be limited to the seven heritage sites. On the learning content, many respondents agreed that the mobile game has the ability to deliver the educational content such as the history of the heritage sites and the views of the heritage sites. Furthermore, they got to know the history and the architectural design of the buildings. In general, the majority of the respondents felt that the interface of the mobile game is user friendly for game playing and can be an excellent guide for them when visiting the city. Finally, some of the respondents proposed that the mobile game should be made available on other platform such as Android, Symbian, Window Mobile, Blackberry and not only on iOS platform. An example of the simulation run of the game is given in Appendix E.

Table 5.21: Result of the Evaluation of Mobile Game

Rating Scale	**Questions**			
	1) Is the game interesting?	2) Have you learnt any basic information of the heritage sites?	3) Have you gained any new knowledge from the game?	4) Would the information be sufficient to guide you when you make a visit to George Town?
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	4	0	0	6
5	14	6	0	24
6	20	20	14	10
7	2	14	26	0
Total	**40**	**40**	**40**	**40**

5.6 Summary

This chapter has focused on evaluating three research questions namely :1) examine and validate whether the knowledge or context can help to improve the retrieval effectiveness in Information Retrieval System, 2) Examine the query modification (global change) and the matching adaptation (local change), and 3) examine and investigate whether mobile game with instant access to information is able to help the user to access, learn, and explore the information on cultural heritage. In summary, it can be concluded that Research Question 1 and Research Question 2 have been supported by the results as shown in the Section 5.3 and Section 5.4. In addition, Section 5.3.8 also has proved that matching adaptation outperforms the query modification. Finally, the Research Question 3 have been supported by questionnaire as given in the Section 5.5.

Part VI

CONCLUSION AND FUTURE WORDS

Chapter 6

Conclusion and Future Work

This chapter presents the conclusion, contribution, and future works in this research. Section 6.1 shows the overall conclusion of this research while Section 6.2 shows the future work or future direction in this research area.

6.1 Conclusion

The main reason to integrate the knowledge into the matching function is to solve the shortcomings of term intersection approach where it only returns the document if the term from the query appeared in the document. One of the ways to solve the shortcomings of term intersection approach is through Query Expansion with knowledge. The most fundamental idea of Query Expansion is executing the first run (or initial ranking) before expanding the terms to the query according to the knowledge. The second run (called final ranking) is required in order to complete the whole process. The whole process of Query Expansion is considered as global adaptation because the query transforms Q to Q' after expanding the terms. Then, Q' will do the process of matching again with all the documents.

In this research, the results showed that the proposed solutions are able to solve the problems arising from the documentation process and the background of the users of digital cultural heritage. A model on Information Retrieval System that allows the user to access, learn, and explore through free-text query has been developed. A model of context search based on the preference of a user by quantifying the level of preference of the user has also been developed. A case study to investigate the mobile game as a tool to help the user with casual behaviour to access, learn, and explore the information on cultural heritage has also been carried out. In conclusion, it can be summarized that the contribution of this research are as follows:

- **Information access - Information Retrieval System:** A framework based on Dirichlet Smoothing to exploit the non-matching terms during retrieval by using external knowledge. Then, two models have been introduced based on this framework which are Extended Dirichlet Smoothing with Increase of Document Size (EDS+) and Extended Dirichlet Smoothing with Decrease of Document Size (EDS-).
- **Information access - Context search:** A Preference Model based on the Theory of Charging a Capacitor that is able to quantify the cognitive context and integrated into the model of EDS+.

- **Information access - Mobile Game:** An environment with education and entertainment that is able to provide access, learn and explore the information on cultural heritage.

In addition, it can be concluded that Research Questions of this research have been answered as follows:

- **Research Question 1:** Knowledge or context can help to improve the retrieval results.
- **Research Question 2:** Matching adaptation outperforms the query modification.
- **Research Question 3:** Mobile game is necessary to support learning activities in the information on cultural heritage to user.

Figure 6.1 shows the overall concept of this research. Basically, the works stand in between the Information Retrieval System, Context Search, and Mobile Game. Two scenarios that can show that Mobile Game can be integrated with Information Retrieval System are as follows:

- Through the Treasure Hunt
- On site visit

In general, the user context can be collected when the user plays the Mobile Game. Besides, the user can operate the Information Retrieval System when the user plays the Mobile Game especially in the Treasure Hunt. As mentioned before, the Treasure Hunt requires the user to answer some Questions (inquiry learning) in order to proceed with the game. So, the user can choose to operate the Information Retrieval System to retrieve the relevant documents. In addition, the Mobile Game is able to support situated learning where the user can operate the Mobile Game especially in the Gallery module when the user is on site at the heritage site. If the user wants to have more information, the user is able to use the Information Retrieval System to find more information. Therefore, this is the main motivation of this research in Information Retrieval System and Context Search.

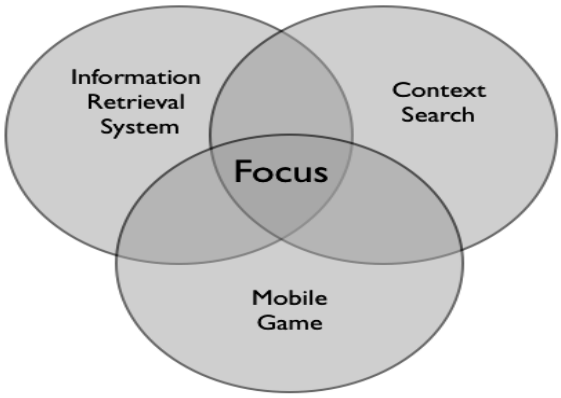


Figure 6.1: Integration between Information Retrieval System, Context Search, and Mobile Game

6.2 Future Work

6.2.1 Information Retrieval System

The domain of cultural heritage involves several other domains of research such as education, genealogy, history, and archaeology. In Chapter 5, it has been shown that knowledge tends to improve the retrieval result. Therefore, the proposed framework and models are able to adapt in the domain of cultural heritage which involves several researchers of various domains as shown in Figure 6.2. For example, a dedicated Information Retrieval System can be provided which only uses the domain knowledge from "history" to extract the term links, and apply this knowledge into the retrieval model. So, the historian can use the dedicated Information Retrieval System to retrieve the documents. It is necessary to highlight that the terms used in each domain is different although the term is the same. For example, in the medical domain, "X-ray" means an image obtained by the used of X-ray radiations, whereas in Physics domain, "X-ray" means the radiations only. Therefore, this approach can be applied to other domains as well. Since the domain of Cultural Heritage involved several parties of researcher, then it is good to integrate their own knowledge into the retrieval model.

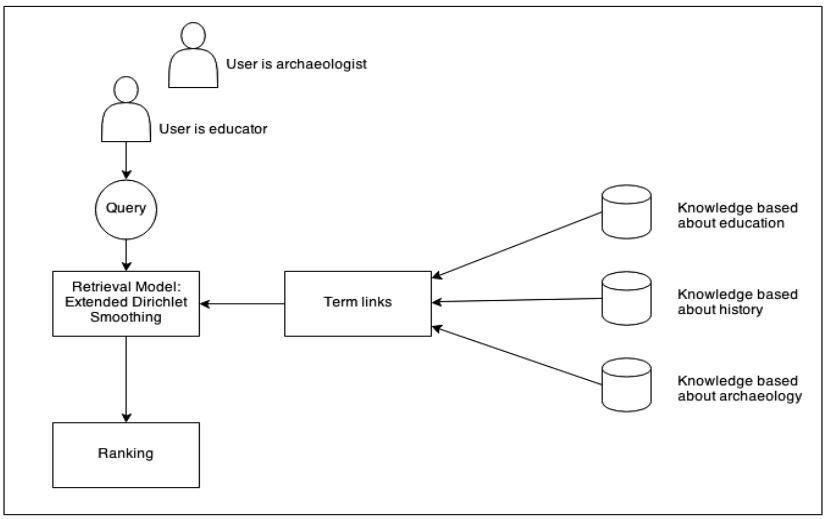


Figure 6.2: Multiple Users from different Domains to Access the Information on Cultural Heritage

Ranking in social media poses different challenges than Web search ranking, one of which is that Microblog messages are really short. For example, Twitter messages are limited to 140 characters. This limit causes the user to write in an abbreviated form which increases the likelihood of having the vocabulary gap (such as term mismatch problem). The vocabulary mismatch/gap problem means that the user query uses one term and the relevant documents use another term and there is a need to match the query with documents. So, the future direction is to study whether and how the proposed framework and models can be used to solve this problem in Microblog. In the domain of medicine, the concept can be extracted from the domain knowledge based which is Unified Medical Language System (*UMLS*) based on a hierarchical approach.

Then, we can integrate the link of hierarchy into the matching model rather than term links as proposed in this research.

For future work, it would be interesting to propose some other similarity measure methods especially in knowledge-based and corpus-based approach. In addition, it would also be interesting to integrate the Term Similarity Matrix in other smoothing methods such as Jelinek-Mercer smoothing. Another interesting direction is to study how to transfer the knowledge learned from one collection to another collection.

6.2.2 Context Search

For the future work, the context search can be integrated with software agent especially "Interface Agent" to collect various type of context from user. Maes & Kozierok [1993] proposed an Interface Agent where the agents can learn and observe the user behaviour by embedding it in some application such as database or spreadsheet. Furthermore, Maes & Kozierok [1993] reported the following four key ways in which the agent can perform learning:

- The agent can learn by observing and imitating the action of the user.
- The agent can learn through receiving feedback from the user. The positive feedback refers to correct prediction while negative feedback refers to incorrect prediction.
- The agent can learn by receiving explicit instruction from the user.
- The agent can learn by asking other agents for advice.

Nwana [1996] showed that an informal explanation of interface agent should emphasise autonomy and learning which can perform a task such as "personal assistants" for their owners and emphasise less on inter-agent collaboration. Therefore, it is believed that the agent is able to learn the user's context when the user is playing the mobile game. The agent can collect several types of context such as the device context (the characteristics of the device), spatio-temporal context (location of the user through such as the location of the user through the Global Positioning System, *GPS*), and user's context. After gathering the user's context from the mobile game, the user's context can be exploited during the retrieval time if the user submits a query.

6.2.3 Mobile Game

With the rise of gamification, it can be argued that Mobile Game can be integrated with Information Retrieval System or Context Search. The concept of gamification is to use the elements of thinking and mechanics from game to engage the users to solve certain problems when playing the game. The core for gamification is to reward the players who accomplish the desired tasks required from the game and the rewards can be points, achievement badges, progress bar, or virtual currency (bitcoin[1]). Gamification has been widely applied in marketing and most of the companies listed by Forbes Global 200[2] plan to use gamification for the purpose of marketing

[1] http://bitcoin.org/en/

[2] http://www.forbes.com/sites/jeannemeister/2012/05/21/gamification-three-ways-to-use-gaming-for-recruiting-training-and-health-amp-wellness/

and customer retention in the year of 2013. Furthermore, gamification can be used to generate some new idea in the company as studied by Modern Survey[1] (human capital measurement and analysis firm). In addition, Microsoft also released the game *Ribbon Hero 2* [2] as an add-on to *Microsoft Office* to help the people to use it effectively. However, the adoption of gamification in Information Retrieval is still in the stage of infancy. In general, gamification can be applied in the domain of Information Retrieval especially in the task of manual annotation of documents in the evaluation of Information Retrieval. In addition, gamification also can be applied in Context Search (Interactive Information Retrieval System) where it can collect the various type of user context when the user is playing the game.

