# MINING SOCIAL NETWORK OF IT-RELATED RESEARCHES AND THAI RESEARCHERS

**USANEE SRIPIROM**

**A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE) FACULTY OF GRADUATE STUDIES MAHIDOL UNIVERSITY 2010**

Research Project
entitled

# MINING SOCIAL NETWORK OF IT-RELATED RESEARCHES AND THAI RESEARCHERS

…….
Miss Usanee Sripirom
Candidate

…….
Lect. Songsri Tangsripairoj, Ph.D.
Major-advisor

…….
Lect. Sudsanguan Ngamsuriyaroj, Ph.D.
Co-advisor

…….
Asst. Prof. Charnyote Pluempitiwiriyawej, Ph.D.
Co-advisor

…….
Prof. Banchong Mahaisavariya, M.D., Dip Thai Board of Orthopedics
Dean
Faculty of Graduate Studies
Mahidol University

…….
Assoc. Prof. Supachai Tangwongsan, Ph.D.
Program Director
Master Program in Computer Science
Faculty of Information and Communication Technology
Mahidol University

Research Project
entitled

# MINING SOCIAL NETWORK OF IT-RELATED RESEARCHES AND THAI RESEARCHERS

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science (Computer Science)
on
March 5, 2010

……
Miss Usanee Sripirom
Candidate

……
Lect. Benjarath Pupacdi, Ph.D.
Chair

……
Lect. Songsri Tangsripairoj, Ph.D.
Member

……
Lect. Sudsanguan Ngamsuriyaroj, Ph.D.
Member

……
Asst. Prof. Charnyote Pluempitiwiriyawej, Ph.D.
Member

……
Prof. Banchong Mahaisavariya, M.D., Dip Thai Board of Orthopedics
Dean
Faculty of Graduate Studies
Mahidol University

……
Assoc. Prof. Jarernsri L. Mitrpanont, Ph.D.
Acting Dean
Faculty of Information and Communication Technology
Mahidol University

# ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and deep appreciations to my advisor, Dr. Songsri Tangsripairoj, for her deep interest, inspiring guidance and continued encouragement which provided the impetus to complete this research project successfully.

I am also grateful to Dr. Sudsanguan Ngamsuriyaroj, Dr. Charnyote Pluempitiwiriyawej and Dr. Benjarath Pupacdi who are my research project committee and provide me their insightful comments, technique review and suggestions on this research project.

My appreciation goes to Faculty of Information and Communication Technology for giving me valuable knowledge and experience throughout my study.

Finally, I wish to specifically thank my father, my mother and my brother, for their understanding and inspiration throughout my life. Their continued love and encouragement throughout this research project is the reason behind its success.

Usanee Sripirom

MINING SOCIAL NETWORK OF IT-RELATED RESEARCHES AND THAI RESEARCHERS

USANEE SRIPIROM 4737988 SCCS/M

M.Sc.(COMPUTER SCIENCE)

RESEARCH PROJECT ADVISORY COMMITTEE: SONGSRI TANGSRIPAIROJ, Ph.D, SUDSANGUAN NGAMSURIYAROJ, Ph.D, CHARNYOTE PLUEMPITIWIRIYAWEJ, Ph.D

ABSTRACT

This research focuses on the study of a social network of Thai researchers that conducted research in only the computer science and information technology fields and submitted their works to 2 conferences, The International Joint Conference on Computer Science and Software Engineering (JCSSE) and the National Computer Science and Engineering Conference (NCSEC) between 2005 and 2008. The bibliographical data which appeared in their research was extracted for database construction. Subsequently, the principle of Social Network Analysis was applied for analysis of the extracted data in order to find relationships between researchers, research topics, conferences and years in which they were organized.

The data analysis enables users to know which researchers in Thailand conducted their research in what area, to which academic institutes they were attached, and who were their co-authors. In addition, the analysis also allows users to see trends of Thai researchers research topics. It was hoped that as a result of this analysis, the derived database will be useful for people who wish to search for research data in areas in which they are interested. Similarly, government agencies working on the promotion and support of research work will be informed of research trends in Thailand so as to speed up work in support of conducting research in significant areas or those in which the country is lacking.

KEY WORDS: SOCIAL NETWORK ANALYSIS / IT-RELATED RESEARCHES/ THAI RESEARCHERS / DATA MINING / KNOWLEDGE DISCOVERY

140 pages.

การทำเหมืองข้อมูลเครือข่ายทางสังคมของงานวิจัยด้านเทคโนโลยีสารสนเทศและนักวิจัยไทย
MINING SOCIAL NETWORK OF IT-RELATED RESEARCHES AND THAI RESEARCHERS

อุษณี ศรีภิรมย์ 4737988 SCCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการที่ปรึกษาโครงงานวิจัย : ทรงศรี ตั้งศรีไพโรจน์, Ph.D.,
สุดสงวน งามสุริยะโรจน์, Ph.D., ชาญยศ ปลื้มปีติวิริยะเวช, Ph.D.

บทคัดย่อ

งานวิจัยนี้ได้มุ่งเน้นที่จะศึกษาเครือข่ายทางสังคมของนักวิจัยคนไทยโดยเน้นเฉพาะนักวิจัยที่ทำงานวิจัยด้านที่เกี่ยวข้องกับ Computer science และ Information technology ซึ่งงานวิจัยนี้ได้นำเอาข้อมูลที่มีอยู่ในผลงานวิจัยของนักวิจัยคนไทยที่ส่งผลงานวิจัยเข้าร่วมการประชุมทางวิชาการ The International Joint Conference on Computer Science and Software Engineering (JCSSE) และ The National Computer Science and Engineering Conference (NCSEC) ตั้งแต่ปีค.ศ. 2005 ถึงปีค.ศ. 2008 โดยข้อมูลส่วน bibliography ของงานวิจัยจะถูกดึงออกมาแล้วนำมาสร้างเป็นฐานข้อมูล หลังจากนั้นจึงนำหลักการของการวิเคราะห์ข้อมูลเครือข่ายทางสังคม(Social Network Analysis) มาใช้ในการวิเคราะห์ข้อมูล เพื่อหาความสัมพันธ์ที่มีอยู่ระหว่างผู้ทำงานวิจัย หัวข้องานวิจัย การประชุมทางวิชาการและปีที่จัดการประชุม

ผลจากการวิเคราะห์ด้วยวิธีการของการวิเคราะห์ข้อมูลเครือข่ายทางสังคมทำให้ผู้ใช้ทราบว่าในประเทศไทยมีนักวิจัยคนใด ทำงานวิจัยด้านใดอยู่บ้าง และอยู่ที่สถาบันการศึกษาหรือองค์กรอิสระใด และทราบว่านักวิจัยคนใดทำงานวิจัยร่วมกันบ้าง นอกจากนั้นยังสามารถทราบได้ว่าหัวข้องานวิจัยใดที่มีนักวิจัยคนไทยให้ความสนใจทำงานวิจัยด้านนั้นเป็นจำนวนมากและมีจำนวนมากน้อยเพียงใด ซึ่งผลลัพธ์ที่ได้นี้จะเป็นประโยชน์แก่ผู้ที่สนใจทั่วไปๆ ที่ต้องการค้นหาข้อมูลผลงานวิจัยด้านที่ตนเองสนใจ นอกจากนั้นหน่วยงานของรัฐที่เกี่ยวข้องกับการส่งเสริมและสนับสนุนการทำงานวิจัยสามารถทราบถึงทิศทางของการทำงานวิจัยในประเทศไทย เพื่อเร่งให้การส่งเสริมในด้านการสร้างเครือข่ายของนักวิจัย ตลอดจนการส่งเสริมการทำงานวิจัยในสาขาที่สำคัญรวมถึงสาขางานวิจัยที่ขาดแคลนต่อไป

140 หน้า

# CONTENTS


| | | Page |
|---|---|---|
| **ACKNOWLEDMENTS** | | iii |
| **ABSTRACT (ENGLISH)** | | iv |
| **ABSTRACT (THAI)** | | v |
| **LIST OF TABLES** | | ix |
| **LIST OF FIGUERS** | | xi |
| **CHAPTER I** | **INTRODUCTION** | |
| | 1.1 Motivation | 1 |
| | 1.2 Problem Statements | 2 |
| | 1.3 Research Objectives | 3 |
| | 1.4 Scope of the Research Project | 3 |
| | 1.5 Research Contributions | 4 |
| **CHAPTER II** | **LITERATURE REVIEW** | |
| | 2.1 Thailand Researches Database and The Conferences on Computer Science and Information Technology | 5 |
| | 2.1.1 Thailand Researches Database | 5 |
| | 2.1.2 The Conferences on Computer Science and Information Technology | 6 |
| | 2.2 Information Retrieval | 9 |
| | 2.2.1 Document Representation | 11 |
| | 2.2.2 Query Representation | 13 |
| | 2.2.3 Query and Document Matching | 14 |
| | 2.2.4 Relevance Feedback | 16 |
| | 2.2.5 Evaluation of IR system | 16 |
| | 2.3 Data Mining | 17 |
| | 2.4 Social Network Analysis | 19 |

# CONTENTS (cont.)


| | | Page |
|---|---|---|
| | 2.4.1 Adjacency Matrix | 21 |
| | 2.4.2 Measurement of Social Network Analysis | 22 |
| | 2.4.3 Two-mode Centrality | 25 |
| | 2.4.4 Social Network Analysis Software | 28 |
| | 2.5 Related Work | 28 |
| | 2.5.1 Researches Using Social Network Analysis to Analyze Various Types of Data | 28 |
| | 2.5.2 Researches Using Social Network Analysis to Analyze Data on Published Research Works | 32 |
| **CHAPTER III** | **SYSTEM DESIGN** | |
| | 3.1 System Architecture | 34 |
| | 3.2 Data Preparation | 35 |
| | 3.2.1 Data Sources | 36 |
| | 3.2.2 Lexical Analysis | 43 |
| | 3.2.3 Data Transformation | 46 |
| | 3.2.4 Database Design | 48 |
| | 3.2.5 Adjacency Matrix Construction | 54 |
| | 3.3 Data Analysis by Social Network Analysis | 61 |
| | 3.3.1 Data Analysis of Expert Finding | 61 |
| | 3.3.2 Data Analysis of Co-author Finding | 63 |
| | 3.3.3 Data Analysis of Research Topic Category Finding | 64 |
| | 3.4 User Interface Design | 67 |
| | 3.4.1 Expert Searching | 67 |
| | 3.4.2 Co-author Searching | 70 |
| | 3.4.3 Research Topic Category Searching | 73 |

# CONTENTS (cont.)


| | | Page |
|---|---|---|
| **CHAPTER IV** | **SYSTEM IMPLEMENTATION** | |
| | 4.1 Hardware and Software for Implementation | 76 |
| | 4.1.1 Hardware Specification | 76 |
| | 4.1.2 Software Specification | 76 |
| | 4.2 Prototype Implementation | 76 |
| | 4.2.1 Data Preparation Tool Implementation | 77 |
| | 4.2.2 Social Network Analysis Tool Implementation | 80 |
| | 4.2.3 Searching Tool Implementation | 82 |
| **CHAPTER V** | **SYSTEM TESTING** | |
| | 5.1 Objectives and Method | 95 |
| | 5.2 Test Scenarios | 95 |
| | 5.2.1 Expert Searching | 95 |
| | 5.2.2 Co-author Searching | 96 |
| | 5.2.3 Research Topic Category Searching | 96 |
| | 5.3 Testing Result | 96 |
| | 5.3.1 Expert Searching Result | 97 |
| | 5.3.2 Co-author Searching Result | 111 |
| | 5.3.3 Research Topic Category Searching Result | 119 |
| | 5.4 Summary and Discussion | 132 |
| **CHAPTER VI** | **CONCLUSION AND FUTURE WORK** | |
| | 6.1 Conclusion | 133 |
| | 6.2 Future Work | 134 |
| **REFERENCES** | | 136 |
| **BIOGRAPHY** | | 140 |

# LIST OF TABLES


| Table | | Page |
|---|---|---|
| Table 2.1: | Example of stopwords | 11 |
| Table 2.2: | The contingency table for a test set | 17 |
| Table 3.1: | Newly defined paper categories | 38 |
| Table 3.2: | A summary of overall initial data used in this research | 39 |
| Table 3.3: | Summarized data of number of research papers classified by research topic categories for each year | 40 |
| Table 3.4: | An example of transformation rules in this system | 47 |
| Table 3.5: | Data structure of table: paper | 50 |
| Table 3.6: | Data structure of table: author | 50 |
| Table 3.7: | Data structure of table: conference | 51 |
| Table 3.8: | Data structure of table: email | 51 |
| Table 3.9: | Data structure of table: category | 51 |
| Table 3.10: | Data structure of table: sub_category | 52 |
| Table 3.11: | Data structure of table: author_defined_keyword | 52 |
| Table 3.12: | Data structure of table: title_keyword | 52 |
| Table 3.13: | Data structure of table: abstract_keyword | 53 |
| Table 3.14: | Data structure of table: base_university | 53 |
| Table 3.15: | Data structure of table: base_faculty | 53 |
| Table 3.16: | Data structure of table: base_department | 54 |
| Table 3.17: | Data structure of table: base_organization | 54 |
| Table 3.18: | An example of data to be constructed as an adjacency matrix | 56 |
| Table 3.19: | An example of an adjacency matrix for analysis of expert finding | 56 |
| Table 3.20: | An example of data to be constructed as an adjacency matrix for analysis of co-author finding | 58 |
| Table 3.21: | An example of an adjacency matrix for analysis of co-author finding | 58 |

# LIST OF TABLES (cont.)


| Table | | Page |
|---|---|---|
| Table 3.22: | An example of data to be constructed as an adjacency matrix for analysis of research topic category finding | 60 |
| Table 3.23: | An example of an adjacency matrix for analysis of research topic category finding | 61 |
| Table 5.1: | Details of test scenario 1 of expert searching | 97 |
| Table 5.2: | Details of test scenario 2 of expert searching | 100 |
| Table 5.3: | Details of test scenario 3 of expert searching | 102 |
| Table 5.4: | Details of test scenario 4 of expert searching | 105 |
| Table 5.5: | Details of test scenario 5 of expert searching | 108 |
| Table 5.6: | Details of test scenario 1 of co-author searching | 111 |
| Table 5.7: | Details of test scenario 2 of co-author searching | 114 |
| Table 5.8: | Details of test scenario 3 of co-author searching | 117 |
| Table 5.9: | Details of test scenario 1 of research topic category searching | 120 |
| Table 5.10: | Details of test scenario 2 of research topic category searching | 123 |
| Table 5.11: | Details of test scenario 3 of research topic category searching | 126 |
| Table 5.12: | Details of test scenario 4 of research topic category searching | 129 |

# LIST OF FIGURES


| Figure | | Page |
|---|---|---|
| Figure 2.1: | The website “thairesearch” | 6 |
| Figure 2.2: | An IR system | 10 |
| Figure 2.3: | Vector of document and query | 15 |
| Figure 2.4: | An example of social network | 20 |
| Figure 2.5: | An example of an adjacency matrix | 22 |
| Figure 2.6: | A social network of an adjacency matrix | 22 |
| Figure 2.7: | An example of a two-mode data set | 25 |
| Figure 2.8: | An example of an academic research network in research of Jie Tang et al. | 30 |
| Figure 2.9: | An example of terrorism network of Jammu and Kashmir | 31 |
| Figure 3.1: | System Architecture | 34 |
| Figure 3.2: | The Structure Chart of the system | 35 |
| Figure 3.3: | Call for paper for JCSSE 2007 | 36 |
| Figure 3.4: | Call for paper for NCSEC 2007 | 37 |
| Figure 3.5: | An example of PDF document in JCSSE conference | 41 |
| Figure 3.6: | An example of PDF document in NCSEC conference | 42 |
| Figure 3.7: | Data Preparation process | 43 |
| Figure 3.8: | An example of text file from pdf file | 44 |
| Figure 3.9: | Processes of Lexical Analysis and Index term selection | 46 |
| Figure 3.10: | An example of data transformation | 47 |
| Figure 3.11: | Data Transformation process | 48 |
| Figure 3.12: | An E-R diagram of the system database | 49 |
| Figure 3.13: | Adjacency matrix construction for expert finding | 55 |
| Figure 3.14: | Adjacency matrix construction for analysis of co-author finding | 57 |
| Figure 3.15: | Adjacency matrix construction for analysis of research topic category finding | 59 |

# LIST OF FIGURES (cont.)


| Figure | | Page |
|---|---|---|
| Figure 3.16: | Author, paper, category relation | 62 |
| Figure 3.17: | Author's relation | 64 |
| Figure 3.18: | Research papers and category relation | 65 |
| Figure 3.19: | An example of a network diagram constructed from an adjacency matrix of expert finding | 66 |
| Figure 3.20: | An example of network diagram constructed from an adjacency matrix of co-author finding | 67 |
| Figure 3.21: | Expert searching process | 68 |
| Figure 3.22: | An example of an expert searching interface | 69 |
| Figure 3.23: | An example of an interface of expert searching results (social network view) | 70 |
| Figure 3.24: | An example of an interface of expert searching results (data view) | 70 |
| Figure 3.25: | Co-author searching process | 71 |
| Figure 3.26: | An example of a co-author searching interface | 72 |
| Figure 3.27: | An example of an interface of co-author searching results (social network view) | 72 |
| Figure 3.28: | An example of an interface of co-author searching results (data view) | 73 |
| Figure 3.29: | A research topic category searching process | 74 |
| Figure 3.30: | An example of a research topic category searching interface | 74 |
| Figure 3.31: | An example of an interface of research topic category searching results (social network view) | 75 |
| Figure 3.32: | An example of an interface of research topic category searching results (data view) | 75 |
| Figure 4.1: | Pseudo code of data preparation tool implementation | 77 |
| Figure 4.2: | Pseudo code of data preparation tool implementation (cont.) | 78 |
| Figure 4.3: | Pseudo code of data preparation tool implementation (cont.) | 79 |
| Figure 4.4: | An interface for users to load a pdf file | 79 |

# LIST OF FIGURES (cont.)


| Figure | | Page |
|---|---|---|
| Figure 4.5: | Pseudo code of adjacency matrix construction for expert finding | 80 |
| Figure 4.6: | Pseudo code of adjacency matrix construction for co-author finding | 81 |
| Figure 4.7: | Pseudo code of adjacency matrix construction for research topic category finding | 81 |
| Figure 4.8: | Pseudo code of data analysis via the Social Network Analysis Technique | 82 |
| Figure 4.9: | The system's main interface | 83 |
| Figure 4.10: | Help menu | 83 |
| Figure 4.11: | Help sub-menu: Searching | 84 |
| Figure 4.12: | Help sub-menu: Organization/university | 85 |
| Figure 4.13: | Help sub-menu: Scope/year | 85 |
| Figure 4.14: | An expert searching interface | 87 |
| Figure 4.15: | An example of filtering faculty and department based on a university chosen | 88 |
| Figure 4.16: | An example of filtering research topic sub-category based on research topic category selected | 88 |
| Figure 4.17: | An example of expert searching results (social network view) | 89 |
| Figure 4.18: | An example of expert searching results (data view) | 89 |
| Figure 4.19: | A co-author searching interface | 91 |
| Figure 4.20: | An example of co-author searching results (social network view) | 91 |
| Figure 4.21: | An example of co-author searching results (social network view) | 92 |
| Figure 4.22: | A research topic category searching interface | 93 |
| Figure 4.23: | An example of research topic category searching results (social network view) | 94 |

# LIST OF FIGURES (cont.)


| Figure | Page |
|---|---|
| Figure 4.24: An example of research topic category searching results (data view) | 94 |
| Figure 5.1: Details of expert searching conditions of test scenario 1 | 98 |
| Figure 5.2: An example of results of expert searching based on test scenario 1 (social network view) | 98 |
| Figure 5.3: An example of results of expert searching based on test scenario 1 (data view) | 99 |
| Figure 5.4: Details of expert searching conditions of test scenario 2 | 100 |
| Figure 5.5: An example of results of expert searching based on test scenario 2 (social network view) | 101 |
| Figure 5.6: An example of results of expert searching based on test scenario 2 (data view) | 101 |
| Figure 5.7: Details of expert searching conditions of test scenario 3 | 103 |
| Figure 5.8: An example of results of expert searching based on test scenario 3 (social network view) | 104 |
| Figure 5.9: An example of results of expert searching based on test scenario 3 (data view) | 104 |
| Figure 5.10: Details of expert searching conditions of test scenario 4 | 106 |
| Figure 5.11: An example of results of expert searching based on test scenario 4 (social network view) | 107 |
| Figure 5.12: An example of results of expert searching based on test scenario 4 (data view) | 107 |
| Figure 5.13: Details of expert searching conditions of test scenario 5 | 109 |
| Figure 5.14: An example of results of expert searching based on test scenario 5 (social network view) | 110 |
| Figure 5.15: An example of results of expert searching based on test scenario 5 (data view) | 110 |
| Figure 5.16: Details of co-author searching conditions of test scenario 1 | 112 |

# LIST OF FIGURES (cont.)


| Figure | | Page |
|---|---|---|
| Figure 5.17: | An example of results of co-author searching based on test scenario 1 (social network view) | 113 |
| Figure 5.18: | An example of results of co-author searching based on test scenario 1 (data view) | 113 |
| Figure 5.19: | Details of co-author searching conditions of test scenario 2 | 115 |
| Figure 5.20: | An example of results of co-author searching based on test scenario 1 (social network view) | 115 |
| Figure 5.21: | An example of results of co-author searching based on test scenario 2 (data view) | 116 |
| Figure 5.22: | Details of co-author searching conditions of test scenario 3 | 117 |
| Figure 5.23: | An example of results of co-author searching based on test scenario 3 | 118 |
| Figure 5.24: | An example of results of co-author searching based on test scenario 3 (social network view) | 118 |
| Figure 5.25: | An example of results of co-author searching based on test scenario 3 (data view) | 119 |
| Figure 5.26: | Details of research topic category searching conditions of test scenario 1 | 121 |
| Figure 5.27: | An example of results of research topic category searching based on test scenario 1 (social network view) | 121 |
| Figure 5.28: | An example of results of research topic category searching based on test scenario 1 (data view) | 122 |
| Figure 5.29: | Details of research topic category searching conditions of test scenario 2 | 124 |
| Figure 5.30: | An example of results of research topic category searching based on test scenario 2 (social network view) | 124 |
| Figure 5.31: | An example of results of research topic category searching based on scenario 2 (data view) | 125 |

# LIST OF FIGURES (cont.)


| Figure | | Page |
|---|---|---|
| Figure 5.32: | Detail of research topic category searching condition of test scenario 3 | 127 |
| Figure 5.33: | An example of results of research topic category searching based on test scenario 3 (social network view) | 127 |
| Figure 5.34: | An example of results of research topic category searching based on test scenario 3 (data view) | 128 |
| Figure 5.35: | Details of research topic category searching conditions of test scenario 4 | 130 |
| Figure 5.36: | An example of results of research topic category searching based on test scenario 4 (social network view) | 130 |
| Figure 5.37: | An example of results of research topic category searching based on test scenario 4 (data view) | 131 |

# CHAPTER I
# INTRODUCTION

## 1.1 Motivation

The Office of the National Research Council of Thailand (Vor Chor) [1] is a state agency established to support in 3 majors aspects as the following. First, to promote research affecting the country's strategies. Secondly, to focus on support in the development of new innovation generating the strength of factors contributing to sustainable and quality economic growth. Finally, to concentrate on the development of human resources and creation of potential in the use of information technology to support of the government sectors.

As a consequence of a workshop on the strategies of small and medium enterprise development, the agencies concerned were of the opinion that there should be a data source collecting Thailand's research works, including providing services on the retrieval of research works data. Accordingly, there has been a collection of researches' data and the country's researches works in Thai research database. This aims to disseminate the data to interested people, apart from rending services on data retrieval and searching from one point on the Internet through the website (www.thairesearch.in.th) [2] , which is a search engine where users can search for data they need. Nevertheless, this type of data searching via a search engine may not be able to meet the users's demand in the case the users require analytical information, which is more than general search for research documents or researchers. For instance, when users want to know a name of expertise leading research in different areas by viewing from the number of research works which have been worked by those researchers, including which institute or organization of the expertise are attached to. This also applied to when user need to know whether each researcher work with other researchers, including the trend of research that are hot issue among researchers in Thailand, etc.

If there is a system or a data source which can collecting data concerning researchers, which is aimed at disseminating research works inside the country, when those data are further analyzed, it will provide data that are useful for users themselves because they are informed of research topics in which they are interested. Furthermore, for agencies or institutes, this will be a promotion for coordination among themselves, which working research in the same areas. Meanwhile, it will allow users to know the trend of research topics in Thailand, and area in which tremendous or little research has been conducted so that there will be an acceleration for the promotion of working of research in the area in which there is a shortage. For the overall picture of the country, the government will be able to follow up the consequences of research and development of the country in different fields, and subsequently use these data for the country administration and national policy.

## 1.2 Problem statements

The problems of this research include the following

1) A problem of collecting research documents data in different formats, which is a combination of data in different structural patterns or presentation formats in order to obtain more complete data [5]. For example, research document data that will be propagated via such media as web pages, and research document data in a pdf file format, etc. These data may have to be merged so as to get more complete research document data.

2) A problem of information extraction, which aims to automatically extract such thing as meaning from a data source with a clear domain or from documents without an obvious structural format. In addition, the things extracted have also to maintain their original meaning of the content [3],[4]. Here, the data domain is data of researches and researchers from various data sources while these data must be able to be automatically extracted.

3) A problem of searching for research documents or experts in different fields, as required by users [6],[7]. A search by the current search engine can not give an answer as to who are experts or in which area they conduct research.

4) A problem of using researcher's data and their research works for the analysis in order to find the relationship in various aspects among the researchers themselves and the area of the research topics [8],[9]. However, the results of data searching by a search engine can not indicate the relationship among researchers, research documents and research topics.

## 1.3 Research Objectives

This research has the following objectives.

1) To collect data and create a research works database with respect to IT of Thai researchers.

2) To analyze data to find the relationship among groups of researchers and the relationship between researchers and their research works through an implementation of the Social Network Analysis technique.

3) To design and develop a prototype system that can search for researchers in different fields as user need, and research topic category, as well as can display the relationship of each researcher.

## 1.4 Scope of the Research Project

The scope of this research project includes the following.

1) The data source used in this research consists of Thai researchers' data and data on research works published for Academic conferences organized in Thailand. Particularly, the two following academic conferences are focused.

- The National Computer Science and Engineering Conference (NCSEC) from 2005 to 2008
- The International Joint Conference on Computer Science and Software Engineering (JCSSE) from 2005 to 2008

The data are in a pdf file format and only research document with English title and abstracts are selected.

2) The data mining technique of Social Network Analysis is used to find

the existing relationship among groups of researchers, and how researchers and their researches are related and in what aspect.

## 1.5 Research Contributions

This research is anticipated to make the following contributions.

1) Be able to search who are experts doing research in various fields related to IT, by considering their research works and to search for their research documents.

2) Support to create a network of researchers from numerous agencies across the country, which regarded as the promotion of co-operations among agencies, institutes and organizations inside the country.

3) The government or agencies concerned with the promotion of research and development of science and technology in the country can follow up the consequences of research and development in this area. Besides, the knowledge derived from this can be applied to the administration and formulation of policies of the country in the area concerned.

# CHAPTER II
# LITERATURE REVIEW

This chapter consists of five sections. Section one discusses the data source collecting research works conducted in Thailand and those obtained from conferences held in the country. Section two describes information retrieval. Sections three and four explain the principles of Data Mining and Social Network Analysis, respectively. Section five, the final section, discusses related works.

## 2.1 Thailand Researches Database and The Conferences on Computer Science and Information Technology

### 2.1.1 Thailand Researches Database

Following the workshop on "Strategies of development of small and medium enterprise" held between 8-9 June, 2001 at Queen Sirikit National Convention Center, Bangkok, the government thought that Thailand should have a data source collecting research works of the country so as to be disseminated to people, including providing services on data searching through the Internet. As a consequence, the National Science and Technology Development Agency, in co-operation with the office of the Thailand Research Fund, the National Institute of Public Health System Research and the National Research Council of Thailand, has established a pilot project on database system for research works of each institute. The data were disseminated on the Internet while the database of different systems could be searched from this service and at a single point. The pilot project for searching the database of research works of Thailand via the Internet has been initiated since September 2001.

Agencies concerned with the promotion and development of research of Thailand have been of the collective opinion that there should be a data source collecting research works of Thailand in one place and providing services on research work data searching on the Internet via the website www.thairesearch.in.th [2] illustrated in Figure 2.1.

The website “thairesearch” provides services on a search for the following data.

- Research projects
- Patents
- Researchers
- Academic work published

Data on published research works collected in the website “thairesearch” are those published in the ECTI annual conference (2004 to present) and International Technical Conference on Circuit/System, Computer and Communications : ITC-CSCC (2006 to present).

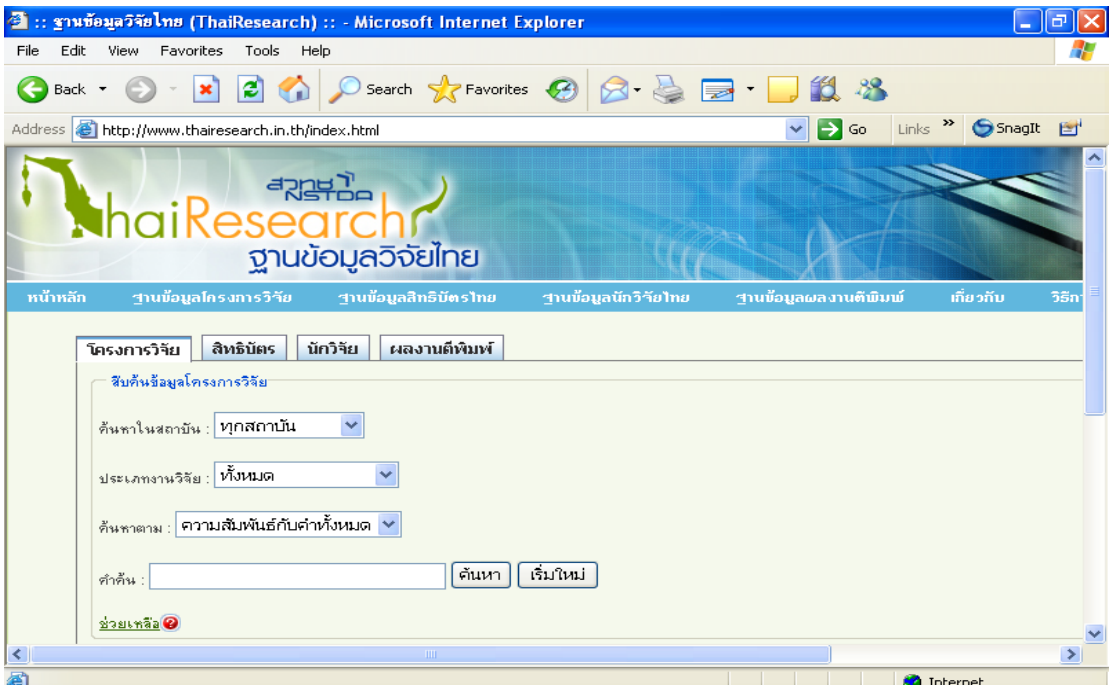



Figure 2.1 The website “thairesearch” [2]

### 2.1.2 The Conferences on Computer Science and Information Technology

Each year Thailand has organized a tremendous number of technical conferences hosted by educational institutes or agencies that alternate with one another. The objectives of most conferences are similar, which are to give an opportunity to

researchers, and to participants to ask questions and express their opinions in accounts related to those research works.

Conferences held in Thailand include the following.

- CIT (National Conference on Computer Information Technologies)
- ECTI-CON (Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology-Conference)
- JCSSE (The International Joint Conference on Computer Science and Software Engineering)
- NCCIT (National Conference on Computing and Information Technology)
- NCIT (National Conference on Information Technology)
- NCSEC (National Computer Science and Engineering Conference)

There are a number of conferences organized in Thailand. Here, details of 4 major conferences are given.

1) **The National Computer Science and Engineering Conference (NCSEC)**

It is a technical conference organized annual, beginning in 1997 and has continued to present (2009) which was the 13$^{th}$ conference. The conference aims to propagate research works and give an opportunity to a new generation of researchers to present their researches. Generally, the researches presented at this conference focus on those related to computer science and engineering. Researchers will submit their researches in one of the 7 following major topics to the NCSEC, identified by the NCSEC committee.

- Computer System
- Artificial Intelligence
- Computer Software
- DSP and Image Processing
- Information Technology
- Computing Theory
- Computer Network

2) **The International Joint Conference on Computer Science and Software Engineering (JCSSE)**

It is a conference held each year, starting for the first time in 2004 and has continued to present (2009), which was the 6th conference. The conference aims to disseminated research works and give an opportunity to a new generation of new researchers to present their researches. Normally, researches concerned with computer science and software engineering. Researchers will submit their researches to the JCSSE in one of the 10 following topics classified by the JCSSE committee.

- Computational Science and Engineering
- Computer Graphics and Computer Animation
- Compute System
- Computer Network
- Information & Communication Technology
- Computer Software
- Computer Intelligence
- Digital Library
- Software Engineering
- Software Application & Development

3) **ECTI Annual Conference (ECTI-CON)**

It is a conferences organized annually by Thailand Technical Electrical Engineering, Electronics, Computers, Telecommunications and Information Technology Association. The conference began for the first time in 2004 and has continued to the present (2009), which was the 6th conference. This conference normally concentrates on presenting research works related to electrical engineering, electronics, computers, telecommunications and information technology. Researchers will submit their researches to the ECIT-CON in one of the 7 following major topics divided by the ECIT committee.

- Computer and Information
- Communication Systems
- Signal Processing
- Circuits and Systems

- Electrical Power
- Control Engineering
- ECTI Management and Policy

4) **The National Conference on Computing and Information Technology (NCCIT)**

It is a conference held each year, starting in 2005 and has continued to the present (2009), which was the 5th conference. It focuses on presenting research works and exchanging knowledge of Computing and Information Technology. Researchers will submit their researches to the NCCIT in one of the following major topics classified by the NCCIT committee.

- Information Technology
- Computer Science
- Computer Education
- Computer Engineering
- Computer Business
- Computer Technology
- Information Science and Management
- Educational Technology and Communication
- Data Communication and Network Technology
- Bioinformatics

## 2.2 Information Retrieval

The Information Retrieval (IR) is a principle related to analysis, data source storage and data searching. The IR structure is favorable to an access to data in which each user is interested. The data that users need will be converted into a query format and subsequently implemented by a search engine. The target of IR is to retrieve data in which users are interested [7].

The objective of creating an IR system is to facilitate a search for a library's documents. But due to numerous advantages and capabilities of computers' calculation, there is widespread use of documents in an electronic format.

On the whole, users of an IR system pays more attention to information retrieval than data retrieval since the latter aims to search for what is corresponding with the required condition only [7], for example, searching for data in the database that have definite structure and meaning. But in information retrieval, it is a search for data related to what users are interested in while the data may not have definite structure.

An IR system can be divided into 3 main parts, which are data needed by users, a group of documents and the results derived from the IR system's performance, as required by users. The results obtained will be displayed in order of relevance to the conditions that users need [6].

In an IR system, documents are displayed by extracting the relationship between words in a sentence and the meaning of data. These two things are used to search for data pertinent to what is required by users. The IR system must be able to extract as much data as possible, which is relevant to what users need, out of data groups. An IR system [7] is shown in Figure 2.2.

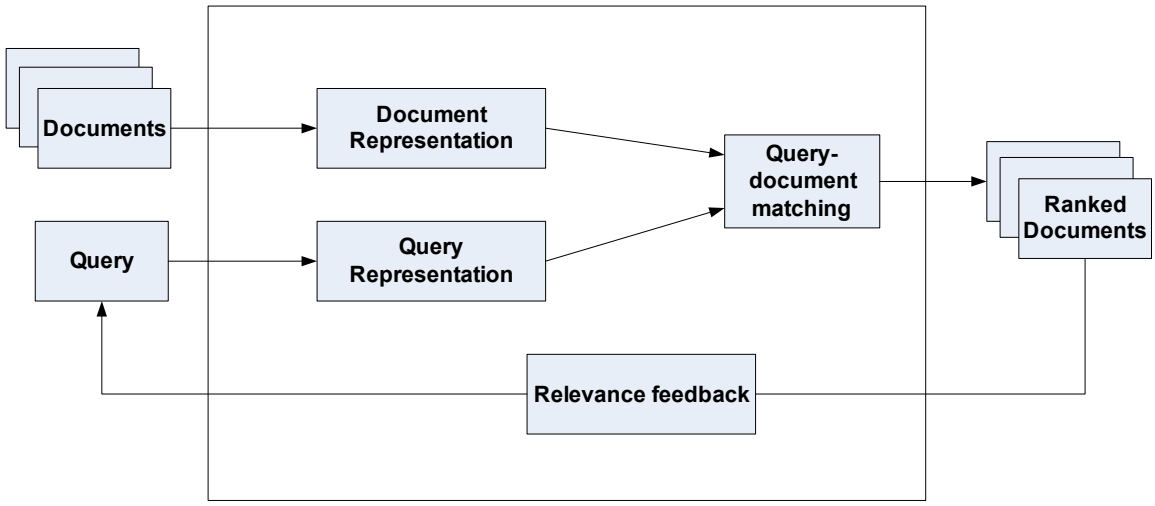



Figure 2.2 An IR system [7]

Section 2.2.1 describes a technique of document representation and what users need. Section 2.2.2 explains a technique for a comparison between document and what users require. In section 2.2.3, the query and document matching is mentioned. Finally, sections 2.2.4 and 2.2.5 discuss relevance feedback and IR system evaluation, respectively.

### 2.2.1 Document Representation

In an IR system the documents to be implemented are not in an original format but they will be implemented through the text operation technique [6] first, which comprise 4 following stages.

**(1) Lexical analysis**

It is a process of separating all texts in a document into an individual word. The objective of the process is to separate words from all texts so that these words can be considered to find whether each word should be a representative of that document. In the English language, lexical analysis is made by segmenting words of a document using white space. This makes it easier to separate each word, compared to Thai words, which are more complicate to be separated than English words.

**(2) Elimination of stopwords**

In a document there is generally a group of words appearing in all documents. In an IR system, these words are called stopwords, which are conjunctions or some verbs. There are about 300 words regarded as stopwords, some of which are given as examples [7] in Table 2.1.

Table 2.1 Example of stopwords

| | | | |
|---|---|---|---|
| A | An | The | On |
| Am | Are | In | Does |
| Was | Were | Do | At |
| Has | Have | Had | Did |
| Be | Is | Will | Shall |

**(3) Stemming**

Stem is part of word to which a word may be added to its beginning or end. The word added to the beginning of a word is called prefix, and added to the end call suffix. Parts of word added to the beginning and end of a stem are call affix. For instance, communicate is the stem of communication. Typically, stemming helps increase the efficiency of an IR system since it reduces the size of an indexing file.

Mostly, to get an affix out is to eliminate a suffix because in most case adding a suffix will create a new word, which is different from an original word. The principles for taking an affix from the stem are numerous but a popular principle is a plotter algorithm, which is an easy principle and yields a good result [6].

A plotter algorithm uses a suffix list to segment a suffix from the stem. Though stemming helps enhance an IR system's efficiency, it also has a disadvantage in that the word's meaning may be lost and some IR systems may require space to store stems and words of which stems have not been segmented.

**(4) Index term weight**

An index term is generally a word representing each document. An index term may be composed of such words as noun, verb, adjective etc. Typically, most of index terms are nouns.

Each word selected from a document to become an index term may also be an index term of other documents. Therefore, the weight value of an index term, call index term weight, is normally defined. It can be calculated from term frequency (tf) and inverted document frequency (idf). For the term frequency (tf), it is the frequency of the number of times for which that index term appears in the document, which can be computed from equation (2.1) [8].

$$tf_{ij} = f_{ij} / \max\{f_{ij}\} \quad \text{---------- (2.1)}$$

Where

$f_{ij}$ = the frequency of term at i in the document No. j

$\max\{f_{ij}\}$ = the maximum frequency of all terms appearing in the document No. j

For the inverted document frequency (idf), it is the value telling whether that index term also appears in others documents of the document collection. The idf term is represented in equation (2.2) [8]

$$idf_i = \log_2 \frac{N}{n_i} \quad \text{---------- (2.2)}$$

Where

$N$ = the number of all documents in the collection

$n_i$ = the number of documents with the terms No. i appearing on them

Mostly, in an IR system, the term weight used is a combined value of tf weight and idf weight, revealed in equation (2.3) [8]

$$W_{ij} = tf_{ij} * idf_i \quad \text{---------- (2.3)}$$

Where

$W_{ij}$ = value of term weight of term i in document No. j

$tf_{ij}$ = frequency (value) of term appearing in document

$idf_i$ = frequency (value) of term appearing in other documents in the collection

Afterwards, the weight value of each index term of documents will be stored in an inverted file.

### 2.2.2 Query Representation

A query is what displays the data that users require. The characteristic of queries is different depending on the model used. Typically, a query consists of keywords. A popular query in an IR system is a keyword-based query because data can be searched and arranged in order rapidly. A keyword-based query may be classified into 4 categories, namely, single-word query, context query, boolean query and natural language query. Here, only a boolean query and a natural language query, which are popular techniques, are discussed [6].

A boolean query has to have an extra operator such as or, not, and to be used for data searching, which is a limitation of an IR system using this type of query. It is not appropriate for users who do not have much knowledge of a boolean query and it normally provides a poor set of answers. Thus, an IR system using a boolean query is replaced by a natural language query. An example of a natural language query is (the sentence) “I want to know about database system”. For this kind of query, users

who do not want knowledge of using an extra operator can search for data required. When data are converted into a query format, an IR system will search for documents or data most relevant to data that user requires by calculating the similarity value between the query and each document in document collection, which will be discussed in section 2.2.3.

Before an IR system deals with a query, there are processes of lexical processing and term weighting. In addition, there is query expansion to enhance the IR system's efficiency, which may be done by the IR system itself while in the second way, it is commanded by users. For query expansion, a term close to one in the query may be selected from a thesaurus. The term chosen from the thesaurus is added to the query before the IR system deal with it (query).

### 2.2.3 Query and Document Matching

For a technique of searching for data relevant to what users require, in an IR system a comparative principle, between what user need and all existing data, is applied. So the consequence of a comparison is the level of relevance of what users require and data, which will be displayed in the format of data item arranged in order of similarity to what users need, from most to least.

A boolean model typically comprises a group of documents and a group of operators. The boolean model will search for documents exactly matching the query, which gives a group of documents that match the query in large number and can not arrange them in order of similarity between the query and documents. As a result, a boolean model needs a technique that can find documents similar to the query and display the level of similarity between query and each document, which helps reduce the number of documents similar to the query.

In a vector space model, only some document similar to a query can be searched by computing the value of similarity between each documents and the query and then arranging in order the level of similarity of the documents and query, from most to least. In a vector space model, both the documents and query are represented by the vector on multidimensional space, which depends on the number of terms in a group of all documents [8].

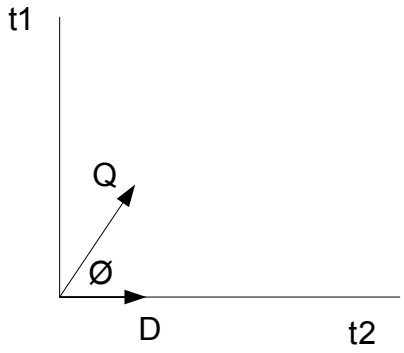



Figure 2.3 Vector of document and query [8]

Figure 2.3 illustrates the representation of document and query by the vector in a vector space model. Document D1 is shown using the term weight of each document while the query is represented in the same way as documents D1 is represented. The angles between the query and documents D1 is represented by $Ø_1$.

$$\text{Sim (D,Q)} = \frac{D * Q}{\|D\| \|Q\|} \quad \text{----------(2.4)}$$

Where

D = Document vector

Q = Query vector

|| Q|| = magnitude of query vector

|| D|| = magnitude of document vector

There are a number of techniques used to measure the similarity between a query and documents, for example, cosine coefficient, which is widely used in an IR system. Cosine similarity is to calculate the angle between the vector of a document and the vector of a query. The cosine similarity value is between 0 and 1. if the similarity value is 0, it indicates that the document and query are not similar. But if the similarity value is 1, it means that the document and query are similar. The calculation of cosine similarity value is shown in equation (2.4).

### 2.2.4 Relevance Feedback

Apart from the query expansion technique using the term approximate or similar to the term in a query, another popular technique for query expansion is relevance feedback. In this technique, an IR system will select documents or data relevant to the data which users need, and then has users consider which document is relevant to what they require. Subsequently, the IR system will use the terms or data in the document that users have considered that they are relevant to data the users need, in order to improve the query so that the newly improved query is identical with or similar to the data that users have chosen to be most identical with data they need.

### 2.2.5 Evaluation of IR system

The evaluation techniques widely used to measure an IR system's efficiency include the precision and recall [6], each of which is described below.

**- Precision (P)**

Precision is an indicator telling how many documents retrieved are relevant to a query, among all documents retrieved. This indicator is defined in equation (2.5)

$$P = \frac{The\_number\_of\_relevant\_documents\_retrieved}{The\_total\_number\_of\_documents\_retrieved} \quad \text{------ (2.5)}$$

**- Recall (R)**

It is an indicator telling how many documents, out of those relevant to a query, are retrieved by an IR system. The recall is defined in equation (2.6)

$$R = \frac{The\_number\_of\_relevant\_documents\_retrieved}{The\_number\_of\_relevant\_documets\_in\_a\_collection} \quad \text{------ (2.6)}$$

The precision and recall can be displayed in a format of contingency table revealed in Table 2.2

Table 2.2 The Contingency table for a test set

| | predicted positive | predicted negative |
|---|---|---|
| actual positive | a | b |
| actual negative | c | d |

a: the number of testing examples correctly assigned to this category.

b: the number of testing examples incorrectly assigned to this category.

c: the number of testing examples incorrectly rejected from this category.

d: the number of testing examples correctly rejected from this category.

So, from the contingency table, the precision and recall values may be computed using equation (2.7) and (2.8), respectively.