[1] http://www.ams-inc.com/newsletter/voices042.asp
[2] http://www.microsoft.com/en-us/download/details.aspx?id=26531

Part VII

APPENDIXES

Appendix A

A Statistical Box Plot

Box plot or boxplot is is the most useful statistical tool when comparing two or more sets of sample data. In other words, box plot is a way to summarize a set of data that are measured on an interval scale. Box plot is usually used in exploratory data analysis and it is a type of statistical tool that can be used it to show the state of the distribution, the central value, and the spread of the value. The diagram of box plot as shown in Figure A.1 usually consists of several values in the dataset which are;

- maximum,
- minimum,
- first quartile (Q1),
- second quartile (Q2) or median,
- third quartile (Q3),
- mean,
- interquartile range,

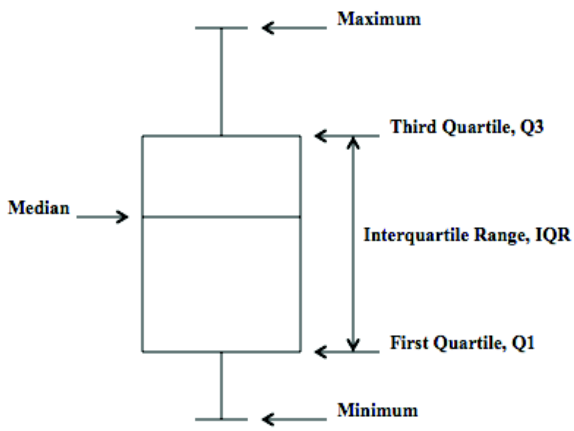


Figure A.1: A Statistical Box Plot

Furthermore, box plot is also very useful when the size of the data (also known as the number of observations) is large and when two or more datasets are being compared. Box plot is useful to indicate whether the distribution is skewed or the data set contains unusual observation such as "outliers" and "extremes". For example, a distribution with a positive skew would have a longer whisker in the positive direction than in the negative direction. Furthermore, a larger mean than the median would also indicate a positive skew. Let assume n is the number of observations where $x_1, x_2, x_3, ..., x_n$ are the values sorted from the lowest to the highest, then we can find the values for maximum, minimum, median, first quartile, third quartile, mean and interquartile range as follows;

Maximum: The highest number in the dataset.

Minimum: The lowest number in the dataset.

Median: The value of the median is the middle value in the dataset after the dataset was sorted. If the dataset has even number of values then the median is the average of the two middle values in the dataset. The median is a measure of central tendency and it represents the value for which 50% of the observations are lower and 50% are higher. In other words, it is the value at the centre of the sorted observations.

When the number of observations is odd, then we can use the following equation to find the position of the number:

$$\frac{x_{n+1}}{2} \tag{A.1}$$

When the number of observations is even, then average of the two middle numbers would be the median.

Quartiles: Quartiles by definition are to separate a quarter of data points from the rest. This means that the first quartile is the value under which 25% of the dataset are found and the third quartile is the value over which 75% of the data are found.

First Quartile, Q1: First quartile is the median of the lower half of the data. If the number of data points is odd, the lower half includes the median. In other words, Q1 is the number for which 25% of the values in the dataset are smaller than $Q1$.

$$\frac{1}{4}(n+1) \tag{A.2}$$

Third Quartile, Q3: Third quartile is the median of the upper half of the data. If the number of data points is odd, then the upper half of the data includes the median. In other words, Q3 is the value such that 75% of the observations are less than $Q3$.

$$\frac{3}{4}(n+1) \tag{A.3}$$

Mean: Mean is commonly known as the average of a set of numerical values. It is a measure of central tendency or a value that estimates the center of a set of numbers. In other words, it is the sum of all values and divided by the total number of values.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i) \tag{A.4}$$

Interquartile Range: The interquartile range (IQR) is the distance between the upper and lower quartiles and contains 50% of the data points around the median. It is a measure of dispersion

and can be calculated IQR as follows:

$$IQR = Q_3 - Q_1 \tag{A.5}$$

IQR is used as a metric to denote two ranges of the data that is shown in Figure A.2. The "inner fence" is 1.5 times the IQR (1.5*IQR) beyond the first and third quartiles such as beyond the box. The "outer fence" extends 1.5 times of the IQR beyond the inner fence such as 3 times IQR (3*IQR) beyond the box. These ranges allow classifying the data points beyond the IQR as "outliers" or "extremes". Outliers are denoted by an "x" symbol while extreme values are denoted by "o" symbol as shown in Figure A.3.

Outliers: An outlier is any data point that is beyond the inner fence but within the outer fence.

Extreme: An extreme is any data point that lies beyond the outer fence.

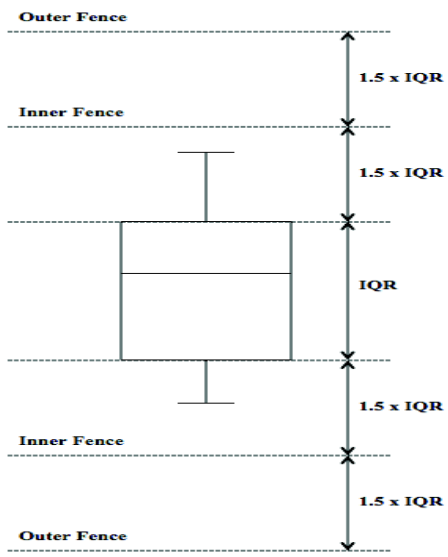


Figure A.2: Landscape of A Statistical Box Plot

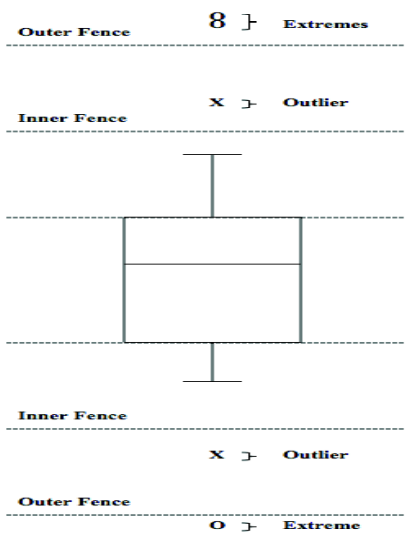


Figure A.3: Outliers and Extremes of a Statistical Box Plot

Table B.1: Query 7 - “schlesien maps”

Number relevant documents:	12
Number relevant documents that retrieved by IRS:	From 1 (without TSM) to 2

Appendix B

Analysis in CHiC 2012

B.1 Analysis for Query 7 “schlesien maps”

Table B.1 shows the total number of relevant documents and total number of relevant document that retrieved by Information Retrieval System. Basically, baseline result only can retrieve 1 relevant document while proposed models which are EDS+ and EDS- managed to retrieve 2 documents. The reason we analysed this query is because we achieved a huge improvement (average precision) for this query. In general, only 1 relevant document contains the term “schlesien” and “map” (in bold) which shown in Table B.2. Please be aware that the rest of the relevant documents for this query does not contain the term of “schlesien”. The main reason the proposed models can retrieve extra one document compared to baseline is because the term “silesia” is inside the Term Similarity Matrix. Therefore, the proposed models able to retrieve the document which contains the term “silesia” (in bold) as shown in Table B.3.

B.2 Analysis for Query 25 “syriac”

As mentioned before, Query 25 contains 80 relevant documents as shown in Table B.4. For this query, we would like to remark that almost 41 relevant documents are using the same title and description as shown in Table B.5. The main reason that 41 relevant documents are using the same title and description is because 41 relevant documents (descriptive representation) refer to the same prayer book (physical representation).

Table B.2: One of the Relevant Document for Query 7

Description	Concept: propaganda, cartography, portrait / personification, advertisement, eating / drinking, business / trading / markets, animals, military personnel, caricature, humorous, royalty, aircraft, motor vehicles Description: whole: the image occupies the whole, held within a narrow green border. The title is integrated and positioned in the top left corner, in green. The text and a small image are separate and positioned lower right, in green, held within a scroll design. image: a map of Europe, with each country represented by depictions of soldiers, animals or caricatured personifications. Britain is represented by John Bull, a lion, and a spider whose legs reach across the English Channel to France. in **Schlesien** Grosste Cognac-Brennerei Deutschlands. Die besondere Genehmigung zur Fuhrung des Namens “Hindenburg” ist von Sr. Exzellenz dem Herrn General-feldmarshal von Hindenburg erteilt worden. I Hindenburg of the same quality and vintages as was supplied to the Officers’ mess at the main Headquarters of His Majesty Kaiser Wilhelm II. Hindenburg Brandy Sales Co. Ltd. Grunberg in **Schlesien** [now Zielona Gora] Grunberg in **Schlesien** Brandy Distillery. Biggest Brandy Distillery in Germany. Special permission to use the name ’Hindenburg’ was granted by His Excellency Field Marshal General von Hindenburg. In restaurants and delicatessens ask expressly for German ’Hindenburg’ brand brandy and liqueur. Intaglio printing by Knackstedt and Co., Hamburg 22. Censorship Altona No. 831] Object: **maps**, steamroller, car.”

Table B.3: One of the Relevant Document for Query 7 which Contains the Term of “silesia” from Term Similarity Matrix

Description	Concept: cartoons, propaganda, cartography, animals, historical reference Description: whole: the image occupies the majority, held within a black border. The title is separate and positioned across the top edge, in black. The text is separate and positioned in the upper right, also in black, and held within a black border. Further text is integrated and positioned throughout the image, in black. All set against a white background. image: a map of Europe, on top of which sit two octopi, one representing Prussia, the other representing Austria-Hungary. The tentacles of each octopus spread out to grab other countries which the two powers have invaded or annexed. text: LA SERBIA invasa e devastata dalla Germania e dall’ Austria-Ungaria nel 1915. IL MONTENEGRO invaso e devastata nel 1915. H. AND C. GRAHAM LTD., LITHRS., LONDON. S E [THE PRUSSIAN OCTOPUS. ’We do not threaten small nations,’ declared the German Chancellor on December 10th, 1915; ’we do not wage war which has been forced upon us in order to subjugate foreign peoples, but for the protection of our life and freedom.’ This pictorial map is a commentary on his words. It shows how Prussia has stolen one province after another from her neighbours and, like a baleful octopus, is still stretching out her tentacles to grasp further acquisitions. The territories included in the original Kingdom of Prussia are marked thus [hatched lines]. The territories since absorbed by negotiation, force, or fraud are marked thus [diagonal lines]. SCHLESWIG HOLSTEIN AND LAUENBURG Wrested from Denmark in 1864. ANNEXED BY PRUSSIA IN 1866. POLISH TERRITORY STOLEN BY PRUSSIA IN 1772, 1793 and 1795. BELGIUM invaded and occupied by Germany in defiance of her treaty obligations, in 1914. ACQUIRED by PRUSSIA in 1813. ANNEXED BY PRUSSIA IN 1866. **Silesia** seized by Prussia from Austria in 1740. ALSACE LORRAINE torn from France in 1871. FEDERATED WITH PRUSSIA SINCE 1870. AUSTRIA HUNGARY SUBJUGATED BY PRUSSIA IN 1866. ALLIED WITH PRUSSIA SINCE 1879. BOSNIA AND HERZEGOVINA annexed by Austria-Hungary in defiance of her treaty obligations in 1908. SERBIA invaded and devastated by Germany and Austria-Hungary in 1915. MONTENEGRO invaded and devastated in 1915] Object: map

Table B.4: Query 25 - “syriac”

Number relevant documents:	80
Number relevant documents that retrieved by IRS:	From 1 (without TSM) to 12

Table B.5: 41 of the Total Relevant Documents are Using the Same Terms inside the Title and Description for Query 25 (“syriac”)

Title	The form of daily prayers according to the custom of the german and polish jews
Description	Prayer-book in Hebrew, Aramaic and English with hard smooth ivory-coloured covers and decorated on front cover with gold- and silver-coloured metalwork, ivory and velvet. Repository/location jewish museum london

Appendix C

Three Distinct Phases that Exploit the User Profile

In context search, three types of the component of the user's profile can affect the retrieval namely:

- Part of Retrieval Process as shown in Figure C.1
- Re-ranking Activity as shown in Figure C.2
- Query Modification as shown in Figure C.3

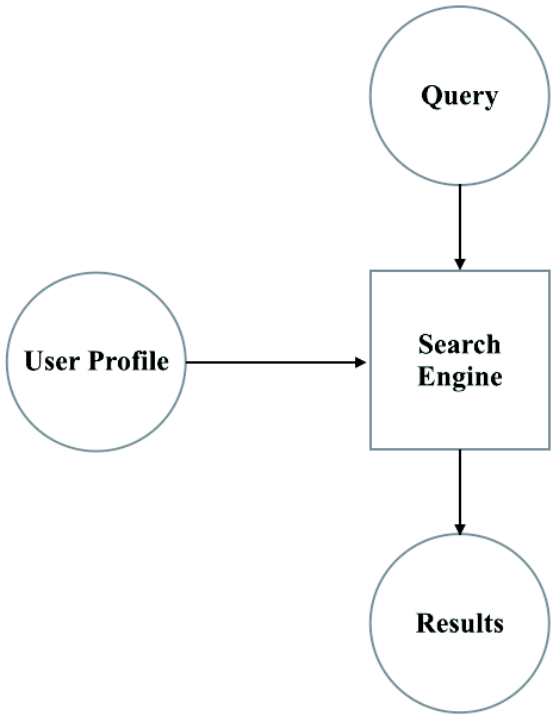


Figure C.1: Context Search: User Profile Occurs during the Process of Retrieval

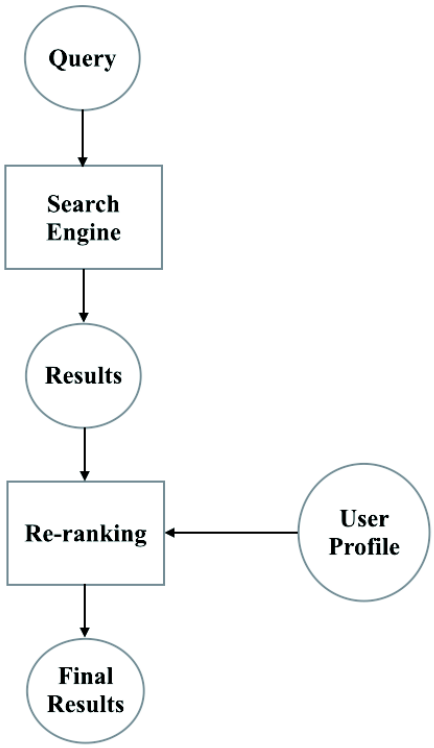


Figure C.2: Context Search: Distinct Re-ranking Activity

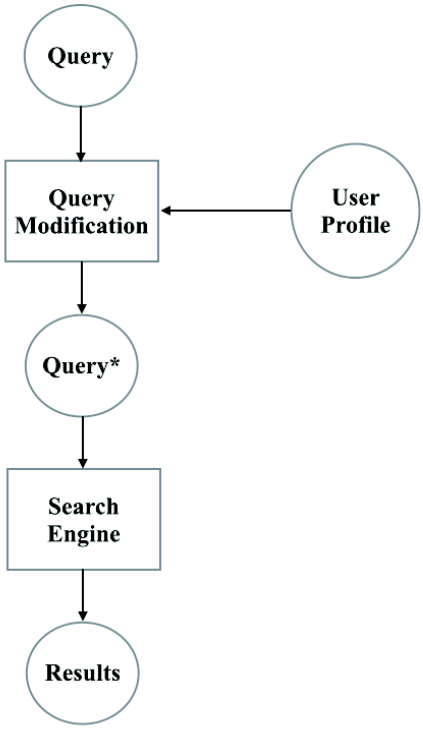


Figure C.3: Context Search: Pre-processing of the User Query

Appendix D

Questionnaire

QUESTIONNAIRE ON
MOBILE GAME – M-HERITAGE HUNT

Name: _____ Date:_____

Address:

_____

Designation: _____ Area of Expertise: _____

The questionnaire contains four parts and they are general question, inquiry learning, experiential learning, and situated learning. Please rate the general question, inquiry learning, experiential learning, and situated learning.

Please circle the number.

Rating scale: 1 – 3: Lowest, 4 –6 Average, 7 – Highest

General Question

		Please Circle one						
1	Is the game interesting?	1	2	3	4	5	6	7

Inquiry Learning

		Please Circle one						
1	Have you learned any basic information of the heritage sites?	1	2	3	4	5	6	7

Experiential Learning

		Please Circle one						
1	Have you gained any new knowledge from the game?	1	2	3	4	5	6	7

Situated Learning

		Please Circle one						
1	Will the information be sufficient to guide you when you make a visit to George Town?	1	2	3	4	5	6	7

Please write feedback on your satisfaction of the mobile game.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Thank your for your help!

Appendix E

Simulation Run

This appendix describes a typical flow for various menu and functions in M-Heritage Hunt. Figure E.1(a) until Figure E.1(f) show the screenshot of the "How to Start". Figure E.1(b) shows all the steps involved in starting the game while Figure E.1(c) until Figure E.1(f) show the description of the individual step and Figure E.2 shows the further highlight of Figure E.1(f). The top bar of the layout displays the name of the current player (Player's Name), the total amount of money the player has in his/her possession (Cash), and the total number of buildings the player owns (Property). The top bar also indicates if the player owns any medical card (Medical Card). The card is used to discharge the player from the hospital during the game. Besides, the top bar also displays the number of rounds the player has missed his turn that is the number of times the player could not roll the dice (Missed Turn). The bottom bar of the layout enables the player to click on his/her information (Player's Info). A new frame displays a summary of the name of the buildings, the total amount of money and the medical card that the player owns. Besides, it also includes the building information for the player if the player would like to view further information on the building (Building's Info). It will also direct the player to the Gallery component of M-Heritage Hunt. The bottom bar also contains icon that indicates the end of the player's turn that is when the game is passed to the next player (Next Player). It also has an icon for the player to roll the dice and the number of times the dice has been rolled will be shown next to it (Roll Dice).

Figure E.3 shows the information for each player. The player can view their cash, current location, medical insurance and properties. Table E.1 provides the explanation for Figure E.3 icons for (A) to (G) and Table E.2 provides more explanation for Figure E.3 for icons (H) to (R).

Figure E.4(a) until Figure E.4(d) show the steps involved in launching a new game. Figure E.4(b) shows the settings before starting the new game. The explanation on the settings can be found in Figure E.4(c). The player can enter their name and a short summary of the settings is displayed before starting the game as shown in Figure E.4(c). Figure E.4(d) is the screenshot at the start of the game.

Figure E.5 shows the step involved in resuming the game. By selecting "Resume Game" as shown in Figure E.5(a), the player can continue to play the game as shown in Figure E.5(b).

Figure E.6(a) until Figure E.6(d) show the steps involved in launching the "Gallery". "Gallery" lists out all the heritage sites as shown in Figure E.6(b). Figure E.6(c) shows the view that can be found in the particular heritage building and Figure E.6(d) shows the panoramic views after

Figure E.1: Screenshots of "How to Start"

the player has selected the views in Figure E.6(c).

Figure E.7(a) until Figure E.8(k) show the steps involved in launching the "Help" as shown in Figure E.7(a). The "Help" contains three files which are "Game Rules", "Icon Directories" and "Credits" as shown in Figure E.7(b). Figure E.7(c) until Figure E.7(e) list out the "Game Rules" while Figure E.7(f) until Figure E.8(j) show the description of all the icons. Figure E.8(k) shows the "Credits" which is on the acknowledgement for the application.

Figure E.9 shows the steps involved in upgrading of a property during the game. Upgrading the property owned by a player is one of the components of the game. The purpose of upgrading a property is to help the player to earn more rentals if an opponent steps on the property. The property can be upgraded up to level 3 which is the maximum level. Figure E.9(a) shows two players at the start cell which are Malay and Chinese. The players need to touch the dice as pointed to by the green arrow in Figure E.9(a) in order to start the move. Figure E.9(b) shows

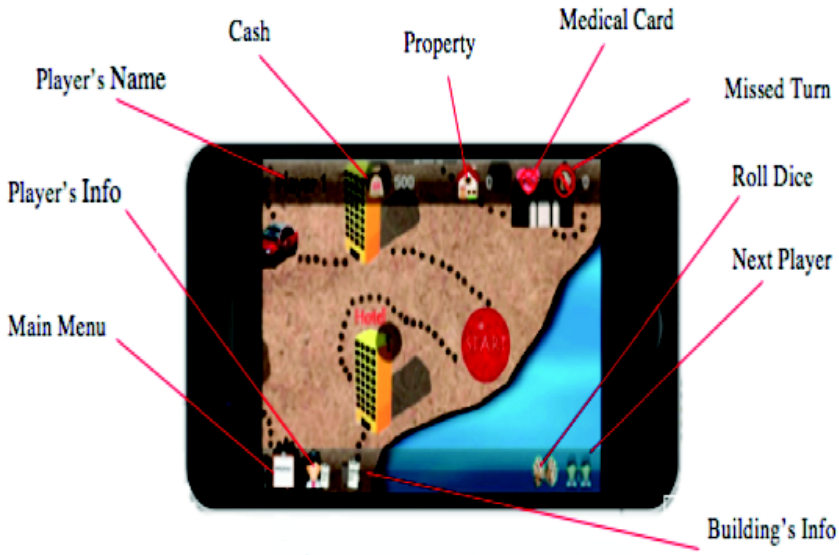


Figure E.2: Further Highlight of Figure E.1(f)

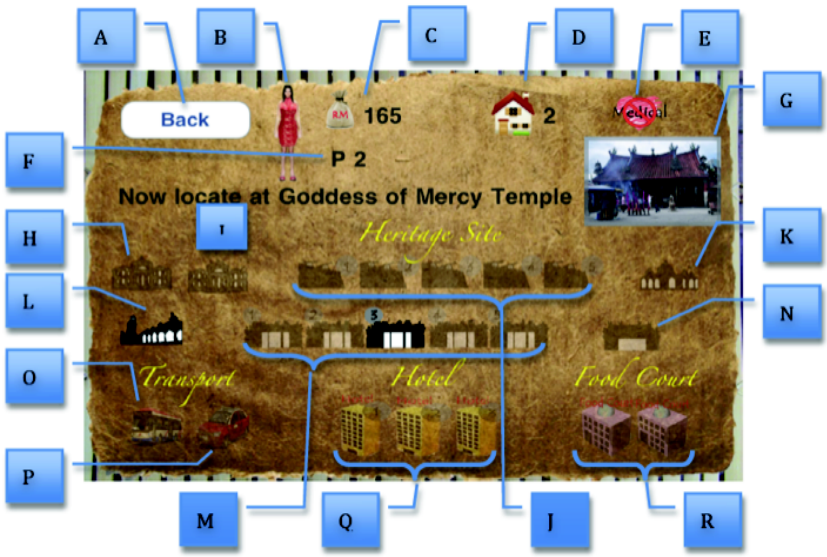


Figure E.3: Information for Each Player

one of the players stepping on the Cabin (Fort Cornwallis) cell and an alert box pops out to inquire the player on whether he or she would like to buy the property. For example, the player would like to buy the property was owned by the Chinese icon indicated by the red flag beside the property. Once the Chinese icon steps on the same cell with the same property again, an alert box will pop out and enquire the player to upgrade the property as shown in Figure E.9(d). The player needs to hunt it in order to upgrade the property. Figure E.9(e) shows the question after the player clicks on "Hunt it" in Figure E.9(d). If the player answers the question correctly,