- Precision

$$P = a/(a + b) \quad \text{---------- (2.7)}$$

- Recall

$$R = a/(a + c) \quad \text{---------- (2.8)}$$

## 2.3 Data Mining

Data Mining, sometimes called knowledge discovery in database-KDD, is a technique of automatically finding for a pattern from an enormous amount of data through use of steps of a statistical approach, machine learning and pattern recognition. Another definitions of data mining is a process acted upon data (mostly in a tremendous amount) in order to find pattern, and relation concealed in that set of data, using statistical principle, recognition, machine learning and mathematical principle [11],[13].

A process of Data Mining is differently divided by theories which each data mining manager uses as a reference. Accordingly, there is no definite working process of data mining but its working process of data mining but its working process may be roughly divided as the following [12].

**1) Problem definition**

It is a process of defining an emerging problem in a comprehensive and distinct manner, including objectives of data mining.

**2) Data exploration**

It is a process of exploring existing data to find its' characteristic.

**3) Data preparation**

It is a process of data management so that data can be put into algorithms of data mining, for instance data selection, data cleaning, data integration, data transformation and data reduction.

**4) Modeling**

It is a process of mining. There are several types of data mining such as database segmentation, predictive modeling, etc. However, each data mining has algorithms to be selected for use, as discussed below.

- Association rule discovery is a search for the relation of events or objects that occur simultaneously. An example of the application of association rules is the analysis of market basket data.
- Database Segmentation (Clustering) is to divide similar data into cluster, for example, division of a group of patients of the same disease based on its symptom in order to the diagnosis of the cause of disease, which may use a k-means algorithm or an unsupervised learning neural network like Kohonen's self organizing map.
- Data Classification and Predictive are to find rules so as to define the type of things from existing properties, for instance, the prediction of characteristic of those who will generate a good or bad debt, in which CART (Classification And

Regression Tree) or supervised learning neural network, may be used.

- Deviation Detection is a technique of finding values different from a standard value or an expected value so as to see how much they are different [13]. On the whole, a statistical approach are visualization is applied. This technique is typically used to verify fake signatures or credit cards, including detection of defects of work in an industrial plant.

**5) Evaluation**

It is a process in which data analysis compile the results of data mining in order for an evaluation and conclusion of the meaning of the results obtained.

**6) Deployment**

It is a final process in which the result of data mining is useful knowledge for supporting in decision making.

## 2.4 Social Network Analysis

Social Network Analysis (SNA) is the structure of a social network. Such structure is constructed from node and line showing the relation between nodes, which are call ties. Nodes in this social network has relation with other nodes in the same network in many characteristics, for instance, the relation of friends, the relation of boss and subordinates, etc. An example of social network is illustrated in Figure 2.4 [15].

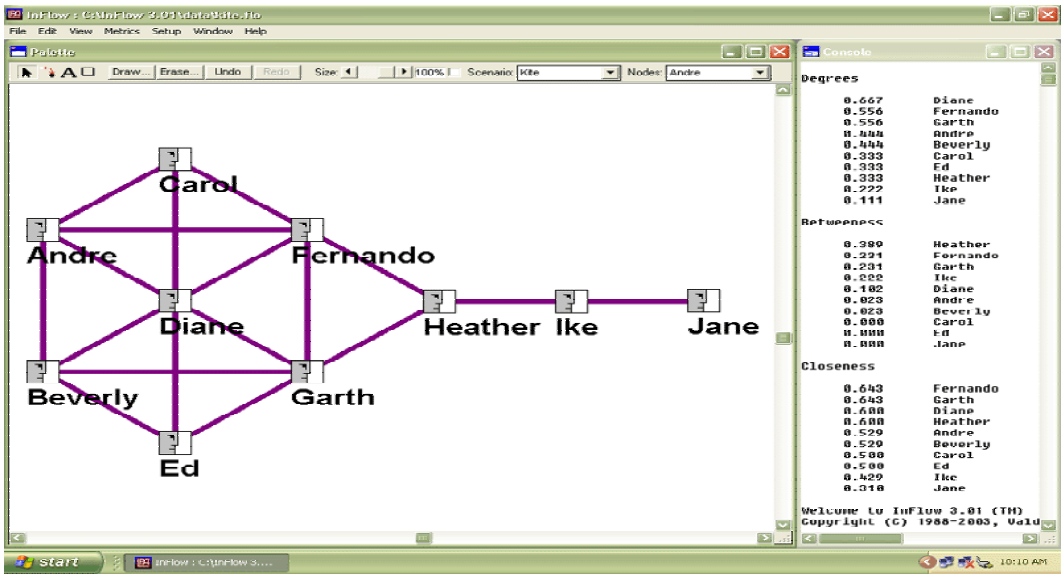



Figure 2.4 An example of social network [15]

Social Network Analysis views the pattern of relation of things in society (in the same network) based on the characteristic of the network theory. According to the concept of a network theory, the components of a network are nodes and their relation (ties).

There are a number of reasons why Social Network Analysis is so popular. First, the basic concept of a social network is easy. A social network is a set of actors which may be spots or nodes or agents. There may be the relation among other nodes within the network (which may be edge or ties). Similarly, there may be several actors or nodes inside the network and several patterns of relation between actors or nodes. An outstanding point that a social network is used is when it (network) is analysed , it enables us to know and explain the relation among actors within that network.

Secondly, when an enormous amount of data used to build a social network is managed in order to construct that network, it enables us to know the complex pattern or social structure hidden inside those data. All works of a social network can be made easier by using mathematical tools to assist in the calculation and display of data result of that pattern in the format of a graph theory.

Thirdly, a mathematical principle is applied. The reason for using mathematics and a graph theory in Social Network Analysis is to display an explanation of the

network. For these works, a computer may be used to assist in storing and managing data rapidly, especially when where is an enormous amount of data.

Fourthly, for Social Network Analysis, the pattern of data is explained in the form of graph , which makes it easier to understand and visualize than to explain by words.

Social Network Analysis used two mathematical theories, namely, matrix and graph to represent data related to the pattern of relation among members in the network. There are various kinds of graphs such as bar chart, pie chart, line and trend graph, etc. For network analysis, a type of graph is used to representing actors and lines or edges for a display of lines of relation.

A graph comprises nodes, called actors, or points connected by edges or ties. Edges or ties reveal the relation between nodes or actors in a graph.

A graph may display only one relation or several relations. Edges or ties showing only one relation between nodes or actors are called simplex relation while those showing several relations between them are called multiplex relation.

### 2.4.1 Adjacency Matrix

A matrix is a rectangle inside of which there are elements. A matrix typically has 2 dimensions or over. The size of a matrix is described by the numbers of rows and columns storing the elements inside it [15].

The matrix size is defined by the numbers of rows and columns. By definition a matrix has i rows and j columns. A matrix with one row is called a row vector while a matrix with one column is called a column vector.

An adjacency matrix is a matrix format used in Social Network Analysis, consisting of rows and columns representing the values of nodes. The values inside the matrix indicate the relation between nodes [16]. An easiest pattern of matrix is a binary matrix, inside of which the values showing the relation between nodes comprise only 2 values of 0 and 1.

| | Bob | Carol | Ted | Alice |
|---|---|---|---|---|
| Bob | --- | 1 | 1 | 0 |
| Carol | 0 | --- | 1 | 0 |
| Ted | 1 | 1 | --- | 1 |
| Alice | 0 | 0 | 1 | --- |

Figure 2.5 An example of an adjacency matrix [15]

Based on Figure 2.5, the adjacency matrix can be constructed as a social network seen in Figure 2.6.

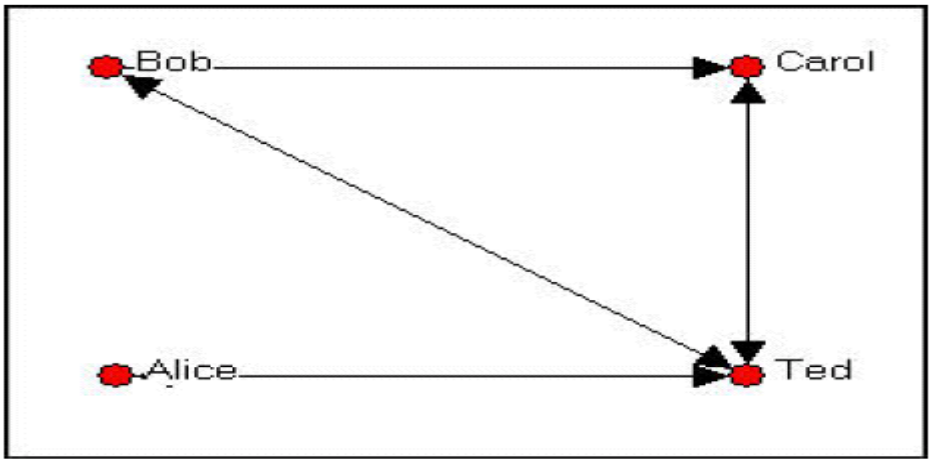



Figure 2.6 A social network of an adjacency matrix [15]

The distance between pairs of actors or nodes in the network may be measured as the value of geodesic distance. This value is the counting of the number of relations in the possible shortest distance of walking from an actor to other actors within the network.

### 2.4.2 Measurement of Social Network Analysis

The analysis of most networks is an explanation of the location of each node or actor. It explains the character of role of each actor or node. The values explaining the role character of nodes can be measured in the form of centrality value.

There are 4 categories of measurements of centrality values, namely, degree, closeness, betweenness and eigenvector [17].

In the graph theory and network analysis, there is measurement of centrality values of actors or nodes. These values are related to the significance of actors or nodes inside the network.

- **Degree Centrality**

It is the number of links of each node. The value of degree centrality indicates the influence of that node inside the network. There are 2 kinds of degrees, indegree and outdegree. The former is the number of links coming to each node while the latter is the number of links coming out of each nodes. Typically, when we speak of the degree. The degree centrality can be defined in equation (2.9).

$$C_D(x) = \text{degree of unit X} \quad \text{---------- (2.9)}$$

Where

degree of unit X = number of tie for each node inside the network

- **Closeness Centrality**

Each node in a network may not have direct link with each other but may be linked with another node through the other. The value of closeness centrality explains whether such a node is the center of the network or not. The closeness centrality value is the sum of all distances from an interesting point to other points in the network. The closeness centrality may be defined in equation (2.10).

$$C_c(x) = \frac{1}{\sum_{y \in U} d(x,y)} \quad \text{---------- (2.10)}$$

Where

d(x,y) = geodesic distance (the number of shortest path) from x to other nodes (y) in the network

U = set of nodes in network

- **Betweenness Centrality**

It is a value telling whether such a node is a passage way which is the shortest path, and how many routes that node is on the shortest path. The betweenness value can be defined in equation (2.11).

$$c_B(x) = \sum_{i,j} \frac{g_{ixj}}{g_{ij}} \qquad \text{---------- (2.11)}$$

Where

$g_{ij}$ = the number of geodesic distance from i to j

$g_{ixj}$ = the number of shortest path between i and j which passes x

- **Eigenvector Centrality**

It is a value used to measure the significance of nodes in a network. This value is defined for each node in the network based on the principle that if a node has many connections, it is the node with a high value of eigenvector. The calculation of eigenvector value may be made through equation (2.12).

$$\lambda v = Av \qquad \text{---------- (2.12)}$$

Where

$\lambda$ = value of constant

$v$ = value of eigenvector centrality

A = an adjacency matrix

From the measurement of different values in a social network which have been discussed earlier, they are measurements of relation values of only one type of data, which is called measurements of relation values in the pattern of one-mode centrality. But in the real world, there may be some groups of data that comprise many categories of data. The measurement of values of relation between many types of data is called two-mode centrality. An approach to the computation of these values differs

from that the calculation in one-mode centrality pattern. Details of two-mode centrality are discussed in section 2.4.3.

### 2.4.3 Two-mode Centrality

With respect to data used for a social network, the relation is mostly measured at a small level and data are normally analysed at one level only. Likewise, those data have only one kind of actor but there are some sets of data in which each data has a wide variety of relations at several levels such as the relation between 2 kinds of actors. This pattern of data is called a two-mode data set [18].

An example of two-mode data set is in Davis et al (1941), which is data showing the relation between women and events that happened. Data are arranged in a format of matrix of relations of women and events they encountered. Data in each element of that matrix are 1 if the element in the position $e_{ij}$ has a woman i encounter the event j. Otherwise, the data in the element in the position $e_{ij}$ are 0, as reveal in Figure 2.7.

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | E1 | E1 | E1 | E1 | E1 |
| 1 | EVELYN | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | LAURA | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | THERESA | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 4 | BRENDA | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | CHARLOTTE | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | FRANCES | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | ELEANOR | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | PEARL | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 9 | RUTH | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 10 | VERNE | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 11 | MYRNA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| 12 | KATHERINE | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 13 | SYLVIA | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 14 | NORA | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 15 | HELEN | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 16 | DOROTHY | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| 17 | OLIVIA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 18 | FLORA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |

Figure 2.7 An example of a two-mode data set

#### 2.4.3.1 Degree Centrality

For the measurement of the degree centrality value of a two-mode data set, it is defined as the number of links of that nodes, the same as in a one-mode data set, which can be defined in equations (2.13) and (2.14)

$$d_i = \frac{d_i}{n_2} \quad \text{----------} \quad (2.13)$$

$$d_j = \frac{d_j}{n_1} \quad \text{----------} \quad (2.14)$$

Where

$d_i$ = the number of links of node *i* for data type 1

$d_j$ = the number of links of node *j* for data type 2

$n_1$ = the maximum number of links for data type 2

$n_2$ = the maximum number of links for data type 1

#### 2.4.3.2 Closeness Centrality

It is defined in the form of the sum of geodesic distances from an interesting node to all other nodes in a network. In the case of a two-mode data set, to calculate the closeness value, the data are divided into 2 groups based on the number of data types. The closeness centrality value of each group may be defined in equations (2.15) and (2.16).

$$c_i = \frac{n_2 + 2(n_1 - 1)}{c_i} \quad \text{----------} \quad (2.15)$$

$$c_j = \frac{n_1 + 2(n_2 - 1)}{c_j} \quad \text{----------} \quad (2.16)$$

Where

$c_i$ = the geodesic distance (number of shortest path ) from node *i* to other nodes that are node of data type 1

$c_j$ = the geodesic distance (number of shortest path ) from node $j$ to other nodes that are nodes of data type 2

$n_1$ = the maximum number of ties for data type 2

$n_2$ = the maximum number of ties for data type 1

#### 2.4.3.3 Betweenness Centrality

The value of betweenness centrality can be computed from the sum of the shortest paths that pass an interesting node. But for a two-mode data set, the betweenness centrality value can be calculated via equations (2.17) and (2.18).

$$b_i = \frac{b_i}{b_{v_1 \max}}, \text{ for } i \in V_1 \quad \text{---------- (2.17)}$$

$$b_j = \frac{b_j}{b_{v_2 \max}}, \text{ for } j \in V_2 \quad \text{---------- (2.18)}$$

Where

$b_i$ = the number of shortest path that passed node $i$ for data type 1

$b_j$ = the number of shortest path that passed node $j$ for data type 2

$b_{v2}$ = the maximum shortest path for data type 2

$b_{v1}$ = the maximum shortest path for data type 1

#### 2.4.3.4 Eigenvector Centrality

The value of eigenvector centrality may be calculated from principle that if a node has many connections, it is the node with a high value of eigenvector. The calculation of eigenvector value for two-mode data set may be made through equation (2.19).

$$\lambda v = Av \quad \text{---------- (2.19)}$$

Where

$\lambda$ = value of constant

$v$ = value of eigenvector centrality

A = an adjacency matrix

### 2.4.4 Social Network Analysis Software

Each type of Social Network Analysis software typically has different functions. But in most case, this software has such core functions as an ability to analyse, illustrate or model the shape and appearance of nodes and ties in a network. In addition some categories of softwares are able to record the consequences of the analysis in several formats of output files. Beside, some kinds of Social Network Analysis softwares may be tools in charge of data analysis only. Thus, it has to depended on other tools to assist in generating the image of relation or modeling the shape and appearance of nodes and ties [15],[16

Examples of Social Network tools include the following.

- Non – Commercial Software : PAJEK , NetDraw, JUNG etc.
- Commercial Software : UCINET, NetMiner, Dynet, INFLOW etc.

## 2.5 Related Works

Related works related to this research are divided into 2 parts. The first part includes researches in which a Social Network Analysis used to analyse various types of data, such as data from e-mail, data on personal history, a search for experts, etc. The second part consists of researches using Social Network Analysis to search a researchers network, in which data on published research works are used for analysis.

### 2.5.1 Researches Using Social Network Analysis to Analyze Various Types of Data

The consist of researches using data that are not in the format of database, such as data from e-mails, data on working record, ect to analyse by a social network

in order to search for the persons having skills, knowledge or experiences related to topics which users need. There are many researches that have made a study on this subject. Examples are given below.

Research of Campbell et al. [19] studied a search for experts within agencies from an e-mail network. They had hypothesized whether we could tell who were experts when viewing e-mail or patterns used for communications. Furthermore, whether or not an e-mail could tell the location of experts by viewing the pattern of communications between each other thorough the e-mail. In Campbell et al's works, there was a comparative experiment between 2 approaches used to search for experts. The first approach was a content-based one, in which e-mail data that were texts were used for mining. The second approach used a graph-based ranking algorithm (HITS)[20]. This algorithm applied both e-mail texts and communication patterns for the analysis so as to find experts. A comparative outcome revealed that a HITS algorithm yielded a better result than when the e-mail, only in a text part, was used in the analysis.

In Culotta et al.'s research [21], they proposed a strategy that could extract data on persons and the relation between them by viewing an e-mail which was in the inbox of each person, and seeing data derived from personal homepage. The system could generate contact information of each person. It would automatically insert contact address data, using the Condition Random Fields [22],[23] technique, which was a probabilistic model used for information extraction. Additionally, the system would analyse data to find the relation between persons using Social Network Analysis

Research of Jie Tang et al., [24],[25],[26], had studied a search for experts by using a propagation-based approach, in which both personal data and data on the relation between persons in the same network were considered. Personal data were calculated for expert score for each person and each person was arranged in order of expert score would be constructed into a sub-graph. Then, a propagation-based approach was used to create the relation between persons by considering expert score. An example of social network work of Jie Tang's research was illustrated in Figure 2.8

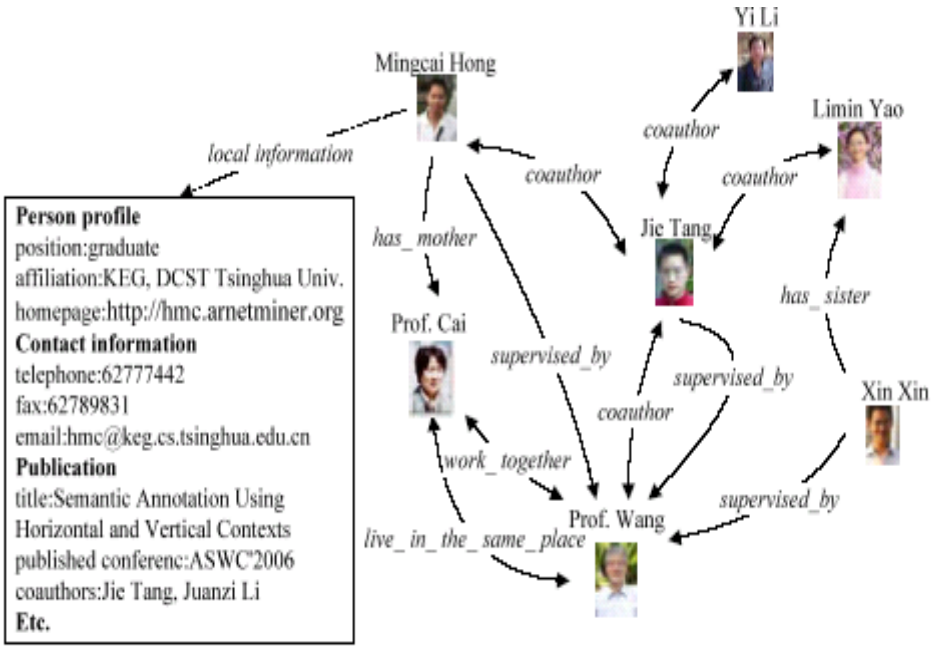



Figure 2.8 An example of an academic research network in research of Jie Tang et al. [24],[25],[26]

Balog et al.'s research [27],[28] investigated a search for experts as part of information retrieval and proposed 2 strategies for expert searching on a group of documents assigned. The first strategy was expert searching by viewing from documents in which they were involved. The second strategy was an arrangement of documents based on their topics and then finding the persons involved with those documents. An experimental result revealed that the second strategy yielded a better consequence than the first one.

In Adamic and Adar's research [29], they were interested in the problem of searching for related people by using an e-mail network. It was proposed that educational level, position inside an agency and address of each person could enhance the efficiency of a search for related people through use of e-mails.

Since 2000, the Asian region has been regarded as the biggest center of different groups of terrorists in the world. These terrorist groups have scattered in such countries as Afganistan, Pakistan, India, Bangladesh or Myanmar. After the incident of September 11,2001 when the world Trade Towers were blown up, more people have become interested in studying the network and linkage of different groups of

terrorists. In this connection, there have been researches using an approach of Social Network Analysis to find the relation and network of those terrorist groups. Examples of researches concern with a search for terrorist networks re given below.

The research of Sudhir Saxena et al. [30] used Social Network Analysis with terrorism's database, which was T2 of IDSA. T2 compiled data on terrorism from various sources, whether newspaper, TV or www. The result of applying Social Network Analysis to T2 database revealed the relation among groups of terrorists residing in Jamu and Kashmir, using Visone as an Social Network Analysis tools. An example of a social network of the research using Social Network Analysis with terrorism's database is illustrated in Figure 2.9

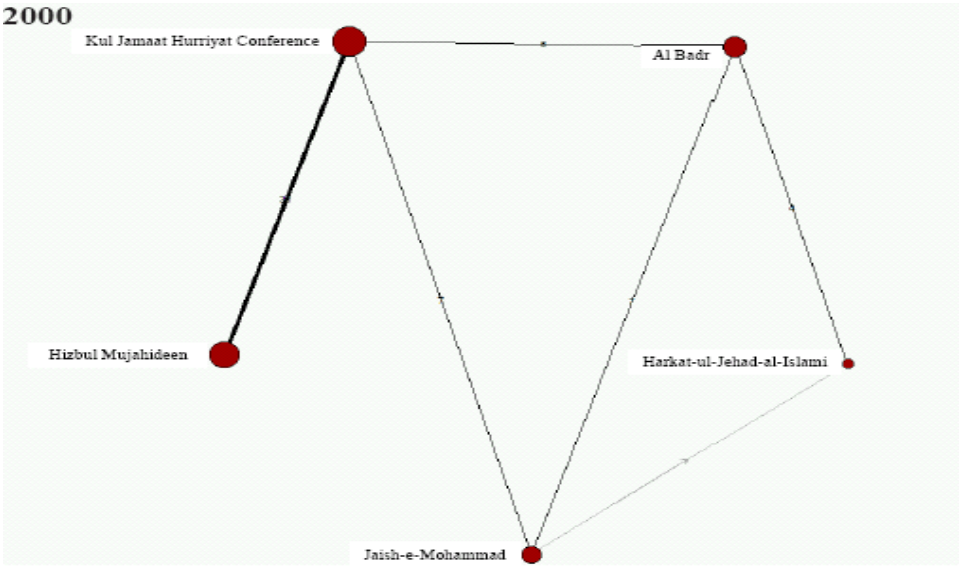



Figure 2.9 An example of terrorism network of Jammu and Kashmir

Aparna Basu's research [31] used Social Network Analysis to find the network and inter-linkage of various groups of terrorists existing in Asia, including the relation among terrorist groups in other regions across the world, focusing mainly on the groups carrying out acts of violence in India. In his research, Basu used the terrorist database of IDSA between 2001-2003. the research scrutinized the betweenness value

for the analysis of relation. The outcome of this research revealed which terrorist group was a huge one and played the maximum role in the Asian region.

### 2.5.2 Researches Using Social Network Analysis to Analyze Data on Published Research Works

Research of Jie Tang et al.[32] was the research concentrating on searching for experts. Data used to analyse to find experts were those on research works from data sources on the internet. Then, these data were constructed into a social network displaying the relation between the researchers and his/her co-workers. Subsequently, the data derived from the social network were used to find experts, utilizing the relevancy propagation-based expert finding algorithm, which functioned in 2 stages. In the first stage, the value of relevancy between each researcher and research topic was computed using each researcher's data. In the second stage, the value of topic relevancy between each researcher and other researchers was calculated.

Another research on a search for a researcher network, applying Social Network Analysis, is that of Duo Zhang et al.[33], which focused on a study of researcher data extraction from data sources on world wide web (www), using the CRF (Condition Random Field) technique. Subsequently, the extracted data were combined with bibliography data of research works obtained from the DBLP database. Later, the constraint was defined to be used in the constrain-based probabilistic model for a solution to the problem of repetitious or similar names of researchers.

Takeru Miki et al.'s research [34] utilized bibliographical citation data from CiteSeer's database and data on world wide web for the analysis to find the relation between researchers through defining the ontology, which defined several types of relation character among researchers, for example, research collaborator, academic committee, educational relationship, etc. a study outcome reveals a research network showing, in groups of different research areas, which researchers work in those areas. N addition, it also reveals a co-author network.

Finally, the research of Guillermo Rueda et al.[35] revealed the extraction of bibliometrics data, which was a technique of capturing text and data in academic journal citation. The data received were analysed via Social Network Analysis. In this research, academic journal data in the field of Nanotechnology, between 1992-2006

were utilized. The consequences enable us to know of various statistics such as those having a large number of researches on Nanotechnology, and the countries where their people study Nanotechnology. Besides, the results allow us to know about the relation among researchers in this area and who play a significant role in Nanotechnology.

# CHAPTHER III
# SYSTEM DESIGN

This chapter discusses the design of a system to be developed for the analysis of the relation of data related to researchers. This chapter consists of 4 parts. The first part is the system architecture. The second part is the preparation of data on researchers. The third part is the analysis of the relation among various parts of data. Finally, the forth part is directly related to user which is a search for relation data. Details of each part are described below.

## 3.1 System Architecture

Figure 3.1 illustrates all components of the system to be developed

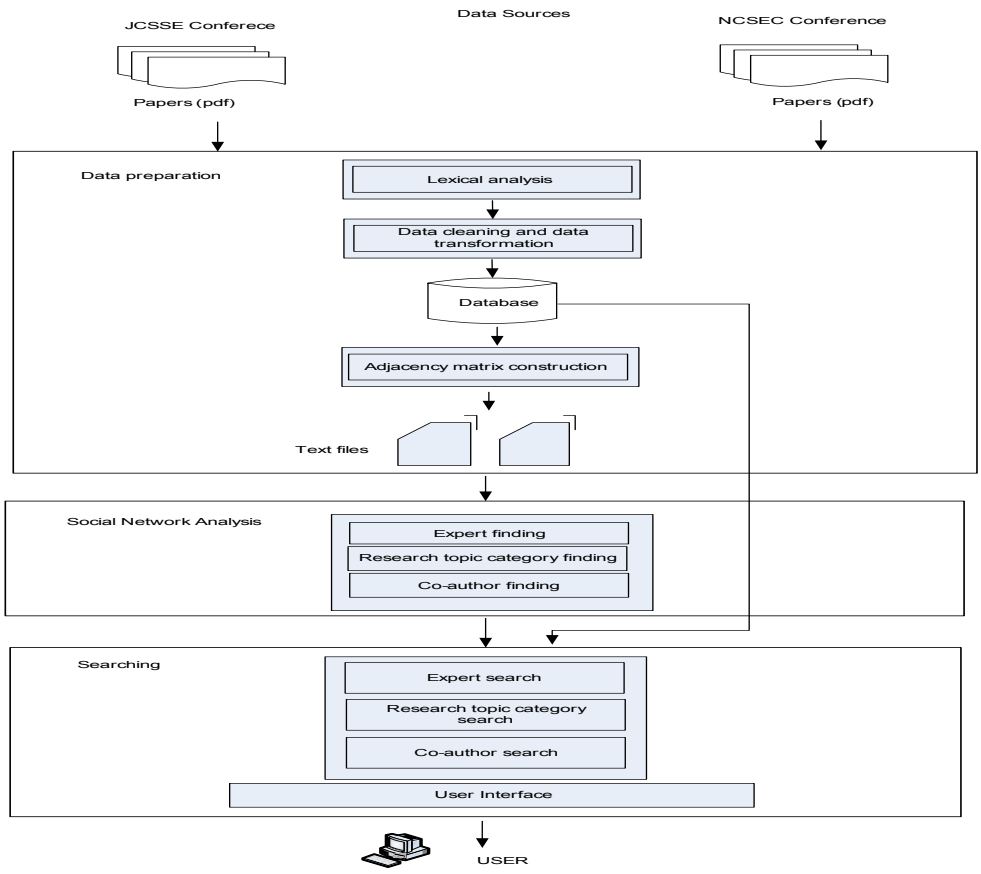



Figure 3.1 System Architecture

Figure 3.1 reveals that the prototype system to be developed comprises 3 majors components, namely, data preparation, Social Network Analysis and searching. Details of each component are given in sections 3.2, 3.3 and 3.4, respectively.

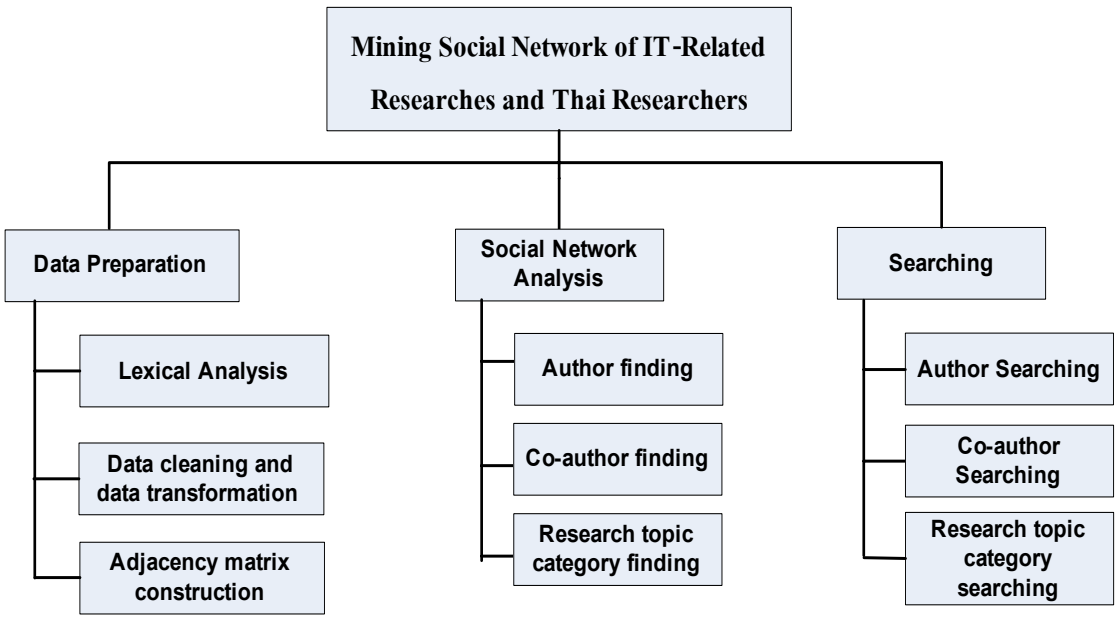



Figure 3.2 The Structure Chart of the system

Figure 3.2 displays that the prototype system is composed of 3 functional parts. The first part is data preparation so that data are in a format appropriate for analysis through Social Network Analysis. The second part is data analysis via Social Network Analysis. Finally, the third part is a search for data analysed by Social Network Analysis.

## 3.2 Data Preparation

Data preparation is a stage of data compilation and preparation for subsequent analysis. This section discusses data sources used in the research, including stages of data preparation process consisting of lexical analysis, data transformation and adjacency matrix construction.

### 3.2.1 Data Sources

Data sources used in this research are research data disseminated in conferences held in Thailand, for instance, The Joint Conference on Computer Science and Software Engineering (JCSSE) and The National Computer Science and Engineering Conference (NCSEC). The research data are in a pdf file format.

For research to be submitted to the JCSSE conference in 2007, there were classified into 10 categories of research altogether, as shown in Figure 3.3 and The NCSEC conference in the same year, there were classified research works into 7 categories as shown in Figure 3.4.


**The 4th Joint conference on Computer Science and Software Engineering**

**May 2nd – 4th, 2007**

JCSSE is an international conference held annually in Thailand. It provides a central forum for experts and developers to promote, share, and discuss various issues and developments in the broad field of computer science and software engineering. JCSSE will provide an opportunity for young researchers to demonstrate their talents and interesting research ideas. The conference will benefit people who are actively involved in research related to computer science, computer engineering, software engineering and digital library in Thailand.

**1.Computational Science and Engineering**
- 1.1 Bioinformatics
- 1.2 DNA Computing
- 1.3 Molecular Computing
- 1.4 Nano Technology
- 1.5 Numerical Methods

**2.Computer Intelligence**
- 2.1 Artificial Intelligence
- 2.2 Evolutionary Computing
- 2.3 Fuzzy and Neural Computing
- 2.4 Genetic Algorithms
- 2.5 Intelligent Agents
- 2.6 Machine Intelligence and Applications
- 2.7 Natural Language Processing
- 2.8 Neural Networks
- 2.9 Robotics and Automation

**3.Computer Graphics**
- 3.1 Computer Vision
- 3.2 Multimedia Technology
- 3.3 Visualization

**4.Computer Systems**
- 4.1 Computer Architecture
- 4.2 Computer Simulation and Modeling
- 4.3 Computer Security
- 4.4 Embedded Control Systems
- 4.5 Fault-Tolerant Computing
- 4.6 High Performance Computing
- 4.7 Human-Computer Interface
- 4.8 Optimization
- 4.9 Parallel and Distributed Computing
- 4.10 VLSI Design and Applications

**5.Computer Networks**
- 5.1 Computer Networks
- 5.2 Communication and QoS
- 5.3 Data Communications
- 5.4 Mobile Computing

**6.Information Technology**
- 6.1 e-Business Technologies and Applications
- 6.2 Information & Communication Technologies
- 6.3 Internet Technology and Applications
- 6.4 Semantic Web & Ontology
- 6.5 Web Services

**7.Computer Software**
- 7.1 Computer Aided Instruction
- 7.2 Data Mining
- 7.3 Database and Information Retrieval
- 7.4 Decision Support Systems
- 7.5 Distributed Database Systems
- 7.6 Image and Signal Processing
- 7.7 Geographical Information Systems
- 7.8 Object Oriented Computing
- 7.9 Programming Languages and Compilers

**8.Digital Library**
- 8.1 Multimedia DLDB Engine
- 8.2 Circulation Management
- 8.3 Automatic Resource Category
- 8.4 Digital Archive Management
- 8.5 Digital Right Management (DRM)
- 8.6 Automatic Indexing
- 8.7 Smart Search Engine
- 8.8 Automatic Collection Tools
- 8.9 Ontology
- 8.10 DL Architecture
- 8.11 User Interface Service: speech, content delivery

**9.Software Engineering**
- 9.1 Formal Methods
- 9.2 Programming Techniques
- 9.3 Software Requirements and Specifications
- 9.4 System Modeling
- 9.5 Software Components
- 9.6 Software Prototyping
- 9.7 Software Designs
- 9.8 Software Validation
- 9.9 Software Management
- 9.10 Software Testing

**10.Software Application & Development**
- 10.1 Tools & Products
- 10.2 Experiences, Surveys & Reports
- 10.3 Education & Training

Figure 3.3 Call for paper for JCSSE 2007

## The 11th National Computer Science and Engineering Conference

## (NCSEC 2007) Call for Papers

NCSEC is one of the most successful conferences held annually in Thailand. It provides a central forum for experts to promote, share, and discuss various issues and developments in the broad field of information and computer technologies. NCSEC will provide an opportunity for young researchers to demonstrate their talent and interesting research ideas. The conference will benefit people who are actively involved in research related to computer science and engineering in Thailand.

NCSEC2007 invites submissions from all areas related to, but not limited to,

1. **Computer Systems**
   - Computer System Design
   - Computer Simulation & Modeling
   - Embedded Systems
   - Fault Tolerant Computing
   - High Performanace Computing
   - VLSI Design & Application
   - Optimization
2. **Artificial Intelligence**
   - Intelligent Agents
   - Neural Networks
   - Pattern Recognition
   - Decision support System
   - Evolutionary Computation
   - Fuzzy Logics
   - Natural Language Processing
   - Robotics & Automation
3. **Computer Software**
   - Multimedia & Animation
   - Human Computer Interaction
   - Object Oriented Programming
   - Programming Language & Compilers
   - Software Engineering
4. **Computer Networks**
   - Data Communication
   - Computer Security
   - Mobile Computing
   - Wireless & Ad hoc Networks
   - Internetworking
   - Quality-of-Service Routing
   - Sensor Networks
5. **DSP and Image Processing**
   - Image & Signal Processing
   - Speech & Audio Processing
   - Biomedical Signal Processing
6. **Information Technology**
   - Data Mining
   - Database & Information Retrieval
   - MIS/AIS/DSS/GIS Application
   - SOA & Web Service
   - Bioinformatics
   - e-Learning
   - Knowledge Management
   - Cyber Security
7. **Computing Theory**
   - Formal Methods
   - Numerical Methods
   - Computation Geometry

Figure 3.4 Call for paper for NCSEC 2007

Figure 3.3 and 3.4 reveal that the determination of research categories for JCSSE 2007 and NCSEC 2007 consists of both similar and different categories defined. Accordingly, when data coming from both conferences have to be analysed, it is necessary to manually define new research categories. The newly defined research categories will use those submitted to the JCSSE and NCSEC 2007 as a prototype for classifying new research categories. These new categories will cover both research categories of JCSSE 2007 and NCSEC 2007. Research categories newly defined comprise 11 categories shown in Table 3.1.

Table 3.1 Newly defined paper categories

| NCSEC | JCSSE | New defined category |
|---|---|---|
| Computer Systems | Computer Systems | Computer Systems |
| Artificial Intelligence | Artificial Intelligence | Artificial Intelligence |
| Computer Networks | Computer Networks | Computer Networks |
| Computer Software | Computer Software | Computer Software |
| DSP and Image Processing | Computational Science and Engineering | Software Engineering |
| Information Technology | Computer Graphics and Computer Animation | Computer Graphics and Computer Animation |
| Computing Theory | Information and Communication Technology | Computational Science and Engineering |
| | Digital Library | Information Technology |
| | Software Engineering | DSP and Image Processing |
| | Software Application and Development | Software Application and Development |
| | | Digital Library |

For initial data, research data submitted to conferences namely, JCSSE and NCSEC from 2005 to 2008 will be used. Table 3.2 displays the number of overall research data used in this research while Table 3.3 reveals data summarizing the number of researches classified by research topic categories for each year (2005-2008).

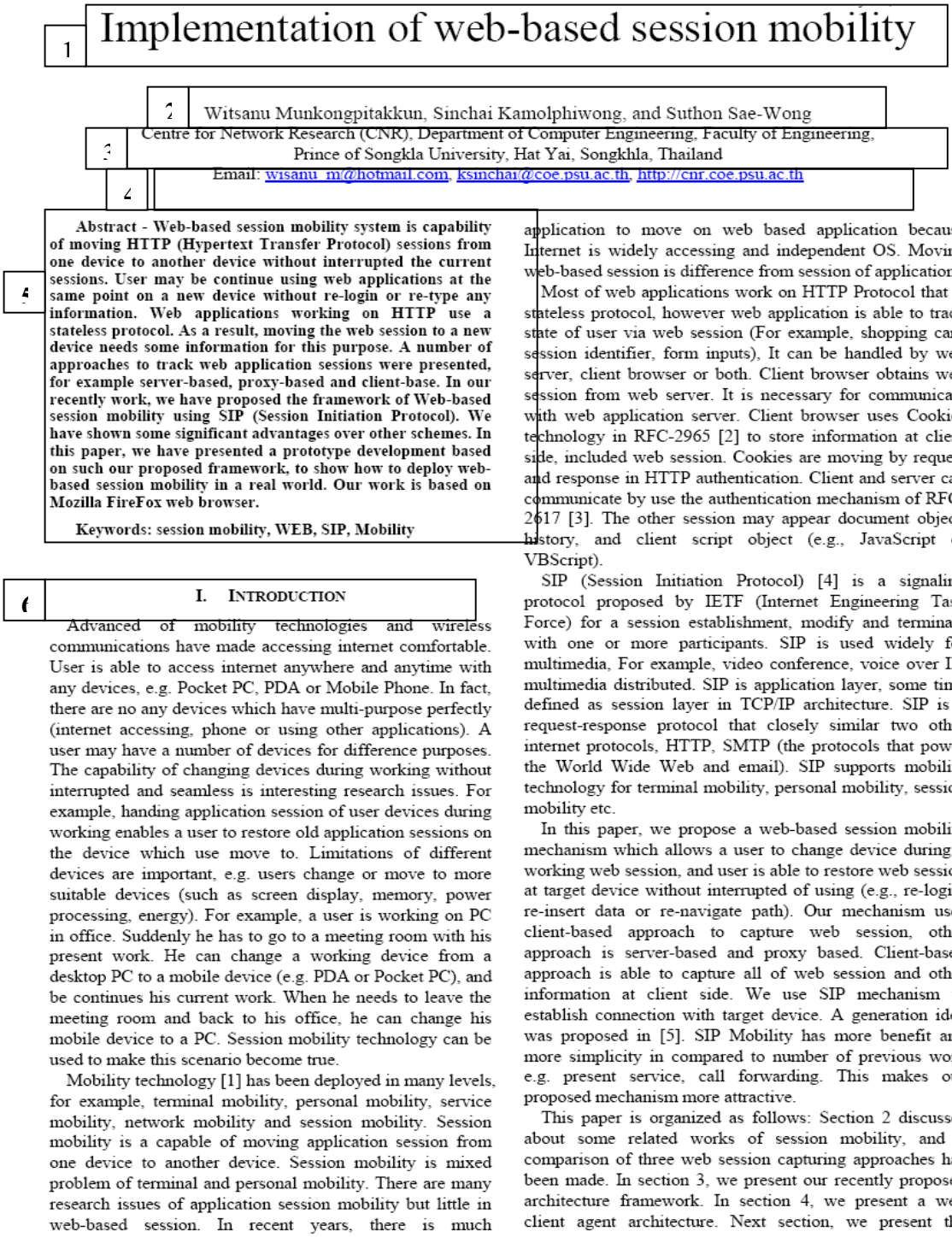
Table 3.2 A summary of overall initial data used in this research

| **Characteristics** | **JCSSE** | | | | | **NCSEC** | | | | | **Grand Total** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2005 | 2006 | 2007 | 2008 | total | 2005 | 2006 | 2007 | 2008 | total | |
| No. of all research (papers) | 53 | 39 | 108 | 120 | 320 | 35 | 79 | 98 | 90 | 302 | 622 |
| No. of researches used (papers) | 46 | 33 | 102 | 102 | 283 | 31 | 70 | 89 | 89 | 279 | 562 |
| No. of researches in National session (papers) | 39 | 25 | 51 | 64 | 179 | 22 | 47 | 58 | 48 | 175 | 354 |
| No. of researches in International session (papers) | 14 | 14 | 57 | 56 | 141 | 13 | 32 | 40 | 42 | 127 | 268 |
| No. of all researchers (persons) | 89 | 71 | 226 | 215 | 601 | 63 | 101 | 128 | 136 | 428 | 1029 |
| No. of Thai researchers (persons) | 81 | 65 | 212 | 201 | 559 | 58 | 94 | 113 | 124 | 389 | 948 |
| No. of foreign researchers (persons) | 8 | 6 | 18 | 14 | 46 | 5 | 7 | 15 | 12 | 39 | 85 |
| No. of all educational institutes/Agencies | 19 | 25 | 39 | 45 | 128 | 18 | 31 | 38 | 47 | 134 | 262 |

Table 3.3 Summarized data of number of research papers classified by research topic categories for each year

| No | category | 2005 | 2006 | 2007 | 2008 | total |
|---|---|---|---|---|---|---|
| 1. | Artificial Intelligence | 25 | 29 | 43 | 20 | 117 |
| 2. | Computer Graphics and Computer Animation | 4 | 2 | 7 | 10 | 23 |
| 3. | Computer Networks | 10 | 5 | 20 | 20 | 55 |
| 4. | Computational Science and Engineering | 5 | 4 | 15 | 14 | 38 |
| 5. | Computer Software | 3 | 0 | 2 | 4 | 9 |
| 6. | Computer Systems | 5 | 13 | 16 | 16 | 50 |
| 7. | Digital Library | 0 | 1 | 0 | 2 | 3 |
| 8. | Digital Signal Processing and Image Processing | 9 | 10 | 14 | 26 | 59 |
| 9. | Information and Communication Technology | 16 | 29 | 49 | 57 | 151 |
| 10. | Software Engineering | 0 | 9 | 20 | 17 | 46 |
| 11. | Software Application and Development | 0 | 1 | 5 | 4 | 10 |

These research papers will be downloaded in advance. The files will be retrieved for needed data later. Example of research to be used as initial data for this experiment are given in Figure 3.5 and 3.6.



# Implementation of web-based session mobility

Witsanu Munkongpitakkun, Sinchai Kamolphiwong, and Suthon Sae-Wong
Centre for Network Research (CNR), Department of Computer Engineering, Faculty of Engineering,
Prince of Songkla University, Hat Yai, Songkhla, Thailand
Email: wisanu_m@hotmail.com, ksinchai@coe.psu.ac.th, http://cnr.coe.psu.ac.th

**Abstract - Web-based session mobility system is capability of moving HTTP (Hypertext Transfer Protocol) sessions from one device to another device without interrupted the current sessions. User may be continue using web applications at the same point on a new device without re-login or re-type any information. Web applications working on HTTP use a stateless protocol. As a result, moving the web session to a new device needs some information for this purpose. A number of approaches to track web application sessions were presented, for example server-based, proxy-based and client-base. In our recently work, we have proposed the framework of Web-based session mobility using SIP (Session Initiation Protocol). We have shown some significant advantages over other schemes. In this paper, we have presented a prototype development based on such our proposed framework, to show how to deploy web-based session mobility in a real world. Our work is based on Mozilla FireFox web browser.**

**Keywords: session mobility, WEB, SIP, Mobility**

## I. INTRODUCTION

Advanced of mobility technologies and wireless communications have made accessing internet comfortable. User is able to access internet anywhere and anytime with any devices, e.g. Pocket PC, PDA or Mobile Phone. In fact, there are no any devices which have multi-purpose perfectly (internet accessing, phone or using other applications). A user may have a number of devices for difference purposes. The capability of changing devices during working without interrupted and seamless is interesting research issues. For example, handing application session of user devices during working enables a user to restore old application sessions on the device which use move to. Limitations of different devices are important, e.g. users change or move to more suitable devices (such as screen display, memory, power processing, energy). For example, a user is working on PC in office. Suddenly he has to go to a meeting room with his present work. He can change a working device from a desktop PC to a mobile device (e.g. PDA or Pocket PC), and be continues his current work. When he needs to leave the meeting room and back to his office, he can change his mobile device to a PC. Session mobility technology can be used to make this scenario become true.