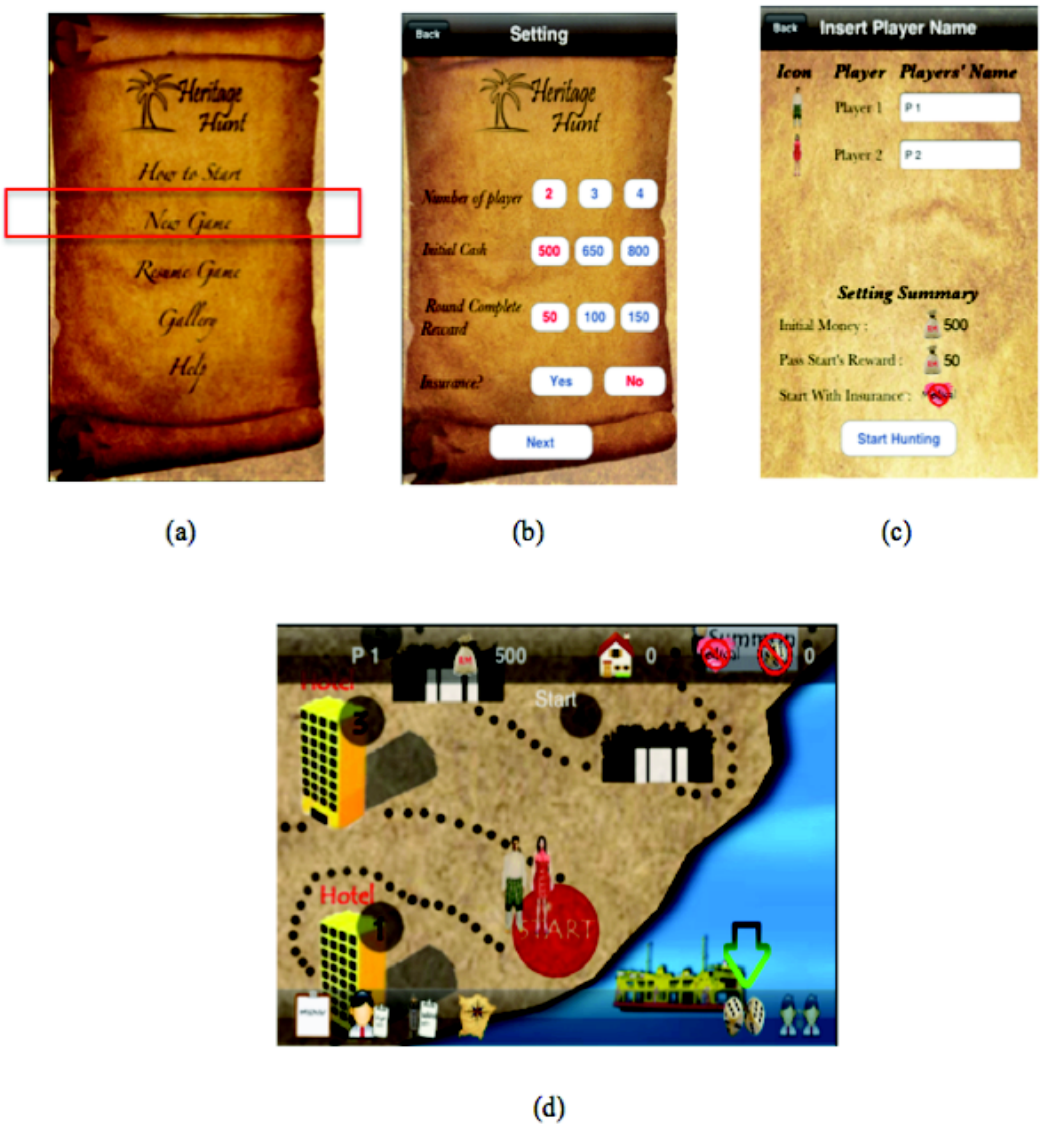
Table E.1: Description of Games Icon for Figure E.3

Icon	Representation	Description
A	Back	Tap her to resume the game
B	Icon	Tab here to show the icon owned by the player
C	Cash	Tab here to show the cash owned by the player
D	Property Owned	Tab here to show the property owned by the player
E	Medical Insurance	Tab here to show the insurance owned by the player
F	Player Name	Tab here to show the name of the player
G	Current Location	Tab here to show the current location of the player

Table E.2: Description from Figure E.3

Icon	Description
H	City Hall
I	Town Hall
J	Fort Cornwallis
K	Kapitan Keling Mosque
L	Acheh Mosque
M	Khoo Kongsi
N	Goddess of Mercy Temple
O	Rapid Penang Bus
P	Taxi
Q	Hotel
R	Food Court

then the player can upgrade the property as shown in Figure E.9(e).



(a) (b) (c)

(d)

Figure E.4: Screenshots of Starting a “New Game”

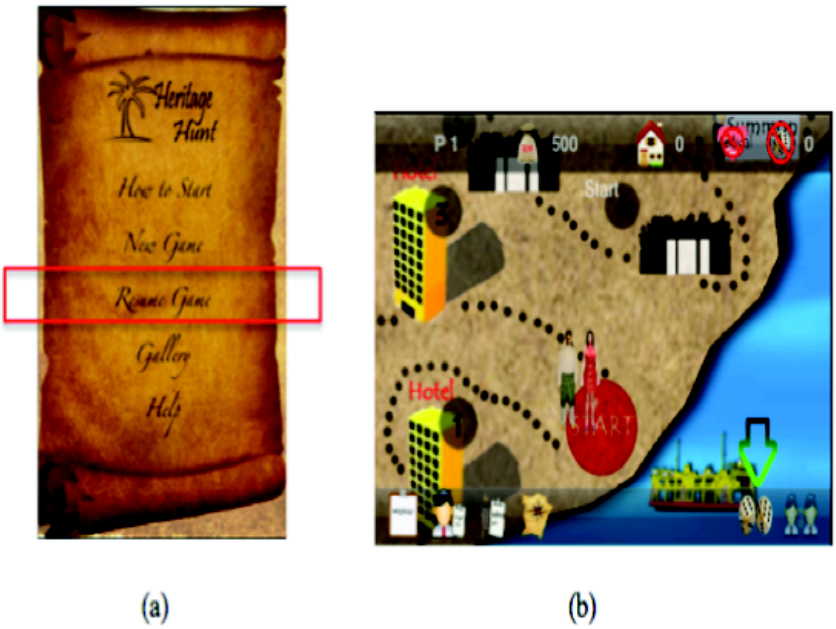


Figure E.5: Screenshots of “Resume Game”

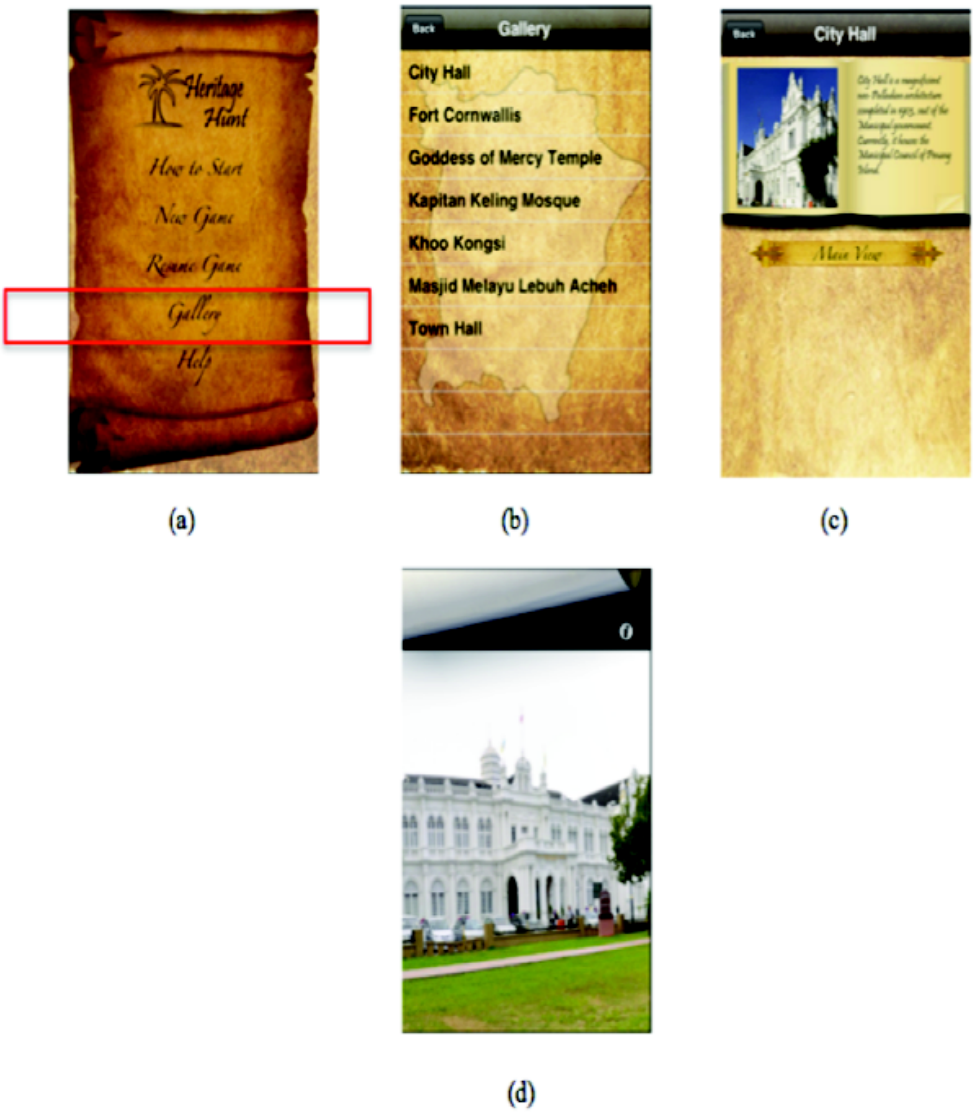


Figure E.6: A Sample Screenshot of the “Gallery”

Figure E.7: Screenshots of “Help”

(g)

(h)

(i)

(j)

(k)