Mobility technology [1] has been deployed in many levels, for example, terminal mobility, personal mobility, service mobility, network mobility and session mobility. Session mobility is a capable of moving application session from one device to another device. Session mobility is mixed problem of terminal and personal mobility. There are many research issues of application session mobility but little in web-based session. In recent years, there is much application to move on web based application because Internet is widely accessing and independent OS. Moving web-based session is difference from session of application.

Most of web applications work on HTTP Protocol that is stateless protocol, however web application is able to track state of user via web session (For example, shopping cart, session identifier, form inputs), It can be handled by web server, client browser or both. Client browser obtains web session from web server. It is necessary for communicate with web application server. Client browser uses Cookies technology in RFC-2965 [2] to store information at client side, included web session. Cookies are moving by request and response in HTTP authentication. Client and server can communicate by use the authentication mechanism of RFC-2617 [3]. The other session may appear document object, history, and client script object (e.g., JavaScript or VBScript).

SIP (Session Initiation Protocol) [4] is a signaling protocol proposed by IETF (Internet Engineering Task Force) for a session establishment, modify and terminate with one or more participants. SIP is used widely for multimedia, For example, video conference, voice over IP, multimedia distributed. SIP is application layer, some time defined as session layer in TCP/IP architecture. SIP is a request-response protocol that closely similar two other internet protocols, HTTP, SMTP (the protocols that power the World Wide Web and email). SIP supports mobility technology for terminal mobility, personal mobility, session mobility etc.

In this paper, we propose a web-based session mobility mechanism which allows a user to change device during a working web session, and user is able to restore web session at target device without interrupted of using (e.g., re-login, re-insert data or re-navigate path). Our mechanism uses client-based approach to capture web session, other approach is server-based and proxy based. Client-based approach is able to capture all of web session and other information at client side. We use SIP mechanism to establish connection with target device. A generation idea was proposed in [5]. SIP Mobility has more benefit and more simplicity in compared to number of previous work e.g. present service, call forwarding. This makes our proposed mechanism more attractive.

This paper is organized as follows: Section 2 discusses about some related works of session mobility, and a comparison of three web session capturing approaches has been made. In section 3, we present our recently proposed architecture framework. In section 4, we present a web client agent architecture. Next section, we present the

Figure 3.5 An example of PDF document in JCSSE conference

**Elementary Discourse Unit Segmentation for Thai Using Discourse Cues and Syntactic Information**

Jirawan Charoensuk, Thana Sukvakree, Asanee Kawtrakul
The Specially Research Unit of Natural Language Processing and
Intelligent Information System Technology,
Department of Computer Engineering, Faculty of Engineering, Kasetsart University
Bangkok, 10900, Thailand
Email: jirawan@src.ku.ac.th, thana@hotmail.com, ak@vivaldi.cpe.ku.ac.th

**Abstract**

*We present an approach to segment full text into minimal discourse units that are clause or clause-like units, which are called elementary discourse unit (EDU). The main cause of EDU boundary ambiguity is that Thai has no punctuation marks or special symbol to signal EDU boundary and embedded EDU that occurs in the middle of another EDU. This paper presents an approach for Thai EDU segmentation by using decision tree learning. The precision and recall of the system are 0.82 and 0.78*

**Key Words:** elementary discourse unit, discourse cue, syntactic information, EDU boundary ambiguity

**1. Introduction**

Discourse segmentation is a part of discourse processing which separates full text into discourse units. The minimal discourse unit that was produced from discourse segmentation process is called elementary discourse unit (EDU) [1,2]. Mann and Thompson [13] defined boundary of EDU as clause or clause-like units. Many applications, such as text summarization [1], discourse parsing [7,8] and machine translations, usually use EDUs as an input because sentences might be long discourse segments for these applications. For example, if we derive discourse structure of text using Rhetorical structure theory [13] for text summarization, we will need to construct EDUs that are determined rhetorical relations to understanding semantics of this text. If we select the most important units of text to be summarization, we select these units from EDUs-lists that are clause-like units. Therefore, correct and precise EDUs segmentation process is a significant process for text summarization.

EDUs segmentation has been claimed that boundaries of EDUs has been segmented using discourse cues [1,2,5] punctuation marks [2,5] , and syntactic information [2,10,8]. Two approaches for EDUs segmentation are rule-based and machine learning approaches.

Rule-based approach defines EDU segmentation rules using discourse cues, punctuation marks [2] and syntactic information [8]. Daniel Marcu [2] developed shallow parser that uses cue phrase and punctuation marks to delimit EDU boundaries. However, this parser cannot identify EDU boundaries in complex sentence correctly. Livia Polanyi [7,8] defined segmentation rules by using syntactic information from parser in sentence level. This approach needs high efficiency parser to identify EDU boundaries. Moreover, the efficiency of this approach is depended on a set of segmentation rules that was defined manually.

Machine learning approach defines features to identify EDU boundaries using Part of Speech (POS), discourse cues, punctuation marks [1], and syntactic tree [2,10] defined features using discourse cues and POS tags that consist of windows of size 5. Marcu used the C4.5 program in order to learn decision trees for EDU segmentation. Radu Soricut [10] used syntactic tree from Penn Treebank and Charniak parser and POS tags as features. He used probability from the position of a word in syntactic tree and WORD POS to identify EDU boundaries. The efficiency of this approach is mainly depended on learning features and corpus sizes.

However, all of this system can not be directly applied to Thai language, since Thai EDU segmentation is more complex than other languages. The following are causes of this problem.

- Thai does not have punctuation marks, such as full stop (“.”), comma (“,”), semicolon (“;”) to identify EDUs boundaries.

Figure 3.6 An example of PDF document in NCSEC conference

Based on Figure 3.5, data in a pdf file are divided into 6 following part.

(1) Title

(2) Authors

(3) Address

(4) E-mail

(5) Abstract

(6) Paper Keyword

A system administrator will deal with the data on researcher papers and researches in a PDF document file through the system, which extracts these data by using PDFBOX, which is a tool for transforming a PDF document into a text file format. Figure 3.7 shows various stages of the data preparation process, which starts with reading data that is a PDF document and then transforming it into a text file format using PDFBOX. Subsequently, data in a text file format are put thorough the process of lexical analysis and data transformation, after which data are stored in the database.

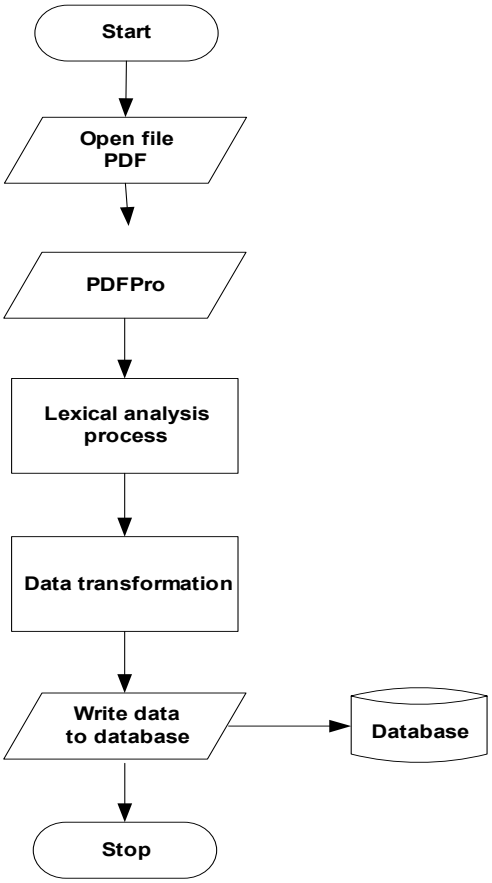
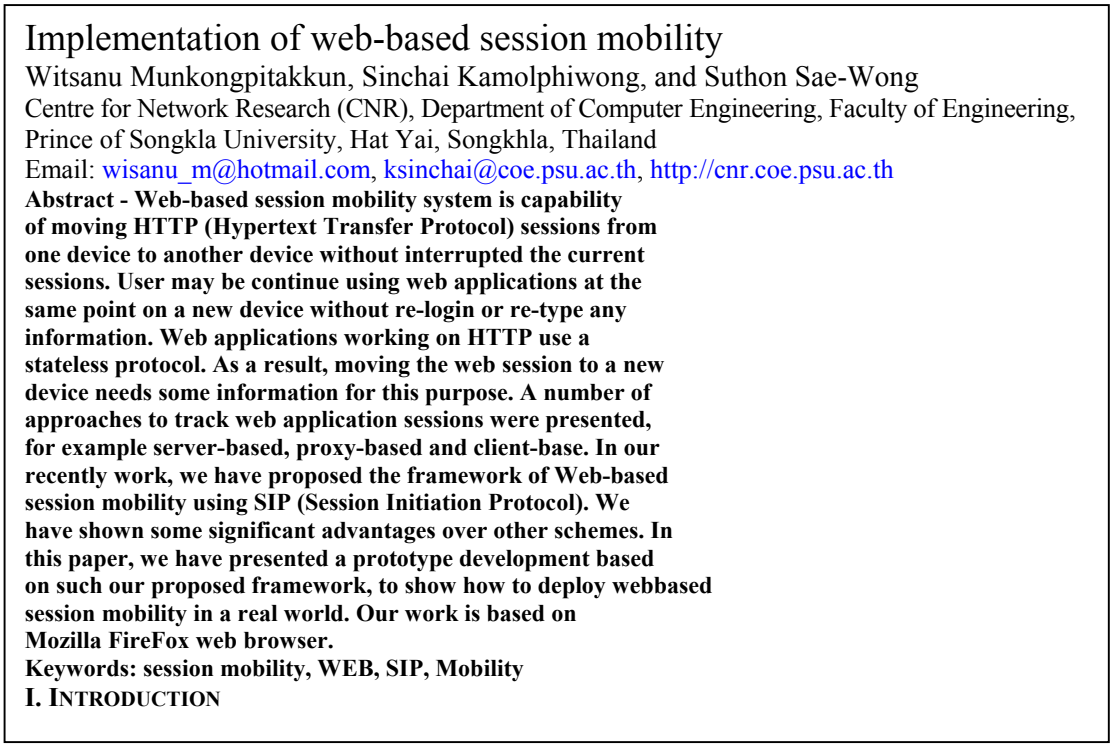



Figure 3.7 Data Preparation process

### 3.2.2 Lexical Analysis

It is a technique of separating texts from documents in the pattern of character sets. The system will read data in a pdf file and use a tool, or PDFBOX, to transform the data from a pdf format into a text file then separate text to token by used

white space. In index term selection, tokens will be selected to representing for documents.



Implementation of web-based session mobility

Witsanu Munkongpitakkun, Sinchai Kamolphiwong, and Suthon Sae-Wong

Centre for Network Research (CNR), Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University, Hat Yai, Songkhla, Thailand

Email: wisanu_m@hotmail.com, ksinchai@coe.psu.ac.th, http://cnr.coe.psu.ac.th

**Abstract - Web-based session mobility system is capability of moving HTTP (Hypertext Transfer Protocol) sessions from one device to another device without interrupted the current sessions. User may be continue using web applications at the same point on a new device without re-login or re-type any information. Web applications working on HTTP use a stateless protocol. As a result, moving the web session to a new device needs some information for this purpose. A number of approaches to track web application sessions were presented, for example server-based, proxy-based and client-base. In our recently work, we have proposed the framework of Web-based session mobility using SIP (Session Initiation Protocol). We have shown some significant advantages over other schemes. In this paper, we have presented a prototype development based on such our proposed framework, to show how to deploy webbased session mobility in a real world. Our work is based on Mozilla FireFox web browser.**

**Keywords: session mobility, WEB, SIP, Mobility**

**I. INTRODUCTION**

Figure 3.8 An example of text file from pdf file

From Figure 3.8, data in a pdf file which are converted into a text file are divided into 6 following parts.

(1) Title

(2) Authors

(3) Address

(4) E-mail

(5) Abstract

(6) Paper keyword

Based on a text file shown in Figure 3.8, each part of data can be separated as follows:

(1) Title : The data read on the first line are the title which is

"Implementation of web-based session mobility"

(2) Authors : the data read on the following line are the authors' names, which is

"Witsanu Munkongpitakkun, Sinchai Kamolphiwong, and Suthon Sae-Wong"

(3) Address : the data read on the next line are the authors's address, which is

"Centre for Network Research (CNR), Department of Computer Engineering, Faculty of Engineering,Prince of Songkla University, Hat Yai, Songkhla, Thailand"

(4) E-mail : next are the author's e-mail, which is

"wisanu_m@hotmail.com, ksinchai@coe.psu.ac.th"

(5) Abstract: the abstract depend on the keyword "abstract" to indicate that data after this word are abstract, which is

"Web-based session mobility system is capability of moving HTTP (Hypertext Transfer Protocol) sessions from one device to another device without interrupted the current sessions. User may be continue using web applications at the same point on a new device without re-login or re-type any information. Web applications working on HTTP use a stateless protocol. As a result, moving the web session to a new device needs some information for this purpose. A number of approaches to track web application sessions were presented, for example server-based, proxy-based and client-base. In our recently work, we have proposed the framework of Web-based session mobility using SIP (Session Initiation Protocol). We have shown some significant advantages over other schemes. In this paper, we have presented a prototype development based on such our proposed framework, to show how to deploy web based session mobility in a real world. Our work is based on Mozilla FireFox web browser."

(6) Paper keywords: data are segmented using the word "keywords". Word after this word are keywords that authors have defined. From the example, keywords are the following.

"session mobility, WEB, SIP, Mobility"

Based on Figure 3.9, which explains a lexical analysis process and index term selection for each document. All 6 parts of data are stored in a temporary table first. Then data on the tile and abstract of each document are read and segmented into token,

using white space as well as examining the obtained token with a list of stopwords. If the token or that word is found in the list of stopwords, then skip that token and simultaneously check each token against the term already existing in the database. if the term is existing in the database, then count the frequency of an additional tokens. Repeat doing this until end of text file. When the token or words which are stopwords have been segmented, keywords representing that document will be gained, which are called **Index terms**.

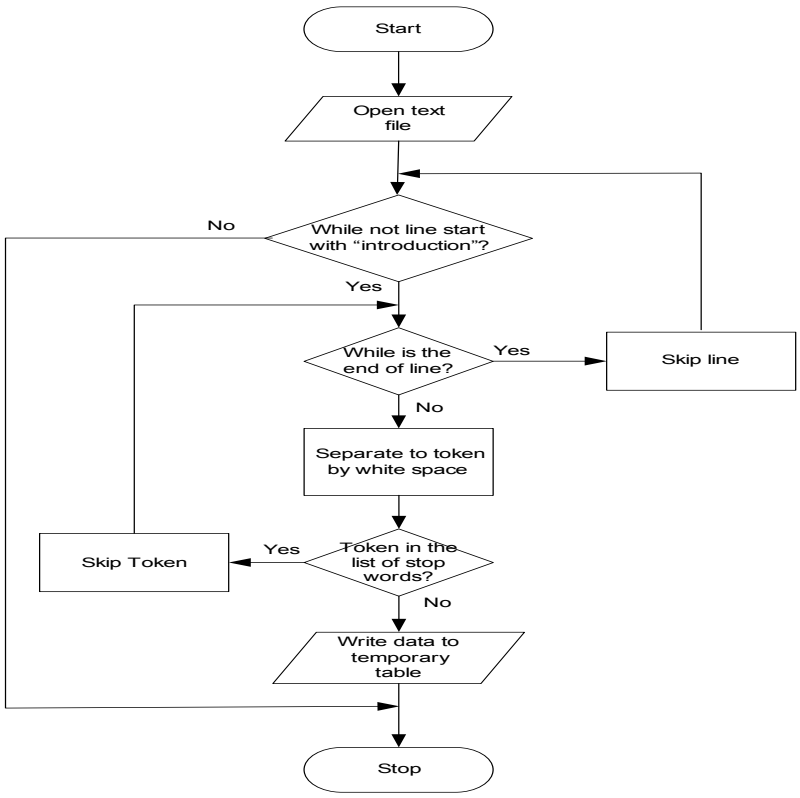



Figure 3.9 Processes of Lexical Analysis and Index term selection

### 3.2.3 Data Transformation

When data on research documents have been stored in the system's database, but since data are in different formats, it is necessary to transform these data first so that they are in the same format and appropriate for analysis and searching.

For data transformation, names of departments, faculties, universities and provinces are adjusted to be in the same format for data containing names of the same department, faculty, university and province. An example of data transformation is

given in Figure 3.10 while the data transformation process is illustrated in Figure 3.11 and Table 3.4 shown an example of transformation rules for this system.

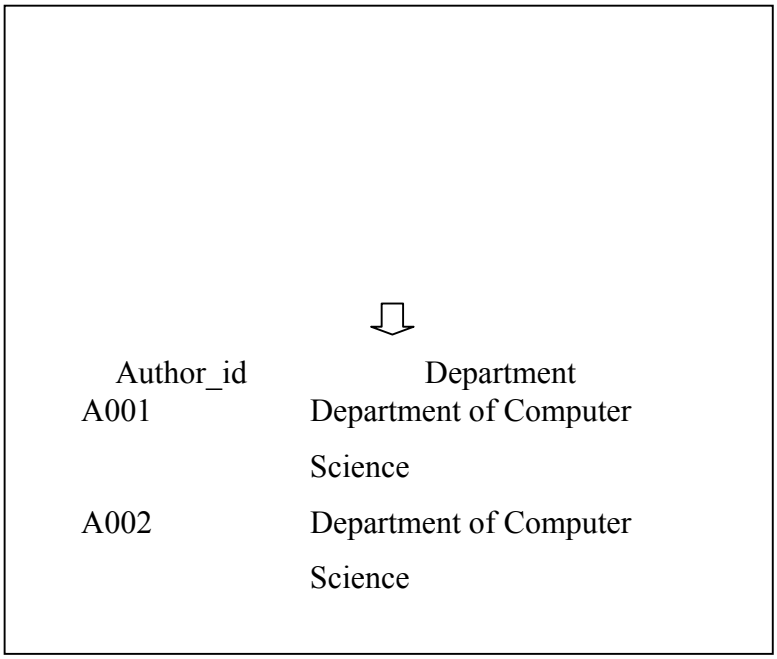



Figure 3.10 An example of data transformation

Table 3.4 An example of transformation rules in this system

| Transformation | Example of Input | Example of Output |
|---|---|---|
| University | มหาวิทยาลัยธรรมศาสตร์ | Thammasat University |
| Faculty | คณะวิทยาศาสตร์ | Faculty of Science |
| Department | ภาควิชาวิทยาการคอมพิวเตอร์ | Department of Computer Science |
| Province | กรุงเทพมหานคร | Bangkok |
| Organization | ศูนย์พันธุวิศวกรรมแห่งชาติ | BIOTEC |

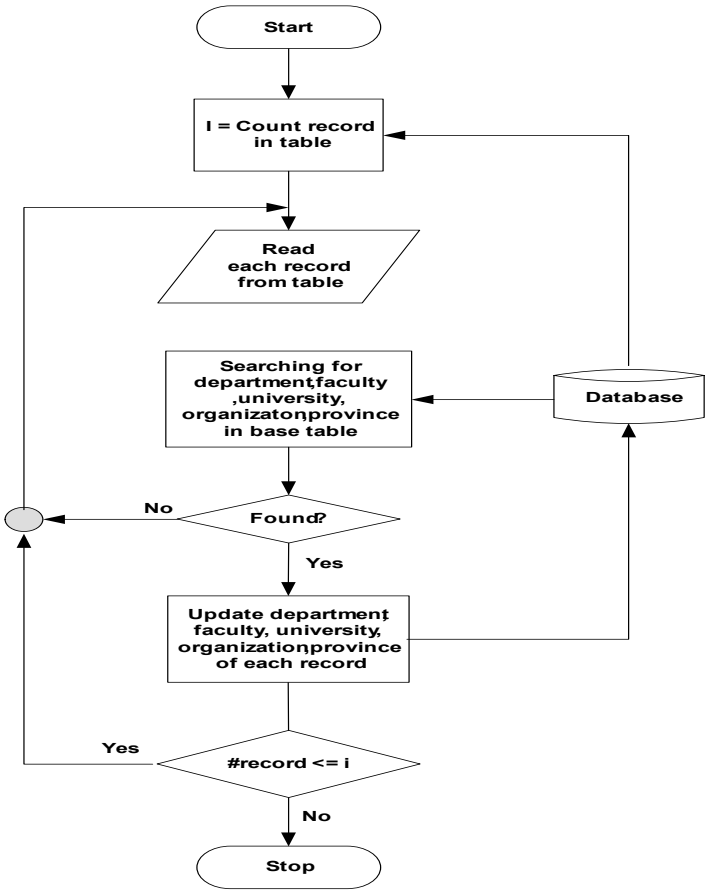



Figure 3.11 Data Transformation process

### 3.2.4 Database Design

There are altogether 13 tables of the database for storing of data to be used in the analysis of various types of relations between researchers and research papers. All tables are listed below.

(1) paper
(2) author
(3) conference
(4) email
(5) category
(6) sub_category
(7) author_defined_keyword
(8) title_keyword
(9) abstract_keyword
(10) base_university
(11) base_faculty
(12) base_department
(13) base_organization

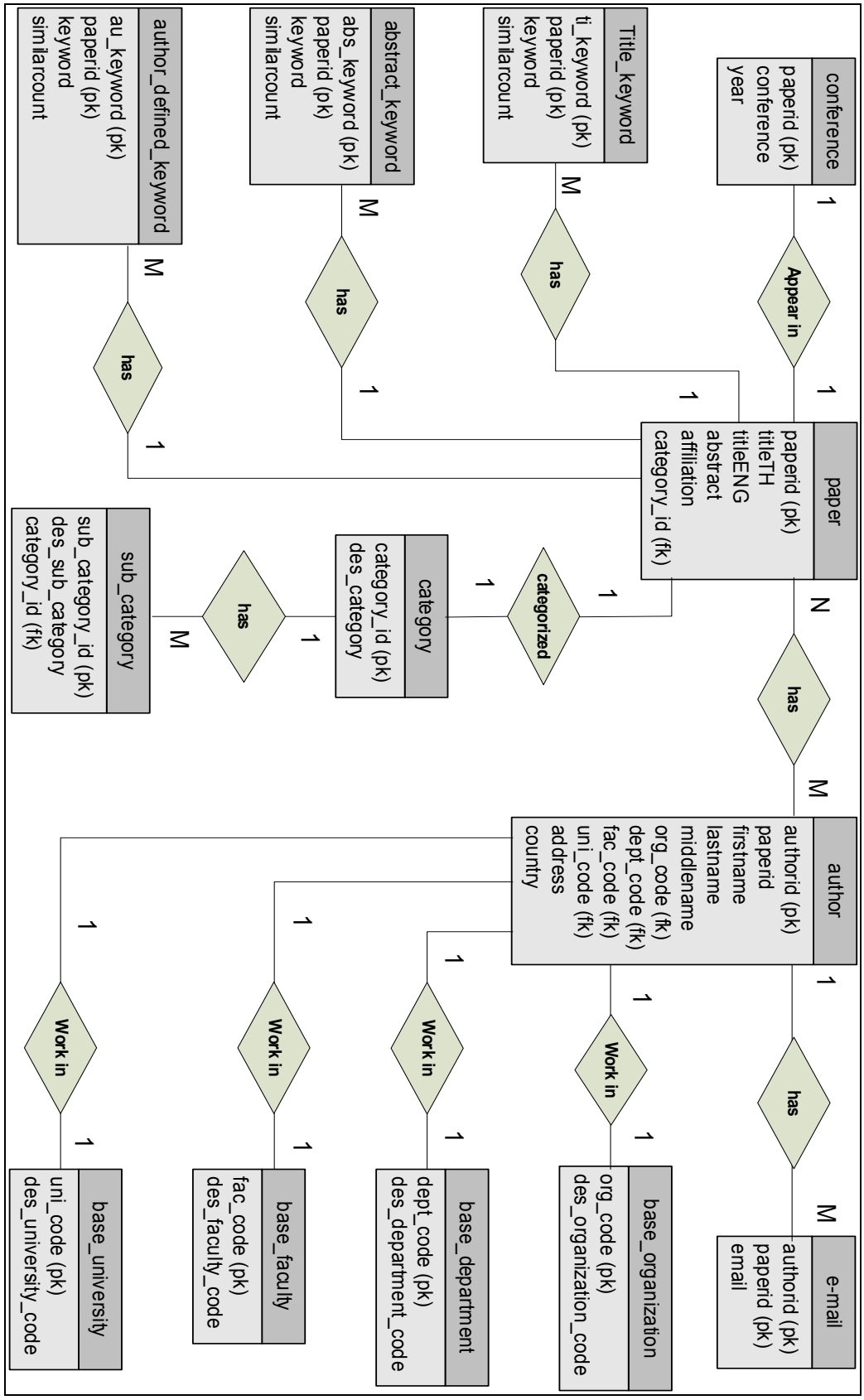



Figure 3.12 An E-R diagram of the system database

Table 3.5 Data structure of table : paper

| **Table name** | paper | | | |
|---|---|---|---|---|
| **Description** | Store data of paper | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | paperid | Paper id | Char | PK |
| 2. | titleTH | Title in thai | Char | |
| 3. | titleENG | Title in English | Char | |
| 4. | abstract | Abstract of paper | Char | |
| 5. | affiliation | Organization | Char | |
| 6. | category_id | Category id | Char | FK |

Table 3.6 Data structure of table : author

| **Table name** | author | | | |
|---|---|---|---|---|
| **Description** | Store detail data of author | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | authorid | Author id | Char | PK |
| 2. | paperid | Paper id | Char | |
| 3. | firstname | First name of author | Char | |
| 4. | lastname | Last name of author | Char | |
| 5. | middlename | Middle name of author | Char | |
| 6. | org_code | Organization of author | Char | FK |
| 7. | dept_code | Department of author | Char | FK |
| 8. | fac_code | Faculty of author | Char | FK |
| 9. | uni_code | University of author | Char | FK |
| 10. | address | Address of author | Char | |
| 11. | country | Country of author | Char | |

Table 3.7 Data structure of table : conference

| **Table name** | conference | | | |
|---|---|---|---|---|
| **Description** | Store detail data of which conference each paper was submitted | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | paperid | Paper id | Char | PK |
| 2. | conference | Conference of paper | Char | |
| 3. | year | Publication year | Char | |

Table 3.8 Data structure of table : email

| **Table name** | e-mail | | | |
|---|---|---|---|---|
| **Description** | Store data of author's e-mail | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | authorid | Author id | Char | PK |
| 2. | paperid | Paper id | Char | PK |
| 3. | email | Author 's email | Char | |

Table 3.9 Data structure of table : category

| **Table name** | category | | | |
|---|---|---|---|---|
| **Description** | Store details of categories of research | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | category_id | Paper category id | Char | PK |
| 2. | des_category | Category name | Char | |

Table 3.10 Data structure of table : sub_category

| **Table name** | sub_category | | | |
|---|---|---|---|---|
| **Description** | Store details of sub-categories of research | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | sub_category_id | Sub category id | Char | PK |
| 2. | des_sub_category | Sub category name | Char | |
| 3. | category_id | Paper category | char | FK |

Table 3.11 Data structure of table : author_defined_keyword

| **Table name** | author_defined_keyword | | | |
|---|---|---|---|---|
| **Description** | Store keywords that authors defined | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | au_keyword | Predefine keyword id | Char | PK |
| 2. | paperid | Paper id | Char | PK |
| 3. | keyword | Keyword name | Char | |
| 4. | similarcount | Frequency of keyword | Char | |

Table 3.12 Data structure of table : title_keyword

| **Table name** | title_keyword | | | |
|---|---|---|---|---|
| **Description** | Store keywords received from title of paper | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | ti_keyword | keyword in title id | Char | PK |
| 2. | paperid | Paper id | Char | PK |
| 3. | keyword | Keyword name | Char | |
| 4. | similarcount | Frequency of keyword | Char | |

Table 3.13 Data structure of table : abstract_keyword

| **Table name** | abstract_keyword | | | |
|---|---|---|---|---|
| **Description** | Store keywords received from abstract | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | abs_keyword | Abstract keyword id | Char | PK |
| 2. | paperid | Paper id | Char | PK |
| 3 | keyword | Keyword name | Char | |
| 4. | similarcount | Frequency of keyword | Char | |

Table 3.14 Data structure of table : base_university

| **Table name** | base_university | | | |
|---|---|---|---|---|
| **Description** | Store detail of university | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | uni_code | University code | Char | PK |
| 2. | des_university_code | University name | Char | |

Table 3.15 Data structure of table : base_faculty

| **Table name** | base_faculty | | | |
|---|---|---|---|---|
| **Description** | Store detail of faculty | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | fac_code | faculty code | Char | PK |
| 2. | des_faculty_code | faculty name | Char | |

Table 3.16 Data structure of table : base_department

| **Table name** | base_department | | | |
|---|---|---|---|---|
| **Description** | Store detail of department | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | dept_code | department code | Char | PK |
| 2. | des_department_code | department name | Char | |

Table 3.17 Data structure of table : base_organization

| **Table name** | base_organization | | | |
|---|---|---|---|---|
| **Description** | Store detail of organization | | | |
| **No** | **Field** | **Description** | **Data type** | **Remark** |
| 1. | org_code | organization code | Char | PK |
| 2. | des_organization_code | organization name | Char | |

### 3.2.5 Adjacency Matrix Construction

Concerning data preparation for analysis using Social Network Analysis, the adjacency matrix has to be constructed, consisting for rows and columns representing data to be analysed in order to find the relation. Data within the matrix are data displaying relation between things of which relation is to be found.

In data preparation for analysis of their relation, 3 relation matrices will be constructed for this research, comprising relation matrix analysis to find experts, relation matrix analysis to find co-authors, and relation matrix analysis to find research topic categories and sub categories.

#### 3.2.5.1 Relation matrix construction for analysis of expert finding

The adjacency matrix construction for Social Network Analysis (SNA), to find who are experts conducting research in various fields, is shown in Figure 3.13.

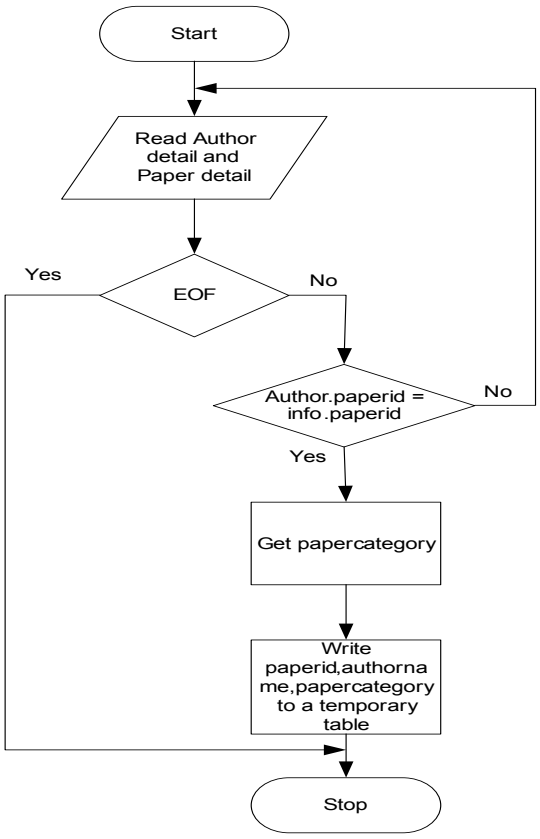



Figure 3.13 Adjacency matrix construction for expert finding

Based on Figure 3.13, an explanation of adjacency matrix construction for analysis of expert finding is given below.

(1) Read data from table: author and paper

(2) Then match data of table: author and paper having a corresponding paperid, and simultaneously count the number of papers of each author in each group of conferences.

(3) Data will be obtained on what paper each author wrote, and into which group of conferences (JCSSE and NCSEC) those papers are classified.

(4) The data of author name, paper category and count_paper are used to create a temporary table .

(5) Expert data from the temporary table as a excel file.

(6) Load the excel file into the SNA visualization tool : NetDraw in order to generate a network diagram showing relation between author name and paper category.

An example of data used to construct an adjacency matrix and an example of an adjacency matrix for analysis are given in Table 3.18 and 3.19, respectively.

Table 3.18 An example of data to be constructed as an adjacency matrix

| count_paper | Authorname | Papercategory |
|---|---|---|
| 1 | A. Jirapornanan | Computational Science and Engineering |
| 1 | A. Surarerks | Computer Systems |
| 1 | Adsak Pongpullponsak | Computer Intelligence |
| 1 | Akekachai Tangsuksant | Software Engineering |
| 1 | Ali Selamat | Computer Software |
| 1 | Anongnart Srivihok | Computer Systems |

Table 3.19 An example of an adjacency matrix for analysis of expert finding

| authorname | Computational Science and Engineering | Computer Graphics | Computer Intelligence | Computer Networks | Computer Software |
|---|---|---|---|---|---|
| A. Jirapornanan | 1 | 0 | 0 | 0 | 0 |
| A. Surarerks | 0 | 0 | 0 | 0 | 0 |
| Adsak Pongpullponsak | 0 | 0 | 1 | 0 | 0 |
| Akekachai Tangsuksant | 0 | 0 | 0 | 0 | 0 |
| Ali Selamat | 0 | 0 | 0 | 0 | 1 |
| Anongnart Srivihok | 0 | 0 | 0 | 0 | 0 |
| Apinetr Unakul | 0 | 0 | 0 | 0 | 0 |
| Assadaporn Sapsomboon | 0 | 0 | 0 | 0 | 0 |
| Athasit Surarerks | 0 | 0 | 0 | 0 | 0 |

#### 3.2.5.2 Relation matrix construction for analysis of co- author finding

The adjacency matrix construction for SNA to find co-authors in conferences may be done as shown in Figure 3.14

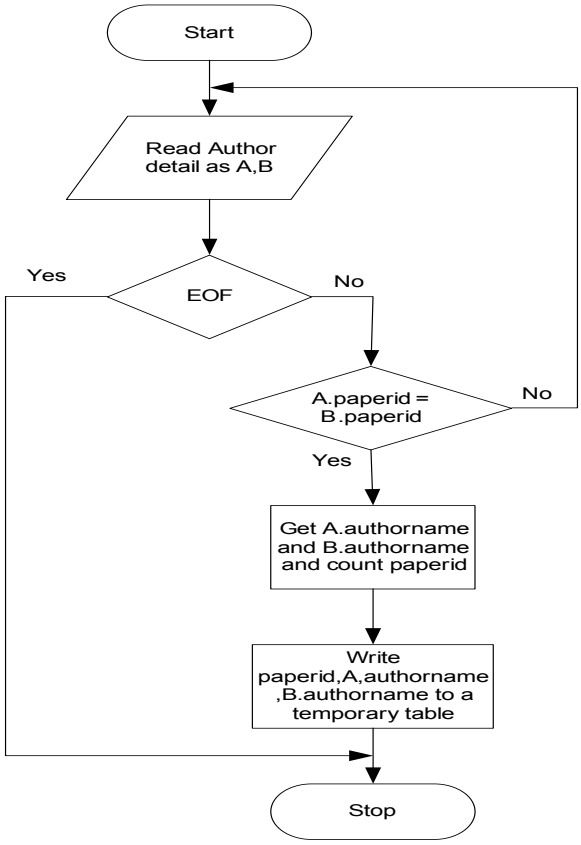



Figure 3.14 Adjacency matrix construction for analysis of co-author finding

From Figure 3.14, an explanation of adjacency matrix construction for analysis of co-author finding is listed below.

(1) Read data from table: author while having it in the format of alias names of A and B.

(2) Match data of authorname from tables A and B having corresponding paperid.

(3) Then authormame writing the same paper is derived, and the number of paperids of co-authors is counted.

(4) Store the number of paperids, A.authorname and B.authorname in a temporary table.

(5) Export data from the temporary table as a excel file.

(6) Load the excel file into the SNA Visualization tool : NetDraw to generate a network diagram displaying the relation between authorname and authorname.

An example of data used to construct adjacency matrix and an example of an adjacency matrix for analysis of co-author finding are revealed in Table 3.20 and Table 3.21, respectively.

Table 3.20 An example of data to be constructed as an adjacency matrix for analysis of co-author finding

| **a_authorname** | **b_authorname** | **paperid** |
|---|---|---|
| Sirimavadee Srinat | Kosin Chamnongthai | 1 |
| Pinit Kumhom | Kosin Chamnongthai | 2 |
| Supaporn Kiattisin | Kosin Chamnongthai | 2 |
| Sornthep Vannarat | Jedsada Phengsuwan | 1 |

Table 3.21 An example of an adjacency matrix for analysis of co-author finding

| a_authorname | Sirimavadee Srinat | Pinit Kumhom | Supaporn Kiattisin | Sornthep Vannarat | Kosin Chamnongthai |
|---|---|---|---|---|---|
| Sirimavadee Srinat | 0 | 0 | 0 | 0 | 1 |
| Pinit Kumhom | 0 | 0 | 0 | 0 | 1 |
| Supaporn Kiattisin | 0 | 0 | 0 | 0 | 1 |
| Sornthep Vannarat | 0 | 0 | 0 | 0 | 0 |

#### 3.2.5.3 Relation matrix construction for analysis of research topic category finding

The adjacency matrix construction for SNA to find research topic category in conferences may be done as shown in Figure 3.15

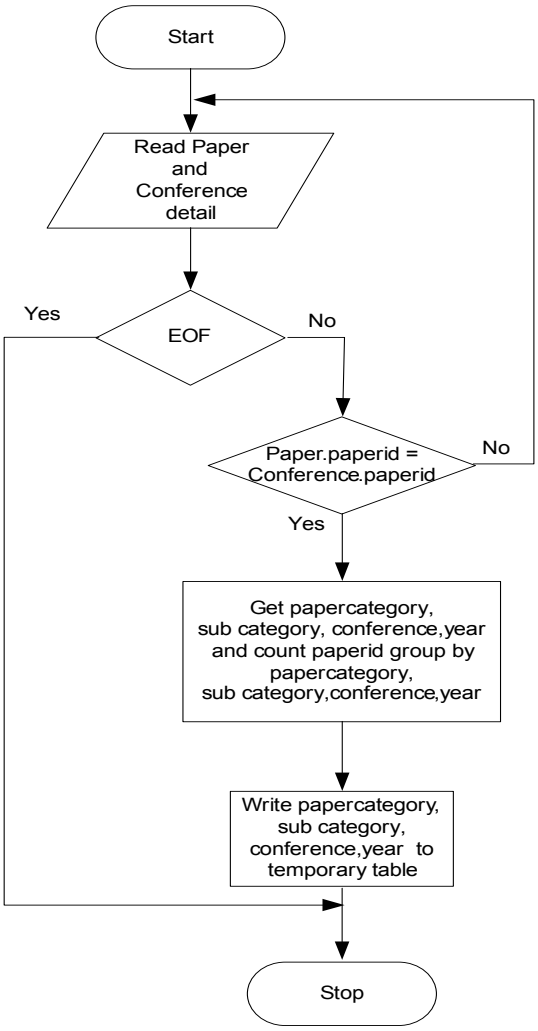



Figure 3.15 Adjacency matrix construction for analysis of research topic category finding

From Figure 3.15, an explanation of adjacency matrix construction for analysis of research topic category finding is listed below.

(1) Read data from table : paper, conference

(2) Then match data of table : paper and conference having

corresponding papered. Data will be obtained into which group of conferences, between JCSSE and NCSEC, each paper is classified, to which conferences it was submitted, and in what year. Meanwhile, the number of papers in each category and sub category is counted based on conferences, category of research and year in which research was submitted.

(3) Data on paper category, sub category, count_paper, conference and year are used to construct as a temporary table.

(4) Export data from the temporary table as a excel file

(5) The excel file is loaded into SNA Visualization tool: NetDraw to generate a network diagram showing the relation between research topic and sub category, which can tell in which category the researches submitted their research conferences, JCSSE and NCSEC, the most, including the number of papers in each research category.

An example of data used to construct an adjacency matrix and an example of an adjacency matrix for analysis of research topic category and sub category finding are given in Table3.22 and 3.23, respectively.

Table 3.22 An example of data to be constructed as an adjacency matrix for analysis of research topic category finding

| category | sub_category | paperid | count_p | count_sub |
|---|---|---|---|---|
| Artificial Intelligent | Genetic Algorithm | 1 | 2 | 1 |
| Artificial Intelligent | Neural network | 2 | 2 | 1 |
| Computer Network | Wireless and Ad hoc Networks | 3 | 2 | 1 |
| Computer Network | Quality of Service Routing | 4 | 2 | 1 |
| Computer software | Computer Accessibility | 5 | 1 | 1 |
| Computer system | Embedded Systems | 6 | 1 | 1 |

Table 3.23 An example of an adjacency matrix for analysis of research topic category finding

| **sub_category** | Artificial Intelligence | Computer Network | Computer software | Computer system |
|---|---|---|---|---|
| Genetic Algorithm | 1 | 0 | 0 | 0 |
| Neural network | 1 | 0 | 0 | 0 |
| Wireless and Ad hoc Networks | 0 | 1 | 0 | 0 |
| Quality of Service Routing | 0 | 1 | 0 | 0 |
| Computer Accessibility | 0 | 0 | 1 | 0 |
| Embedded Systems | 0 | 0 | 0 | 1 |

## 3.3 Data Analysis by Social Network Analysis

### 3.3.1 Data Analysis of Expert Finding

For analysis to finding experts in various fields, it aims to search which researcher has conducted more researches than others in the same field. This is to find who plays a significant role or has much impact on a researcher group. In this research, the degree centrality value is measured in order to search for researchers playing a significant role on a group of all researchers who have submitted their research to the conferences JCSSE and NCSEC.

To measure the degree centrality value of expert finding, the relation between researchers and group of research works is searched for, while the relation has to be represented in the form of matrix showing first the relation between researchers and groups of research works, which is the inobvious relation. However, the relation we can clearly see in example groups comprises 2 relations. The first one is the relation between researchers and their research works while the second one is the relation between research works and their groups. Therefore, we can find the

relation between researchers and groups of research works through the 2 relations mentioned earlier.

In this research, a weight matrix is generated to display the relation between researchers and groups of research works. The weight is the value telling which researcher is related to which group of research works, as defined in equation (3.1).

$$W_{ij} = (A_i C_{j,}, E) \qquad \text{---------- (3.1)}$$

Where

$W_{ij}$ = weight value of researcher i and research topic category j

$A_i C_{j,}$ = set of researchers and research topic categories

E = the number of lines linking the relation between the researcher and research group through which topic that research belongs to.

Figure 3.16 reveals an example between the group of research and researchers. Each square represents research paper while a circle represents a researcher, and a hexagon represents a research topic category.

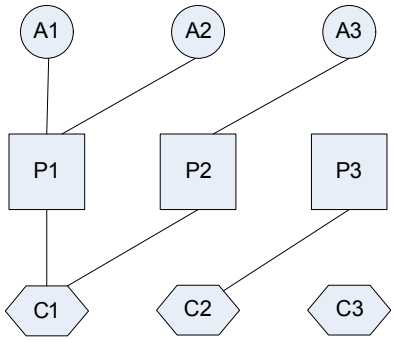



Figure 3.16 Author, paper, category relation

### 3.3.2 Data Analysis of Co-author Finding

Searching for co-authors is to find which author has conducted research with others, which allows us to know the relation among researchers. Researchers who have collectively conducted research are regarded as ones having relation.

To find the relation among researchers, a weight matrix showing the relation among researchers, called author's relationship, has to be constructed. Here, the closeness value is used for measurement. The weight value can be defined in equation (3.2).

$$W_{ij} = (A_i A_{j,}, E) \qquad \text{---------- (3.2)}$$

Where

$W_{ij}$ = weight value of researcher i and researcher j

$A_i A_{j,}$ = set of researchers

E = the number of lines linking the relation among co-author

Apart from the construction of author's relationship displaying the relation among researchers, and using the centrality value to analyse for co-author finding, we can also generate the author's community in order to show a researcher's community. The author 's community reveals each researcher's network, which tells which researcher he/she has collectively conducted research with, on the same topic. The value used for analysis of the author's community finding is measured by the closeness value, which shows the approximateness between networks or groups of researchers.

Figure 3.17 displays an example of the relation between researchers. Based on the figure, a square and a circle represent research paper and researcher, respectively.

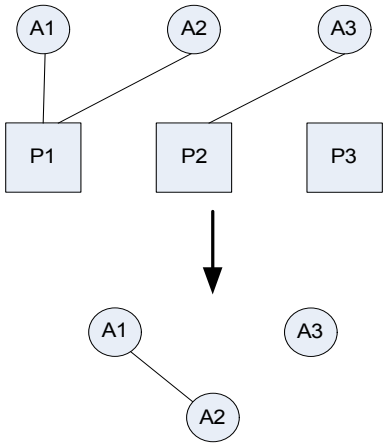



Figure 3.17 Author's relation

### 3.3.3 Data Analysis of Research Topic Category Finding

Searching for research's research topic category is to find which research topic categories and sub categories are the majority ones submitted to the conferences, JCSSE and NCSEC.

To find the relation between research topic category and sub category, a weight matrix, showing the relation between research topic category /sub category and research papers, has to be constructed. Here, the degree centrality value is used for measurement, and the weight value is defined in equation (3.3)

$$W_{ij} = (P_i C_{j,}, E) \quad \text{---------- (3.3)}$$

Where

$W_{ij}$ = weight value of research i and research topic category j

$P_i C_{j,}$ = set of research papers and a research topic category

E = the number of lines linking the relation between research paper and a research topic category, which tells what topic category/sub category that research is about, and in which group it belongs to.

Figure 3.18 reveals an example of the relation between the research papers and research topic category/sub category. From the figure a square and a hexagon represent a research category/sub category and research papers, respectively.

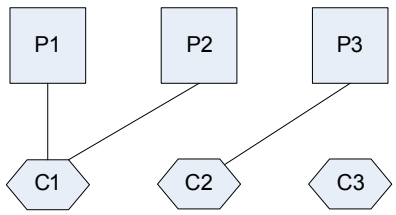



Figure 3.18 Research papers and category relation

In the data analysis process through a Social Network Analysis method, the Social Network Analysis visualization tool, or NetDraw, is used to analyse and show a picture of relation constructed by an adjacency matrix. Here, from the adjacency matrix construction stage, an adjacency matrix in the format of excel file is obtained, consisting of 3 files, namely, an adjacency matrix for expert finding, an adjacency matrix for sub category topic finding and an adjacency matrix for co-author finding. These 3 matrices will be analysed through the Social Network Analysis method. The value used to measure the relation between data comprises the degree centrality value and closeness value, which enables us to know which data are significant for this data group, and makes it possible to create a network diagram for each adjacency matrix, as illustrated in Figure 3.19 and 3.20.

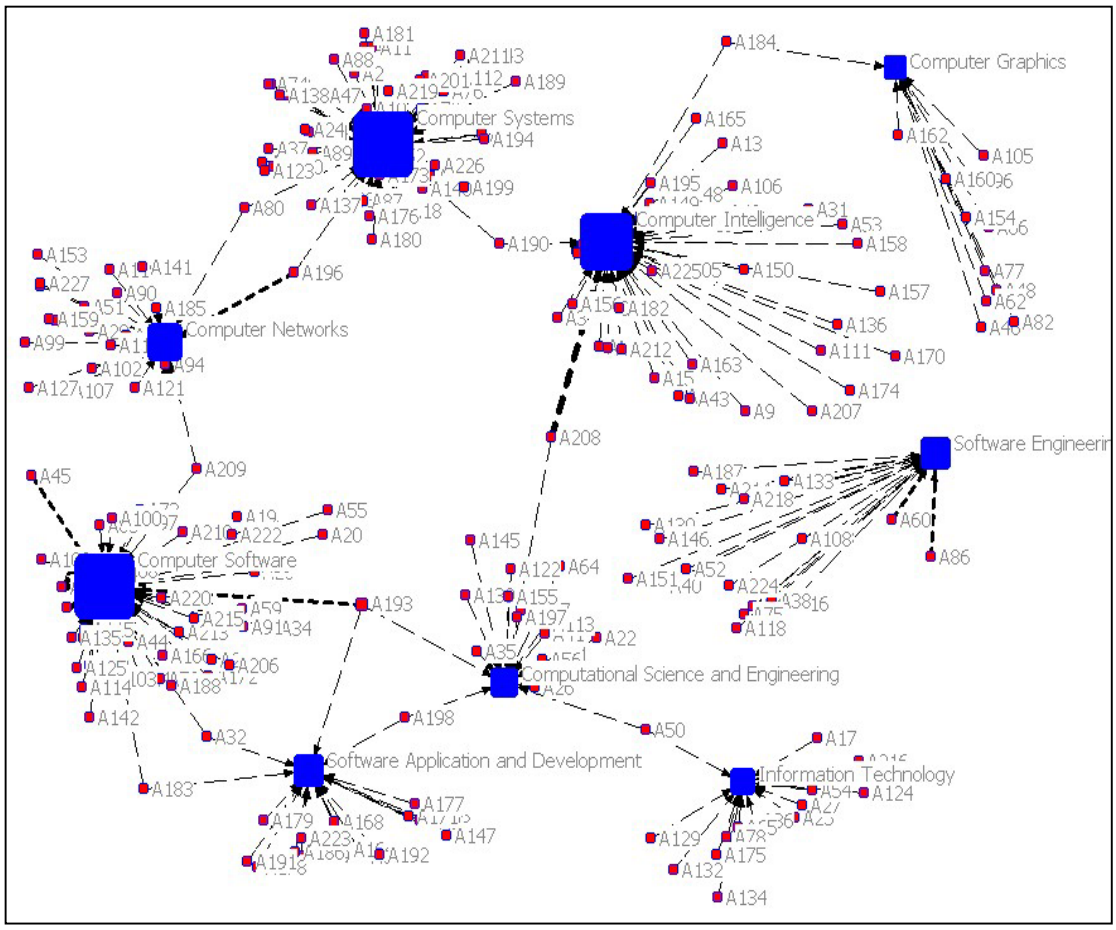



Figure 3.19 An example of a network diagram constructed from an adjacency matrix of expert finding

Figure 3.19 illustrated a network diagram showing the relation between researchers and research topics categories that each researcher has submitted to the conference JCSSE 2007. From the figure, it can be explained that category of research topics submitted to the conference are represented by blue squares, while researchers submitted their research are represented by red circles. Dotted lines reveal which research topics have been submitted by researchers. In this regard, thicker dotted lines indicate that a researcher has submitted more than 1 research topic category. Besides, a network diagram also points out which researcher has submitted his/her research for over 1 research topic category. For instance, the researcher no A196, or Ms.Vipada Vejprasit, has submitted her research in 2 research topic categories, namely, the computer system and computer network.

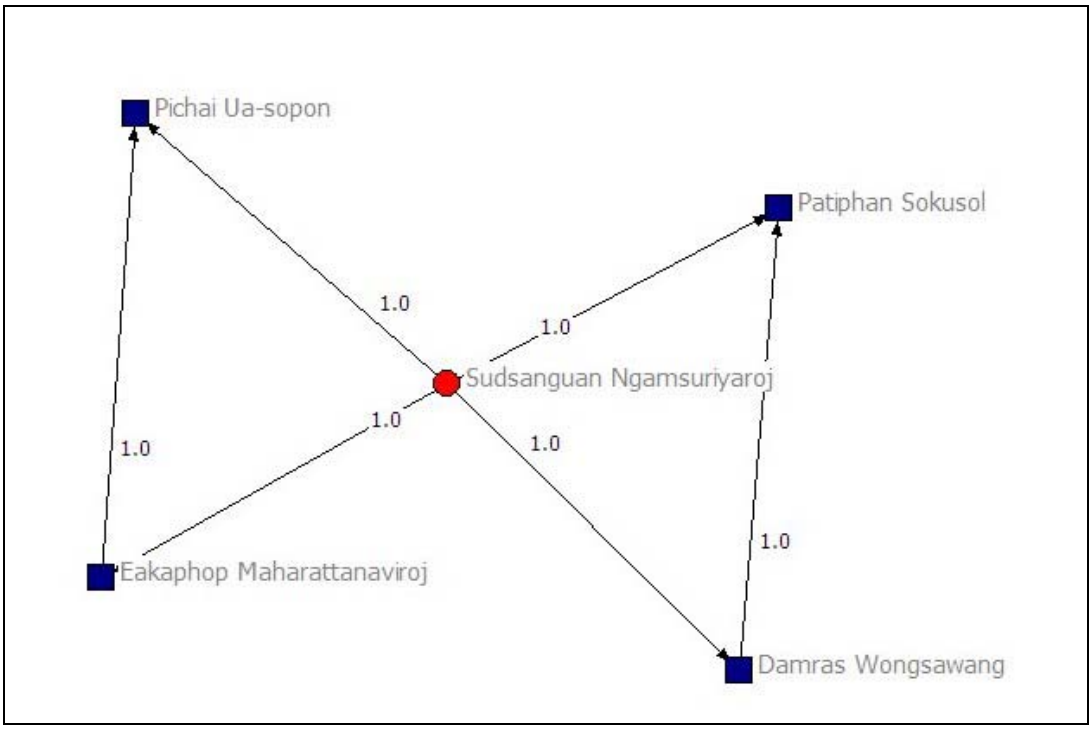



Figure 3.20 An example of a network diagram created from an adjacency matrix of co-author finding

Figure 3.20 illustrates a network diagram displaying the relation between a researcher, Ms. Sudsanguan Ngamsuriyaroj, and her co-authors. Based on the figure, Ms. Sudsangaun has 4 co-authors, comprising Mr. Pichai Ua-Sopon, Mr. Pathiphan Sokusol, Mr. Eakaphop Maharattanaviroj and Mr. Damras Wongsawang. The number attached on each line, showing the relation between Ms. Sudsanguan and her co-authors, signifies the number of research papers each co-author has conducted collectively with Ms. Sudsanguan. Here, each co-author has only 1 research paper.

## 3.4 User Interface Design

The design for a search for data related to researchers is divided into 3 parts based on search function, composed of (1) expert searching, (2) co-author searching and (3) research topic category searching

### 3.4.1 Expert Searching

For expert data searching, in this research, each researcher's research topic will be considered. The expert searching process is illustrated in Figure 3.21.

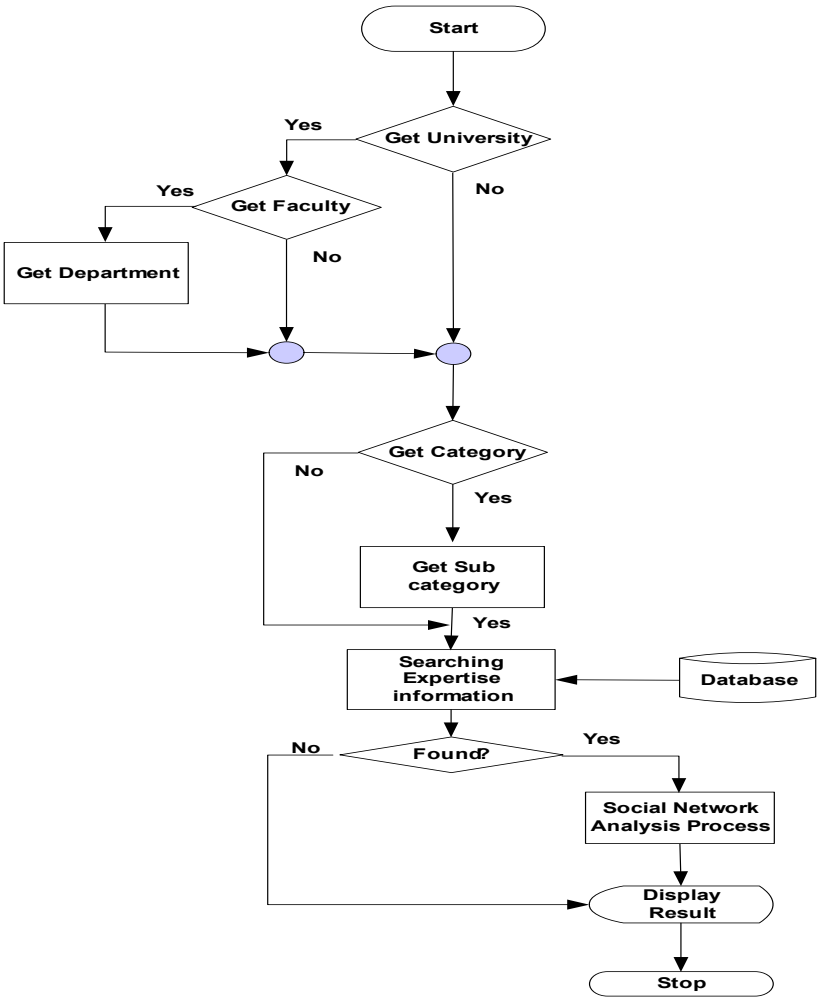



Figure 3.21 Expert searching process

Figure 3.21 illustrates a process of expert data searching. A user has to identify the category of research topic he wants to search, and can also identify a researcher's agency. After the user has already identified the searching condition, the system will search for data of researcher conducting research on the topic identified by the user. When the data are found, the system will deliver researcher data to the Social Network Analysis tool for Analysis, after which a network diagram of each researcher conducting research on the topic identified by the user is sent back to the system. Meanwhile, the part searching data in the database will display data of researchers and research pertinent to the topic being searched for, which are in order of researcher who does research on that topic category, arranged from most to least. An example of a user interface of expert searching is illustrated in Figure 3.22, while searching results are illustrated in Figure 3.23 (social network view) and 3.24 (data view), respectively.

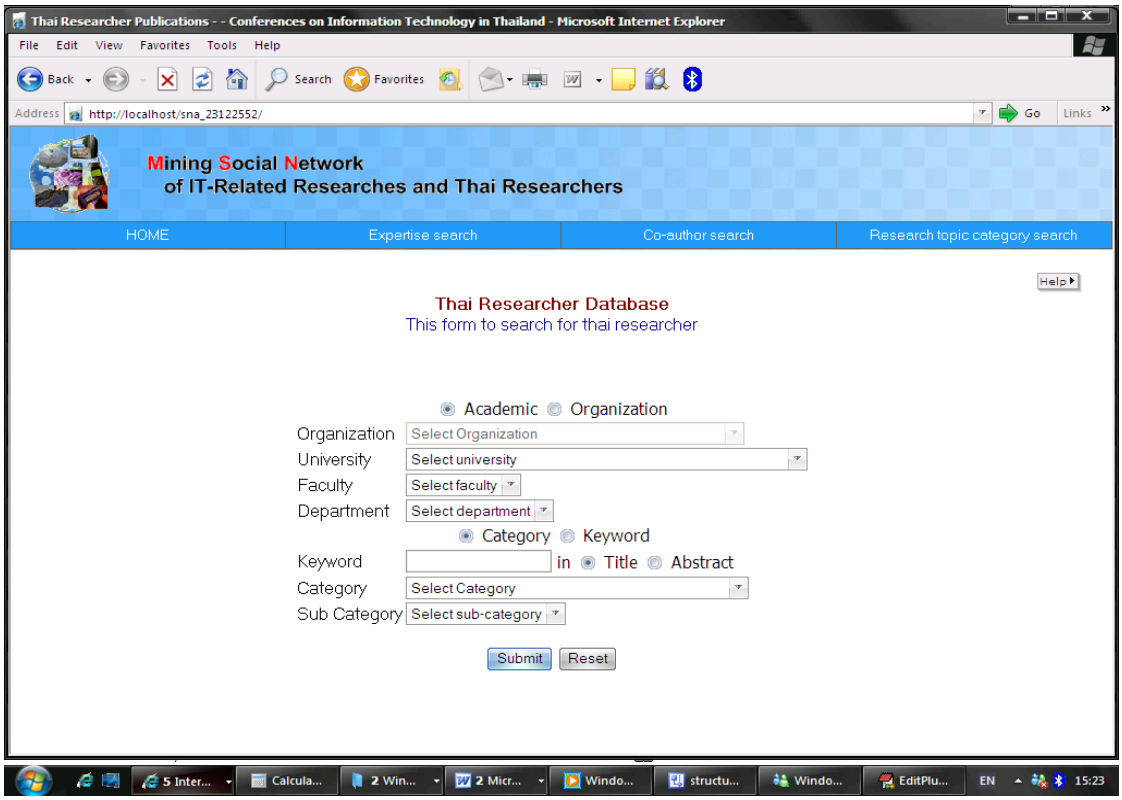



Figure 3.22 An example of an expert searching interface

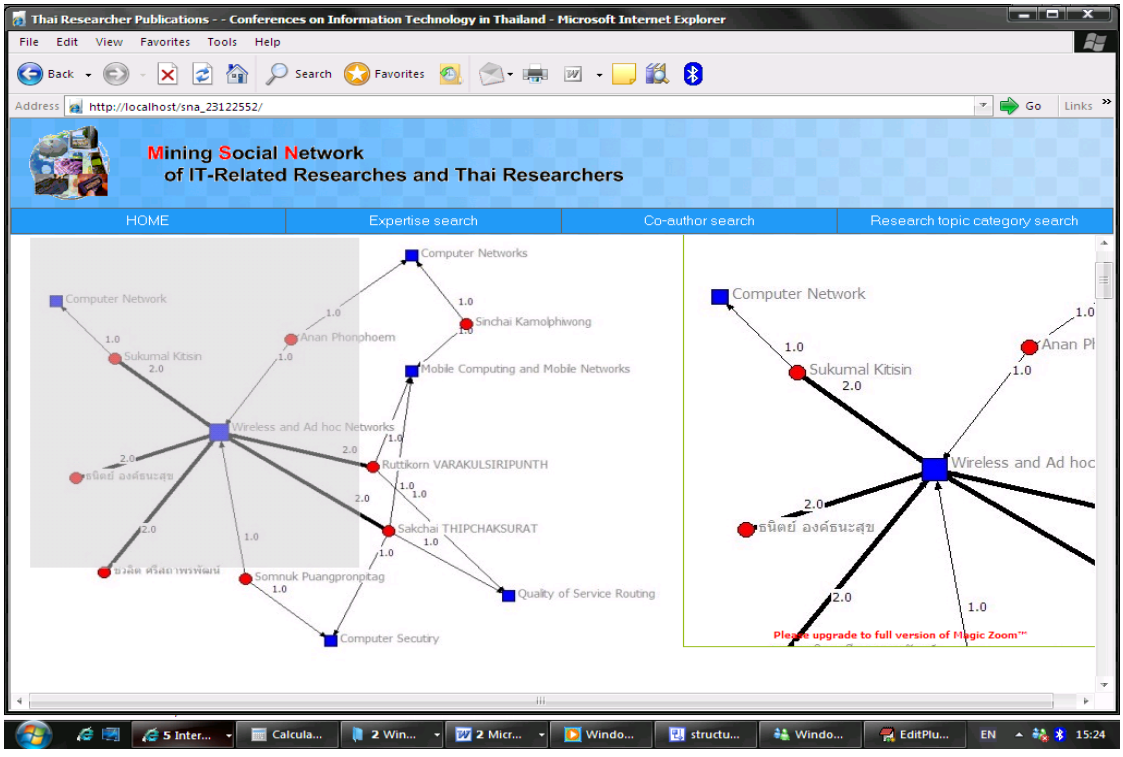



Figure 3.23 An example of an interface of expert searching results (social network view)

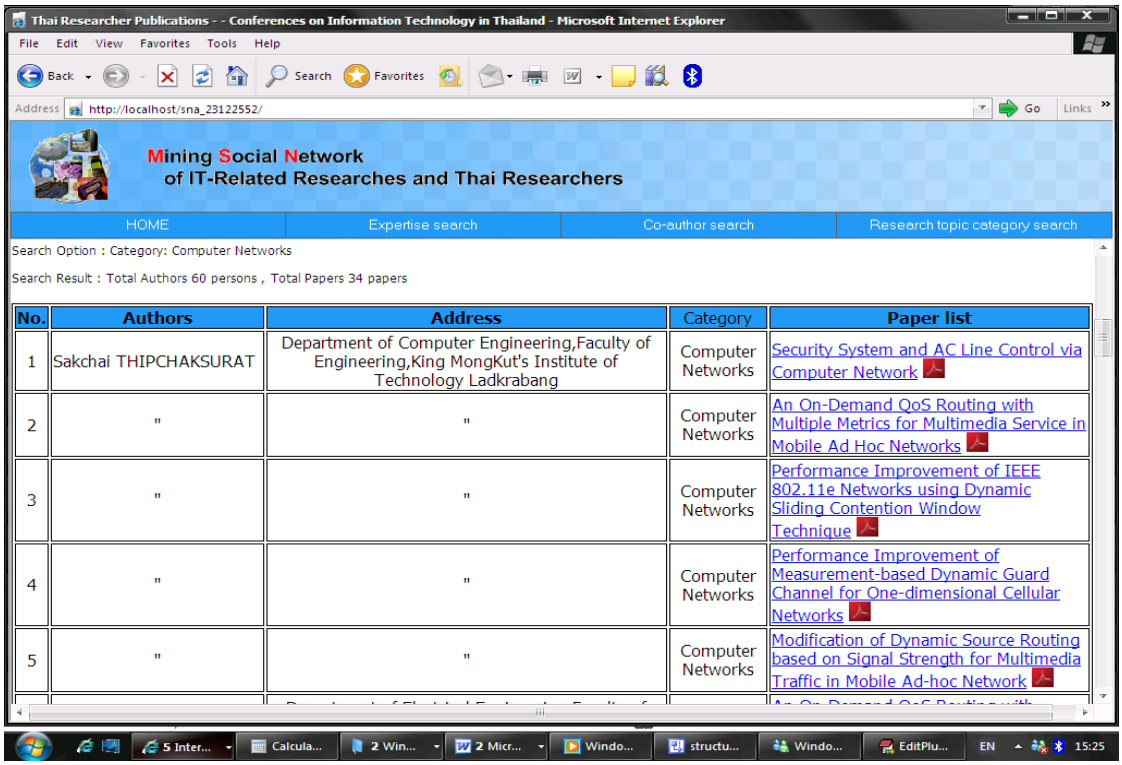



Figure 3.24 An example of an interface of expert searching results (data view)

### 3.4.2 Co-author Searching

Searching for co-authors is to find researchers collectively conducting research. The co-author searching process is illustrated in Figure 3.25

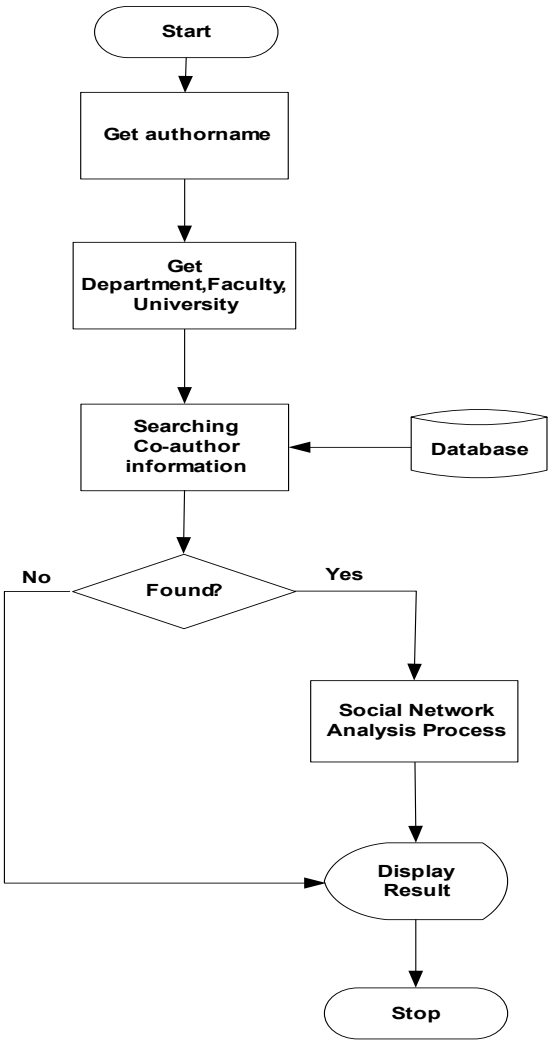



Figure 3.25 Co-author searching process

Figure 3.25 displays a process of co-author data searching. A user has to identify a researcher's name and may identify a co-author's agency. The system will search for researcher data in the database. When co-author data are found, they will be analysed through the Social Network Analysis tool later. After the Social Network Analysis tool has constructed a network diagram for co-authors for whom the user wants to search, the result will be returned to the system for a display along with the outcome obtained from data searching in the database. Figures 3.26, 3.27 and 3.28 illustrate examples of a co-author searching interface and a searching result, respectively.

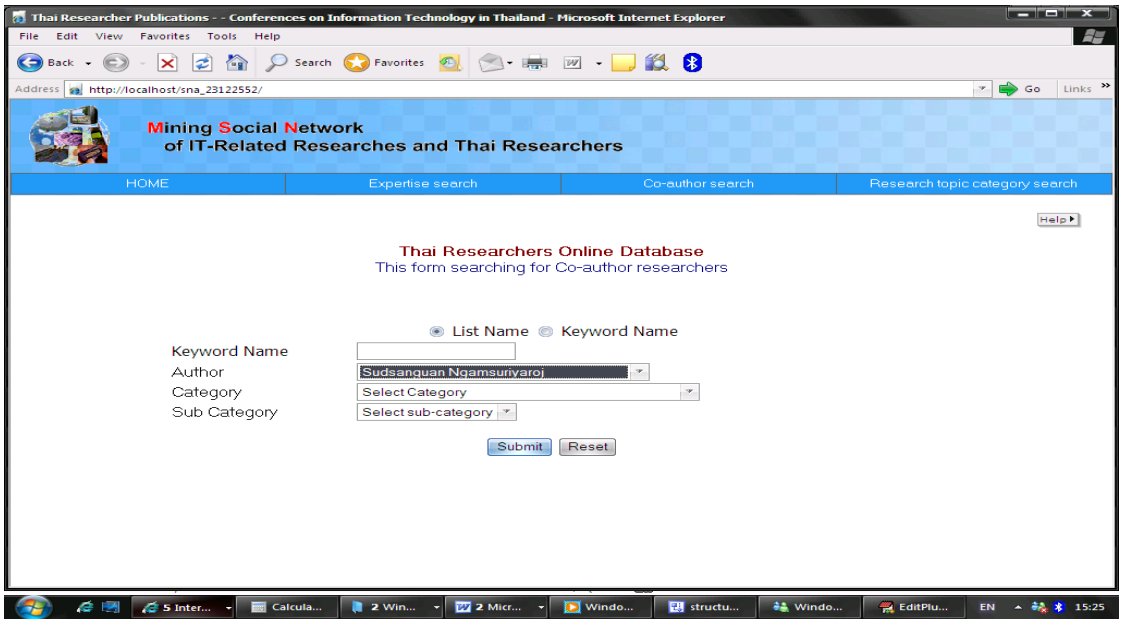



Figure 3.26 An example of a co-author searching interface

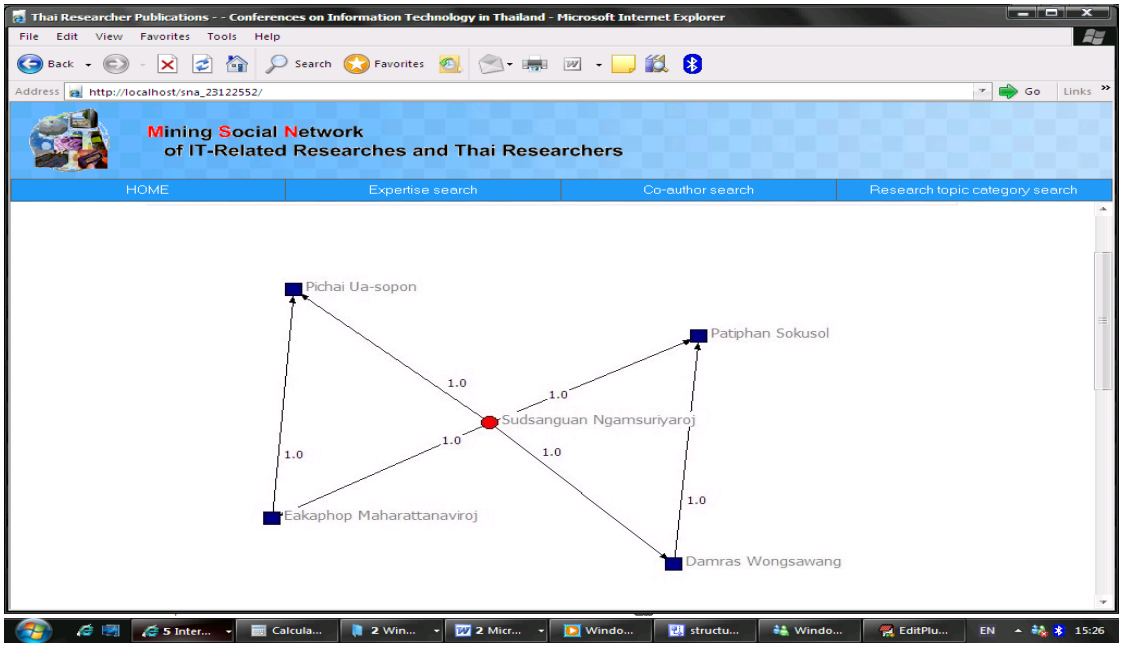



Figure 3.27 An example of an interface of co-author searching results (social network view)

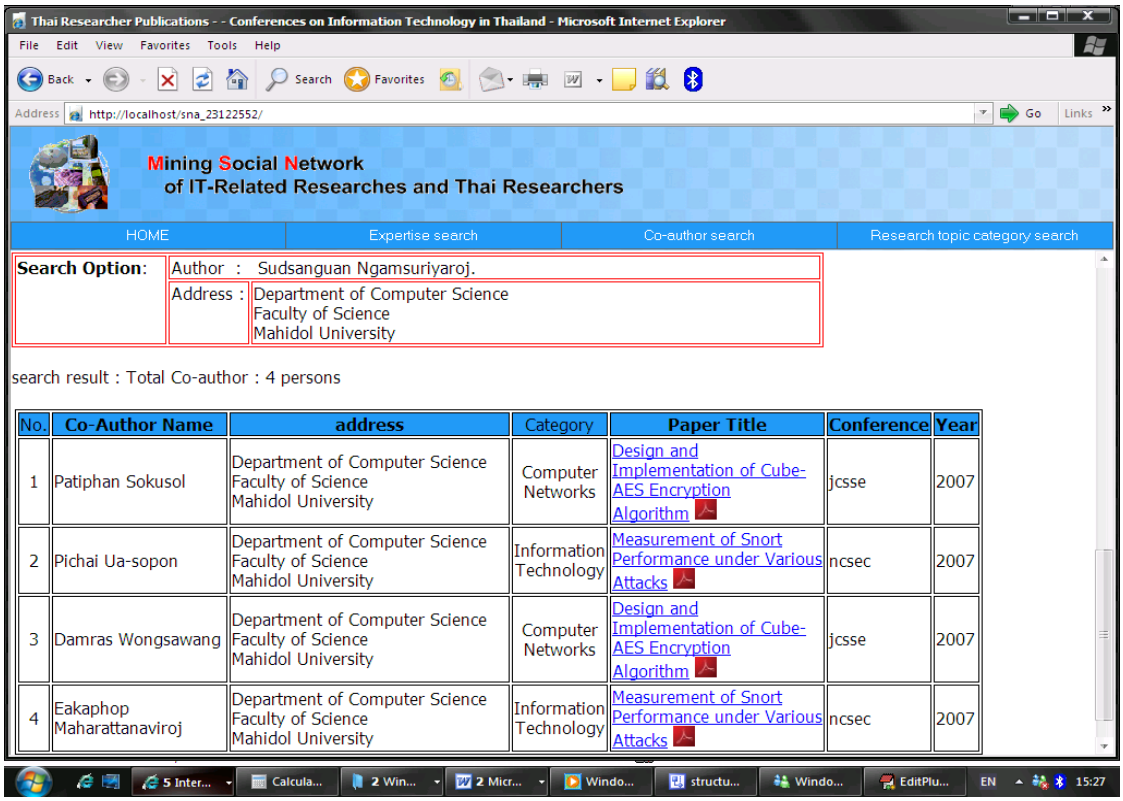



Figure 3.28 An example of an interface of co-author searching results (data view)

### 3.4.3 Research Topic Category Searching

To search for research topic category and sub category and they can be viewed from the number of researches submitted to the conferences, JCSSE and NCSEC, during the time period defined, to find out whether which research topic category and sub category were submitted among a large number of researches, as illustrated in Figure 3.29. An example of research topic category searching interface and example of an interface of research topic category and searching result are illustrated in Figures 3.30, 3.31 and 3.32, respectively.

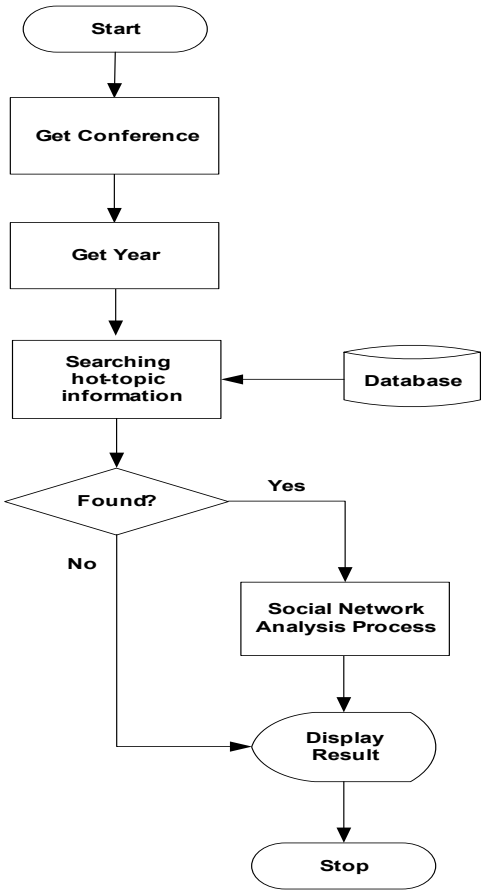



Figure 3.29 Research topic category searching process

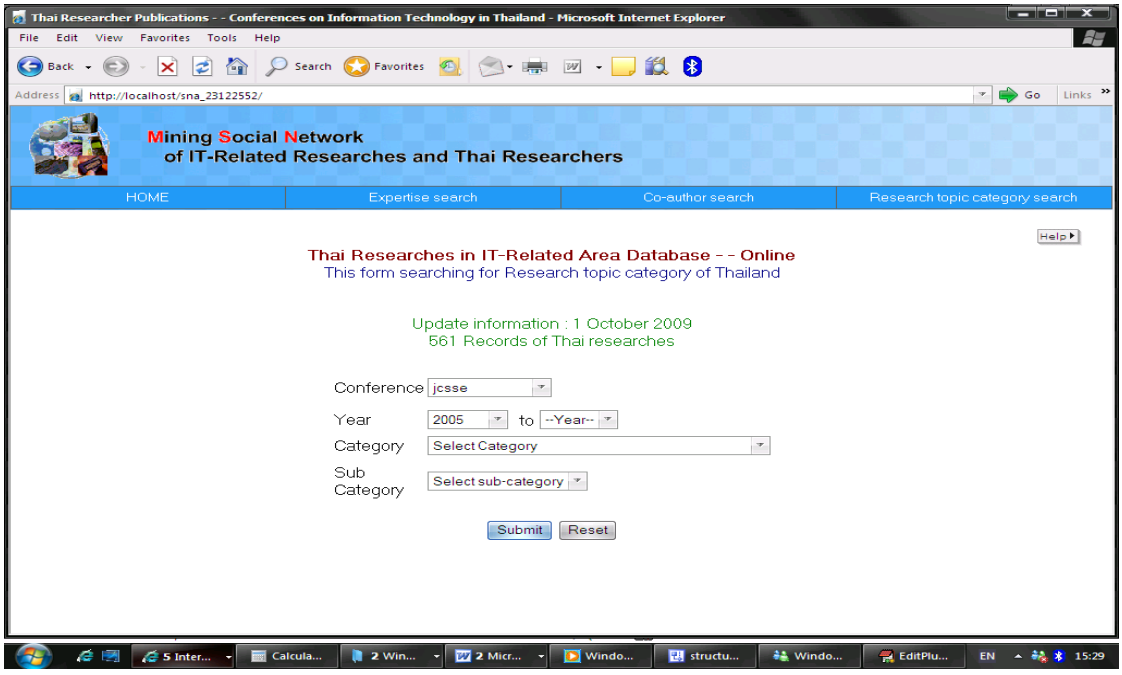



Figure 3.30 An example of a research topic category searching interface

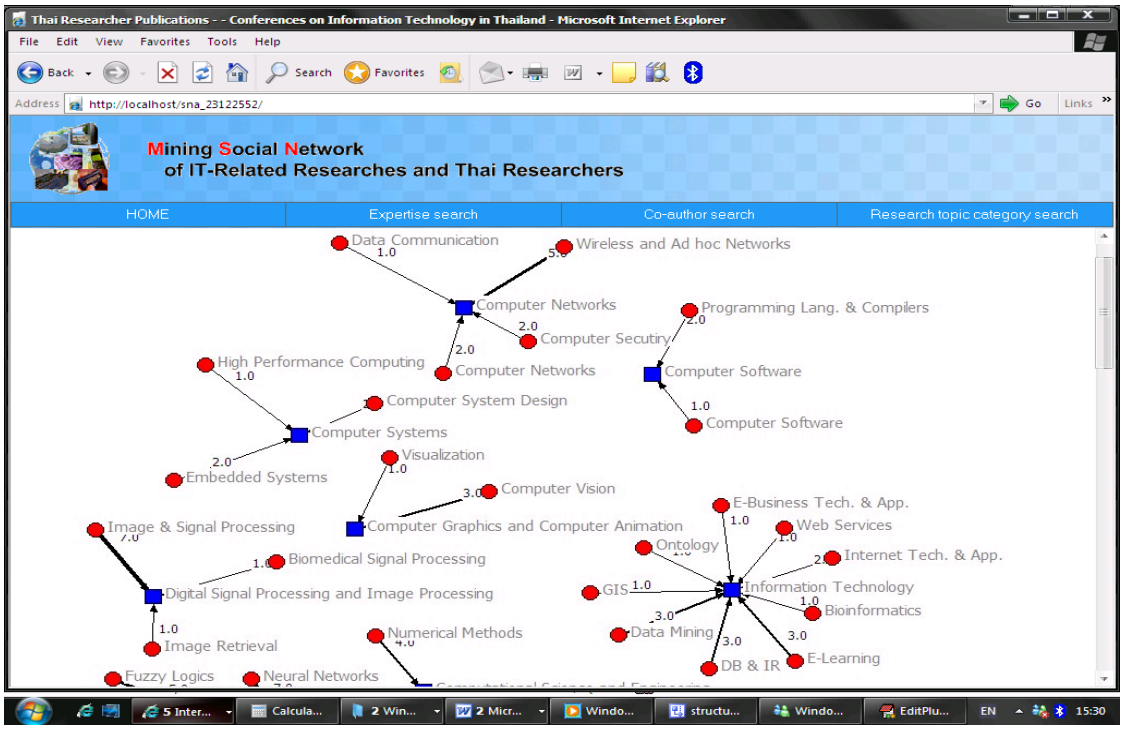
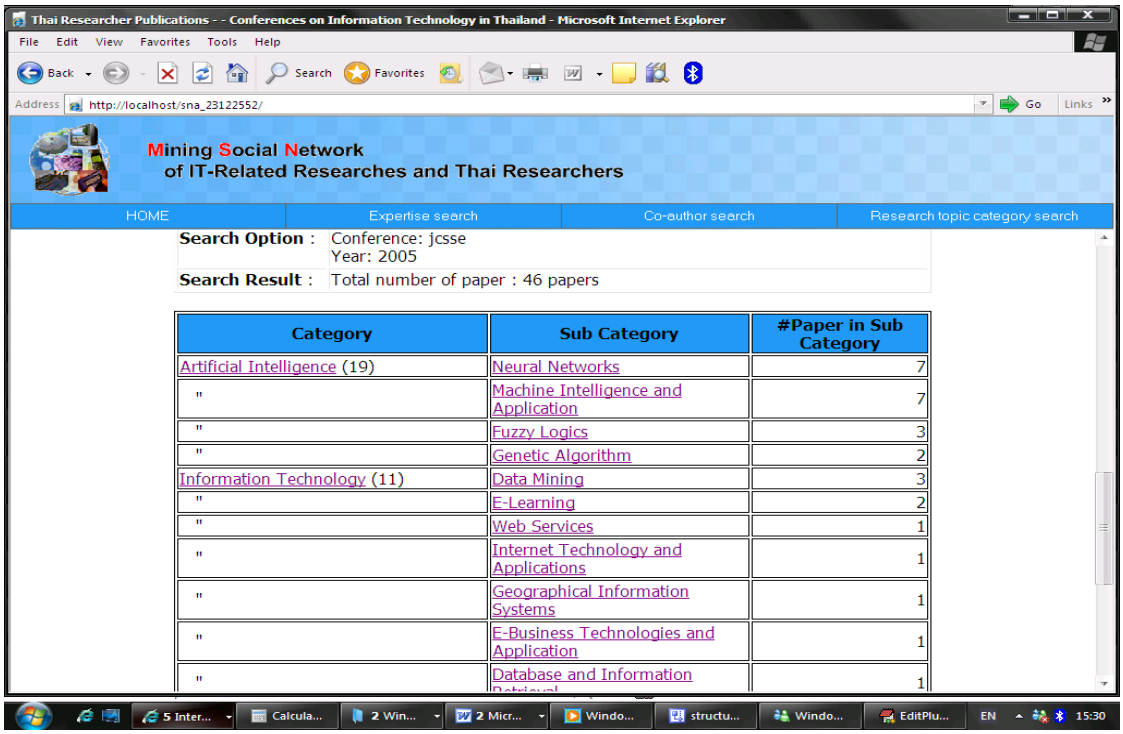



Figure 3.31 An example of a research topic category searching results (social network view)




| Category | Sub Category | #Paper in Sub Category |
|---|---|---|
| Artificial Intelligence (19) | Neural Networks | 7 |
| " | Machine Intelligence and Application | 7 |
| " | Fuzzy Logics | 3 |
| " | Genetic Algorithm | 2 |
| Information Technology (11) | Data Mining | 3 |
| " | E-Learning | 2 |
| " | Web Services | 1 |
| " | Internet Technology and Applications | 1 |
| " | Geographical Information Systems | 1 |
| " | E-Business Technologies and Application | 1 |
| " | Database and Information | 1 |

Figure 3.32 An example of a research topic category searching results (data view)

# CHAPTER IV
# SYSTEM IMPLEMENTATION

This chapter discusses details of the development of a prototype system of Data Mining of Social Network for IT-related researches and Thai researchers. The prototype system is developed by using techniques described in the previous chapter. Details of the prototype system development consist of 2 parts, (1) Hardware and software for implementation, and (2) Prototype implementation.

## 4.1 Hardware and Software for Implementation

### 4.1.1 Hardware Specification

- Processor : Intel Pentium M 1.5 GHz
- RAM : 1 GB
- Hard disk : 40 GB

### 4.1.2 Software Specification

- Operating System : Microsoft Windows XP (service pack 2)
- Software Tool Development : NetDraw , PHP, C#, Microsoft .NET Framework 3.0
- Database Server : MySQL
- Application Server : Apache tomcat version 2.5.10

## 4.2 Prototype Implementation

For the development of a miming social network for IT-related researches and Thai researchers component implementation, it comprises the following parts : (1) Data preparation tool implementation, (2) social network analysis tool implementation, and (3) searching tool implementation. Each part is discussed in detail below.

### 4.2.1 Data Preparation Tool Implementation

The development for data preparation begins with loading data into the system, and data transformation before storing data in the system database.

For data loading, a tool and an interface for extracting data from a pdf file format are developed along with PDFBOX, which is a tool for converting documents form a pdf format into a text format.

PDFBOX is a library for extracting text from pdf documents. It is developed by using the C# language, which has to work on the Microsoft .NET Framework. The development of a user interface is revealed in Figures 4.1, 4.2 and 4.3

```
Get pdf file name
Set count of phase to zero
Set total of author to zero
Set indicator of author to 1
PDFBOX readln text in pdf
Repeat
Check ascii code of first character in first line
Repeat(1)
If ascii code is thai character then
        Save titleTh
        Split each words by using white space
        Save title keyword
Else
        Save titleEng
        Split each words by using white space
        Save title keyword
End if
Read next line
Repeat (2)
Split phase by white space then put each phase[i] in array of phase
```

Figure 4.1 Pseudo code of Data preparation implementation

```
Increment count of phase by 1
Until EOF (2)
Set total of author is divide count of phase by 2
Repeat (3)
Read each element of array
If element [i] is odd position then
        Save author name is phase[i]
Else
        Save author name is author name concatenate with phase[i+1]
End if
Until end of array and total author equal to count of phase divided by 2 (3)
Read next line
Repeat (5)
Split phase of character by white space
Push each phase to array

Repeat (4)
Pop element[i] of array
If found word ‘department’  then
        Save department
Else if found word ‘faculty’  then
        Save faculty
Else if fount word ‘university’ then
        Save university
    Else save organization
End if
Until end of array (4)
Until  EOF (5)
Read next line
```

Figure 4.2 Pseudo code of Data preparation implementation (Cont.)

Repeat (6)
Split phase of character by white space
Push each phase to array
Repeat (7)
Pop element from array
Address = address + element[i]
Until end of array (7)
Save address
Until found 'abstract' (6)
Save abstract
Split phase of character by white space
Repeat(8)
Push each phase into array
Pop each element of array
Save each phase of abstract
Until end of array (8)
Until introduction(1)

Figure 4.3 Pseudo code of Data preparation implementation (Cont.)

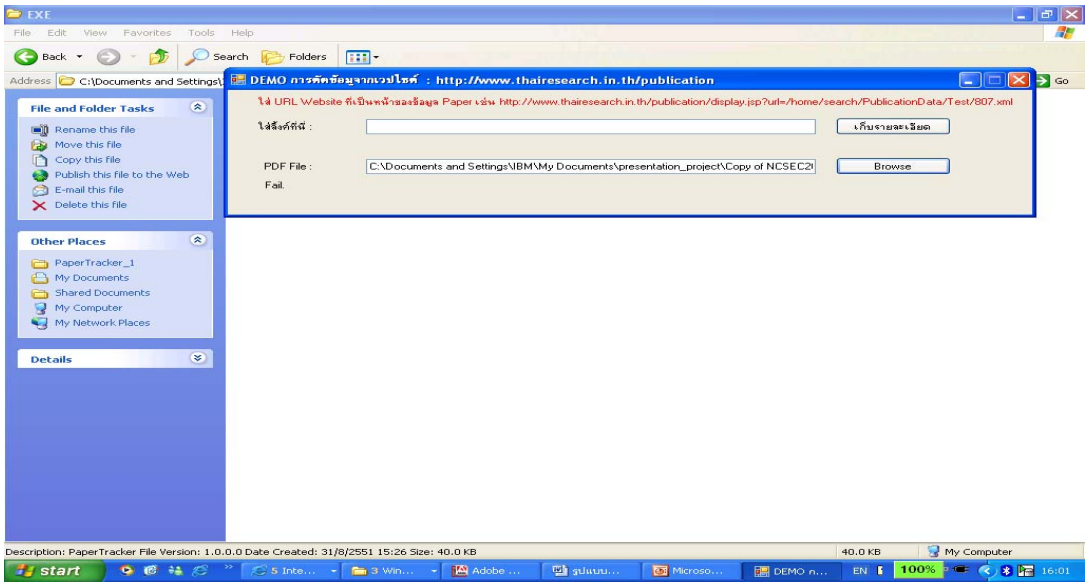



Figure 4.4 An interface for users to load a pdf file

Figure 4.4 illustrates an interface for users to load a pdf file from which information needs to be extracted in order to be used for subsequent analysis.

A user will select a pdf file stored in a directory and then press Enter . Afterwards, the system will process by having PDFBOX extract information as a stream of text. Later, the system will chose only the information required, to get it through the processes of lexical analysis and data transformation and then store data in the database.