Figure E.8: Continued from Figure E.7

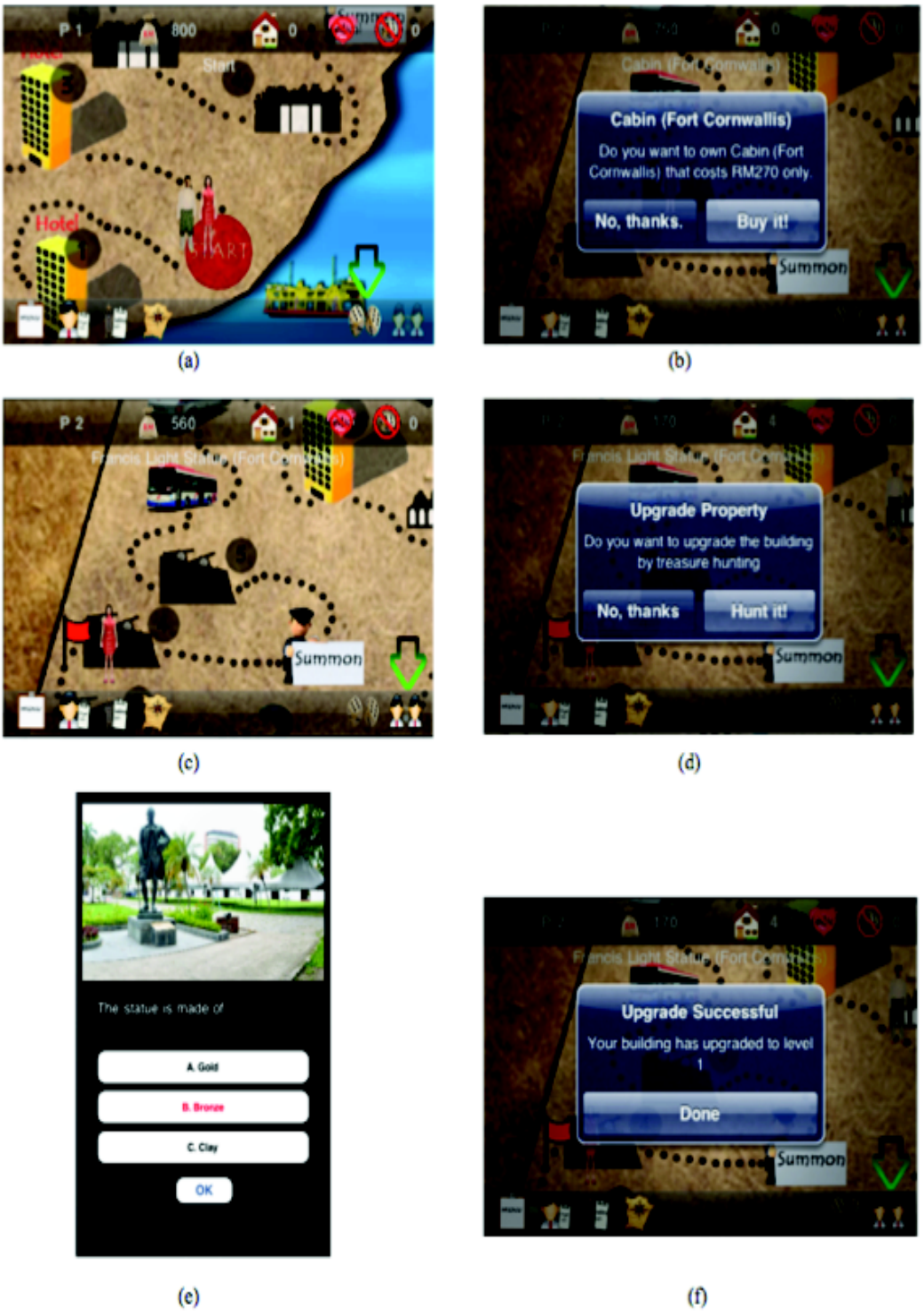


Figure E.9: Screenshots of “Upgrade Property”

Appendix F

Resume etendu en francais

Avec la croissance explosive de la numérisation du patrimoine culturel, de nombreuses institutions s'attaquent à la conversion des objets du patrimoine culturel en une représentation numérique ou une représentation descriptive, ou annotations. Toutefois, cette conversion présente des particularités telles que : 1) ces annotations sont de nature descriptive, 2) ces annotations sont ambigües ou très courtes, 3) un vocabulaire très spécifique est utilisé, et 4), il existe une forte variation dans les termes utilisés dans ces annotations. En outre, l'utilisation de mots-clés incorrects dans ces annotations entraîne une difficulté à résoudre les requêtes courtes. La plupart du temps, ces problèmes sont causés par des erreurs ou des imprécisions des annotations des documents. Le problème des requêtes courtes est causé par un utilisateur non spécialiste qui ne possède pas assez de connaissances et d'expériences dans le domaine du patrimoine culturel. Dans cette recherche, notre objectif principal est de modéliser un Système d'Accès à l'Information capable de surmonter partiellement les questions soulevées par le processus d'annotation des objets du patrimoine culturel et par le manque de connaissances expertes des utilisateurs du patrimoine culturel numérique. Par conséquent, trois types d'outils d'accès aux informations sont proposés dans ce travail à savoir : un Système de Recherche d'Information, une technique de recherche utilisant le contexte de l'utilisateur, ainsi qu'un jeu sur terminal mobile sur le thème du patrimoine culturel. Ces trois techniques permettent à l'utilisateur d'accéder, d'explorer les informations sur le patrimoine culturel et également de s'instruire. Dans ce travail, notre solution principale consiste à étendre le Systèmes de Recherche d'Informations en utilisant un contexte de recherche, et en intégrant des relations entre les termes dans la fonction de correspondance. Nous utilisons une fonction de correspondance basée sur le Modèle de Langue de type Dirichlet. Cette extension modifie le lissage de cette fonction pour résoudre les problèmes que posent à la fois le processus d'indexation et le contexte des connaissances incomplètes des utilisateurs non experts. En outre, nous présentons un modèle de préférence utilisateur basé sur la théorie physique de la charge d'un condensateur. Ce modèle physique permet de quantifier un contexte cognitif de l'utilisateur en utilisant le temps qu'il passe à certaines activités. Nous les intégrons ensuite dans la fonction de correspondance de type Dirichlet. Finalement, nous proposons un jeu sur mobile inspiré des jeux de Monopoly et de chasse au trésor. Ces jeux permettent de palier aux problèmes du faible niveau d'expertise des utilisateurs non spécialistes en leur permettant d'apprendre les termes du domaine au travers d'une activité ludique. Les deux premières approches ont été testées sur la collection de test CLEF (CHIC) dédiée au patrimoine culturel. Cette collection se compose de requêtes résolues et de document de type descriptifs sur des objets des collections des musées européens. Les résultats de notre solution sur cette collection, montrent que l'approche est efficace et donne une meilleure précision dans la résolution des requêtes. Enfin, nous avons mené une évaluation basée sur une enquête utilisateur, pour évaluer la troisième approche. Les résultats positifs nous laissent penser qu'une activité spécifiquement ludique est en mesure d'aider efficacement les participants à explorer et apprendre des informations pertinentes sur le patrimoine culturel. Les participants ont également estimé qu'un outil de type recherche d'information intégré au jeu, peut fournir plus d'informations à l'utilisateur, de manière plus pratique tout en jouant au jeu et en visitant virtuellement les sites du patrimoine culturel. En conclusion, nos résultats montrent que les solutions proposées dans ce travail de thèse, sont en mesure de résoudre les problèmes posés par les annotations numériques des objets du patrimoine culturel.

Appendix G

List of Publications

Lim, C. K., Tan, T. P., Tan, K. L. and Talib, A. Z. (2012). Randomized psychoacoustic model for mobile, panoramic, heritage-viewing applications, *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, VRCAI '12, ACM, New York, NY, USA, pp. 315–322.
URL: *http://doi.acm.org/10.1145/2407516.2407588*

Lim, C., Tan, K. and Talib, A. (2014). A hybrid of fixed-size and dynamic-size tile algorithm for panoramic view on mobile devices, *in* H. Y. Jeong, N. Y. Yen and J. J. J. H. Park (eds), *Advanced in Computer Science and its Applications*, Vol. 279 of *Lecture Notes in Electrical Engineering*, Springer Berlin Heidelberg, pp. 703–708.
URL: *http://dx.doi.org/10.1007/978-3-642-41674-3*$_1$00

Tan, K. L., Almasri, M., Chevallet, J.-P., Mulhem, P. and Berrut, C. (2013). Multimedia informaiton modeling and retrieval (mrim)/laboratoire d'informatique de grenoble (lig) at chic2013, *in* P. Forner, J. Karlgren and C. Womser-Hacker (eds), *CLEF (Online Working Notes/Labs/Workshop)*.
URL: *http://www.clef-initiative.eu/documents/71612/44d25853-ba7d-406c-ba5c-00bb64120f42*

Tan, K. L., Chevallet, J.-P. and Mulhem, P. (2013). Language model: Extension to the similarity of non-matching terms in retreival time, *Atelier Recherche d'Information SEmantique (RISE)*, pp. 4–17.
URL: *https://sites.google.com/site/frenchsemanticir/documents*

Tan, K. L., Lim, C. K. and Talib, A. Z. (2011a). A mobile gaming approach to virtual heritage exploration., *in* E. Ariwa and E. El-Qawasmeh (eds), *DEIS*, Vol. 194 of *Communications in Computer and Information Science*, Springer, pp. 493–500.
URL: *http://dblp.uni-trier.de/db/conf/deis/deis2011.htmlTanLT11*

Tan, K. L., Lim, C. K. and Talib, A. Z. (2011b). Mobile virtual heritage exploration with heritage hunt with a case study of george town, penang, malaysia, *IJEEI* **2**(4): 74–86.
URL: *http://dx.doi.org/10.4018/jeei.2011100106*

Bibliography

ABDULAHHAD, K., CHEVALLET, J.P. & BERRUT, C. (2011). Solving concept mismatch through bayesian framework by extending umls meta-thesaurus. in la huitime dition de la confrence en recherche d'information et applications. 38

ACCARY-BARBIER, T. & CALABRETTO, S. (2008). Building and using temporal knowledge in archaeological documentation. *J. Intell. Inf. Syst.*, **31**, 147–159. 7

ADRIAN, P. (2013). Cea list's participation at the clef chic 2013. In P. Forner, J. Karlgren & C. Womser-Hacker, eds., *CLEF (Online Working Notes/Labs/Workshop)*. 27, 34

AGGARWAL, N. & BUITELAAR, P. (2012). Query expansion using wikipedia and dbpedia. In P. Forner, J. Karlgren & C. Womser-Hacker, eds., *CLEF (Online Working Notes/Labs/Workshop)*. 24

AGICHTEIN, E., BRILL, E. & DUMAIS, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, 19–26, ACM, New York, NY, USA. 26, 43

AGIRRE, E., CLOUGH, P.D., FERNANDO, S., HALL, M., OTEGI, A. & STEVENSON, M. (2012). The sheffield and basque country universities entry to chic: Using random walks and similarity to access cultural heritage. In P. Forner, J. Karlgren & C. Womser-Hacker, eds., *CLEF (Online Working Notes/Labs/Workshop)*. 24, 38, 41, 114

AKASEREH, M., NAJI, N. & SAVOY, J. (2012). Unine at clef 2012. In P. Forner, J. Karlgren & C. Womser-Hacker, eds., *CLEF (Online Working Notes/Labs/Workshop)*. 24, 25, 28, 38, 39

AKTOLGA, E. & ALLAN, J. (2011). Reranking search results for sparse queries. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, 173–182, ACM, New York, NY, USA. 44

ALETRAS, N., STEVENSON, M. & CLOUGH, P. (2013). Computing similarity between items in a digital library of cultural heritage. *J. Comput. Cult. Herit.*, **5**, 16:1–16:19. 61

AMATI, G. & VAN RIJSBERGEN, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transaction on Information Systems*, **20**, 357–389. 25, 58

ANASTOPOULOU, S., SHARPLES, M., WRIGHT, M., MARTIN, H., AINSWORTH, S., BENFORD, S., CROOK, C., GREENHALGH, C. & O'MALLEY, C. (2008). Learning 21st century science in context with mobile technologies. In *Proceedings of the mLearn 2008 Conference*. 47, 53

ARDISSONO, L., KUFLIK, T. & PETRELLI, D. (2012). Personalization in cultural heritage: the road travelled and the one ahead. *User Modeling and User-Adapted Interaction*, **22**, 73–99. 28