### 4.2.2 Social Network Analysis Tool Implementation

NetDraw is a tool used for visualizing Social Network data both in 1-mode and 2-mode types. NetDraw can also cope with data having various relations. In addition, it can tackle nodes with respect to their color, shape and character of lines displaying the relation.

NetDraw is able to read a variety of files such as UCINET DL file, Pajek file,etc. When NetDraw has read those files, it will calculate the centrality values and then display the feature of a Social network of those data groups. The feature of a Social Network graph of a data group will vary with the centrality value that needs to be measured.

Figure 4.5, 4.6 and 4.7 reveal pseudo codes of adjacency matrix construction for analysis of expert finding, co-author finding and sub category topic finding, respectively.

```
Read Author detail from author file and paper detail from info file
Repeat(1)
If (paperid that each author written) then
        Get paper category
        Save  paperid,authorname,paper category  to create
adjacency matrix in excel format for expert analysis
Else
        Read next data
End if
Until EOF (1)
```

Figure 4.5 Pseudo code of adjacency matrix construction for expert finding

```
Read Author detail from author file as alias name A and B
Repeat(1)
If (A.paperid = B.paperid) then
        Get A.authorname , B.authorname  and count paperid
        Save  paperid,A.authorname,B.authorname  to create
adjacency matrix in excel format for co-author analysis
Else
        Read next data
End if
Until EOF (1)
```

Figure 4.6 Pseudo code of adjacency matrix construction for co-author finding

```
Set count paper to 0
Read paper detail from info file as A and read paper detail from
appearin file as B
Repeat(1)
If (A.paperid = B.paperid) then
        Increment Count paperid
        Get A.authorname ,category,sub
category,conference,year  and count paperid
        Save paperid,A.authorname, category,sub
category,conference,year  and count paperid to create adjacency
matrix in excel format for co-author analysis
Else
        Read next data
End if
Until EOF (1)
```

Figure 4.7 Pseudo code of adjacency matrix construction for sub category topic finding

```
Read adjacency matrix  file in text (excel) format
Convert data set in excel file format to data set in Pajek format by
run Excel to Pajek Save to data set in Pajek format file
Read data set in Pajek format
Set parameter for NetDraw execution
NetDraw calculate Network Centrality
NetDraw Visualization Network Diagram of this data set
Present Network diagram from NetDraw
```

Figure 4.8 Pseudo code of data analysis via the Social Network Analysis Technique

When an adjacency matrix in a excel file format for data analysis in each part has been constructed, a excel file has to be converted into a format which NetDraw can read its data. So, a tool called exceltopajek is used to generate a data set in a pajek format out of a excel file.

After a data set in a pajek format has been derived, NetDraw will read data in that data set and then calculate the network centrality value and create a network diagram.

### 4.2.3 Searching Tool Implementation

The development of searching part comprises 3 parts based on the system's core functions of expert searching, co-author searching and sub category topic searching, of which details are discussed in the following sections, as illustrated in Figure 4.9.


Figure 4.9 The system's main interface

Help menu, which will have general information. The user may want to know, including how to find information, the scope of the information contained in the database, and educational institutions or organizations to submit research to participate in technical meetings JCSS and NCSEC all examples of the screen Help is shown in Figure 4.10.

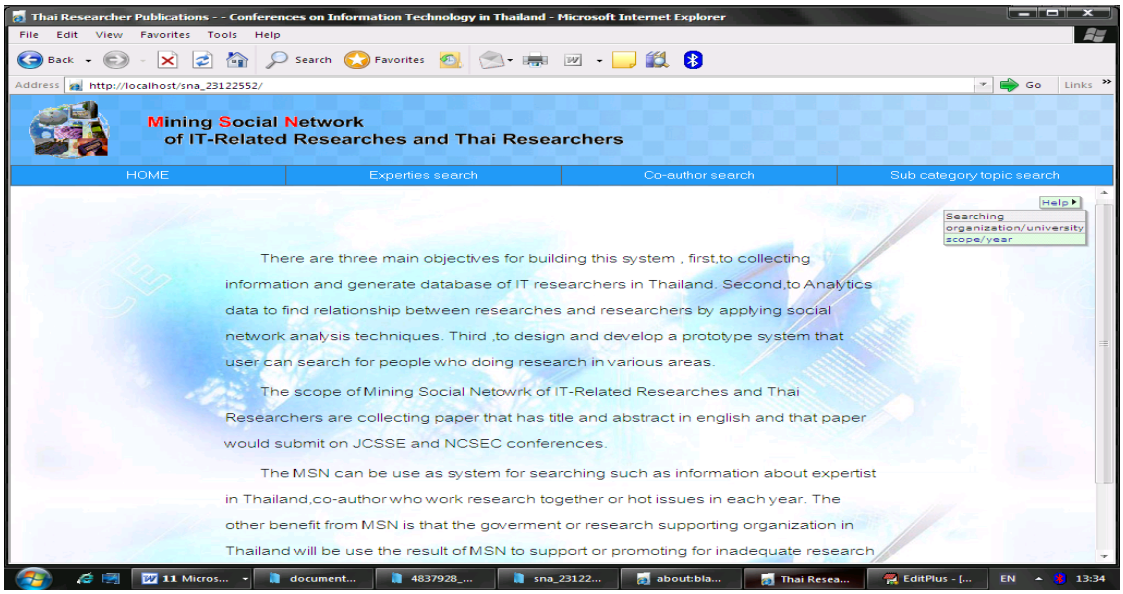



Figure 4.10 Help menu

The Help menu contains sub- menu: Searching, Organization / university and scope / year.

Searching sub-menu will be displays details on how to find information. The sub-menu Organization / university Will shows detail of the organization and institutions that have sent researches to conferences JCSSE and NCSEC and sub-menu Scope / year to provide information about the scope of research is in database system. All sub-menu of help menu is shown in Figure 4.11, 4.12 and 4.13, respectively.

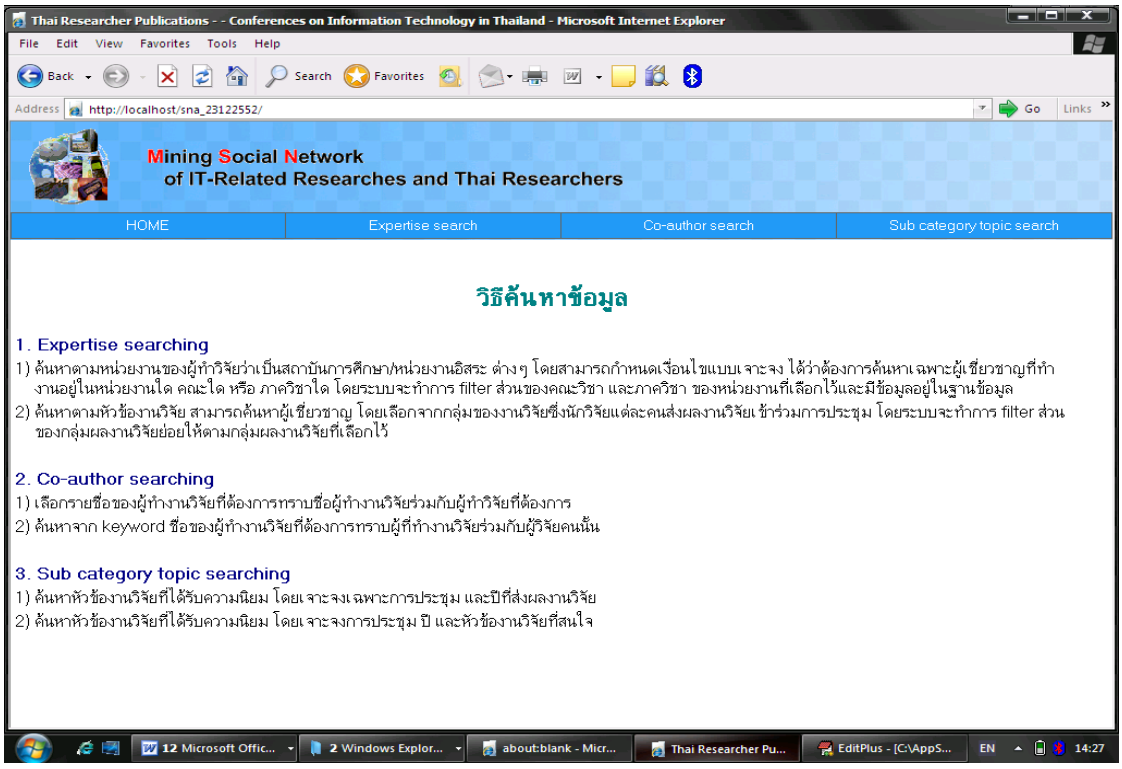



Figure 4.11 Help sub-menu: Searching

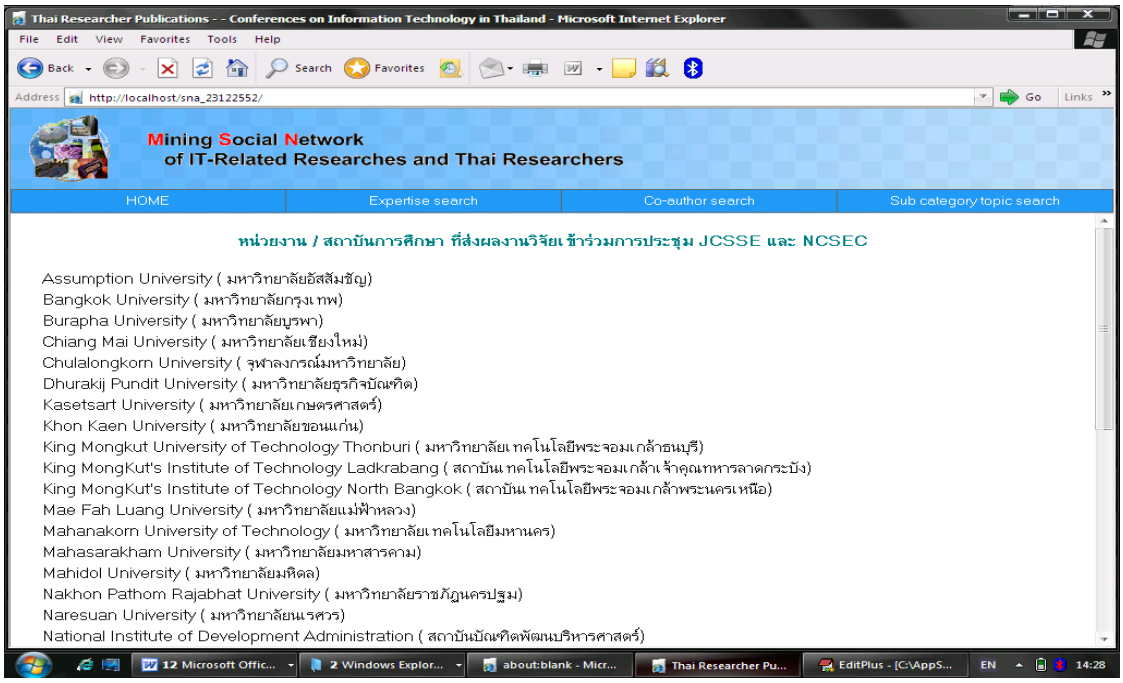



Figure 4.12 Help sub-menu: Organization/university

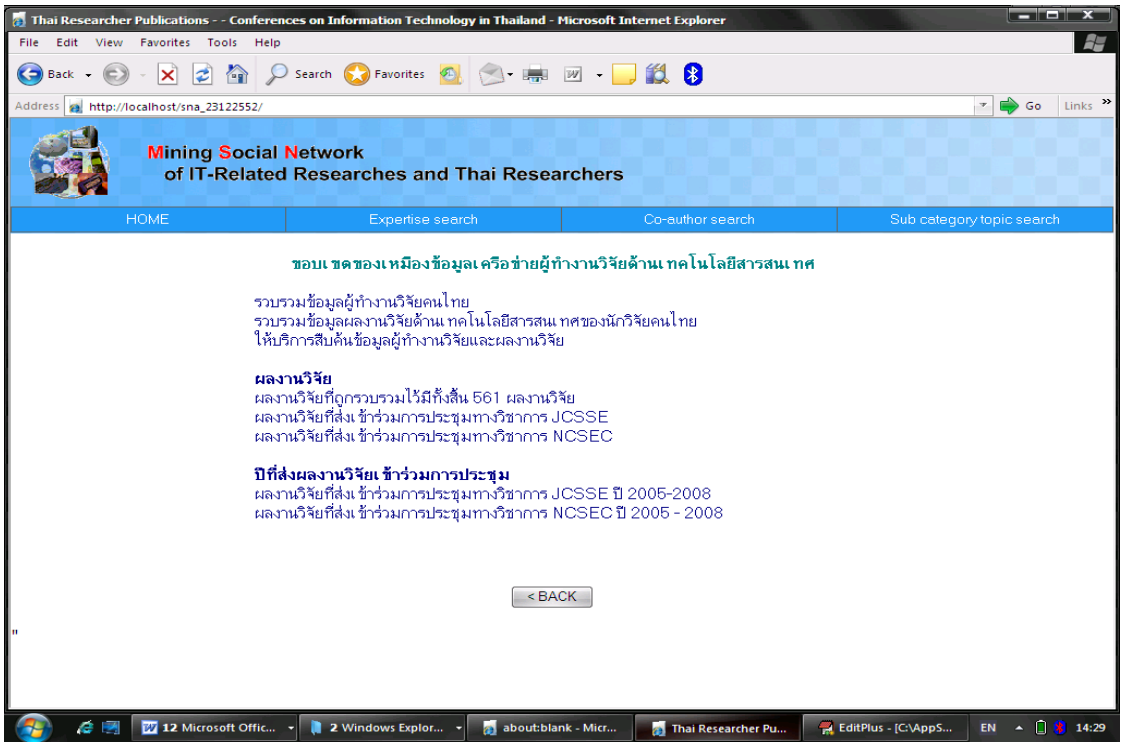



Figure 4.13 Help sub-menu: Scope/year

### 4.2.3.1 Expert Searching

For the development of expert searching part, it is to develop

functions in searching for data of experts in different fields related to information technology, as to which researchers have conducted research in what field.

Expert data searching aims to discover which researcher has done more researches than others in the same field, who have submitted their researches to the conferences JCSSE and NCSEC. An expert searching interface is illustrated in Figure 4.14.

A user can identify data searching conditions which may be general or specific. He or she may search for expert data on the following conditions.

(1) A researcher's agency, either an educational institution or independent agency. A more specific condition may be defined, for example, only experts of which educational institution, faculty or department are required. The system will filter the faculty of particular university, and filter the department of only the faculty selected, and only data existing in the database. For instance, if data of Mahidol University is selected, only data of 2 of its faculties, Science and Engineering Faculties, exist in the system's database. Similarly, departments are filtered corresponding to the faculty chosen, as illustrated in Figure 4.15.

(2) Research topic. An expert may be searched for by picking out from a research group in which each researcher has submitted his/her research for the conference. The system will filter sub category of research based on the research category (group) opted for, for example if a user selects a research category (group) "Artificial Intelligence" as illustrated in Figure 4.16

Subsequently, The system will display a searching result in the form of social network and data view as in Figure 4.17 and 4.18, respectively. The symbols that represent information in network diagram are following:

- Represented the name of researcher.
- Represented researches categories
- 1.0 Represented the relationship between researchers and the research category and the number of researches work.

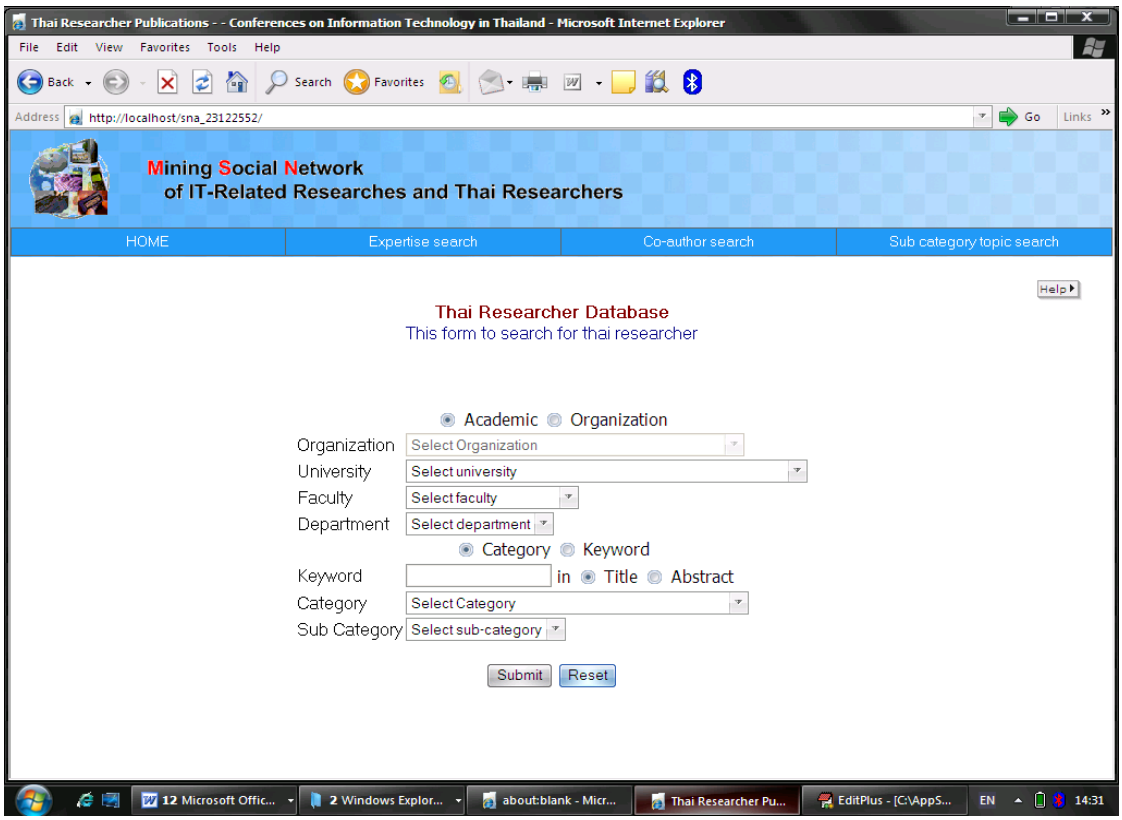



Figure 4.14 An expert searching interface

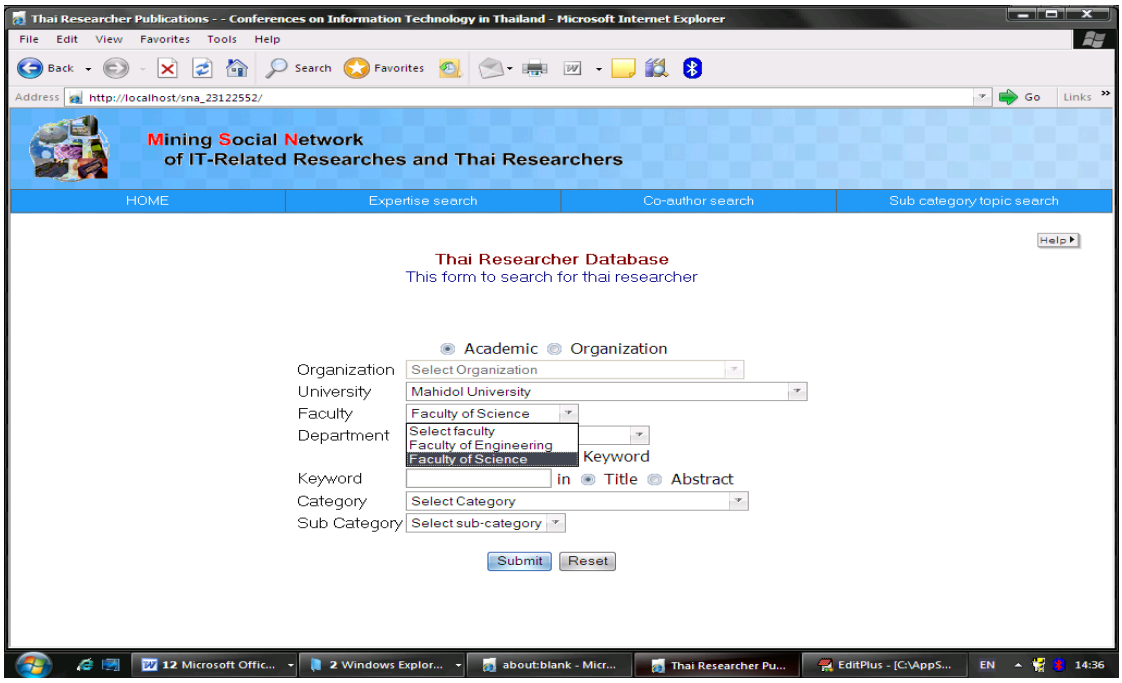



Figure 4.15 An example of filtering faculty and department based on a university Chosen

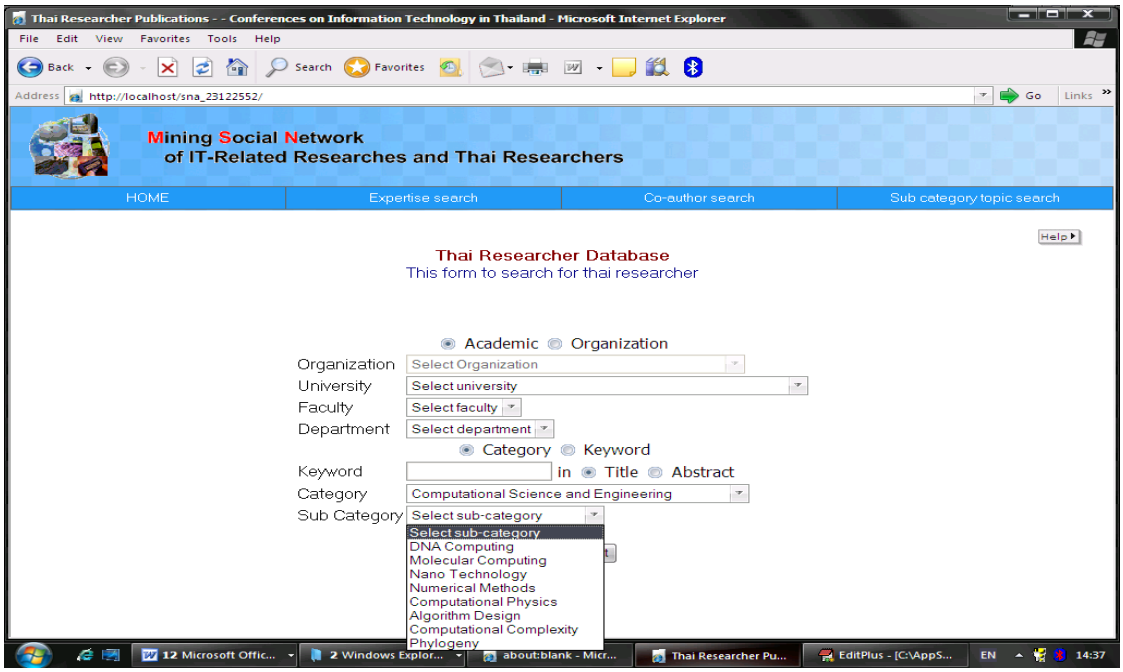



Figure 4.16 An example of filtering research topic sub-category based on research topic category selected

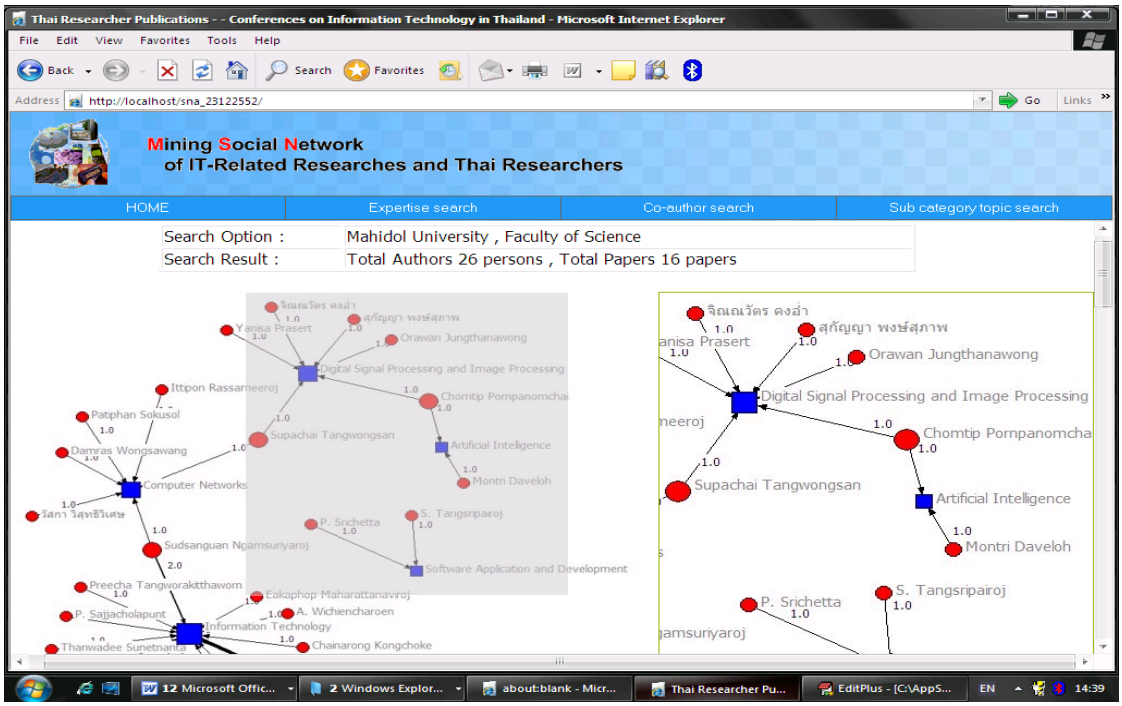
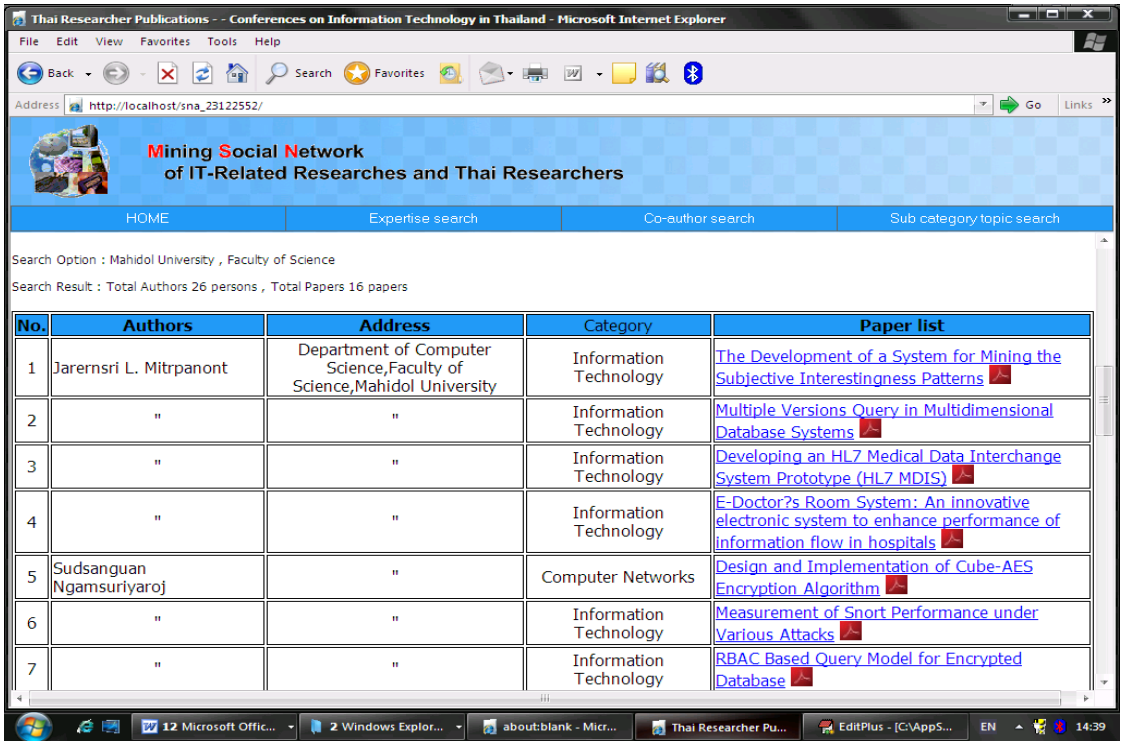



Figure 4.17 An example of expert searching consequence (social network view)




| No. | Authors | Address | Category | Paper list |
|---|---|---|---|---|
| 1 | Jarernsri L. Mitrpanont | Department of Computer Science,Faculty of Science,Mahidol University | Information Technology | The Development of a System for Mining the Subjective Interestingness Patterns |
| 2 | " | " | Information Technology | Multiple Versions Query in Multidimensional Database Systems |
| 3 | " | " | Information Technology | Developing an HL7 Medical Data Interchange System Prototype (HL7 MDIS) |
| 4 | " | " | Information Technology | E-Doctor?s Room System: An innovative electronic system to enhance performance of information flow in hospitals |
| 5 | Sudsanguan Ngamsuriyaroj | " | Computer Networks | Design and Implementation of Cube-AES Encryption Algorithm |
| 6 | " | " | Information Technology | Measurement of Snort Performance under Various Attacks |
| 7 | " | " | Information Technology | RBAC Based Query Model for Encrypted Database |

Figure 4.18 An example of expert searching consequence (data view)

#### 4.2.3.2 Co-author Searching

Concerning the development of co-author searching, it is to develop Functions in searching for data of co-author so as to find out with whom each researcher has conducted research with, and in which field. A co-author searching interface is illustrated in Figure 4.19.

To search for co-author data, a user can search in 2 approaches listed below.

(1) Chose the name of a researcher we want to know his/her co-author. A selection is made from a researcher list.

(2) Identify the keyword of a researcher's name by typing the name of a researcher we need to know his/her co-author.

Then, the system will search for researcher data in the database and when co-author data are found, the outcome will be displayed on an interface shown in Figure 4.20 and 4.21, respectively. In social network view, the symbols that represent information in network diagram are following:

-  Represented a researcher who is interested
-  Represented a co-author of interested researcher
-  Represented the relationship between co-author of interested researcher. It is shown the number of research work.

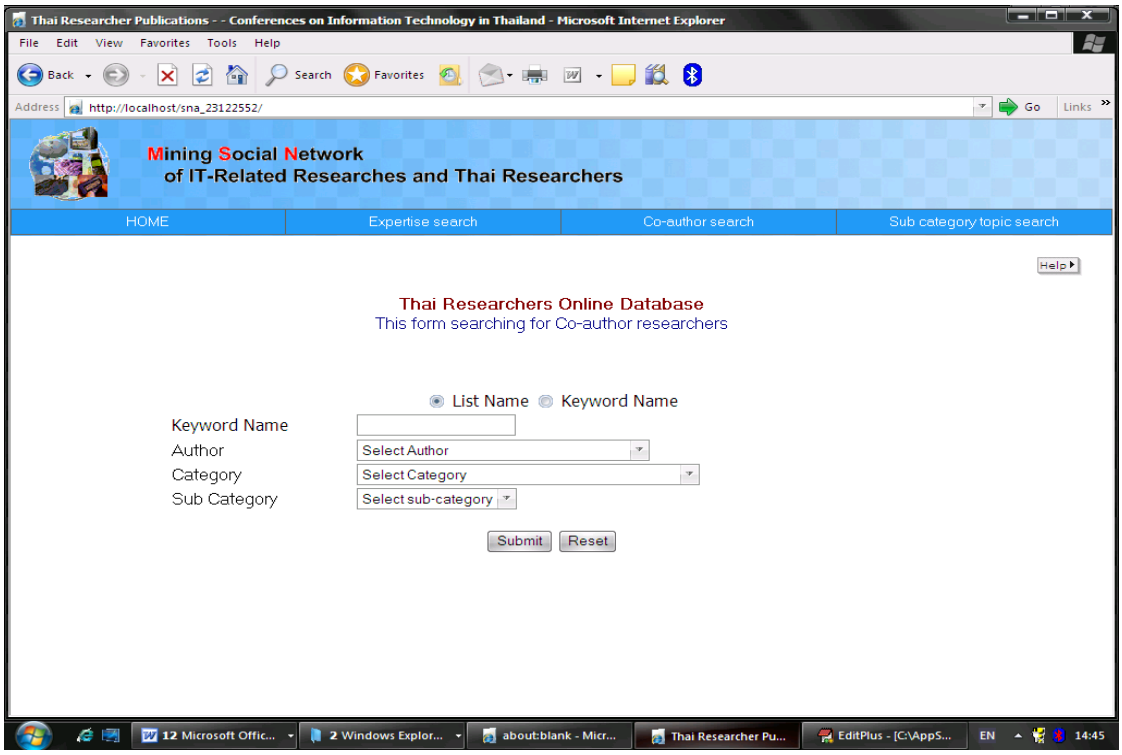



Figure 4.19 A co-author searching interface

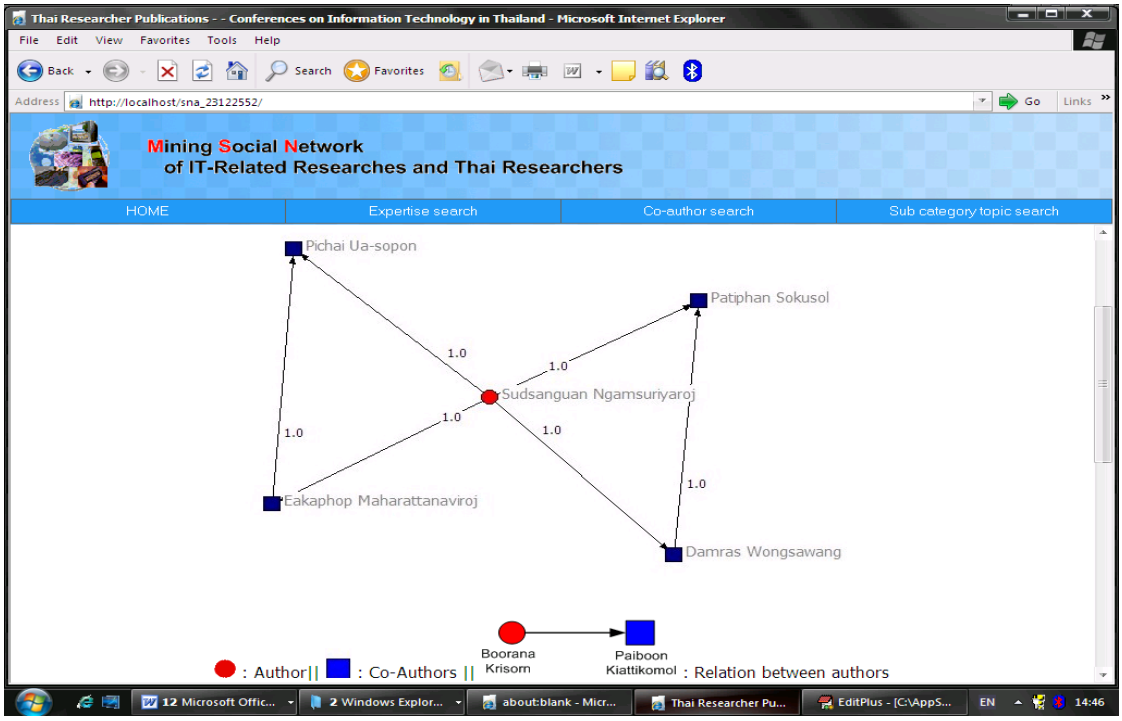



Figure 4.20 An example of co-author searching consequence (social network view)

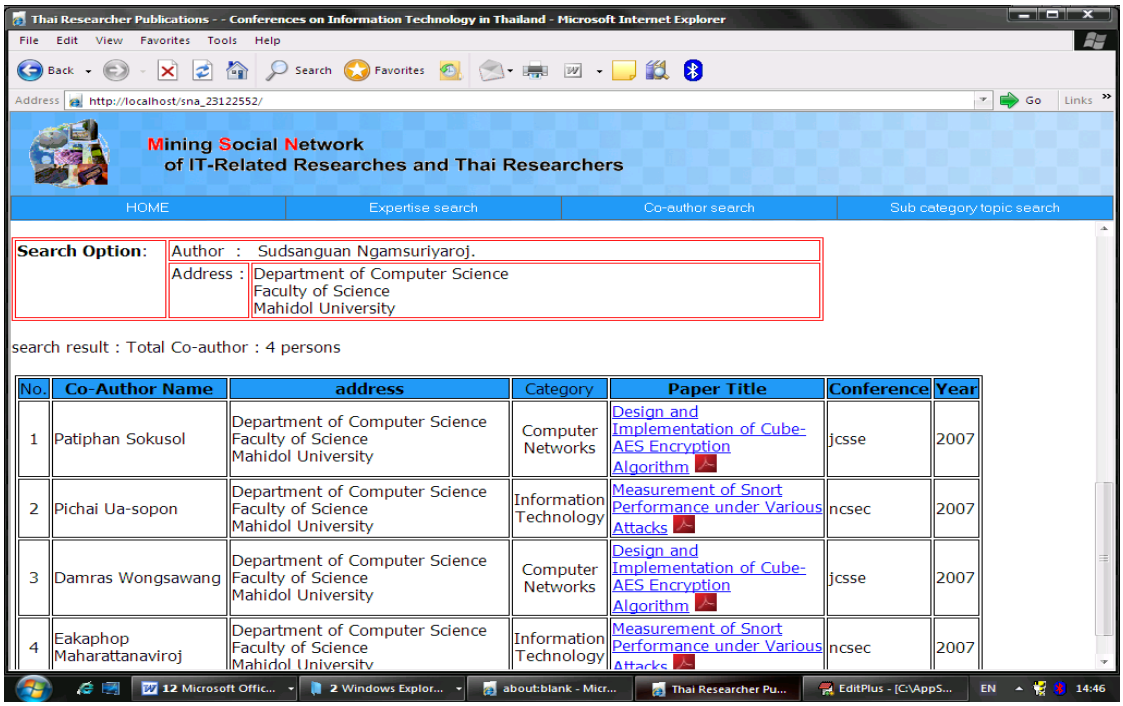



Figure 4.21 An example of co-author searching consequence (social network view)

#### 4.2.3.3 Sub category topic Searching

Regarding, the development of sub category topic searching, it is to develop functions in searching for data of sub category research topics. A sub category topic searching interface is illustrated in Figure 4.22.

To search for sub category topic data, it is to find out, for each conference in each year, what research topics were submitted to the conferences. A user may search through the following conditions.

(1) Identify only the name of the conference, without identifying the year in which the conference was held.

(2) Identify only the year in which the conference was held, without identifying the conference name.

(3) Identify the research topic category and also, research topic sub category.

Later, the system will search for research topic data that have been submitted to the conference, corresponding with the condition chosen by the user, as illustrated in Figure 4.23 and 4.24, respectively. In social network view, the symbols that represent information in network diagram are following:

- Represented research category
- Represented research sub category
- 1.0 Represented the relationship between research category and research sub category. It is shown the number of researches in each sub category, category.

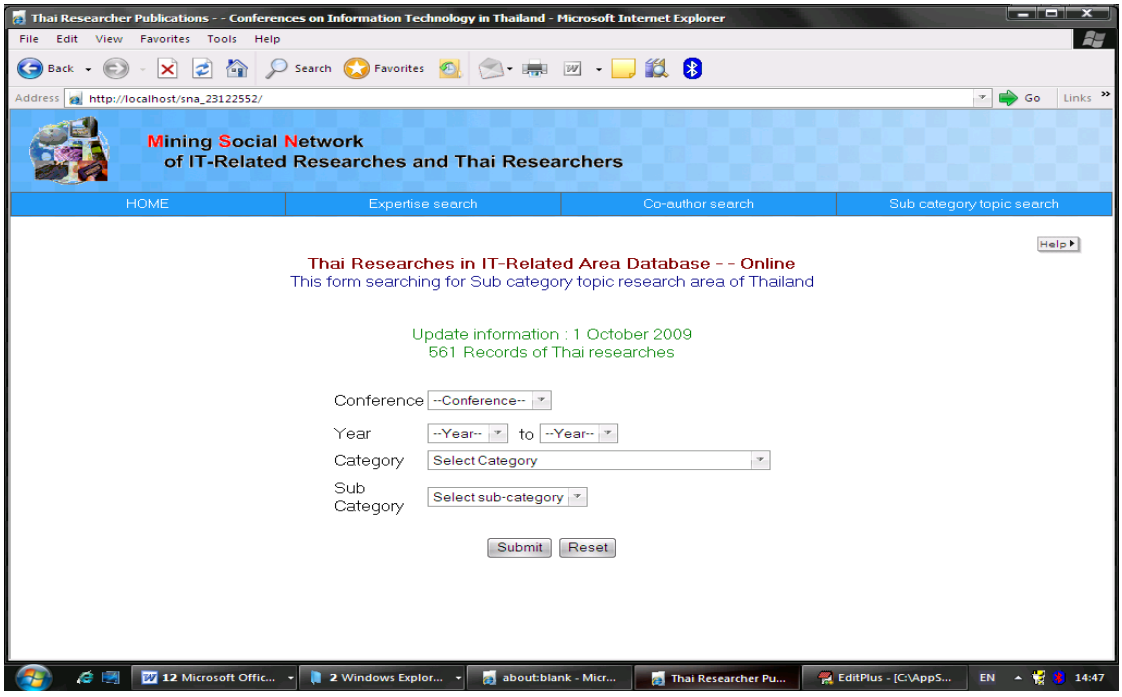



Figure 4.22 A sub category topic searching interface

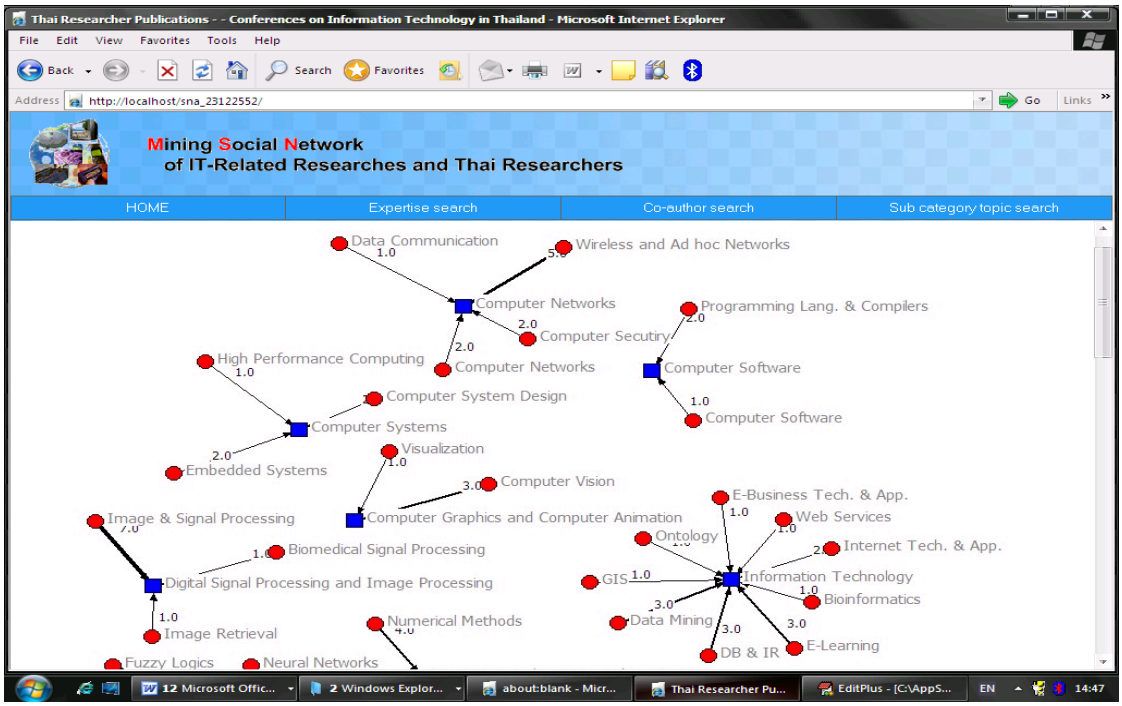
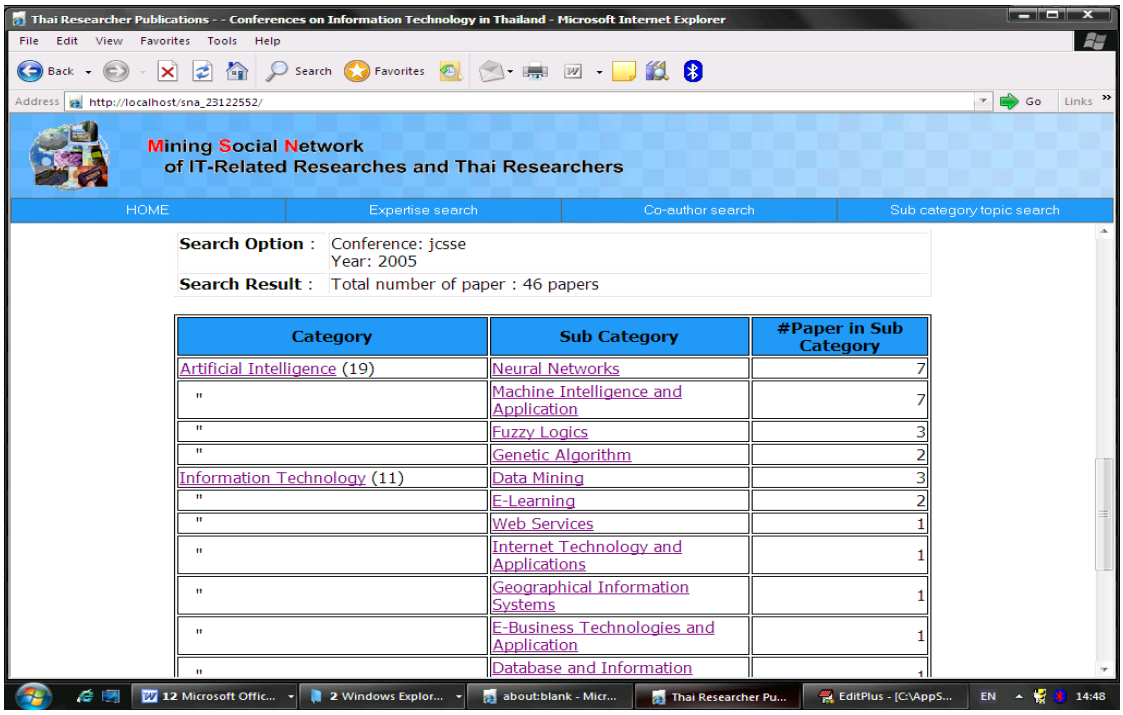



Figure 4.23 An example of sub category topic searching consequence (social network view)




| Category | Sub Category | #Paper in Sub Category |
|---|---|---|
| Artificial Intelligence (19) | Neural Networks | 7 |
| " | Machine Intelligence and Application | 7 |
| " | Fuzzy Logics | 3 |
| " | Genetic Algorithm | 2 |
| Information Technology (11) | Data Mining | 3 |
| " | E-Learning | 2 |
| " | Web Services | 1 |
| " | Internet Technology and Applications | 1 |
| " | Geographical Information Systems | 1 |
| " | E-Business Technologies and Application | 1 |
| " | Database and Information | 1 |

Figure 4.24 An example of sub category topic searching consequence (data view)

# CHAPTER V
# SYSTEM TESTING

This chapter discusses the details of testing the functionality of the prototype system. Details of the testing are divided into 4 parts: (1) Objective and Method (2) Testing Scenarios (3) System Testing Result and (4) Summary and Discussion. Details of each part are described below.