BARAB, S., THOMAS, M., DODGE, T., CARTEAUX, R. & TUZUN, H. (2005). Making learning fun: Quest atlantis, a game without guns. *Educational Technology Research and Devel-*

opment, **53**, 86–107. viii, 49, 50, 53

BELKIN, N.J., COOL, C., KELLY, D., LIN, S.J., PARK, S.Y., PEREZ-CARBALLO, J. & SIKORA, C. (2001). Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Inf. Process. Manage.*, **37**, 403–434. 26, 43

BENDERSKY, M. & CROFT, W.B. (2008). Discovering key concepts in verbose queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, 491–498, ACM, New York, NY, USA. 37, 38

BENFORD, S., CRABTREE, A., FLINTHAM, M., DROZD, A., ANASTASI, R., PAXTON, M., TANDAVANITJ, N., ADAMS, M. & ROW-FARR, J. (2006). Can you see me now? *ACM Trans. Comput.-Hum. Interact.*, **13**, 100–133. vii, viii, 47, 48, 49, 50

BENTIVOGLI, L., FORNER, P., MAGNINI, B. & PIANTA, E. (2004). Revising the wordnet domains hierarchy: semantics, coverage and balancing. In G. Sérasset, ed., *COLING 2004 Multilingual Linguistic Resources*, 94–101, COLING, Geneva, Switzerland. 61

BERGER, A. & LAFFERTY, J. (1999). Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, 222–229, ACM, New York, NY, USA. 37, 39, 40, 41, 42

BERLAND, M., LEE, V. & DUMONT, M. (2010). Small groups, big mistakes: The emergence of faulty rules during a collaborative board game. In *Proceedings of the 9th International Conference of the Learning Sciences - Volume 2*, ICLS '10, 397–398, International Society of the Learning Sciences. 4, 49

BLANCO, A., TORRENTE, J., MARCHIORI, E.J., MARTÍNEZ-ORTIZ, I., MORENO-GER, P. & FERNÁNDEZ-MANJÓN, B. (2010). Easing assessment of game-based learning with <e-adventure> and lams. In *Proceedings of the Second ACM International Workshop on Multimedia Technologies for Distance Leaning*, MTDL '10, 25–30, ACM, New York, NY, USA. 4, 49

BULMER, M. (2004). *Questionnaires*. No. v. 1 in Sage Benchmarks in Social Research Methods, Sage Publications. 33

CAMPBELL, I. (2000). The ostensive model of developing information-needs. 20

CAO, G., ROBERTSON, S. & YUN NIE, J. (2008). Selecting query term alterations for web search by exploiting query contexts. In *Proceedings of Association for Computational Linguistics*, 148–155. 38

CARMAGNOLA, F., CENA, F., CONSOLE, L., CORTASSA, O., GENA, C., GOY, A., TORRE, I., TOSO, A. & VERNERO, F. (2008). Tag-based user modeling for social multi-device adaptive guides. *User Modeling and User-Adapted Interaction*, **18**, 497–538. vii, 28, 29

CARMEL, D., ZWERDLING, N. & YOGEV, S. (2012). Entity oriented search and exploration for cultural heritage collections: the eu cultura project. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, 227–230, ACM, New York, NY, USA. vii, 25, 26, 30

CARPINETO, C., DE MORI, R., ROMANO, G. & BIGI, B. (2001). An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, **19**, 1–27. 38

CARR, D. (2010). Visual computer game features for teaching relativity. *2012 Ninth International Conference on Computer Graphics, Imaging and Visualization*, **0**, 35–40. 4, 48

CHEVALLET, J.P. (2005). X-iota: An open xml framework for ir experimentation. In S. Myaeng, M. Zhou, K.F. Wong & H.J. Zhang, eds., *Information Retrieval Technology*, vol. 3411 of *Lecture Notes in Computer Science*, 263–280, Springer Berlin Heidelberg. 100

CHEVALLET, J.P., LIM, J.H. & LE, D.T.H. (2007). Domain knowledge conceptual intermedia indexing: Application to multilingual multimedia medical reports. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, 495–504, ACM, New York, NY, USA. 34

CHEVERST, K., DAVIES, N., MITCHELL, K. & SMITH, P. (2000). Providing tailored (context-aware) information to city visitors. In *Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, AH '00, 73–85, Springer-Verlag, London, UK, UK. 28

CHIRITA, P.A., NEJDL, W., PAIU, R. & KOHLSCHÜTTER, C. (2005). Using odp metadata to personalize search. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, 178–185, ACM, New York, NY, USA. 46

CHURCH, K. & SMYTH, B. (2009). Understanding the intent behind mobile information needs. In *Proceedings of the 14th international conference on Intelligent user interfaces*, IUI '09, 247–256, ACM, New York, NY, USA. 8, 42

CHURCH, K., SMYTH, B., BRADLEY, K. & COTTER, P. (2008). A large scale study of european mobile search behaviour. In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, MobileHCI '08, 13–22, ACM, New York, NY, USA. 8

CLAYPOOL, M., LE, P., WASED, M. & BROWN, D. (2001). Implicit interest indicators. In *Proceedings of the 6th International Conference on Intelligent User Interfaces*, IUI '01, 33–40, ACM, New York, NY, USA. 43, 81

CLEVERDON, C. (1997). Readings in information retrieval. chap. The Cranfield tests on index language devices, 47–59, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 31, 32

CLOUGH, P., FORD, N. & STEVENSON, M. (2011). Personalizing access to cultural heritage collections using pathways. In *PATCH 2011: 3rd International Workshop on Personalized Access to Cultural Heritage*. vii, 25, 27, 30

COLE, H. & STANTON, D. (2003). Designing mobile technologies to support co-present collaboration. *Personal Ubiquitous Comput.*, **7**, 365–371. viii, 48, 51, 52

CRASWELL, N. & SZUMMER, M. (2007). Random walks on the click graph. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, 239–246, ACM, New York, NY, USA. 44

CRESTANI, F. (2000). Exploiting the similarity of non-matching terms at retrieval time. *Journal of Information Retrieval*, **2**, 25–45. 40, 41, 42, 59, 62

CROFT, W.B. & HARPER, D.J. (1988). Document retrieval systems. chap. Using Probabilistic Models of Document Retrieval Without Relevance Information, 161–171, Taylor Graham Publishing, London, UK, UK. 23

CUI, Y. & ROTO, V. (2008). How people use the web on mobile devices. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, 905–914, ACM, New York, NY, USA. 4

DANG, V. & CROFT, B.W. (2010). Query reformulation using anchor text. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, 41–50, ACM, New York, NY, USA. 38

DAOUD, M., TAMINE-LECHANI, L. & BOUGHANEM, M. (2008). Using a graph-based ontological user profile for personalizing search. In *Proceedings of the 17th ACM Conference*

on Information and Knowledge Management, CIKM '08, 1495–1496, ACM, New York, NY, USA. 46

DEERWESTER, S., DUMAIS, S.T., FURNAS, G.W., LANDAUER, T.K. & HARSHMAN, R. (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, **41**, 391–407. 37, 39

DOOLITTLE, J.H. (1995). Using riddles and interactive computer games to teach problem-solving skills. *Teaching of Psychology*, **22**, 33–36. 49

DOU, Z., SONG, R. & WEN, J.R. (2007). A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, 581–590, ACM, New York, NY, USA. 44

ELSWEILER, D., MANDL, S. & KIRKEGAARD LUNN, B. (2010). Understanding casual-leisure information needs: A diary study in the context of television viewing. In *Proceedings of the Third Symposium on Information Interaction in Context*, IIiX '10, 25–34, ACM, New York, NY, USA. 20

FACER, K., JOINER, R., STANTON, D., REID, J., HULL, R. & KIRK, D. (2004). Savannah: mobile gaming and learning? *Journal of Computer Assisted Learning*, **20**, 399–409, iD number: ISI:000225405800001. 49

FALK, J. & DIERKING, L. (2000). *Learning from Museums: Visitor Experiences and the Making of Meaning*. American Association for State and Local History book series, AltaMira Press. 55

FAW, T. & NUNNALLY, J. (1967). The effects on eye movements of complexity, novelty, and affective tone. *Perception Psychophysics*, **2**, 263–267. 43

FERNANDO, S., HALL, M.M., AGIRRE, E., SOROA, A., CLOUGH, P. & STEVENSON, M. (2012). Comparing taxonomies for organising collections of documents. In M. Kay & C. Boitet, eds., *COLING*, 879–894, Indian Institute of Technology Bombay. 30, 61

FINK, J., KOBSA, A. & NILL, A. (1999). Adaptable and adaptive information provision for all users, including disabled and elderly people. 30

FRIAS-MARTINEZ, E., CHEN, S.Y., MACREDIE, R.D. & LIU, X. (2007). The role of human factors in stereotyping behavior and perception of digital library users: A robust clustering approach. *User Modeling and User-Adapted Interaction*, **17**, 305–337. 27

FROHBERG, D., GÖTH, C. & SCHWABE, G. (2009). Mobile learning projects – a critical analysis of the state of the art. *Journal of Computer Assisted Learning*, **25**, 307–331. 48

FURNAS, G.W., LANDAUER, T.K., GOMEZ, L.M. & DUMAIS, S.T. (1987). The vocabulary problem in human-system communication. *Commun. ACM*, **30**, 964–971. 39

GAO, J., HE, X. & NIE, J.Y. (2010). Clickthrough-based translation models for web search: From word models to phrase models. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, 1139–1148, ACM, New York, NY, USA. 39

GAO, J., TOUTANOVA, K. & YIH, W.T. (2011). Clickthrough-based latent semantic models for web search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, 675–684, ACM, New York, NY, USA. 44

GAUCH, S., SPERETTA, M., CHANDRAMOULI, A. & MICARELLI, A. (2007). The adaptive web. chap. User Profiles for Personalized Information Access, 54–89, Springer-Verlag, Berlin, Heidelberg. 44, 45

GENTILI, G., MICARELLI, A. & SCIARRONE, F. (2003). Infoweb: An adaptive information filtering system for the cultural heritage domain. *Applied Artificial Intelligence*, **17**, 715–744.

45
GOKER, A. & MYRHAUG, H. (2008). Evaluation of a mobile information system in context. *Inf. Process. Manage.*, **44**, 39–65. 25, 26
GOKER, A. & MYRHAUG, H.I. (2002). User context and personalisation. In *ECCBR Workshop on Case Based Reasoning and Personalisation, Aberdeen*. 26
GÖKER, A. & MYRHAUG, H.I. (2008). Evaluation of a mobile information system in context. *Inf. Process. Manage.*, **44**, 39–65. vii, 28
GWIZDKA, J. & CHIGNELL, M. (1999). Towards information retrieval measures for evaluation of web search engines. 32
HALL, M.M., CLOUGH, P.D., FERNANDO, S., GOODALE, P., STEVENSON, M., AGIRRE, E., OTEGI, A., SOROA, A., FERNIE, K., GRIFFITHS, J. & BERGHEIM, R. (2013). Information seeking in digital cultural heritage with paths. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, 1105–1106, ACM, New York, NY, USA. 30
HAMPSON, C., AGOSTI, M., ORIO, N., BAILEY, E., LAWLESS, S., CONLAN, O. & WADE, V. (2012). The cultura project: Supporting next generation interaction with digital cultural heritage collections. In M. Ioannides, D. Fritsch, J. Leissner, R. Davies, F. Remondino & R. Caffo, eds., *EuroMed*, vol. 7616 of *Lecture Notes in Computer Science*, 668–675, Springer. 7
HANCOCK-BEAULIEU, M. & WALKER, S. (1992). An evaluation of automatic query expansion in an online library catalogue. *J. Doc.*, **48**, 406–421. 26
HARMAN, D. (1995). Overview of the fourth text retrieval conference (trec-4). In *TREC*. 32
HARPSTEAD, E., MYERS, B.A. & ALEVEN, V. (2013). In search of learning: Facilitating data analysis in educational games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 79–88, ACM, New York, NY, USA. 4, 49
HENNESSY, S., WISHART, J., WHITELOCK, D., DEANEY, R., BRAWN, R., LA VELLE, L., MCFARLANE, A., RUTHVEN, K. & WINTERBOTTOM, M. (????). *Computers Education*, 137–152. 47
HUANG, P.S., HE, X., GAO, J., DENG, L., ACERO, A. & HECK, L. (2013). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, 2333–2338, ACM, New York, NY, USA. 44
HUIZENGA, J., ADMIRAAL, W., AKKERMAN, S. & DAM, G.T. (2009). Mobile game-based learning in secondary education: engagement, motivation and learning in a mobile city game. *J. Comp. Assisted Learning*, **25**, 332–344. 47, 52, 54
INGWERSEN, P. (1994). Polyrepresentation of information needs and semantic entities: Elements of a cognitive theory for information retrieval interaction. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, 101–110, Springer-Verlag New York, Inc., New York, NY, USA. 27
JANSEN, B.J., BOOTH, D.L. & SPINK, A. (2007). Determining the user intent of web search engine queries. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, 1149–1150, ACM, New York, NY, USA. 27
JIANG, D., PEI, J. & LI, H. (2013). Mining search and browse logs for web search: A survey. *ACM Trans. Intell. Syst. Technol.*, **4**, 57:1–57:37. 44
JING, Y. & CROFT, W.B. (1994). An association thesaurus for information retrieval. In *In RIAO 94 Conference Proceedings*, 146–160. 23, 37
JOACHIMS, T., GRANKA, L., PAN, B., HEMBROOKE, H. & GAY, G. (2005). Accurately

interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, 154–161, ACM, New York, NY, USA. 46

JONES, R., REY, B., MADANI, O. & GREINER, W. (2006). Generating query substitutions. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, 387–396, ACM, New York, NY, USA. 38

KAMVAR, M., KELLAR, M., PATEL, R. & XU, Y. (2009). Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices. In *Proceedings of the 18th international conference on World wide web*, WWW '09, 801–810, ACM, New York, NY, USA. 8

KARIMZADEHGAN, M. & ZHAI, C. (2010). Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, 323–330, ACM, New York, NY, USA. 37, 40, 41, 42, 59

KARIMZADEHGAN, M. & ZHAI, C. (2012). Axiomatic analysis of translation language model for information retrieval. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, 268–280, Springer-Verlag, Berlin, Heidelberg. 42

KARIMZADEHGAN, M., ZHAI, C. & EFRON, M. (2013). Statistical translation language model for twitter search. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ICTIR '13, 26:121–26:124, ACM, New York, NY, USA. 37, 39, 42

KEKÄLÄINEN, J. & JÄRVELIN, K. (2002). Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In *in Proceedings of the CoLIS 4 Conference*, 253–270. 31

KELLY, D. & TEEVAN, J. (2003). Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, **37**, 18–28. 44, 47

KELLY, D., HARPER, D.J. & LANDAU, B. (2008). Questionnaire mode effects in interactive information retrieval experiments. *Inf. Process. Manage.*, **44**, 122–141. 34

KESELMAN, A. (2003). Supporting inquiry learning by promoting normative understanding of multivariable causality. *Journal of Research in Science Teaching*, **40**, 898–921. 54

KIM, H.R. & CHAN, P.K. (2003). Learning implicit user interest hierarchy for context in personalization. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, 101–108, ACM, New York, NY, USA. 25, 27

KIM, K.S. (2008). Effects of emotion control and task on web searching behavior. *Inf. Process. Manage.*, **44**, 373–385. 27

KOOLEN, M. & KAMPS, J. (2010). Searching cultural heritage data: does structure help expert searchers? In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, 152–155, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, France, France. 8, 20

KOOLEN, M., KAMPS, J. & DE KEIJZER, V. (2009). Information retrieval in cultural heritage. *Interdisciplinary Science Reviews*, **34**. 20

KROVETZ, R. (1993). Viewing morphology as an inference process. 191–202, ACM Press. 37, 38, 39

KUMARAN, G. & CARVALHO, V.R. (2009). Reducing long queries using query quality predictors. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, 564–571, ACM, New York, NY, USA. 38

KÜRSTEN, J., WILHELM, T., RICHTER, D. & EIBL, M. (2012). Chemnitz at the chic evaluation lab 2012: Creating an xtrieval module for semantic enrichment. In P. Forner, J. Karlgren

& C. Womser-Hacker, eds., *CLEF (Online Working Notes/Labs/Workshop)*. 24, 25, 38

LACOSTE, C., LIM, J.H., CHEVALLET, J.P. & LE, D.T. (2007). Medical-image retrieval based on knowledge-assisted text and image indexing. *IEEE Trans. Cir. and Sys. for Video Technol.*, **17**, 889–900. 34

LAVRENKO, V. & CROFT, W.B. (2001). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, 120–127, ACM, New York, NY, USA. 23, 37, 113

LEASE, M. (2008). Incorporating relevance and psuedo-relevance feedback in the markov random field model: Brown at the trec'08 relevance feedback track. In *In TREC '08*. 38

LEUTNER, D. (1993). Guided discovery learning with computer-based simulation games: Effects of adaptive and non-adaptive instructional support. *Learning and Instruction*, **3**, 113 – 132. 49

LIEBERMAN, H. (1995). Letizia: An agent that assists web browsing. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, 924–929, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 45

LINN, M. (2003). Technology and science education: Starting points, research programs, and trends. *International Journal of Science Education*, **25**, 727–758. 47

LOVELL, K. (1959). A follow-up study of some aspects of the work of piaget and inhelder on the child's conception of space. *British Journal of Educational Psychology*, **29**, 104–117. 53, 54, 85, 93

LOVINS, J.B. (1968). Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, **11**. 23, 39

MAES, P. & KOZIEROK, R. (1993). Learning interface agents. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, AAAI'93, 459–464, AAAI Press. 126

MARKINES, B., CATTUTO, C., MENCZER, F., BENZ, D., HOTHO, A. & STUMME, G. (2009). Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World wide web*, WWW '09, 641–650, ACM, New York, NY, USA. 62

MCFARLANE, A. & SAKELLARIOU, S. (2002). The role of ict in science education. *Cambridge Journal of Education*, **32**, 219–232. 47

MICARELLI, A., GASPARETTI, F., SCIARRONE, F. & GAUCH, S. (2007). The adaptive web. chap. Personalized Search on the World Wide Web, 195–230, Springer-Verlag, Berlin, Heidelberg. 45

MILLER, G.A., BECKWITH, R., FELLBAUM, C., GROSS, D. & MILLER, K. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, **3**, 235–244. 23

MITRA, M., SINGHAL, A. & BUCKLEY, C. (1998). Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, 206–214, ACM, New York, NY, USA. 23, 38

MORITA, M. & SHINODA, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, 272–281, Springer-Verlag New York, Inc., New York, NY, USA. 43, 81

MOUKAS, A. & MAES, P. (1998). Amalthaea: An evolving multi-agent information filtering and discovery system for the www. *Autonomous Agents and Multi-Agent Systems*, **1**, 59–88. 45

NAJI, N. & SAVOY, J. (2011). Information retrieval strategies for digitized handwritten me-

dieval documents. In *Proceedings of the 7th Asia Conference on Information Retrieval Technology*, AIRS'11, 103–114, Springer-Verlag, Berlin, Heidelberg. 7

NANAS, N., VAVALIS, M. & DE ROECK, A. (2010). A network-based model for high-dimensional information filtering. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, 202–209, ACM, New York, NY, USA. 45

NICHOLAS, D., CLARK, D., ROWLANDS, I. & JAMALI, H.R. (2013). Information on the go: A case study of europeana mobile users. *Journal of the American Society for Information Science and Technology*, **64**, 1311–1322. 8, 42

NICHOLS, D.M. (1997). Implicit rating and filtering. In *IN PROCEEDINGS OF THE FIFTH DELOS WORKSHOP ON FILTERING AND COLLABORATIVE FILTERING*, 31–36. 26, 43, 47

NWANA, H.S. (1996). Software agents: An overview. *Knowledge Engineering Review*, **11**, 205–244. 126

OBERLANDER, J., O'DONNELL, M., KNOTT, A. & MELLISH, C. (1998). Conversation in the museum: Experiments in dynamic hypermedia with the intelligent labelling explorer. 30

OGIER, J.M. (2006). Madonne: Document Image Analysis Techniques for Cultural Heritage Documents. *International Conference on Digital Cultural Heritage*. 7

OOSTENDORP, A. & BERLYNE, D.E. (1978). Dimensions in the perception of architecture:ii. measures of exploratory behavior. *Scandinavian Journal of Psychology*, **19**, 83–89. 43, 81

OTEGI, A., ARREGI, X. & AGIRRE, E. (2011). Query expansion for ir using knowledge-based relatedness. In *IJCNLP*, 1467–1471. 41

PARIS, C., WAN, S., WILKINSON, R. & WU, M. (2001). Generating personal travel guides - and who wants them? In *Proceedings of the 8th International Conference on User Modeling 2001*, UM '01, 251–253, Springer-Verlag, London, UK, UK. 30

PARKHOMENKO, O., LEE, Y. & PARK, E.K. (2003). Ontology-driven peer profiling in peer-to-peer enabled semantic web. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, 564–567, ACM, New York, NY, USA. 45

PAZZANI, M., MURAMATSU, J. & BILLSUS, D. (1996). Syskill & webert: Identifying interesting web sites. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 1*, AAAI'96, 54–61, AAAI Press. 45

PÉDAUQUE, R.T. (2003). Document: Form, sign and medium, as reformulated for electronic documents. 5

PENG, F., AHMED, N., LI, X. & LU, Y. (2007). Context sensitive stemming for web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, 639–646, ACM, New York, NY, USA. 39

PETRAS, V., FERRO, N., GÄDE, M., ISAAC, A., KLEINEBERG, M., MASIERO, I., NICCHIO, M. & STILLER, J. (2012). Cultural heritage in clef (chic) overview 2012. In P. Forner, J. Karlgren & C. Womser-Hacker, eds., *CLEF (Online Working Notes/Labs/Workshop)*. 23, 24, 59

PETRAS, V., BOGERS, T., TOMS, E., HALL, M., SAVOY, J., MALAK, P., PAWŁOWSKI, A., FERRO, N. & MASIERO, I. (2013). Cultural heritage in clef (chic) 2013. In P. Forner, H. Müller, R. Paredes, P. Rosso & B. Stein, eds., *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, vol. 8138 of *Lecture Notes in Computer Science*, 192–211, Springer Berlin Heidelberg. 34, 39

PONTE, J.M. & CROFT, W.B. (1998). A language modeling approach to information retrieval.