## 5.1 Objectives and Method

There are 2 objectives of prototype system testing. First, to show that when data of research submitted to the conferences JCSSE and NCSEC during 2005 and 2008 are analysed through Social Network Analysis, we will get the result revealing an what research topic category and sub category, related to Computer Science and Information Technology, Thai researchers have conducted research, apart from network of their co-authors and a trend of research topics category and sub category.

Secondly, to display that the searching component of the prototype system can search for a consequence, derived from data analysis via Social Network Analysis, pertinent to a condition needed by a user, and can be also a searching result to the user.

For an approach of testing the prototype system functionality, the functionality of a Social Network Analysis component is tested simultaneously with a searching component. A testing scenario will be defined based on the 3 searching functions, namely, expert searching, co-author searching and research topic category searching.

## 5.2 Test Scenarios

The testing is composed of functionality testing of 3 main functions. To test each searching function, 12 test scenarios are defined as the following:

### 5.2.1 Expert Searching

For expert search testing, 5 test scenarios are set as follows:

(1) Scenario 1 : expert searching by identifying a researcher's agency/academic.

(2) Scenario 2 : expert searching by identifying a researcher's agency that is not an academic.

(3) Scenario 3 : expert searching by identifying the category of research submitted to the conferences.

(4) Scenario 4 : expert searching by identifying the category and sub category of research submitted to the conferences.

(5) Scenario 5 : expert searching by identifying a researcher's agency and the category of research submitted to the conferences.

### 5.2.2 Co-author Searching

For co-author search, 3 test scenarios are set as the following:

(1) Scenario 1 : co-author searching by identifying a researcher's name.

(2) Scenario 2 : co-author searching by identifying a researcher's name and the research topic category.

(3) Scenario 3 : co-author searching by identifying a keyword of researcher's name.

### 5.2.3 Research Topic Category Searching

For research topic category search testing, 4 test scenarios are set as follows:

(1) Scenario 1 : research topic category searching by identifying that all conference in each are to be viewed.

(2) Scenario 2 : research topic category searching by identifying a conference and year.

(3) Scenario 3 : research topic category searching by identifying a conference and a period of the year.

(4) Scenario 4 : research topic category searching by identifying a conference and year, including the research topic category.

## 5.3 Testing Result

For system functionality testing by each test scenario based on the 3 core functions, an outcome of each scenario test is discussed below.

### 5.3.1 Expert Searching Result

(1) Scenario 1: expert searching by identifying a researcher's agency/academic. Testing details are shown in Table 5.1 while search conditions and search results are illustrated in Figures 5.1, 5.2 and 5.3.

Table 5.1 Details of test scenario 1 of expert searching

| Testing item | Testing detail | | results that should be derived |
|---|---|---|---|
| Testing of expert searching. | Choose the menu "Expertise search"<br>Condition for expert searching<br>Case: identifying Academic details | | |
| | 1) Identify data of agency/academic | Academic = "Mahidol University" | Can search for data of Mahidol University's researchers and display searching results on the screen. |
| | 2) Identify data of faculty | Faculty = " " | |
| | 3) Identify data of department | Department = " " | |
| | 4) Identity data of research category | Category = " " | |
| | 5) Identify data of research sub category | Sub category = " " | |
| | 6) Search for data | Press button "submit" | |

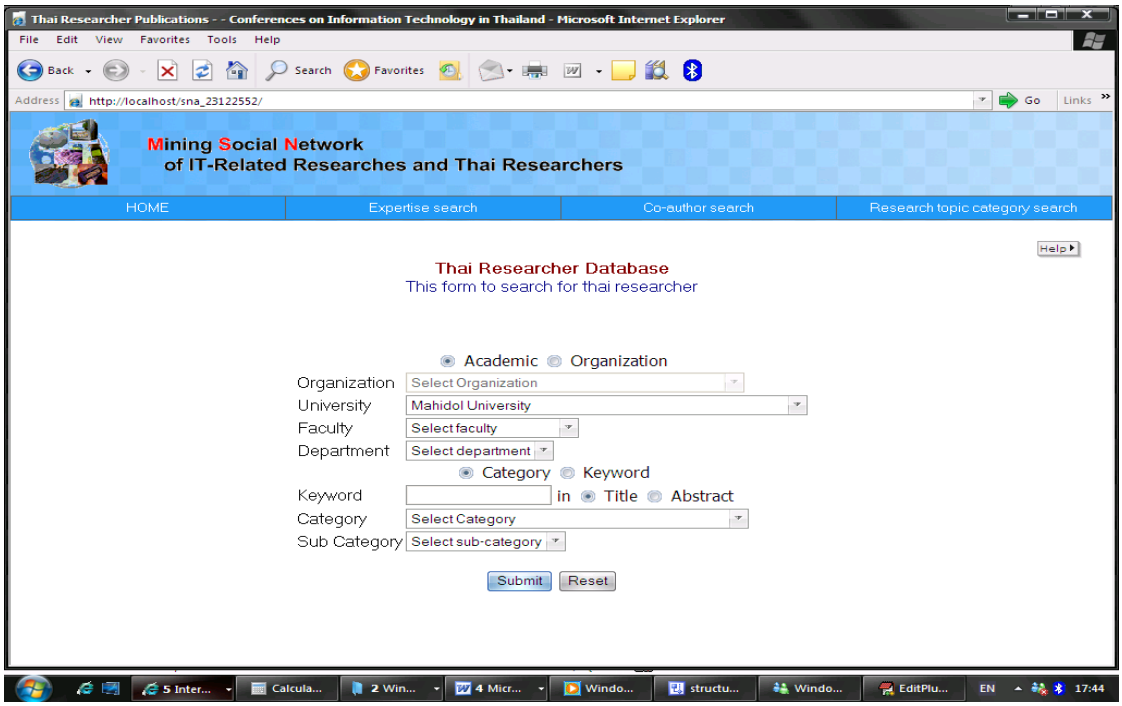



Figure 5.1 Details of expert searching conditions of test scenario 1

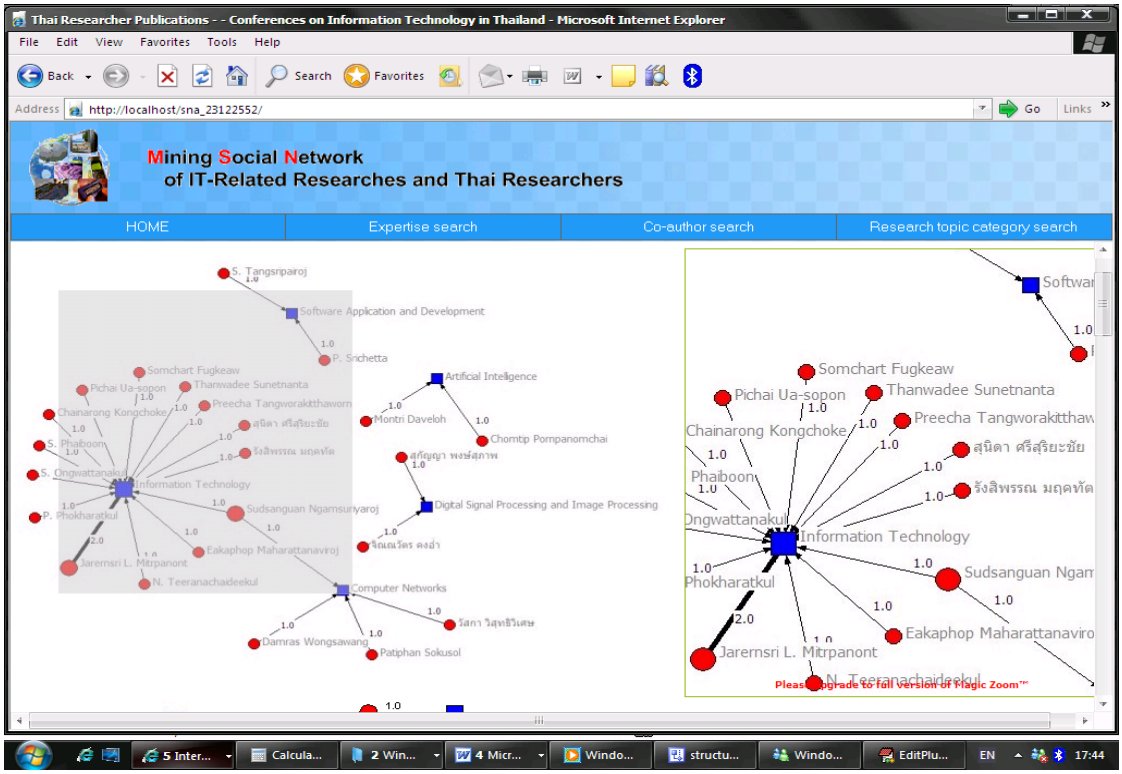
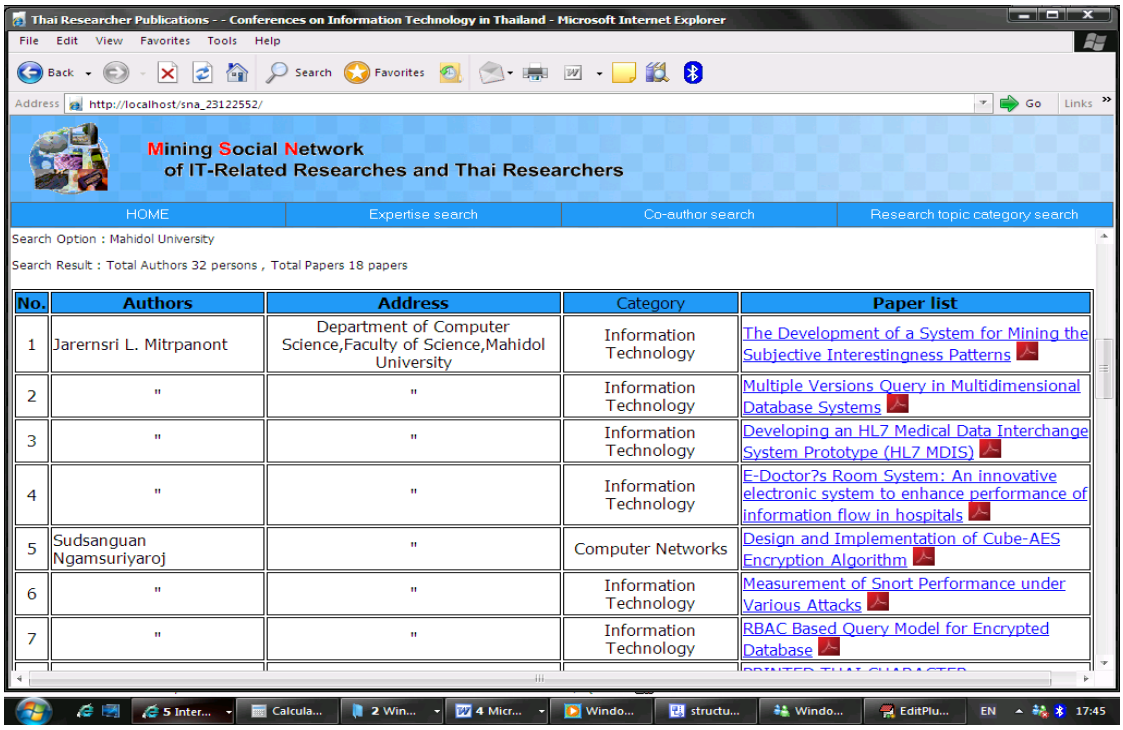



Figure 5.2 An example of results of expert searching based on test scenario 1 (social network view)




| No. | Authors | Address | Category | Paper list |
|---|---|---|---|---|
| 1 | Jarernsri L. Mitrpanont | Department of Computer Science,Faculty of Science,Mahidol University | Information Technology | The Development of a System for Mining the Subjective Interestingness Patterns |
| 2 | " | " | Information Technology | Multiple Versions Query in Multidimensional Database Systems |
| 3 | " | " | Information Technology | Developing an HL7 Medical Data Interchange System Prototype (HL7 MDIS) |
| 4 | " | " | Information Technology | E-Doctor?s Room System: An innovative electronic system to enhance performance of information flow in hospitals |
| 5 | Sudsanguan Ngamsuriyaroj | " | Computer Networks | Design and Implementation of Cube-AES Encryption Algorithm |
| 6 | " | " | Information Technology | Measurement of Snort Performance under Various Attacks |
| 7 | " | " | Information Technology | RBAC Based Query Model for Encrypted Database |

Figure 5.3 An example of results of expert search based on test scenario 1 (data view)

Figures 5.2 and 5.3 illustrate the consequences of expert search testing by identifying a researcher's agency, which is Mahidol University (MU). Figure 5.2 shows a social network diagram of searching results indicating that MU has a total of 32 researchers submitting 18 research papers to the conferences JCSSE and NCSEC. The person possessing the highest number of research papers is Mrs. Jarernsri L. Mitrapanont. MU researchers have conducted research mostly on Information Technology, followed by Computer Network and Artificial Intelligence, respectively. It is also found that Ms. Sudsanguan Ngamsuriyaroj conducts her research in both Information Technology and Computer Network. Figure 5.3 displays a name list of all MU researchers along with each researcher's research topics.

(2) Scenario 2 : expert searching by identifying a researcher's agency which is not an academic. Testing details are revealed in Table 5.2 while searching conditions and outcomes are illustrated in Figures 5.4, 5.5 and 5.6.

Table 5.2 Details of test scenario 2 of expert searching

| Testing item | Testing detail | | results that should be derived |
|---|---|---|---|
| Testing of expert searching. | Choose the menu “Expertise search”<br>Condition for expert searching<br>Case: identifying whether it is an agency or organization that is not an academic. | | |
| | 1) select sub menu | Organization | Can search for data of NECTEC researchers and display searching results on the screen. |
| | 2) Identify an Organization | Organization = “NECTEC” | |
| | 3) Search for data | Press button “submit” | |

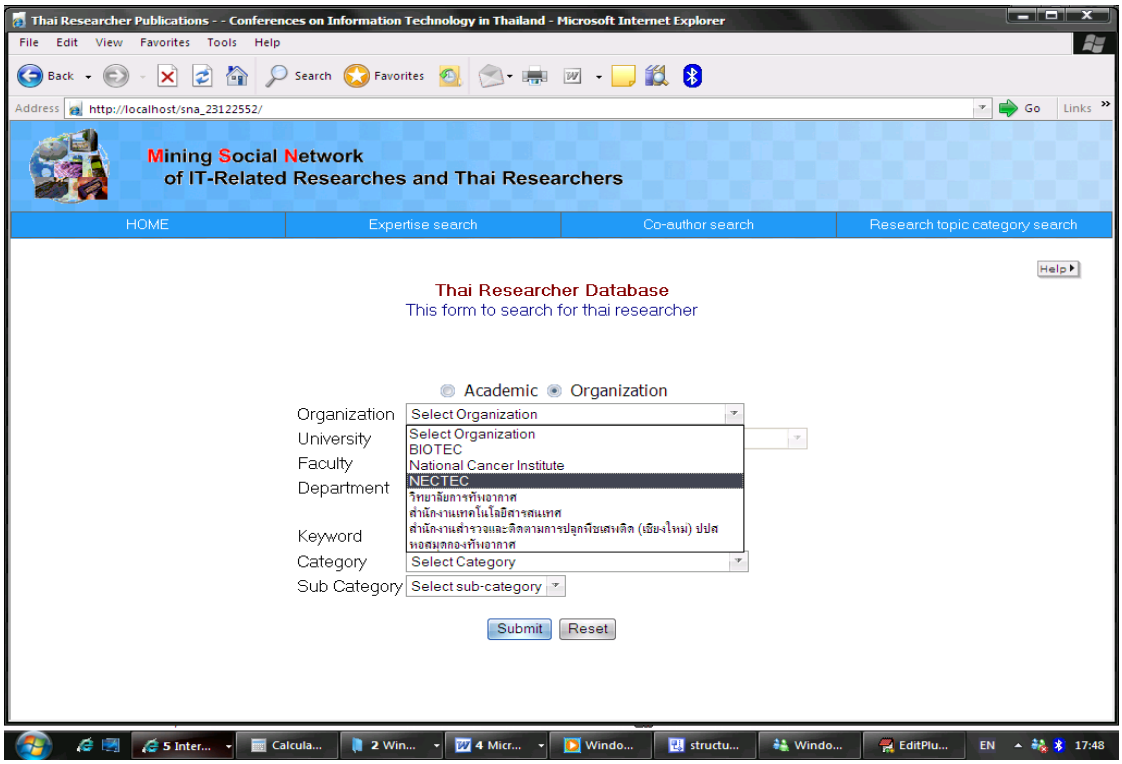



Figure 5.4 Details of expert searching conditions of test scenario 2

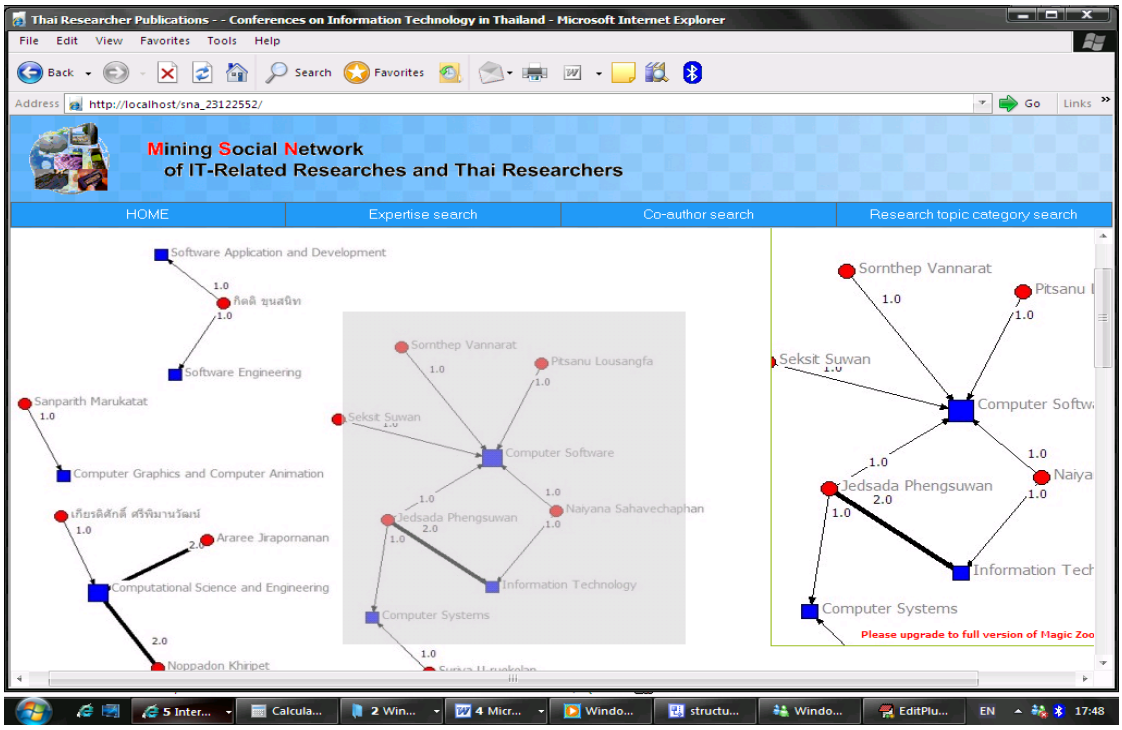
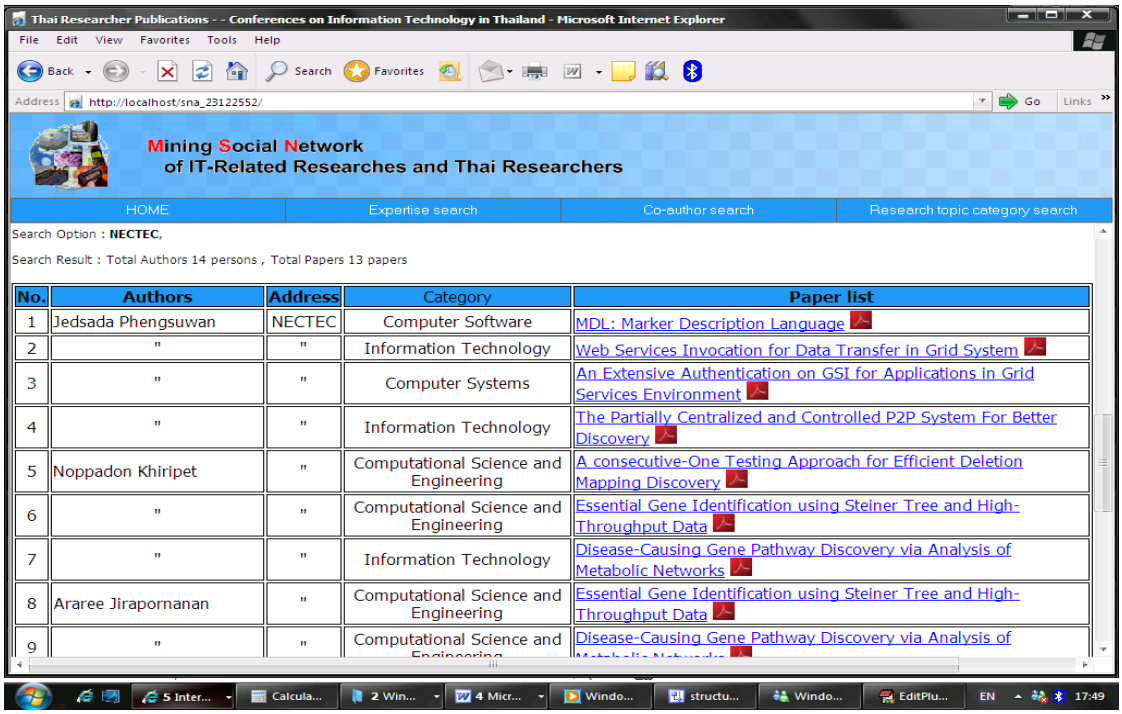



Figure 5.5 An example of results of expert searching based on test scenario 2 (social network view)




| No. | Authors | Address | Category | Paper list |
|---|---|---|---|---|
| 1 | Jedsada Phengsuwan | NECTEC | Computer Software | MDL: Marker Description Language |
| 2 | " | " | Information Technology | Web Services Invocation for Data Transfer in Grid System |
| 3 | " | " | Computer Systems | An Extensive Authentication on GSI for Applications in Grid Services Environment |
| 4 | " | " | Information Technology | The Partially Centralized and Controlled P2P System For Better Discovery |
| 5 | Noppadon Khiripet | " | Computational Science and Engineering | A consecutive-One Testing Approach for Efficient Deletion Mapping Discovery |
| 6 | " | " | Computational Science and Engineering | Essential Gene Identification using Steiner Tree and High-Throughput Data |
| 7 | " | " | Information Technology | Disease-Causing Gene Pathway Discovery via Analysis of Metabolic Networks |
| 8 | Araree Jirapornanan | " | Computational Science and Engineering | Essential Gene Identification using Steiner Tree and High-Throughput Data |
| 9 | " | " | Computational Science and | Disease-Causing Gene Pathway Discovery via Analysis of |

Figure 5.6 An example of results of expert searching based on test scenario 2 (data view)

Figures 5.5 and 5.6 illustrate the outcomes of expert search testing by identifying a researcher's agency, NECTEC, which is an independent one. Figure 5.5 shows a social network diagram of searching outcomes indicating that NECTEC has a total of 14 researchers submitting 13 research papers to the conferences JCSSE and NCSEC. Mr. Jedsada Pengsuwan is NECTEC researcher having the most number of research papers. NECTEC has produced a variety of researches such as Information Technology, Computer System, Computer, Computer Graphic, etc. Nevertheless, research on Computational Science and Engineering and Computer Software ranks the first and second, respectively. Figure 5.6 display a name list of all NECTEC researchers as well as each researcher's research topics.

(3) Scenario 3: expert searching by identifying the category of research submitted to the conferences. Testing details are given in Table 5.3 while searching condition and searching results are illustrated in Figures 5.7, 5.8 and 5.9.

Table 5.3 Details of test scenario 3 of expert searching

| Testing item | Testing detail | | results that should be derived |
|---|---|---|---|
| Testing of expert searching. | Choose the menu "Expertise search"<br>Condition for expert searching<br>Case: identifying category and sub category of research. | | |
| | 1) Identify data of agency/academic | Academic = " " | |
| | 2) Identify data of faculty | Faculty = " " | |

Table 5.3 Details of test scenario 3 of expert searching (cont.)

| Testing item | Testing detail | | results that should be derived |
|---|---|---|---|
| | 3) Identify data of department | Department = " " | Can search for data of researchers working on the research category "Artificial Intelligence", and reveal searching outcome on the screen. |
| | 4) Identity data of research category | Category= "Artificial Intelligence" | |
| | 5) Identify data of research sub category | Sub category = " " | |
| | 6) Search for data | Press button "submit" | |

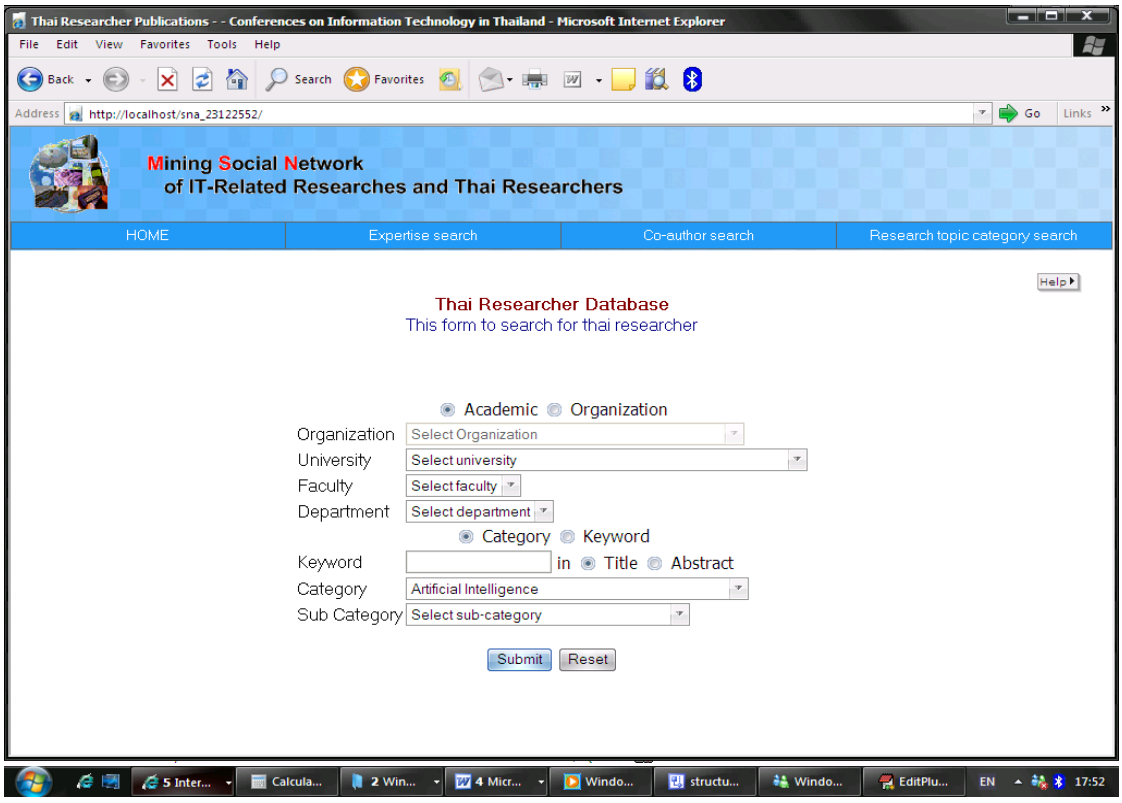



Figure 5.7 Details of expert searching conditions of test scenario 3

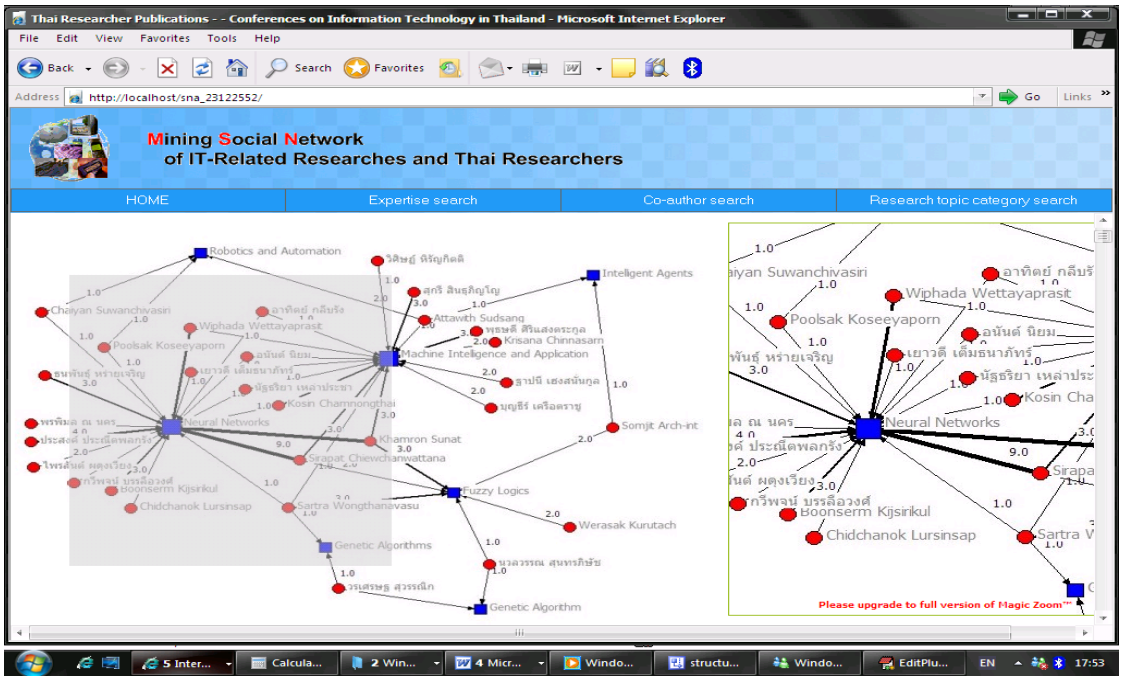



Figure 5.8 An example of results of expert searching based on test scenario 3 (social network view)

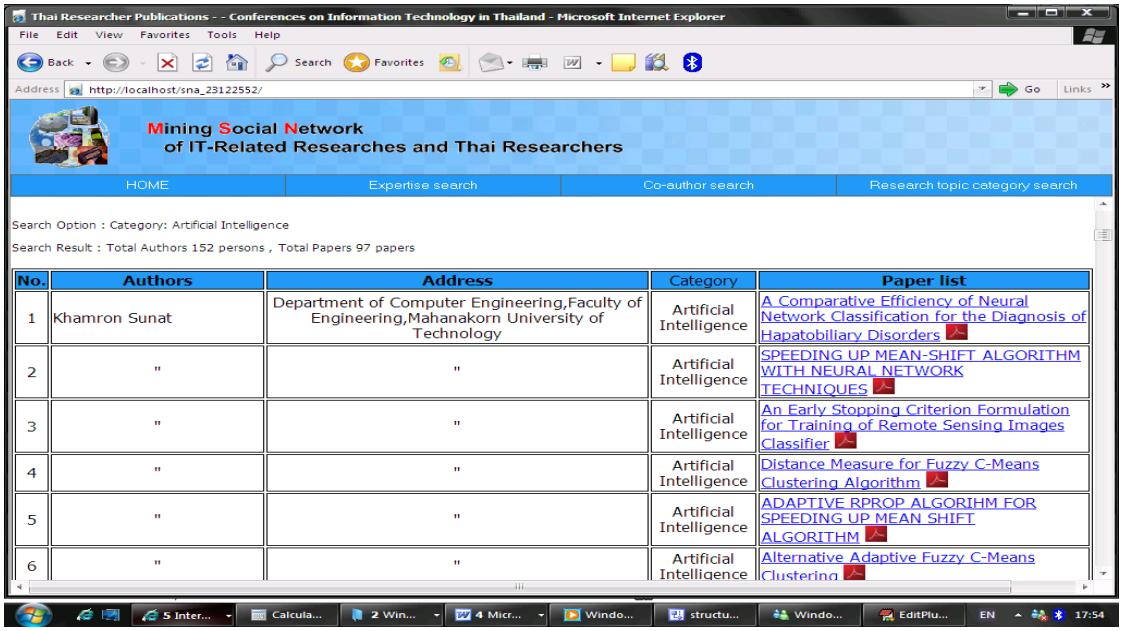



Figure 5.9 An example of results of expert searching based on test scenario 3 (data view)

Figures 5.8 and 5.9 illustrate the results of expert search testing without identifying a researcher's agency, but identifying the research topic category, Artificial Intelligence, only. Figure 5.8 reveals a social network diagram of searching results indicating that there are a total of 152 Thai researchers doing research on Artificial Intelligence while 52 research papers in this field have been submitted to the conferences JCSSE and NCSEC. Out of this number, research on neural networks has been conducted the most, followed by Machine Intelligence and Application. Other research is on Fuzzy logic, Genetic Algorithm, etc. It has also been found that Mr. Khamron Sunat has the highest number of research papers in this field (category). Figure 5.9 shows a name list of researchers doing research on Artificial Intelligence, including each researcher's research topics.

(4) Scenario 4 : expert searching by identifying the research topic category and sub category. Testing details are listed in Table 5.4 while searching conditions and searching outcomes are illustrated in Figures 5.10, 5.11 and 5.12.

Table 5.4 Details of test scenario 4 of expert searching

| Testing item | Testing detail | | results that should be derived |
|---|---|---|---|
| Testing of expert searching. | Choose the menu "Expertise search"<br>Condition for expert searching<br>Case: identifying category and sub category of research. | | |
| | 1) Identify data of agency/academic | Academic = " " | |
| | 2) Identify data of faculty | Faculty = " " | |

Table 5.4 Details of test scenario 4 of expert searching (cont.)

| Testing item | Testing detail | | results that should be derived |
|---|---|---|---|
| Testing of expert searching. | 3) Identify data of department | Department = " " | Can search for data of researchers doing research in the category "Artificial Intelligence" and sub category "Neural Network", as well as display searching consequences on the screen. |
| | 4) Identity data of research category | Category = "Artificial Intelligence" | |
| | 5) Identify data of research sub category | Sub category = "Neural Network" | |
| | 6) Search for data | Press button "submit" | |

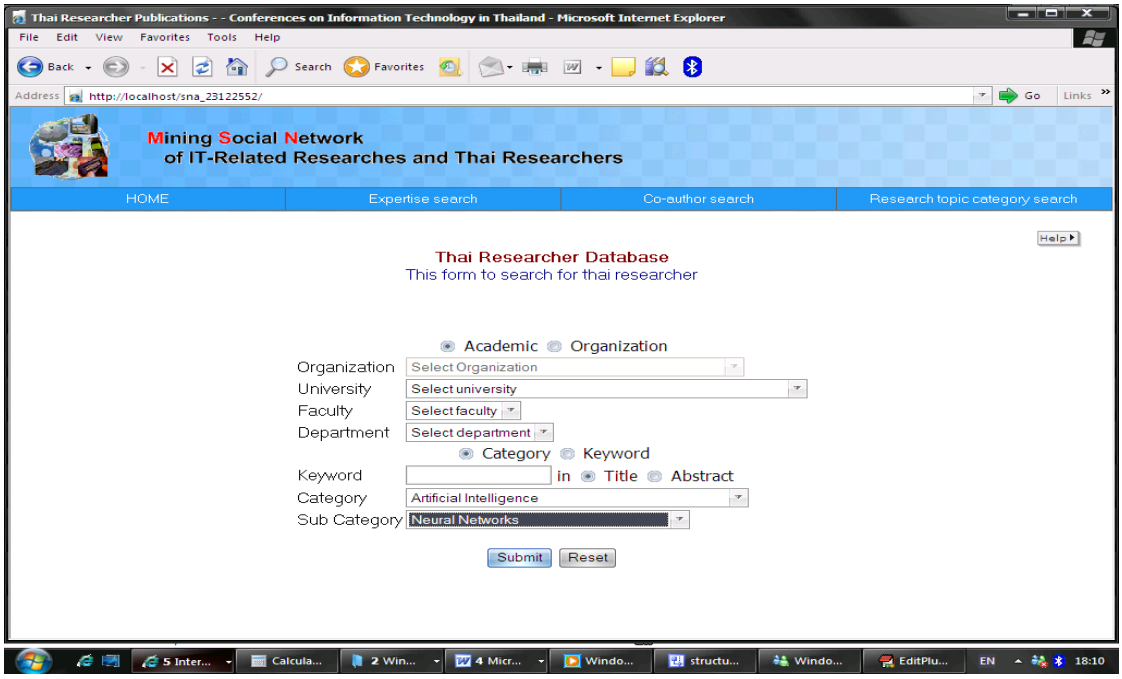



Figure 5.10 Details of expert searching conditions of test scenario 4

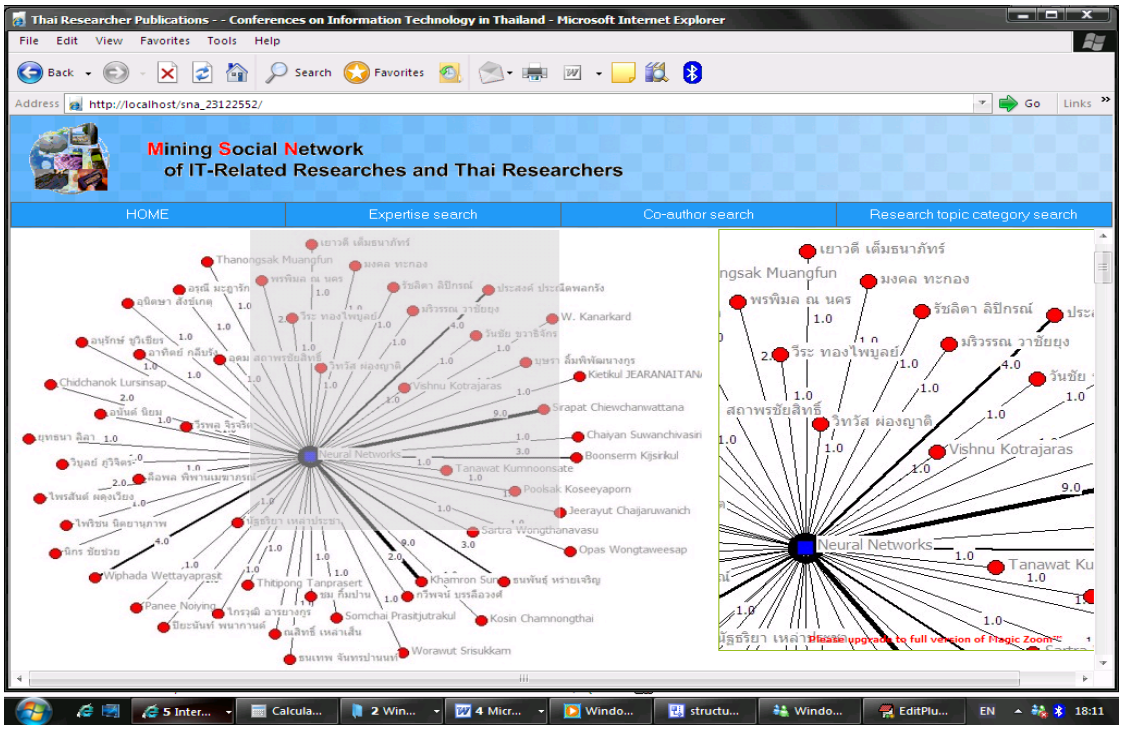
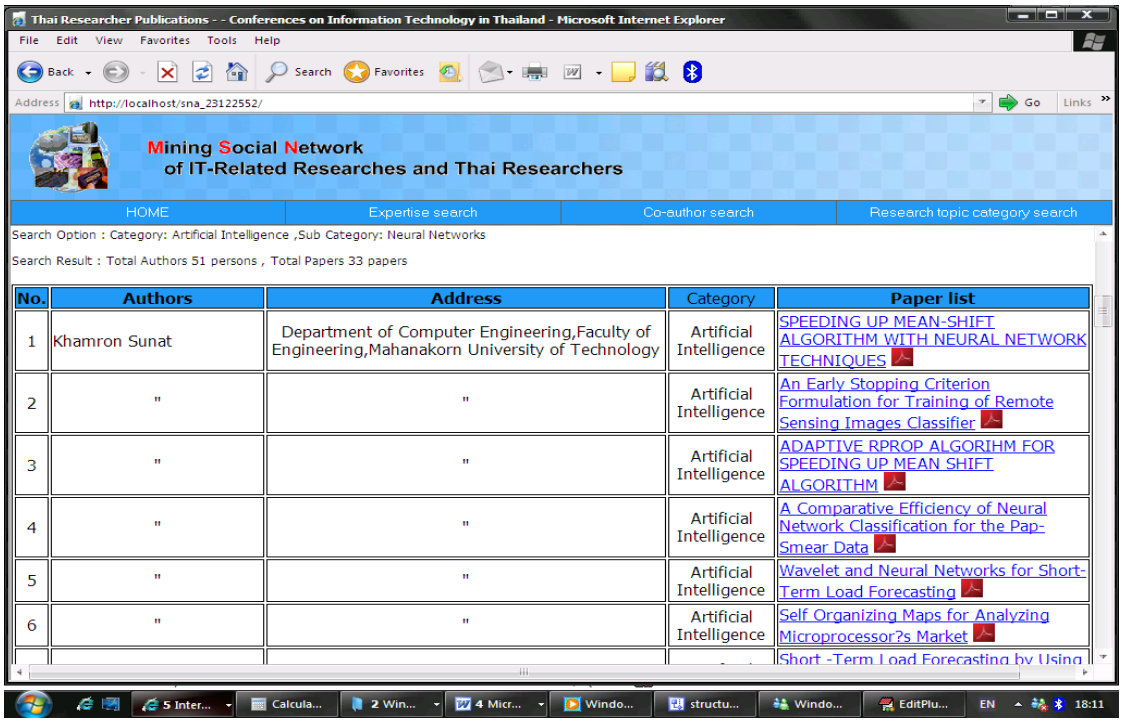



Figure 5.11 An example of results of expert searching based on test scenario 4 (social network view)




| No. | Authors | Address | Category | Paper list |
|---|---|---|---|---|
| 1 | Khamron Sunat | Department of Computer Engineering,Faculty of Engineering,Mahanakorn University of Technology | Artificial Intelligence | SPEEDING UP MEAN-SHIFT ALGORITHM WITH NEURAL NETWORK TECHNIQUES |
| 2 | " | " | Artificial Intelligence | An Early Stopping Criterion Formulation for Training of Remote Sensing Images Classifier |
| 3 | " | " | Artificial Intelligence | ADAPTIVE RPROP ALGORIHM FOR SPEEDING UP MEAN SHIFT ALGORITHM |
| 4 | " | " | Artificial Intelligence | A Comparative Efficiency of Neural Network Classification for the Pap-Smear Data |
| 5 | " | " | Artificial Intelligence | Wavelet and Neural Networks for Short-Term Load Forecasting |
| 6 | " | " | Artificial Intelligence | Self Organizing Maps for Analyzing Microprocessor?s Market |

Figure 5.12 An example of results of expert searching based on test scenario 4 (data view)

Figures 5.11 and 5.12 illustrate the consequences of expert search testing with out identifying a researcher's agency/academic but identifying the research category, Artificial Intelligence, and sub category, Neural Network, only. Figure 5.11 shows a social network diagram of searching consequences indicating that there are a total of 51 researchers submitted 33 research papers on "Artificial Intelligence" in the sub category "Neural Network" to the conferences JCSSE and NCSEC. It is found that Mr. Khamron Sunat of Mahanakorn University has conducted the most research in this field (category). Figure 5.12 displays a name list of all researchers doing research on "Artificial Intelligence" in the sub category "Neural Network", including each researcher's research topics.

(5) Scenario 5 : expert searching by identifying a researcher's agency and the category of research submitted to the conference. Testing details are given in Table 5.5 while searching results are illustrated in Figures 5.13, 5.14 and 5.15.

Table 5.5 Details of test scenario 5 of expert searching

| Testing item | Testing detail | | results that should be derived |
|---|---|---|---|
| Testing of expert searching. | Choose the menu "Expertise search"<br>Condition for expert searching<br>Case: identifying category and sub category of research. | | |
| | 1) Identify data of agency/academic | Academic = "Mahidol University" | |
| | 2) Identify data of faculty | Faculty = "Faculty of Science" | |

Table 5.5 Details of test scenario 5 of expert searching (cont.)

| Testing item | Testing detail | | results that should be derived |
|---|---|---|---|
| | 3) Identify data of department | Department = "Department of Computer Science" | Can search for data of researchers conducting research in the category "Artificial Intelligence" and attached to Mahidol University's department of Computer Science, Faculty of Science, and show searching results on the screen. |
| | 4) Identity data of research category | Category = "Artificial Intelligence" | |
| | 5) Identify data of research sub category | Sub category = " " | |
| | 6) Search for data | Press button "submit" | |

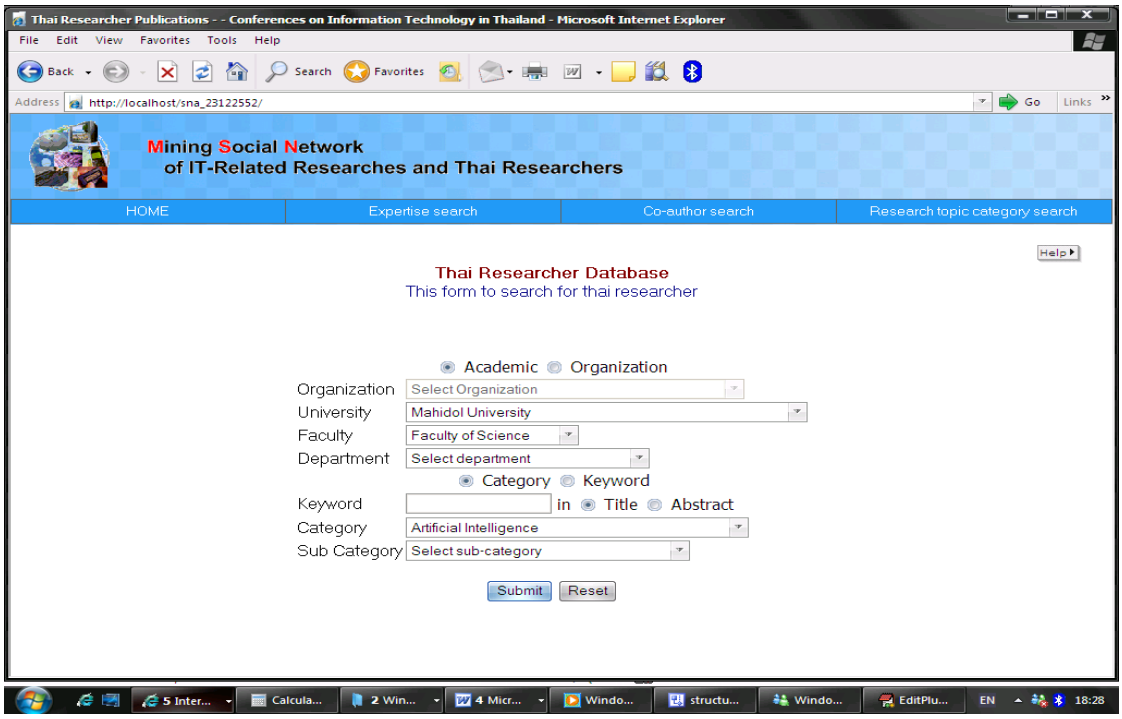



Figure 5.13 Details of expert searching conditions of test scenario 5

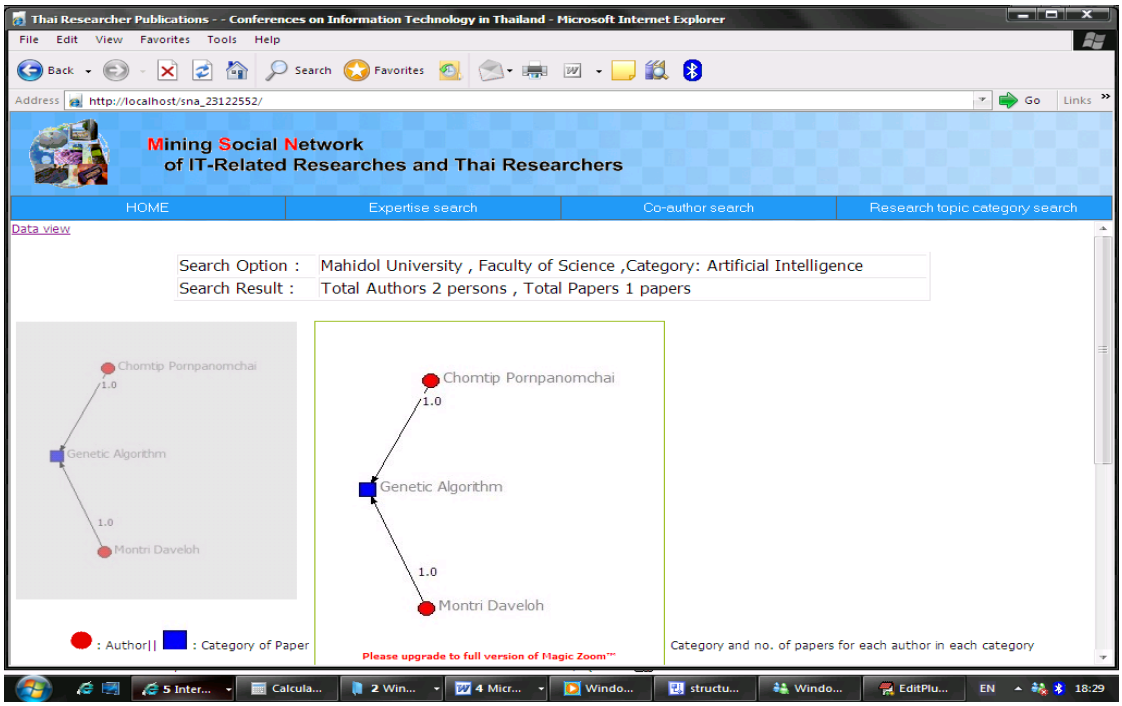



Figure 5.14 An example of results of expert searching based on test scenario 5 (social network view)

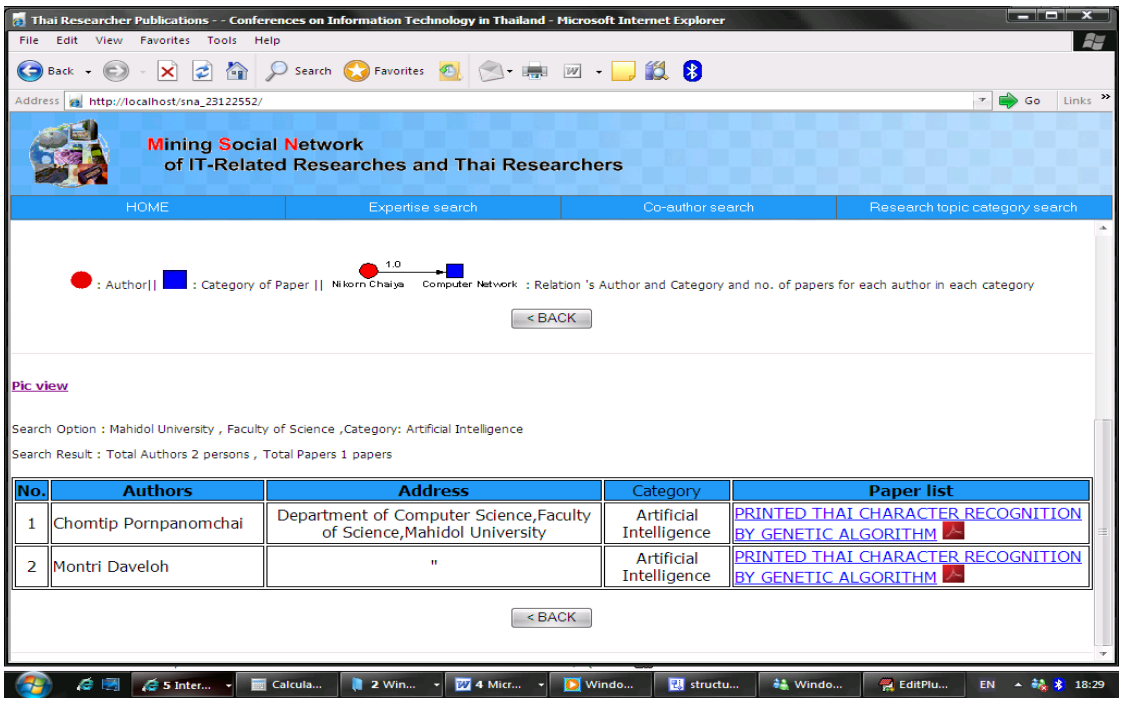



Figure 5.15 An example of results of expert searching based on test scenario 5 (data view)

Figures 5.14 and 5.15 illustrate the results of expert search testing by identifying a researcher's agency, or Mahidol University's Department of Computer Science, Faculty of Science, including the research topic category "Artificial Intelligence" Figure 5.14 demonstrates a social network diagram of searching results indicating that there are a total of 2 researchers of MU's Faculty of Science, Department of Computer Science submitting 1 research on "Artificial Intelligence" to the conferences JCSSE and NCSEC. It is found that Mr. Chomtip Pompanomchai and Mr. Montri Develoh have conducted research in this field. Figure 5.15 shows a name list of MU researchers doing research on "Artificial Intelligence", and each researcher's research topic.

### 5.2.2 Co-author Searching Result

(1) Scenario 1 : co-author searching by identifying a researcher's name. testing details are given in Table 5.6 while searching conditions and searching results are illustrated in Figures 5.16, 5.17 and 5.18.

Table 5.6 Details of test scenario 1 of co-author searching

| Testing item | Testing detail | | results that should be derived |
|---|---|---|---|
| Testing of co-author searching. | Choose the menu "Co-author search"<br>Conditions for co-author searching<br>Case: select a name of researcher from the researcher name list, by selecting option: list name | | |
| | 1) Choose the name of researcher | Researcher name = "Athasit Surarerks" | Can search for data of a co-author of the researcher "Athasit Surarerks"and display searching results on the screen. |
| | 2) Identify data of research category | Category = " " | |

Table 5.6 Details of test scenario 2 of co-author searching (cont.)

| Testing item | Testing detail | | results that should be derived |
|---|---|---|---|
| | 3) Identify data of research sub category | Sub category = “ ” | |
| | 4) Search for data | Press button “submit” | |

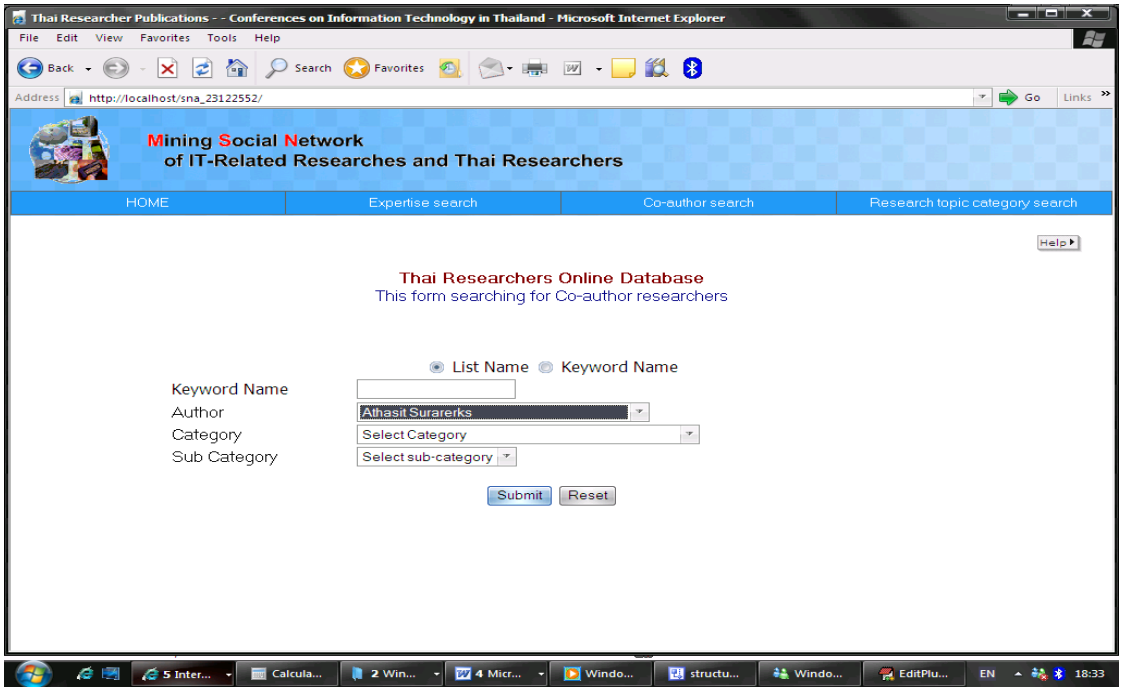



Figure 5.16 Details of co-author searching conditions of test scenario 1

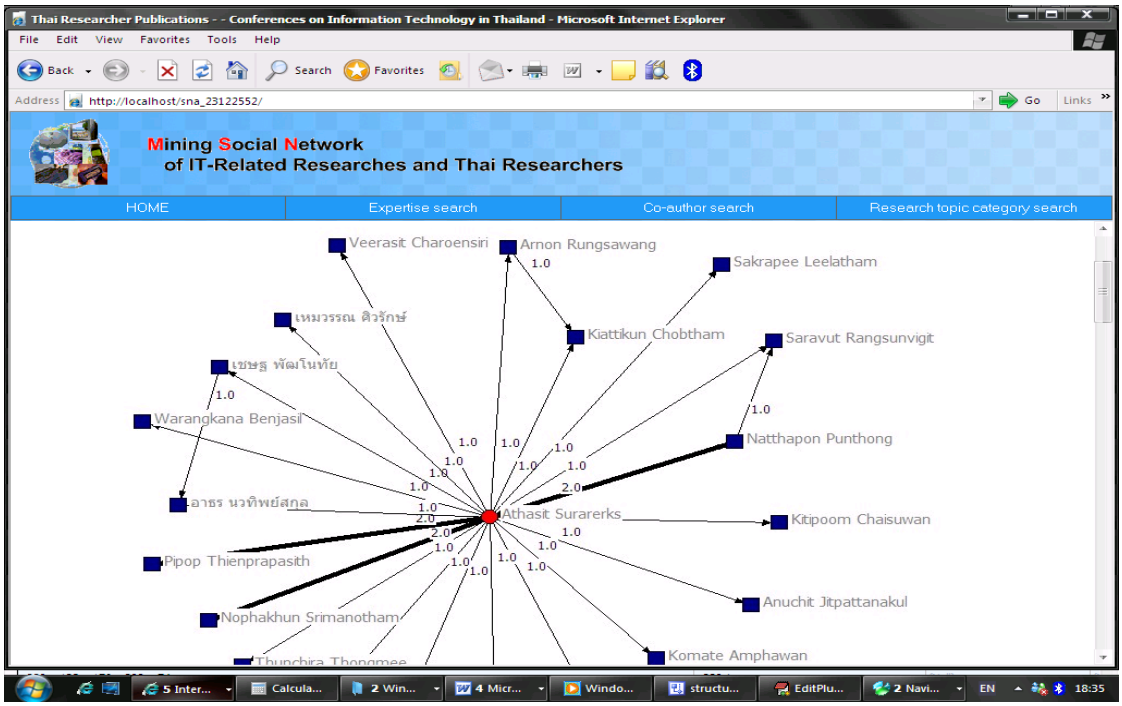
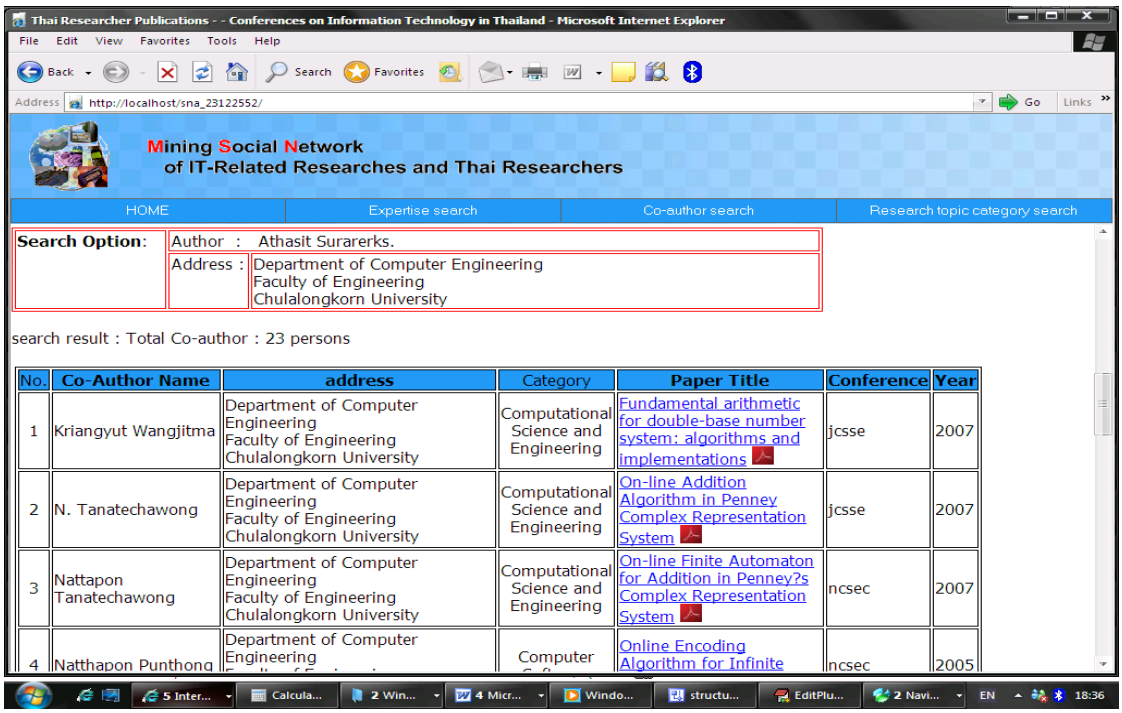



Figure 5.17 An example of results of co-author searching based on test scenario 1 (social network view)




| No. | Co-Author Name | address | Category | Paper Title | Conference | Year |
|---|---|---|---|---|---|---|
| 1 | Kriangyut Wangjitma | Department of Computer Engineering Faculty of Engineering Chulalongkorn University | Computational Science and Engineering | Fundamental arithmetic for double-base number system: algorithms and implementations | jcsse | 2007 |
| 2 | N. Tanatechawong | Department of Computer Engineering Faculty of Engineering Chulalongkorn University | Computational Science and Engineering | On-line Addition Algorithm in Penney Complex Representation System | jcsse | 2007 |
| 3 | Nattapon Tanatechawong | Department of Computer Engineering Faculty of Engineering Chulalongkorn University | Computational Science and Engineering | On-line Finite Automaton for Addition in Penney?s Complex Representation System | ncsec | 2007 |
| 4 | Natthapon Punthong | Department of Computer Engineering | Computer | Online Encoding Algorithm for Infinite | ncsec | 2005 |

Figure 5.18 An example of results of co-author searching based on test scenario 1 (data view)

Figures 5.17 and 5.18 illustrate the results of co-author search testing by identifying the researcher name “Athasit Surarerks”. Figure 5.17 shows a social network diagram of searching results indicating that Mr. Athasit Surarerks has a total of 23 co-authors, who mostly submit 1 research paper to the conferences JCSSE and NCSEC. Nonetheless, there are some co-authors submitting 2 research papers to the conferences, such as Mr. Natthapon Punthong. Figure 5.18 displays a name list of Mr. Athasit Surarerks’s co-authors and their research topics, including the name list of conferences the researches were submitted to.

(2) Scenario 2 : co-author searching by identifying a researcher’s name and the research category. Testing details are shown in Table 5.7 while searching conditions and searching outcomes are illustrated in Figures 5.19, 5.20 and 5.21.

Table 5.7 Details of test scenario 2 of co-author searching

<table>
<tr><th>Testing item</th><th colspan="2">Testing detail</th><th>results that should be derived</th></tr>
<tr><td>Testing of co-author searching.</td><td colspan="2">Choose the menu “Co-author search”<br>Conditions for co-author searching<br>Case: select a name of researcher from the researcher name list and identify research category.</td><td></td></tr>
<tr><td rowspan="4"></td><td>1) Choose the name of researcher</td><td>Researcher name = “ Athasit Surarerks”</td><td rowspan="4">Can search for data of co-authors of the researcher " Athasit Surarerks" working on the research category “Computer Software”, and show searching result on the screen.</td></tr>
<tr><td>2) Identify data of research category</td><td>Category = “Computer Software”</td></tr>
<tr><td>3) Identify data of research sub category</td><td>Sub category = “ ”</td></tr>
<tr><td>4) Search for data</td><td>Press button “submit”</td></tr>
</table>

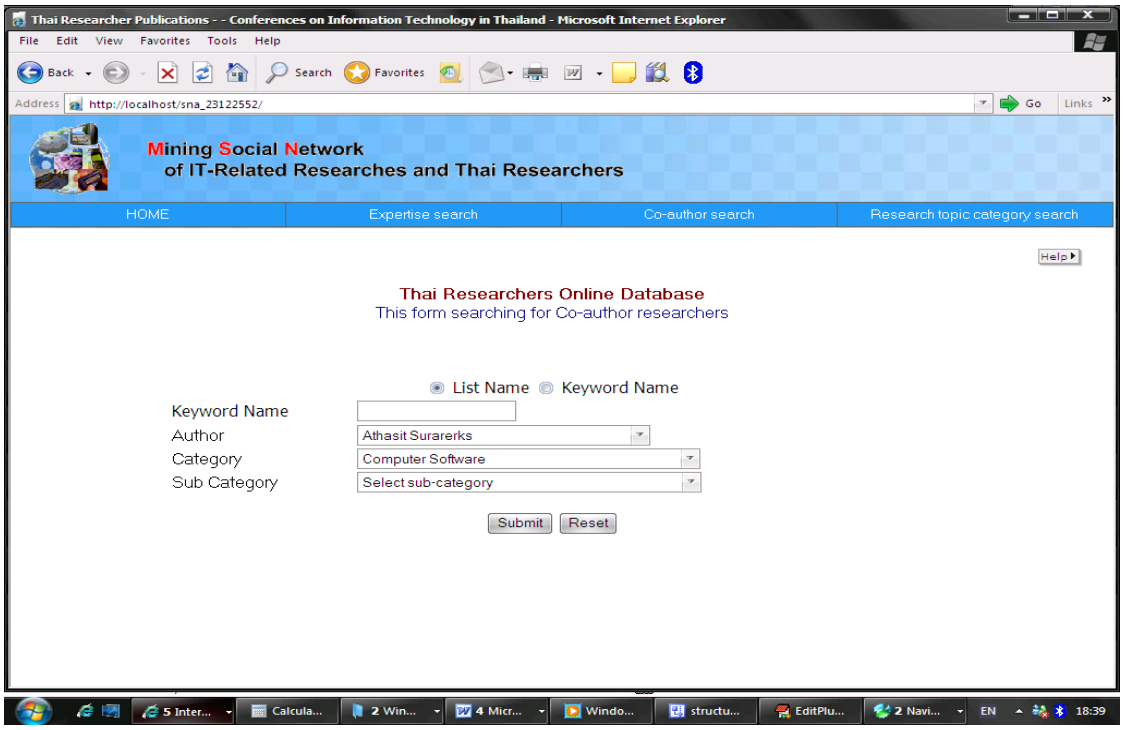



Figure 5.19 Details of co-author searching conditions of test scenario 2

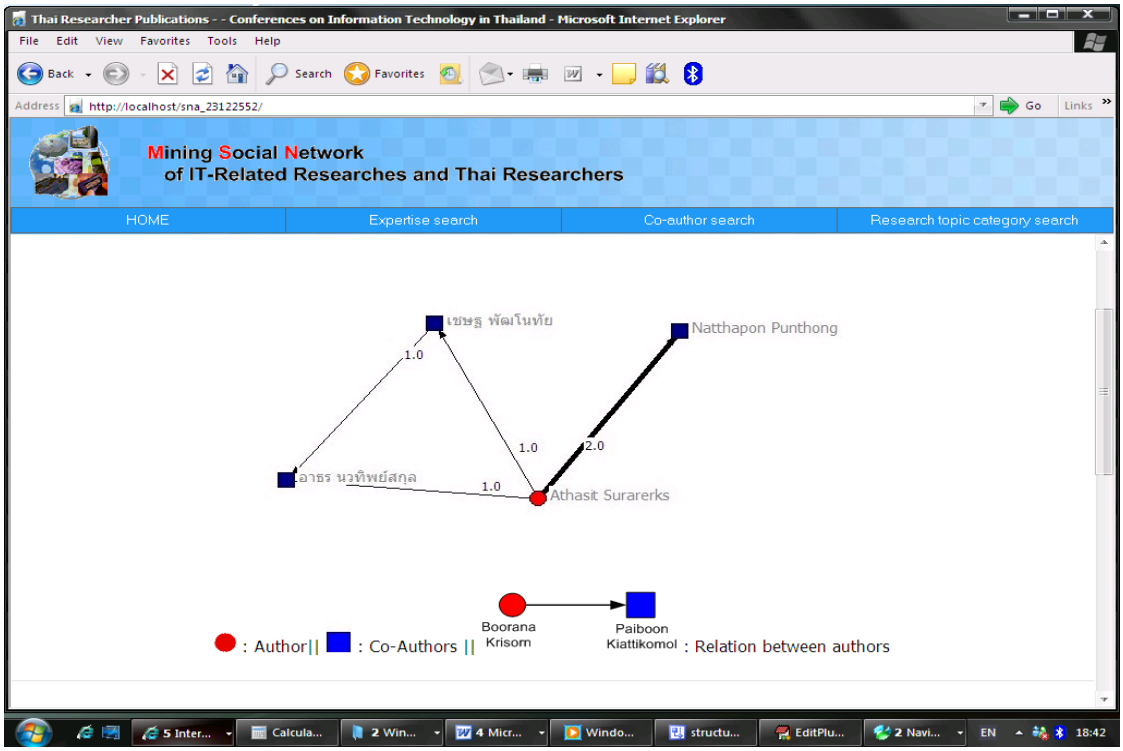



Figure 5.20 An example of results of co-author searching based on test scenario 2 (social network view)

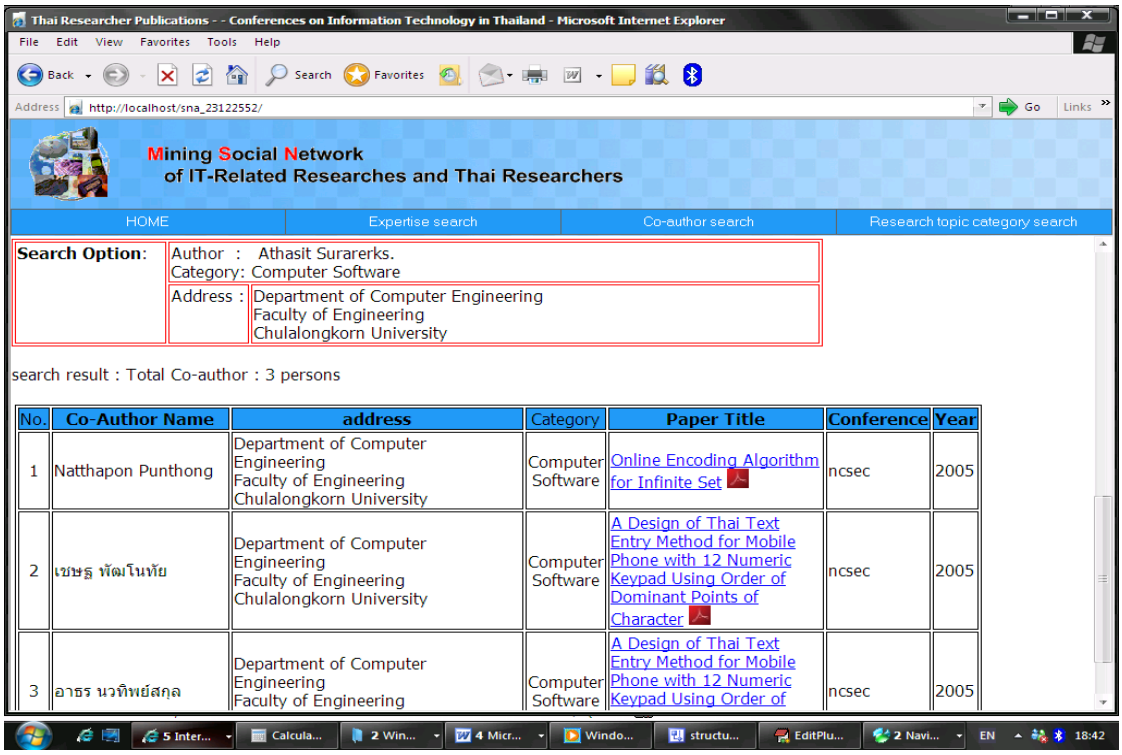



Figure 5.21 An example of results of co-author searching based on test scenario 2 (data view)

Figures 5.20 and 5.21 illustrate the outcomes of co-author search testing by identifying the researcher name “Athasit Surarerks” and the research topic category “Computer Software”. Figure 5.20 displays a social network diagram of searching outcomes indicating that Mr. Athasit Surarerks has a total of 3 co-authors conducting research on Computer Software, who have submitted their research to the conferences JCSSE and NCSEC. Their names are Mr. Natthapon Puthong, Mr. เชษฐ พัฒโนทัย and Mr. อาธร นวทิพย์สกุล. It is also found that Mr. Natthapon Puthong has done 2 research papers in cooperation with Mr. Athasit Surarerks’s co-authors and research topics on Computer Software, including years and names of conferences the researches were submitted to.

(3) Scenario 3 : co-author searching by identifying the keyword of researcher names. Testing details are given in Table 5.8 while searching conditions and searching results are illustrated in Figures 5.22, 5.23, 5.24 and 5.25.

Table 5.8 Details of test scenario 3 of co-author searching

| Testing item | Testing detail | | results that should be derived |
|---|---|---|---|
| Testing of co-author searching. | Choose the menu “Co-author search” Conditions for co-author searching Case: identify the researcher name by choosing option: keyword name. | | |
| | 1) Choose the name of researcher | Researcher name = “กร” | Can search for data of co-authors of the researcher whose name has the part “กร”, including displaying searching outcomes on the screen. |
| | 2) Identify data of research category | Category = “ ” | |
| | 3) Identify data of research sub category | Sub category = “ ” | |
| | 4) Search for data | Press button “submit” | |

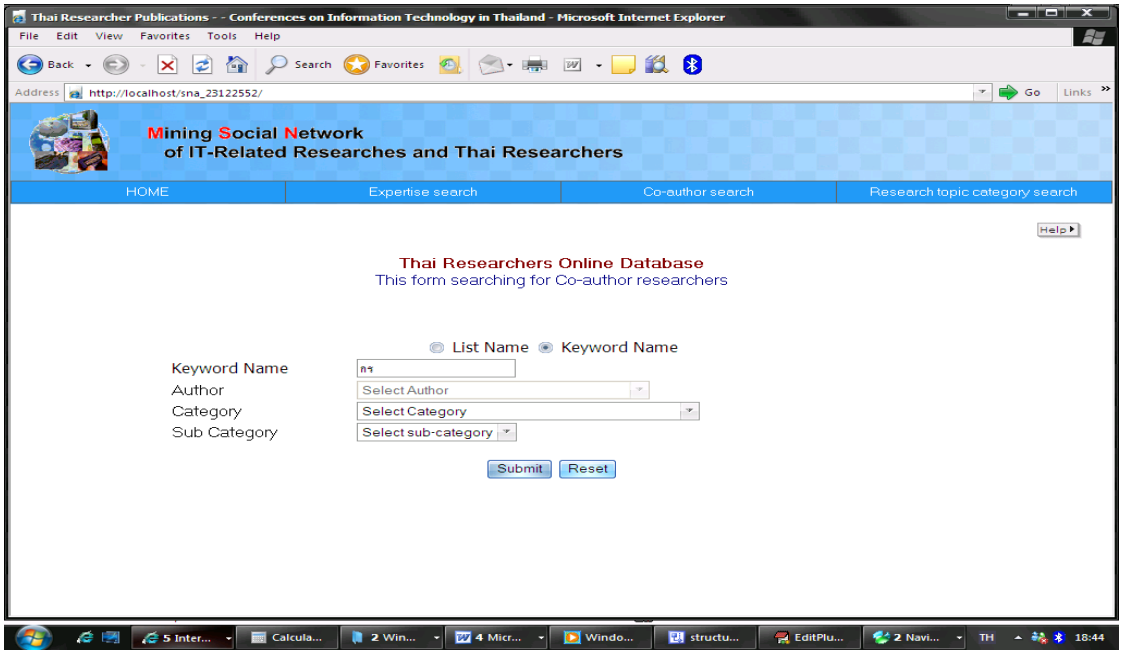



Figure 5.22 Details of co-author searching conditions of test scenario 3

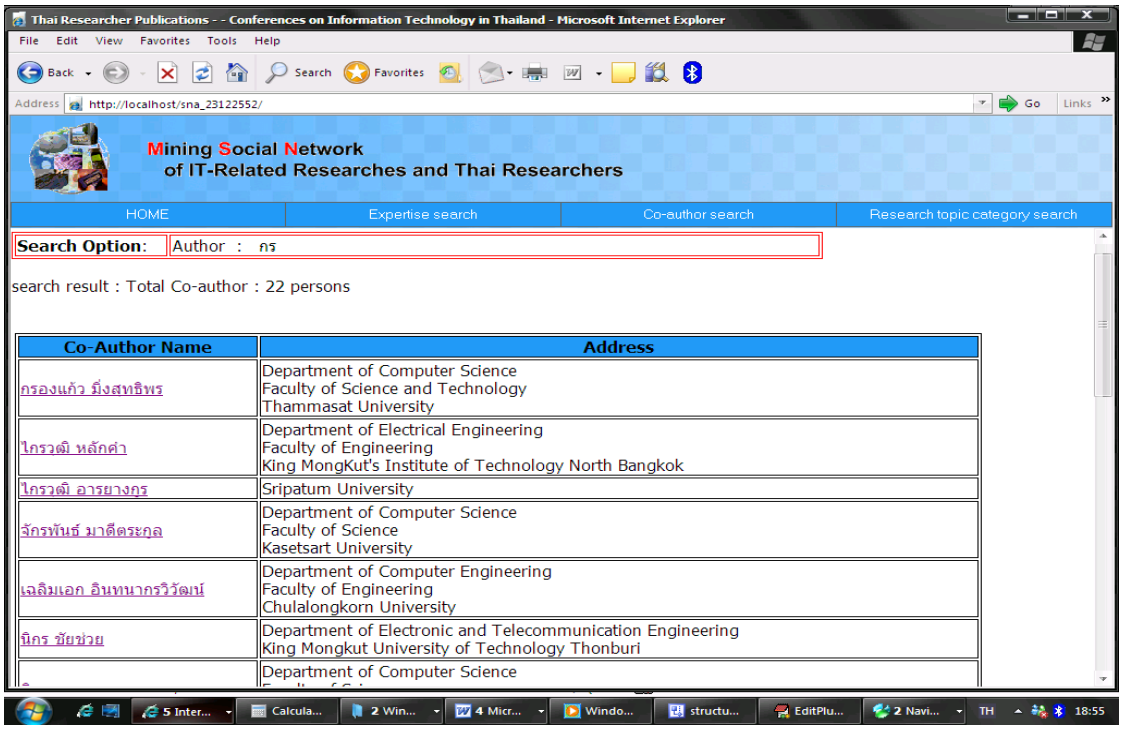



Figure 5.23 An example of results of co-author searching based on test scenario 3

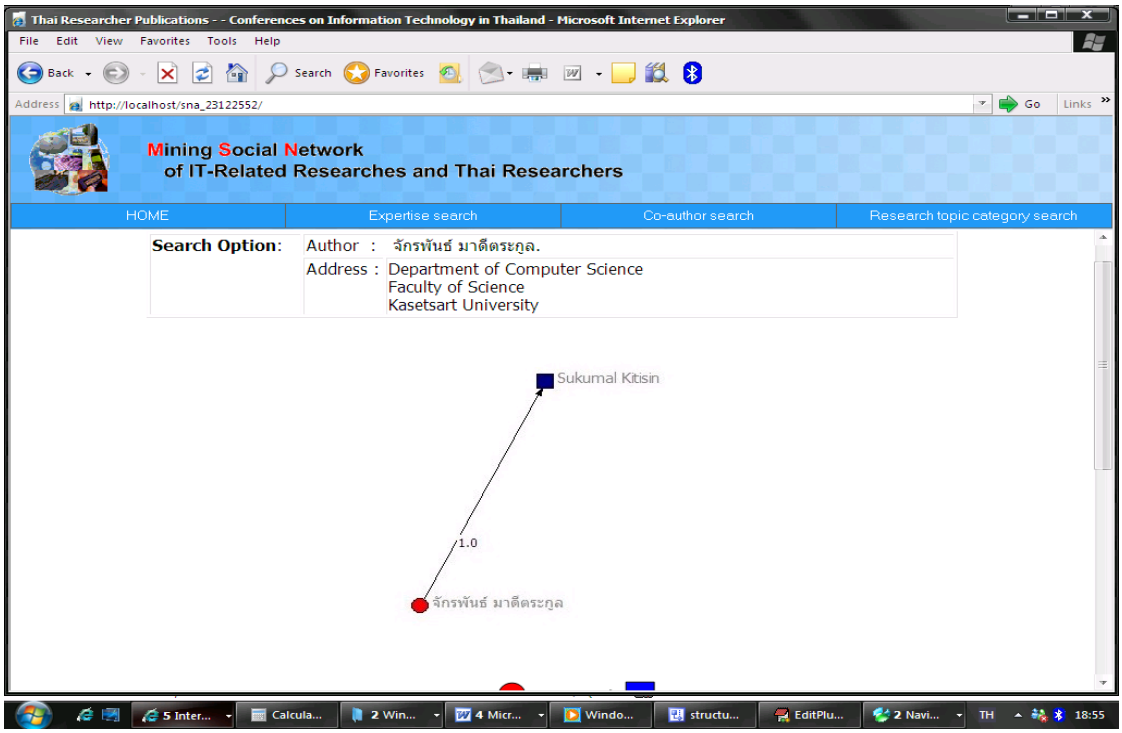



Figure 5.24 An example of results of co-author searching based on test scenario 3 (social network view)

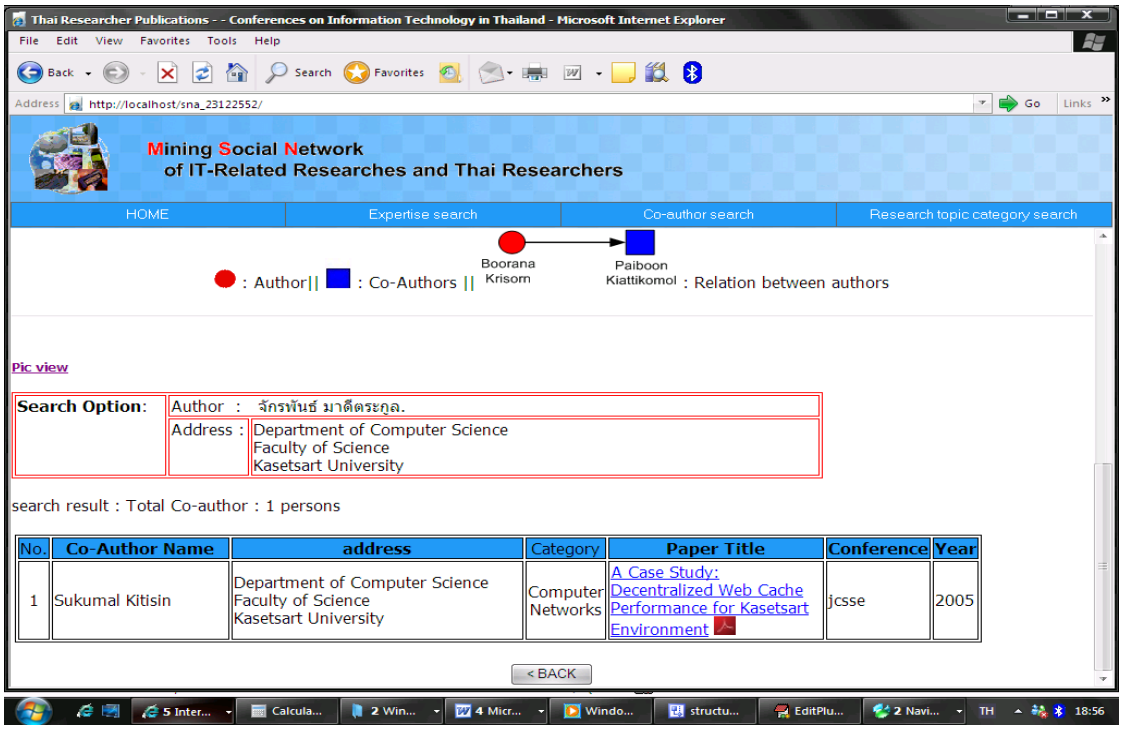



Figure 5.25 An example of results of co-author searching based on test scenario 3 (data view)

Figures 5.23, 5.24 and 5.25 illustrate the results of co-author search testing by identifying the keyword of the researcher name "กร". Figure 5.24 reveals a social network diagram of searching results indicating that there are a total of 22 researchers, with the word "กร" as part of their names, submitting research to the conferences JCSSE and NCSEC. When data of Mr. จักรพันธ์ มาดีตระกูล's co-author are selected, the system will display his/her name, which is Ms. Sukumal Kittisin. Figure 5.25 shows a name list of Mr. จักรพันธ์ มาดีตระกูล's co-author, or Ms. Sukumal Kittisin, who collectively conducted research on Computer Network, including research topics in this field, as well as years and names of conferences the researches were submitted to.

### 5.2.3 Research Topic Category Searching Result

(1) Scenario 1 : research topic category searching by identifying that all conference in each year are to be viewed. Testing details are shown in Table 5.9 while searching conditions and searching consequences are illustrated in Figures 5.26, 5.27 and 5.28.

Table 5.9 Details of test scenario 1 of research topic category searching

<table>
<tr><td>Testing item</td><td colspan="2">Testing detail</td><td>results that should be derived</td></tr>
<tr><td>Testing of research topic category searching.</td><td colspan="2">Choose the menu “ Research topic category search”<br>Condition for sub category searching<br>Case: search for research topic category of all conferences in each year</td><td></td></tr>
<tr><td rowspan="3"></td><td>1) Identify data of conference name</td><td>Conference = “All Conference”</td><td rowspan="3">Can search for data of the research category and the number of researches of all conferences in each year, and reveal searching results on the screen.</td></tr>
<tr><td>2) Identify data of conference year</td><td>Year = “all year”</td></tr>
<tr><td>3) Identify data of research category</td><td>Category = “ ”</td></tr>
<tr><td rowspan="2"></td><td>4) Identify data of research sub category</td><td>Sub category = “ ”</td><td rowspan="2"></td></tr>
<tr><td>5) Search for data</td><td>Press button “submit”</td></tr>
</table>

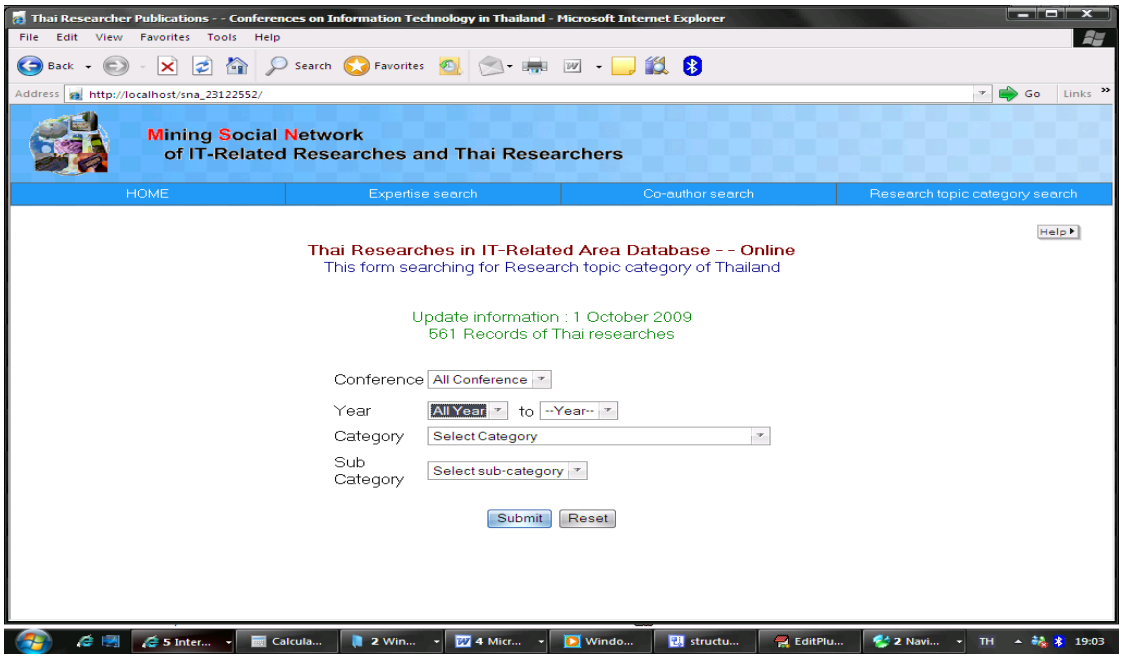



Figure 5.26 Details of research topic category searching conditions of test scenario 1

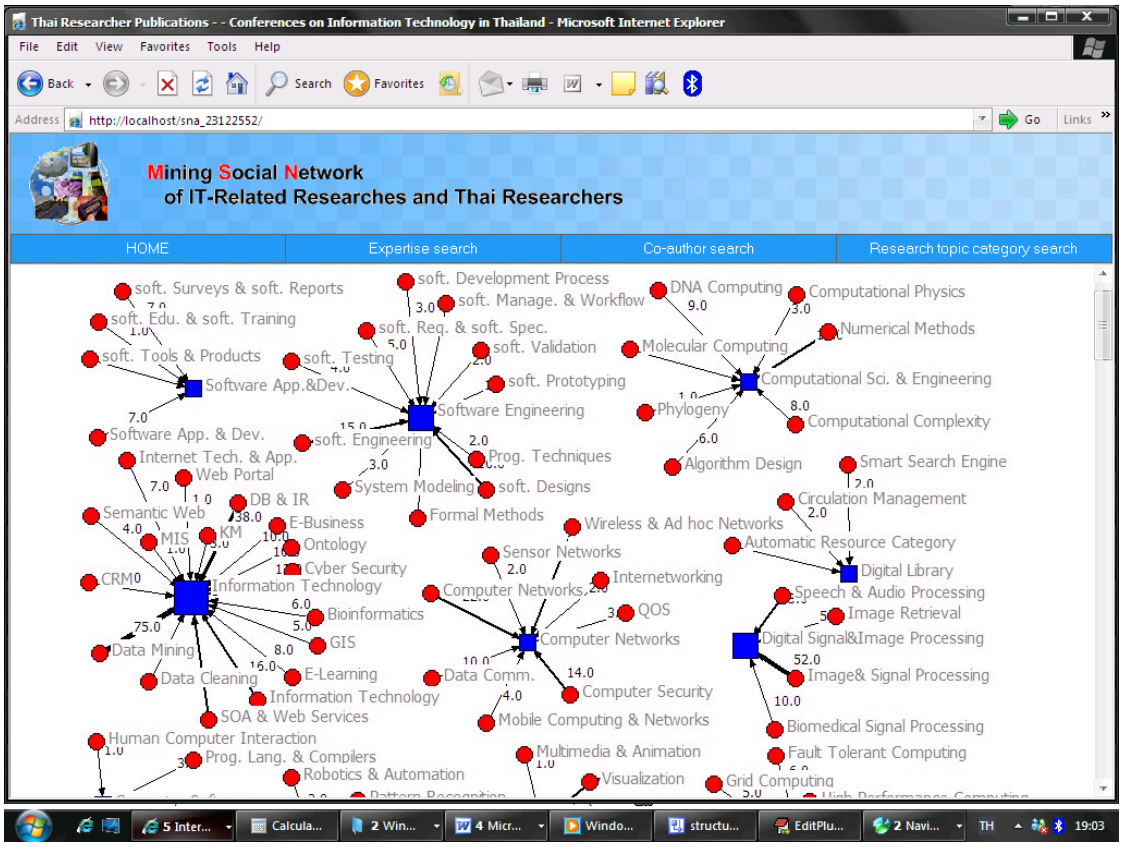



Figure 5.27 An example of results of research topic category searching based on test scenario 1 (social network view)

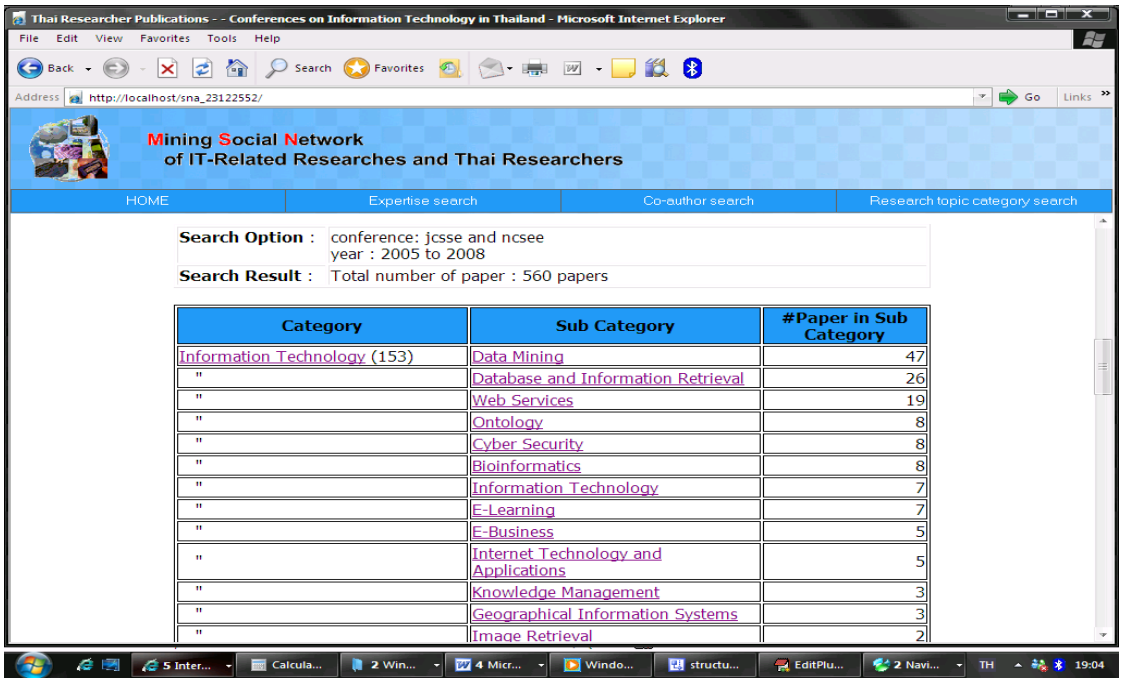



Figure 5.28 An example of results of research topic category searching based on test scenario 1 (data view)

Figures 5.27 and 5.28 illustrate the results of research topic category search testing by identifying that all conferences in each year are to be viewed. Figure 5.18 displays a social network diagram of searching results indicating that there are a total of 561 research papers submitted to the conferences JCSSE and NCSEC between 2005 and 2008. it is found that the research on Information Technology makes up the largest portion, totaling 151 research papers, followed by Artificial Intelligence, allowing us to know a research trend of Thai researchers during 2005 through 2008 that research on Information Technology represented the majority of IT-related research. Figure 5.28 reveals the sub category of research submitted to the conference, arranged in order of the number of researches, from most to least.