In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, 275–281, ACM, New York, NY, USA. 22, 58

PORTER, M.F. (1997). Readings in information retrieval. chap. An algorithm for suffix stripping, 313–316, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 23, 37, 38, 39

PORTIER, P.I., CHATTI, N., CALABRETTO, S., EGYED-ZSIGMOND, E. & PINON, J.M. (2012). Modeling, encoding and querying multi-structured documents. *Inf. Process. Manage.*, **48**, 931–955. 6

PRAYAGA, L. (2005). Game technology as a tool to actively engage k-12 students in the act of learning. In *Proceedings of the 6th Conference on Information Technology Education*, SIGITE '05, 307–310, ACM, New York, NY, USA. 4, 49

RADHOUANI, S. & FALQUET, G. (2006). Using external knowledge to solve multi-dimensional queries. In P. Ghodous, R. Dieng-Kuntz & G. Loureiro, eds., *ISPE CE*, vol. 143 of *Frontiers in Artificial Intelligence and Applications*, 426–437, IOS Press. 34

RADHOUANI, S., FALQUET, G. & CHEVALLETINST, J.P. (2008). Description logic to model a domain specific information retrieval system. In *Proceedings of the 19th international conference on Database and Expert Systems Applications*, DEXA '08, 142–149, Springer-Verlag, Berlin, Heidelberg. 34

RIJSBERGEN, C.J.V. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edn. 23, 62

ROBERTSON, S.E. (1997). Overview of the okapi projects. *Journal of Documentation*, **53**, 3–7. 21, 22, 25, 58

ROCCHIO, J.J. (1971). Relevance feedback in information retrieval. In G. Salton, ed., *The Smart retrieval system - experiments in automatic document processing*, 313–323, Englewood Cliffs, NJ: Prentice-Hall. 25, 42

ROGERS, Y., PRICE, S., HARRIS, E., PHELPS, T., UNDERWOOD, M., WILDE, D., SMITH, H., WEAL, M.T.M.J. & MICHAELIDES, D.T. (2002). Learning through digitally-augmented physical experiences: reflections on the ambient wood project. Tech. rep. 54

ROGERS YVONNE, G.S.S.H.H.E., SCAIFE MIKE (2002). A conceptual framework for mixed reality environments: Designing novel learning activities for young children. *Presence: Teleoper. Virtual Environ.*, **11**, 677–686. 48

ROSCHELLE, J. & PEA, R. (2002). A walk on the wild side: How wireless handhelds may change cscl. In *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community*, CSCL '02, 51–60, International Society of the Learning Sciences. 48

RUTHVEN, I. & LALMAS, M. (2003). A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, **18**, 95–145. 42

SAKAGAMI, H. & KAMBA, T. (1997). Learning personal preferences on online newspaper articles from user behaviors. *Comput. Netw. ISDN Syst.*, **29**, 1447–1455. 45

SALTON, G., ed. (1971). *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, Englewood, Cliffs, New Jersey. 21, 37

SALTON, G. (1991). The smart project in automatic document retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, 356 – 358, Chicago, Illinois, United States. 21, 25, 58

SALTON, G. & BUCKLEY, C. (1988). Term-weighting approaches in automatic text retrieval. In *INFORMATION PROCESSING AND MANAGEMENT*, 513–523. 58

SALTON, G. & BUCKLEY, C. (1997). Readings in information retrieval. chap. Improving Retrieval Performance by Relevance Feedback, 355–364, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 42

SAVILL-SMITH, C. (2005). The use of palmtop computers for learning: a review of the literature. *British Journal of Educational Technology*, **36**, 567–568. 4, 48

SCHAER, P., HIENERT, D., SAWITZKI, F., WIRA-ALAM, A. & LÜKE, T. (2012). Dealing with sparse document and topic representations: Lab report for chic 2012. In *CLEF 2012 Labs and Workshop, Notebook Papers: CLEF/CHiC Workshop-Notes*. 24, 25, 38

SCHÜTZE, H., HULL, D.A. & PEDERSEN, J.O. (1995). A comparison of classifiers and document representations for the routing problem. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, 229–237, ACM, New York, NY, USA. 39

SCHWABE, G. (2005). Mobile learning with a mobile game: design and motivational effects. *JCAL - Journal of Computer Assisted Learning*, **21**, 204 – 216. 4, 49

SEO, Y.W. & ZHANG, B.T. (2000). Learning user's preferences by analyzing web-browsing behaviors. In *Proceedings of the Fourth International Conference on Autonomous Agents*, AGENTS '00, 381–387, ACM, New York, NY, USA. 44

SETTEN, M.V., POKRAEV, S., KOOLWAAIJ, J. & INSTITUUT, T. (2004). Context-aware recommendations in the mobile tourist application compass. In *In Nejdl, W. De Bra, P. (Eds.). AH 2004, LNCS 3137*, 235–244, Springer-Verlag. 28

SHAFFER, D.W., SQUIRE, K.R., HALVERSON, R., GEE, J.P., DAVID, C., SHAFFER, W., SQUIRE, K.R., HALVERSON, R., GEE, J.P., SHAFFER, D.W., SQUIRE, K.R., HALVERSON, R. & GEE, J.P. (2005). Video games and the future of learning. *Phi Delta Kappan*, **87**, 104–111. 48, 49

SHARPLES, M. (2006). *Big Issues in Mobile Learning: Report of a workshop by the Kaleidoscope Network of Excellence Mobile Learning Initiative*. University of Nottingham, Nottingham, sharples (Hg.) 2006 - Big Issues in Mobile Learning file:///Big%20issues%20in%20mobile%20learning.pdf js, 13.05.2007. 55

SHARPLES, M., CORLETT, D. & WESTMANCOTT, O. (2002). The design and implementation of a mobile learning resource. *Personal Ubiquitous Comput.*, **6**, 220–234. 48

SHIRATUDDIN, N. & ZAIBON, S.B. (2010). Mobile game-based learning with local content and appealing characters. *Int. J. Mob. Learn. Organ.*, **4**, 55–82. 4, 48

SIEG, A., MOBASHER, B. & BURKE, R. (2007). Web search personalization with ontological user profiles. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, 525–534, ACM, New York, NY, USA. 46

SONG, M., ELIAS, T., MARTINOVIC, I., MULLER-WITTIG, W. & CHAN, T.K.Y. (2004). Digital heritage application as an edutainment tool. In J.R. Brown & Y. Cai, eds., *VRCAI*, 163–167, ACM. 4

SPINK, A., GREISDORF, H. & BATEMAN, J. (1998). From highly relevant to not relevant: examining different regions of relevance. *Information Processing Management*, **34**, 599 – 621. 42

SQUIRE, K. & JAN, M. (2007). Mad city mystery: Developing scientific argumentation skills with a place-based augmented reality game on handheld computers. *Journal of Science Education and Technology*, **16**, 5–29. viii, 47, 50, 52, 54, 55

SRINIVAS, G., TANDON, N. & VARMA, V. (2010). A weighted tag similarity measure based on a collaborative weight model. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, 79–86, ACM, New York, NY, USA. 62

STAMATINA ANASTOPOULOU, M.W.H.M.S.B.C.C.C.G., MIKE SHARPLES (2008). Learning 21st century science in context with mobile technologies. In *University of Wolverhampton*, 12–19. 54

SUGIYAMA, K., HATANO, K. & YOSHIKAWA, M. (2004). Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, 675–684, ACM, New York, NY, USA. 25, 27

SUMI, Y. & MASE, K. (2001). Agentsalon: facilitating face-to-face knowledge exchange through conversations among personal agents. In *Proceedings of the fifth international conference on Autonomous agents*, AGENTS '01, 393–400, ACM, New York, NY, USA. 28

SUN, J.T., ZENG, H.J., LIU, H., LU, Y. & CHEN, Z. (2005). Cubesvd: A novel approach to personalized web search. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, 382–390, ACM, New York, NY, USA. 44

TAKANO, K., HONDA, H. & LI, K.F. (2012). User preference profiling based on speech recognition for personalized recommendation. In *BWCCA*, 308–315, IEEE. 27

TAMINE-LECHANI, L., BOUGHANEM, M. & DAOUD, M. (2010). Evaluation of contextual information retrieval effectiveness: Overview of issues and research. *Knowl. Inf. Syst.*, **24**, 1–34. 25, 72

TAYLOR, J., SHARPLES, M., O'MALLEY, C., VAVOULA, G. & WAYCOTT, J. (2006). Towards a task model for mobile learning: a dialectical approach. *Int. J. Learn. Technol.*, **2**, 138–158. 48, 53

ULABY, F.T. (1997). *Fundamentals of applied electromagnetics*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 79

VALLET, D., CANTADOR, I. & JOSE, J.M. (2010). Exploiting external knowledge to improve video retrieval. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, 101–110, ACM, New York, NY, USA. 34

VAVOULA, G., SHARPLES, M., RUDMAN, P., MEEK, J. & LONSDALE, P. (2009). Myartspace: Design and evaluation of support for learning with multimedia phones between classrooms and museums. *Comput. Educ.*, **53**, 286–299. viii, 47, 50, 51, 55

VIEIRA, V., TEDESCO, P., SALGADO, A.C. & BRÉZILLON, P. (2007). Investigating the specifics of contextual elements management: The cemantika approach. In *Proceedings of the 6th International and Interdisciplinary Conference on Modeling and Using Context*, CONTEXT'07, 493–506, Springer-Verlag, Berlin, Heidelberg. 25

VILLA, R., CLOUGH, P., HALL, M. & RUTTER, S. (2013). Search or browse? casual information access to a cultural heritage collection. In *Proceedings of the 3rd European Workshop on Human-Computer Interaction and Information Retrieval*, 19–22. 20

VOORHEES, E.M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, 315–323, ACM, New York, NY, USA. 32

VYGOTSKY, L.S. (1978). *Mind and society: The Development of Higher Mental Processes*. Harvard University Press, Cambridge, MA. 52

WANG, S., ISAAC, A., CHARLES, V., KOOPMAN, R., AGOROPOULOU, A. & VAN DER WERF, T. (2013). Hierarchical structuring of cultural heritage objects within large aggregations. *CoRR*, **abs/1306.2866**. 61

WANG, X. & ZHAI, C. (2008). Mining term association patterns from search logs for effective query reformulation. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, 479–488, ACM, New York, NY, USA. 38

WHITE, R.W., RUTHVEN, I. & JOSE, J.M. (2002). Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, 57–64, ACM, New York, NY, USA. 44, 46

WILENSKY, U. & STROUP, W.M. (2000). Networked gridlock: Students enacting complex dynamic phenomena with the hubnet architecture. In *In B. Fishman S. O'Connor-Divelbiss (Eds.), Fourth International Conference of the Learning Sciences*, 282–289. 47

WILSON, K.A., BEDWELL, W.L., LAZZARA, E.H., SALAS, E., BURKE, C.S., ESTOCK, J.L., ORVIS, K.L. & CONKEY, C. (2009). Relationships between game attributes and learning outcomes. *Simul. Gaming*, **40**, 217–266. 49

XU, J. & CROFT, W.B. (1996). Query expansion using local and global document analysis. In *In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 4–11. 23, 37

YANG, Y. & CHUTE, C.G. (1993). An application of least squares fit mapping to text information retrieval. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, 281–290, ACM, New York, NY, USA. 39

YI, J., MAGHOUL, F. & PEDERSEN, J. (2008). Deciphering mobile search patterns: a study of yahoo! mobile search queries. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, 257–266, ACM, New York, NY, USA. 8

YIN, Z., SHOKOUHI, M. & CRASWELL, N. (2009). Query expansion using external evidence. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, 362–374, Springer-Verlag, Berlin, Heidelberg. 38, 60

YOGEV, S., ROITMAN, H., CARMEL, D. & ZWERDLING, N. (2012). Towards expressive exploratory search over entity-relationship data. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, 83–92, ACM, New York, NY, USA. 25, 30

ZHAI, C. & LAFFERTY, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, **22**, 179–214. 22, 58, 59

ZHAO, L. & CALLAN, J. (2012). Automatic term mismatch diagnosis for selective query expansion. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, 515–524, ACM, New York, NY, USA. 37

ZNAIDIA, A., SHABOU, A., POPESCU, A., LE BORGNE, H. & HUDELOT, C. (2012). Multimodal feature generation framework for semantic image classification. In *Proceedings of the 2Nd ACM International Conference on Multimedia Retrieval*, ICMR '12, 38:1–38:8, ACM, New York, NY, USA. 27