(2) Scenario 2 : research topic category searching by identifying the name and year of the conference. Testing details are given in Table 5.10 while searching conditions and searching outcomes are illustrated in Figures 5.29, 5.30 and 5.31.

Table 5.10 Details of test scenario 2 of research topic category searching

<table>
<tr><th>Testing item</th><th colspan="2">Testing detail</th><th>results that should be derived</th></tr>
<tr><td>Testing of research topic category searching.</td><td colspan="2">Choose the menu “ Research topic category”<br>Condition for research topic category searching<br>Case: search for research topic category by identifying conference : NCSEC of 2005</td><td></td></tr>
<tr><td rowspan="5"></td><td>1) Identify data of conference name</td><td>Conference = “ncsec”</td><td rowspan="5">Can search for data of the research topic category and the number of research papers of the NCSEC 2005, and show searching outcomes on the screen.</td></tr>
<tr><td>2) Identify data of conference year</td><td>Year = “2005”</td></tr>
<tr><td>3) Identify data of research category</td><td>Category = “ ”</td></tr>
<tr><td>4) Identify data of research sub category</td><td>Sub category = “ ”</td></tr>
<tr><td>5) Search for data</td><td>Press button “submit”</td></tr>
</table>

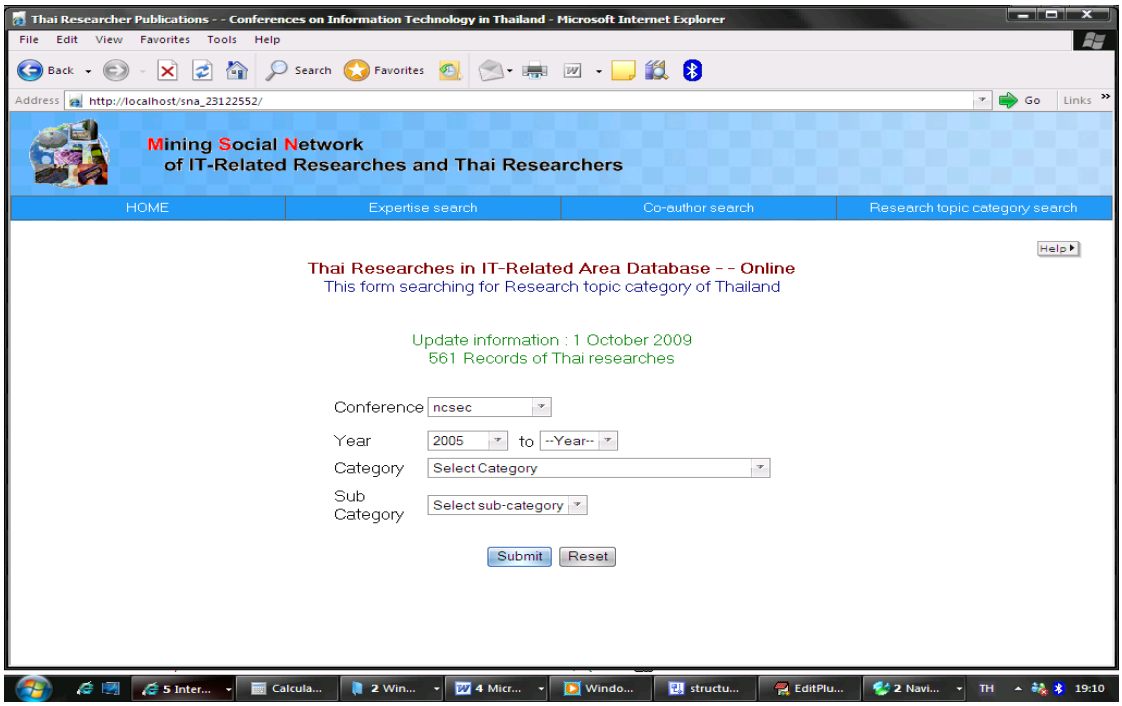



Figure 5.29 Details of research topic category searching conditions of test scenario 2

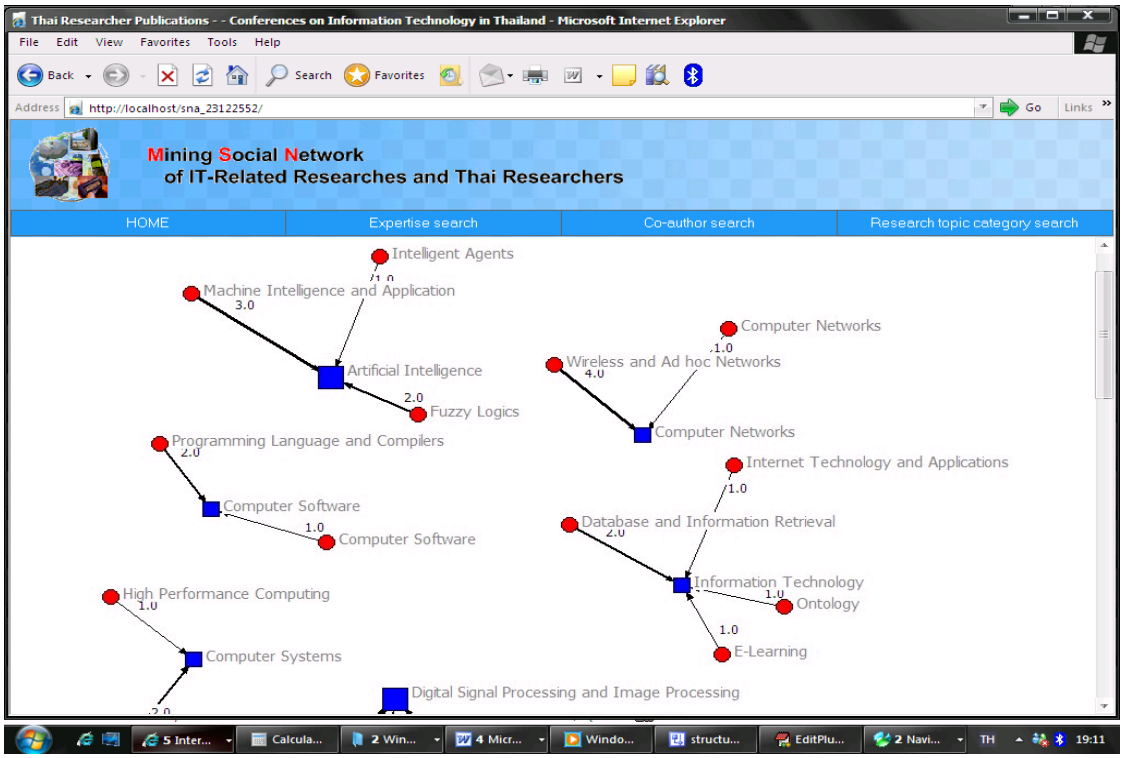



Figure 5.30 An example of results of research topic category searching based on test scenario 2 (social network view)

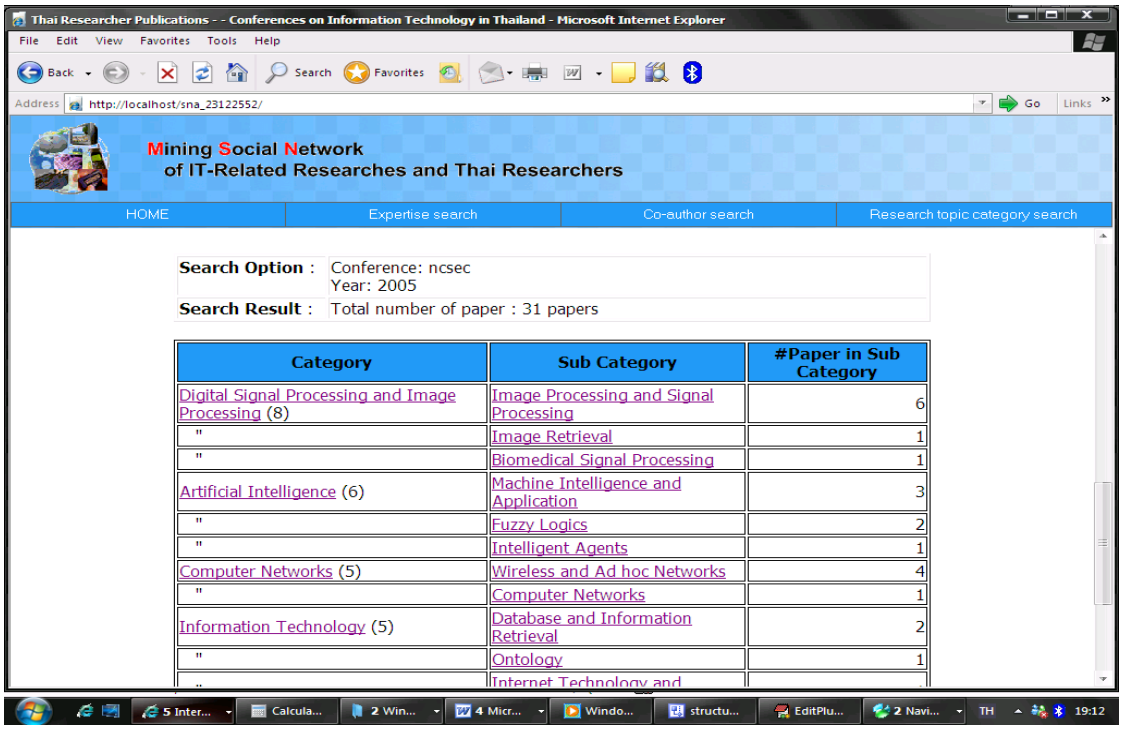



Figure 5.31 An example of results of research topic category searching based on test scenario 2 (data view)

Figures 5.30 and 5.31 illustrate the outcomes of research topic category search testing by identifying the data of NCSEC 2005 to view the conference data in that period. Figure 5.20 shows a social network diagram of searching outcomes indicating that there are a total of 31 research papers submitted to the conference. It has been found that the research on Digital Signal Processing and Image Processing made up the largest portion of IT-related research in 2005, amounting to 8 research papers, followed by 6 research papers on Artificial Intelligence, allowing us to know a research trend of Thai researchers in 2005, who focused mainly on Digital Signal Processing and Image Processing. Figure 5.31 reveals the category and sub category of research submitted to the conference, arrange in order of the number of research, from most to least.

(3) Scenario 3 : research topic category searching by identifying the name and time period (in year) of the conference. Testing details are shown in Table 5.11 while searching conditions and searching results are illustrated in Figures 5.32, 5.33 and 5.34.

Table 5.11 Details of test scenario 3 of research topic category searching

| Testing item | Testing detail | | results that should be derived |
|---|---|---|---|
| Testing of research topic category searching. | Choose the menu “Research topic category search”<br>Condition for research topic category searching<br>Case: search for research topic category by identifying conference : NCSEC between 2005 - 2006 | | |
| | 1) Identify data of conference name | Conference = “ncsec” | Can search for data of the research topic category and the number of research of the NCSEC between 2005 and 2006, including displaying searching results on the screen. |
| | 2) Identify data of conference year | Year = “2005” to “2006” | |
| | 3) Identify data of research category | Category = “ ” | |
| | 4) Identify data of research sub category | Sub category = “ ” | |
| | 5) Search for data | Press button “submit” | |

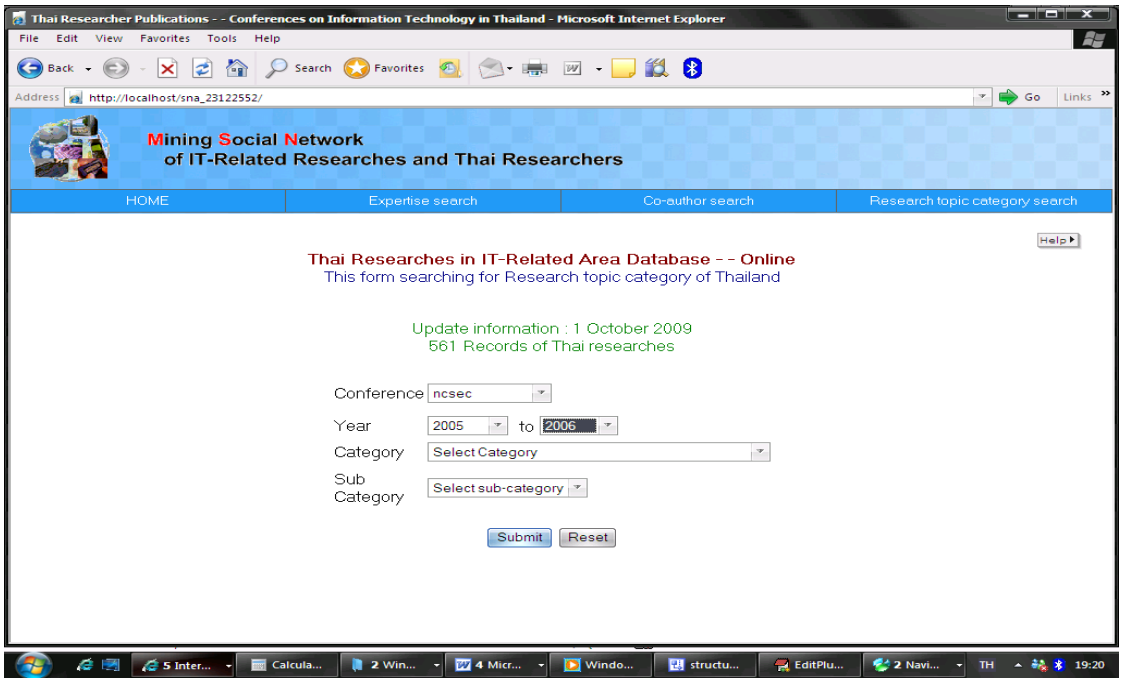



Figure 5.32 Details of research topic category searching conditions of test scenario 3

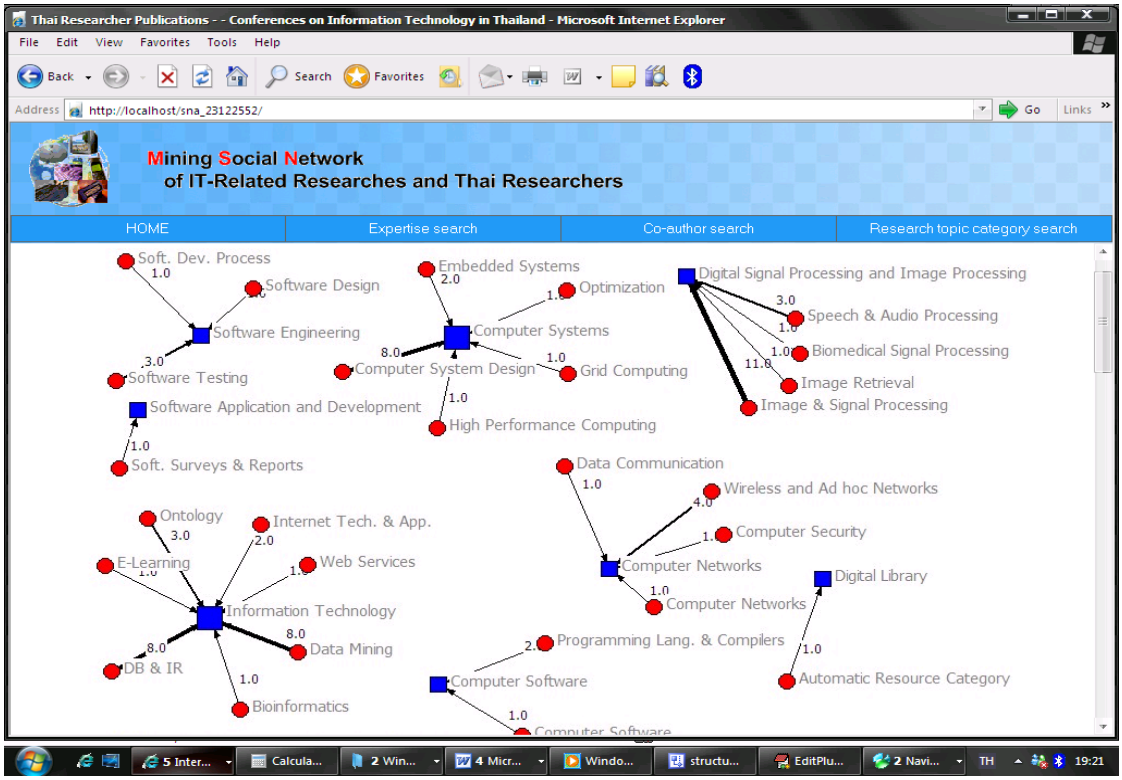



Figure 5.33 An example of results of research topic category searching based on test scenario 3 (social network view)

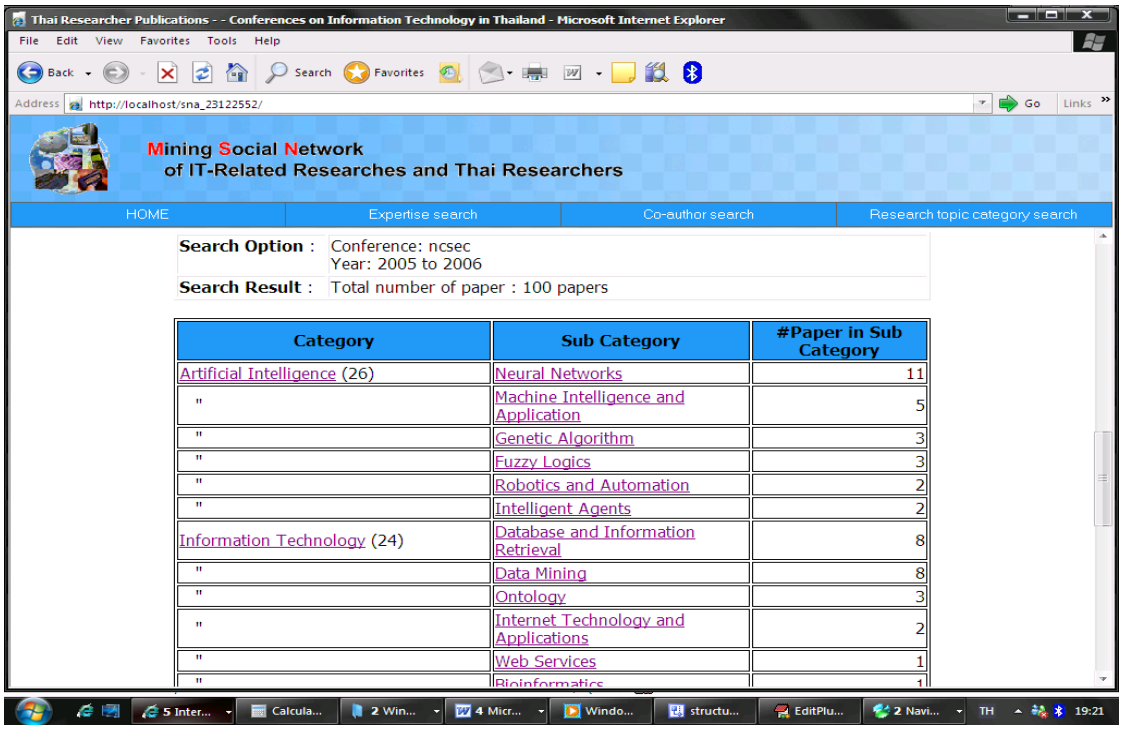



Figure 5.34 An example of results of research topic category searching test based on test scenario 3 (data view)

Figures 5.33 and 5.34 illustrate the results of research topic category search testing by identifying the NCSEC time period between 2005 and 2006 in order to see the conference data in that period. Figure 5.33 displays a social network diagram of searching results indicating that there are a total of 101 research papers submitted to the NCSEC between 2005 and 2006. it is found that the research on Artificial Intelligence makes up the majority, totaling 27 research papers, followed by 24 research papers on Information Technology, allowing us to know a research trend of Thai researchers between 2005 and 2006, who concentrated mainly on Artificial Intelligence and Information Technology. Figure 5.34 shows the sub category of research submitted to the conference, arranged in order of the number of researches, from most to least.

(4) Scenario 4 : research topic category searching by identifying the name and year of the conference, including the research topic category. Testing details are displayed in Table 5.12 while searching conditions and searching consequences are illustrated in Figures 5.35, 5.36 and 5.37.

Table 5.12 Details of test scenario 4 of research topic category searching

<table>
<tr><th>Testing item</th><th colspan="2">Testing detail</th><th>results that should be derived</th></tr>
<tr><td>Testing of research topic category searching.</td><td colspan="2">Choose the menu “Research topic category search”<br>Condition for research topic category searching<br>Case: search for research topic category by identifying conference : NCSEC 2005 and the research category “Computer Network”</td><td></td></tr>
<tr><td rowspan="3"></td><td>1) Identify data of conference name</td><td>Conference = “ncsec”</td><td rowspan="3">Can search for data of the research topic category and the number of research of the NCSEC 2005, in the category “Computer Network”, and show searching results on the screen.</td></tr>
<tr><td>2) Identify data of conference year</td><td>Year = “2005”</td></tr>
<tr><td>3) Identify data of research category</td><td>Category = “ Computer Network”</td></tr>
<tr><td rowspan="2"></td><td>4) Identify data of research sub category</td><td>Sub category = “ ”</td><td rowspan="2"></td></tr>
<tr><td>5) Search for data</td><td>Press button “submit”</td></tr>
</table>

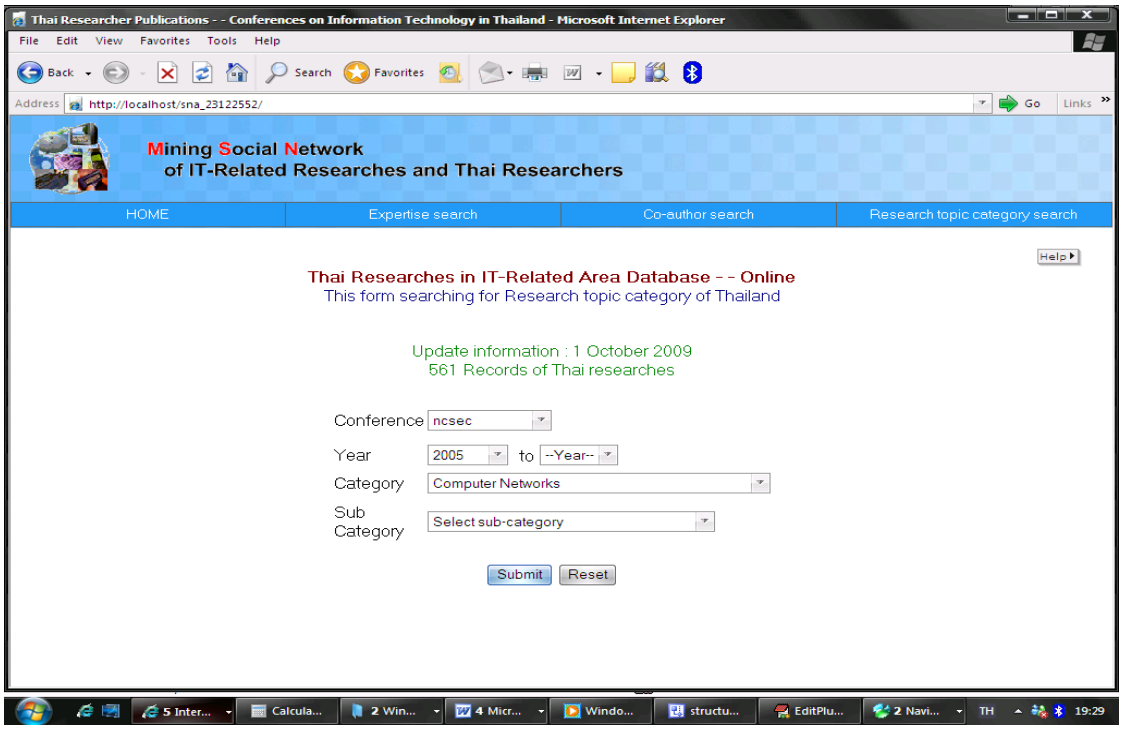



Figure 5.35 Details of research topic category searching conditions of test scenario 4

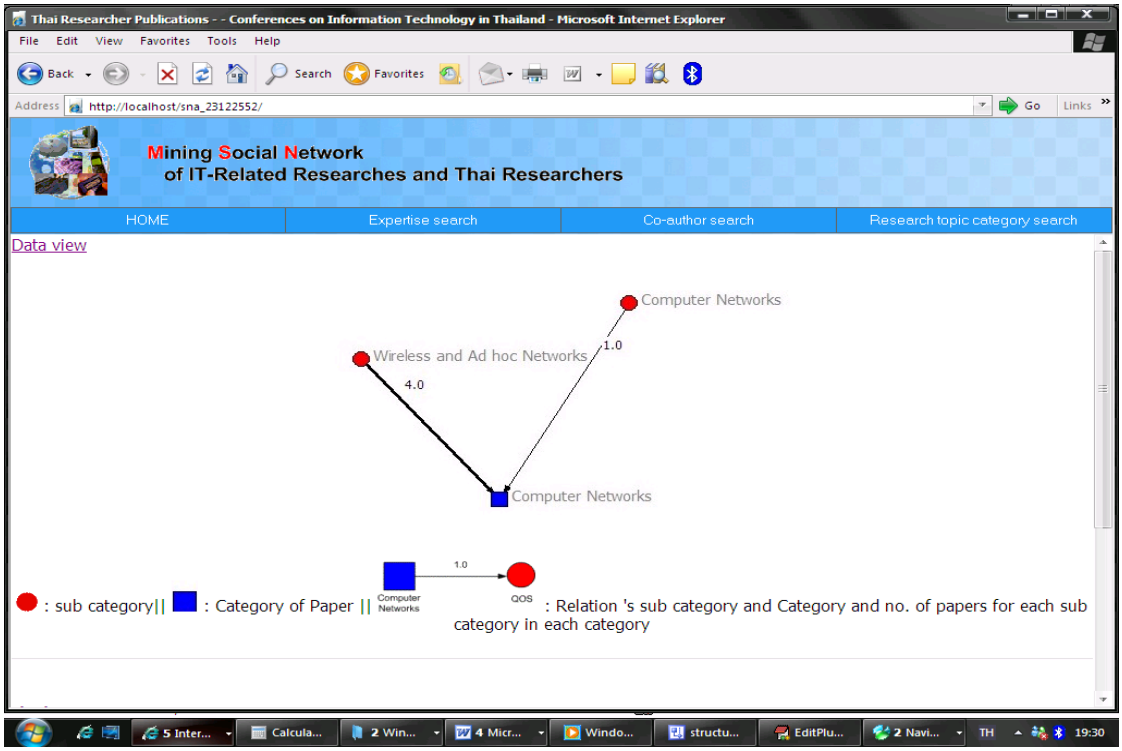



Figure 5.36 An example of results of research topic category searching based on test scenario 4 (social network view)

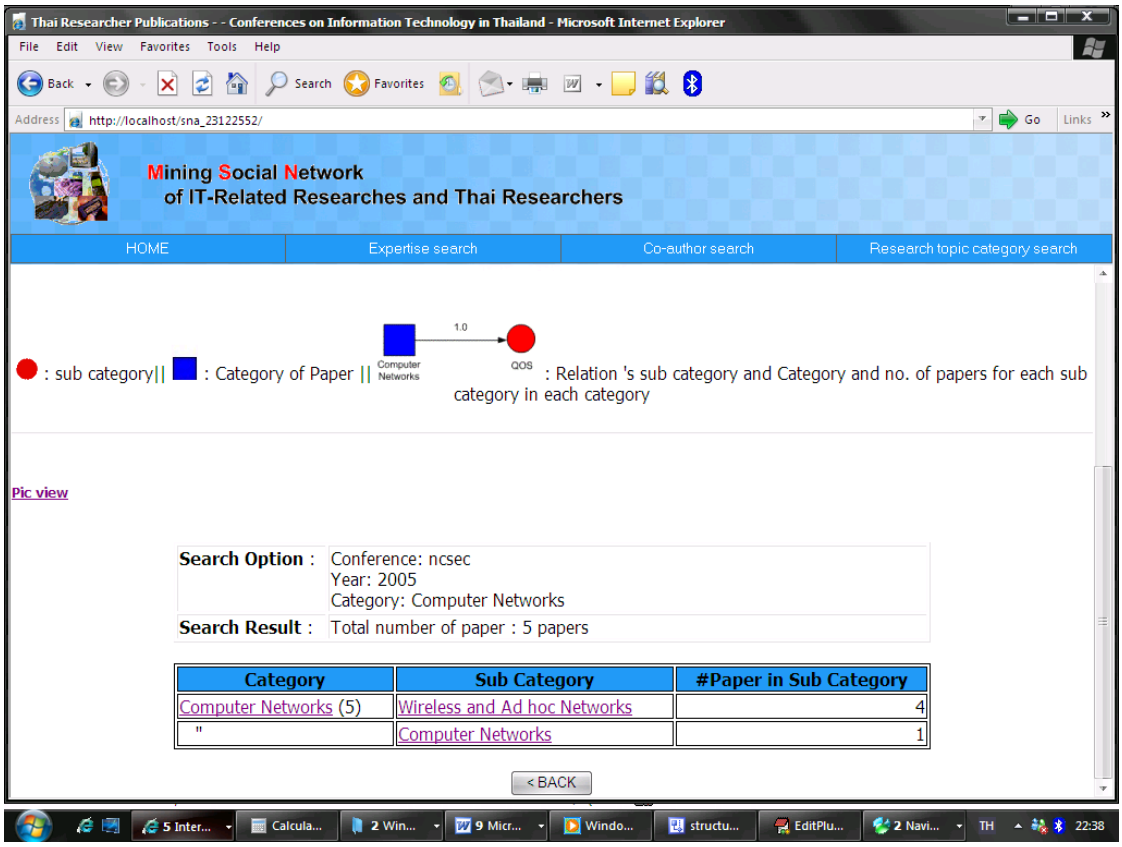



Figure 5.37 An example of results of research topic category searching based on test scenario 4 (data view)

Figures 5.36 and 5.37 illustrate the consequences of research topic category search testing by identifying the NCSEC 2005 and only the research category "Computer Network" so as to see the conference data of that year. Figure 5.36 reveals a social network diagram of searching consequences indicating that there are a total of 5 research papers on Computer Network submitted to the NCSEC 2005. It is found that out of this number, 4 research papers are in the sub category "Wireless and Ad hoc Networks", which shows that most researchers doing research on Computer Network in 2005 focused mainly on Wireless and Ad hoc Networks. Figure 5.37 shows the sub category of research submitted to the conference, arranged in order of the number of research papers, from most to least.

## 5.4 Summary and Discussion

Testing the system to test the functionality of two components is Social Network Analysis component and searching component together by 12 testing scenarios that separate tests of the search function were expertise search, co-author search and research topic category search.

The results of testing three system functions, illustrate who are research expertise in which research category (field), who are each researcher's co-author and their agency or institute. Furthermore, the system also shows that each year a research topic category and sub category which some researchers have focused on the research work a lot.

In research topic category search testing, if the user chooses several years in searching condition, we derived the large social network diagram. It could not be seen clearly, but the overall results of system testing are satisfactory.

# CHAPTER VI
# CONCLUSION AND FUTURE WORK

This chapter discusses the conclusion of the research project of Mining Social Network of IT-related researches and Thai researchers, as well as a work plan to be developed in the future, both in the section 6.1 and 6.2, respectively.

## 6.1 Conclusion

The Mining Social Network of IT-Related Researches and Thai Researchers is a system which compiles research papers and uses data in various papers to analyse and find the relation existing in those data.

Research papers compiled in the Mining Social Network of IT-Related Researches and Thai Researchers include the research papers of Thai researchers submitted to the conferences JCSSE and NCSEC between 2005 and 2008. In addition, the research papers collected must have an English title and abstract.

For the Mining Social Network of IT-Related Researches and Thai researchers, the data of researchers with respect to their titles, authors and their agencies, and keywords are used for analysis of the relation between researchers and research topic categories, between researchers themselves, and between research topic categories, the conferences and years in which they were organized. Data analysis is via Social Network Analysis to find the relation and construct a network diagram to display such relation.

When some parts of data from academic works have been analysed through an approach of Social Network Analysis, the Mining Social Network of IT-Related Researches and Thai Researchers has developed core functions for searching for data obtained from analysis via Social Network Analysis. Those functions include expert searching, co-author searching and research topic category searching, for each conference in each year.

Based on the consequences of the system functionality test applying test scenario, it is revealed that the prototype system developed can display to user who are research expertise in which research category (field), who is each researcher's co-author, and a trend of research in Thailand, which is reflected in the topics of research submitted to the conferences.

## 6.2 Future Work

A work plan to be developed in the future consists of 6 following parts.

**1) References Paper**

This research focuses on the relation between researches and research topic categories, and between researchers themselves. Nevertheless, this research has not discovered the relation among researches used as a reference by other researches. The relation among researches used as reference papers can tell a significant role of research in a particular area if that research is frequently used as a reference paper. Besides, it is useful for researchers in the related area since they will be able to search for researches related to their work.

**2) Paper Categorization by using Ontology**

Basically, in each conference, research topics are not identically classified, for instance, for JCSSE, researches are classified into 10 categories, while for the NCSEC, researches are divided into 7 categories. However in some years the classification may be a bit different.

Based on the above problem, the ontology may be developed for the classification of topics of researches submitted to the conferences JCSSE and NCSEC, which will make the classification more accurate and flexible when the research categories vary for previous ones.

**3) Automatic Data Preparation Tool**

Since the prototype system developed still has limitations in a stage of data preparation in order to load research data into the system. Currently, the system administrator has to load data into the system. In the future, when the proposed system

is developed to a level that it can connect with a database of a digital library of the conferences like JCSSE, this stage will become more convenient and help relieve the burden of the system administrator for system maintenance.

**4) More dataset**

At present the prototype system's database has selected only the research submitted to only the 2 conferences, JCSSE and NCSEC, from 2005 to 2008. Nonetheless, there are still other conferences on Computer Science and Information Technology held inside the country and internationally such as ECTI-CON, NCCIT, etc. If there is a compilation of research submitted to other conferences apart from the JCSSE and NCSEC, a large amount and wider variety of more comprehensive data to be used for analysis will be derived.

Through some Thai researches have collectively conducted research with foreign researchers, the latter's data may not be displayed. Therefore, in the future, foreign researchers's work should be included to make the system more complete.

**5) Statistical data**

The information derive from data analysis via Social Network Analysis technique can demonstrate relationships within the data. It also can be used to display the information in the form of statistical data such as a list of researchers working on each research category (field), the agency that regularly submitted research papers to the conference. In addition, the proportion of Thai and foreign researchers who submitted research papers to the conference in each conference.

**6) Applied to various researches areas**

In the scope of this research, it is clearly defined that a technique of Social Network Analysis will be used to analyse the relation in various patterns of research on Computer Science and Information Technology only. Based on the outcomes of the prototype system testing, they are satisfactory. Thus, the data analysis technique via Social Network Analysis may be effectively applied to research in other areas, for example, medical and public health research, engineering research, etc. Hopefully, the analysis consequences can be utilized in terms of the promotion and support of more conduction of research in different fields.

# REFERENCES


1 สำนักงานคณะกรรมการวิจัยแห่งชาติ, *เอกสารเผยแพร่ของสำนักงานคณะกรรมการวิจัยแห่งชาติ*, 30 กันยายน 2545.

2 สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ, *ความเป็นมาของการพัฒนาฐานข้อมูลนักวิจัย*, 1994.

3 P. M. Andersen, P. J. Hayes, A. K. Heuttner, L. M. Schmandt, and I. B. Nirenberg. Automatic extraction. Proceedings of the Conference of the Association for Artificial Intelligence; Philadelphia, Penn.;1986, pp.1089-1093.

4 H. Cunningham, R. Gaizauskas and Y. Wilks. GATE: a General Architecture for Text Engineering**.** Proceedings of the 16th conference on Computational linguistics, Copenhagen, Denmark, 1996, pp. 1057-1060.

5 Jeffrey D. Ullman. Information Integration Using Logical Views. Lecture Notes In Computer Science 1997; Vol. 1186:19-40.

6 R. Baeza-Yates and B.Ribeiro-Neto. Modern Information Retrieval. Boston : Addison-Wesley; 1999.

7 Doyle, Lauren; Becker, Joseph . Information Retrieval and Processing. Los Angeles : Melville; 1975.

8 C. J. van RIJSBERGEN. Information Retrieval. 2nd ed. London: Butterworths; 1979.

9 D.Yimam-Seid and A.Kobsa. Expert Finding System for Organizations: Problem and Domain Analysis and the Demoir Approach. Journal of Organizational Computing and Electronic Commerce 2003; Vol. 13(1):1-24.

10 D. Mattrox,M.T. Maybury,D.Morey. Enterprise Expert and Knowledge Discovery. Proceedings of the 8th International Conference on Human-Computer Interaction. Human-Computer Interaction: Communication, Cooperation, and Application Design 1999; Vol. 2:303-307.

11 G. Piatetsky-Shapiro, U. M. Fayyad, and P. Smyth. From data mining to knowledge discovery: an overview. Advances in Knowledge Discovery and Data Mining, Menlo Park : American Association for Artificial Intelligence; 1996. pp. 1-34.

12 Peter Cabena, Pablo Hadjnian, Rolf Stadler, Jaap Verhees, Alessandro Zanasi. Discovering Data Mining: From Concept to Implementation. Prentice-Hall; 1997.

13 Kantardzic, Mehmed. Data Mining: Concepts, Models, Methods, and Algorithms. Wiley-IEEE Press; 2002.

14 J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann; 2000.

15 Linton Freeman. The Development of Social Network Analysis. Vancouver: Emperical Press; 2006.

16 Brandes,Ulrik and Thomas Erlebach. Network Analysis:Methodological Foundation. Berlin: Springer Verlag; 2005.

17 K Faust, S Wasserman. Social Network Analysis: Methods and Applications. Cambridge University Press; 1994.

18 Hanneman, Robert A. and Mark Riddle. Introduction to social network Methods. Available at: http://www.faculty.ucr.edu/~hanneman/nettext; Mar 2008.

19 Christopher S. Campbell, Paul P. Maglio, Alex Cozzi, Byron Dom. Expert Identification Using Email Communications. Proceedings of the twelfth international conference on Information and knowledge management, New Orleans; 2003. pp. 528-531.

20 B. Dom, I. Eiron, A. Cozzi, and Y. Zhang. Graph-based ranking algorithms for e-mail expertise analysis. Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery; San Diego, California; 2003. pp. 42-48.

21 A.Culotta, R.Bekkerman, A. McCallum. Extracting social network and contact information from e-mail and the web. Proceedings of the 1st Conference on Email and Anti-Spam. ; 2004.

22 Trausti Kristjansson, Aron Culotta, Paul Viola, and Andrew McCallum. Interactive information extraction with constrained conditional random fields. Proceedings of the 19th national conference on Artificial intelligence ; San Jose, California ; 2004.

23 Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proceedings of the seventh conference on Natural language learning; Edmonton, Canada ; 2003. pp. 188-191.

24 Jing Zhang, Jie Tang, Juanzi Li. Expert Finding in A Social Network. Proceedings of the Twelfth Database Systems for Advanced Applications; 2007. pp. 1066-1069.

25 Jing Zhang, Jie Tang, Juanzi Li. A new Approach to personal network search based on information extraction. Proceedings of the 1st Asian Semantic Web Conference; 2006.

26 Jing Zhang, Jie Tang, Juanzi Li, Hong-Jun Lu, Bangyong Liang, Kehong Wang. iASA:Learning to Annotate the semantic web. Journal on Data Semantic, 2005 Nov; (5):110-145.

27 K.Balog, L.Azzopardi, M. de Rijke. Formal Models for Expert Finding in Enterprise Corpora. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval; Seattle, Washington, USA ; 2006. pp.43-50.

28 K. Balog , M. de Rijke. Searching for People in the Personal Workspace. Proceedings of International Workshop on Intelligent Information Access; 2006.

29 L.A. Adamic and E. Adar. How to search a social network. Social Networks 2005 July; 27(3):187-203.

30 Sudhir Saxena, K. Santhanam, Aparna Basu. Application of Social Network Analysis (SNA) to Terrorist Networks in Jammu & Kashmir. Strategic Analysis 2004 Jan-Mar; 28(1): 84-101.

31 Aparna Basu. Social Network Analysis of Terrorist Organizations in India. Proceedings of the North American Association for Computational Social and Organizational Science conference; 2005.

32 Jie Li, Jie Tang, Jing Zhang, Qiong Luo, Yunhao Liu, and Mingcai Hong. EOS: expertise oriented search using social networks. Proceedings of the 16th international World Wide Web Conference; 2007. pp. 1271-1272.

33 Duo Zhang, Jie Tang, Limin Yao. Social Network Extraction of Academic Researchers. Proceedings of IEEE International Conference on Data Mining; 2007. pp. 292 – 301.

34 Takeru Miki , Saeko Nomura,Toru Ishida. Semantic Web Link Analysis to Discover Social Relationships in Academic Communities. Proceedings of the 2005 Symposium on Applications and the Internet; 2005. pp.38-45.

35 Guillermo Rueda, Pisek Gerdsri, Dundar F. Kocaoglu. Bibliometrics and Social Network Analysis of the Nanotechnology Field. Proceedings of Portland International Center for Management of Engineering and Technology; 2007. pp. 2905-2991.

# BIOGRAPHY

**NAME** Miss Usanee Sripirom

**DATE OF BIRTH** February 10, 1977

**PLACE OF BIRTH** Khon Kaen, Thailand

**INSTITUTIONS ATTENDED**

Silpakorn University, 1995-1998:
Bachelor of Science (Computer Science)
Mahidol University, 2004-2010:
Master of Science (Computer Science)

**POSITION & OFFICE**

2001 – Present,
Computer Technical Officer,
The Revenue Department, Thailand

**ADDRESS**

Bureau of Information Techonology,
Revenue Department,
90 Soi Phaholyothin7,
Phaholyothin Road,
Bangkok 10400, Thailand
Tel. 0-2272-9469
E-mail:Usanee.sr@rd.go.th