

กรอบการทำงานของโปรโตคอลเพื่อการค้นหาสารสนเทศจำเพาะ

นางสาวมัลลิกา วัฒนนะ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

บทคัดย่อและแฟ้มข้อมูลฉบับเต็มของวิทยานิพนธ์ตั้งแต่ปีการศึกษา 2554 ที่ให้บริการในคลังปัญญาจุฬาฯ (CUIR)
เป็นแฟ้มข้อมูลของนิสิตเจ้าของวิทยานิพนธ์ที่ส่งผ่านทางบัณฑิตวิทยาลัย

The abstract and full text of theses from the academic year 2011 in Chulalongkorn University Intellectual Repository (CUIR)
are the thesis authors' files submitted through the Graduate School.

FRAMEWORK OF SPECIFIC INFORMATION SEARCH PROTOCOL

Miss Monlica Wattana

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Computer Science

Department of Mathematics and Computer Science

Faculty of Science

Chulalongkorn University

Academic Year 2011

Copyright of Chulalongkorn University

มัลลิกา วัฒนะ : กรอบการทำงานของโพรโทคอลเพื่อการค้นหาสารสนเทศจำเพาะ.
(FRAMEWORK OF SPECIFIC INFORMATION SEARCH PROTOCOL) อ. ที่
ปริญญาวิทยานิพนธ์หลัก : ผศ. ดร. ภัทรสินี ภัทรโกศล, 88 หน้า.

อินเทอร์เน็ตเป็นแหล่งทรัพยากรที่มีขนาดใหญ่ที่สำคัญต่อมนุษย์ โปรแกรมค้นหาได้ถูกพัฒนาขึ้นเพื่อเข้าถึงทรัพยากรที่ต้องการบนอินเทอร์เน็ต โปรแกรมค้นหาทุกตัวให้ผู้ใช้สามารถใส่คำสำคัญในช่องค้นหาเพื่อทำการค้นหา แต่การค้นหาโดยใช้คำสำคัญไม่ได้ข้อมูลที่จำเพาะ ผลของการค้นหาโดยทั่วไปจะได้ข้อมูลจำนวนมากที่ไม่สัมพันธ์กับคำสำคัญเพราะผู้ใช้ไม่ทราบคำสำคัญที่เหมาะสมในการค้นหา ดังนั้นโพรโทคอลที่มีชื่อว่า โพรโทคอลเพื่อการค้นหาสารสนเทศ ได้ถูกนำเสนอขึ้น ซึ่งผู้ใช้สามารถใช้บางส่วนของยูอาร์แอลและคำสำคัญที่ต้องการในการค้นหาได้ ดังนั้นโพรโทคอลเพื่อการค้นหาสารสนเทศสามารถทำให้ได้ค้นหาแบบลงได้ แต่อย่างไรก็ตามโพรโทคอลเพื่อการค้นหาสารสนเทศไม่มีการทำงานในการค้นหาภาพและยังมีข้อเสียของการค้นหาเส้นทางซึ่งเป็นสาเหตุของการคับคั่งของข้อมูล ดังนั้นงานวิจัยนี้นำเสนอกลไกการค้นหาภาพและการแก้ปัญหาการค้นหาเส้นทางเพื่อประยุกต์ใช้กับโพรโทคอลเพื่อการค้นหาสารสนเทศ เนื่องจากโพรโทคอลเพื่อการค้นหาสารสนเทศถูกปรับปรุงทุกกลไกการทำงาน ดังนั้นโพรโทคอลเพื่อการค้นหาสารสนเทศที่ถูกปรับปรุงใหม่นี้เปลี่ยนชื่อเรียกว่า โพรโทคอลเพื่อการค้นหาสารสนเทศจำเพาะ การสกัดคุณลักษณะของภาพในกลไกการค้นหาภาพถูกสร้างขึ้นโดยใช้เทคนิคสหสัมพันธ์บริเวณสี นอกจากนี้งานวิจัยยังได้เสนออัลกอริทึมที่แก้เกิดตรวจตราเพื่อแก้ปัญหาการค้นหาเส้นทาง ผลการทดลองพบว่าการค้นหาโดยใช้โพรโทคอลเพื่อการค้นหาสารสนเทศจำเพาะมีความจำเพาะในการค้นหาและแก้เกิดตรวจตราสามารถหลีกเลี่ยงการเกิดความคับคั่งของข้อมูลได้ ดังนั้นโพรโทคอลเพื่อการค้นหาสารสนเทศจำเพาะเป็นโพรโทคอลที่มีประสิทธิภาพ

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์ ลายมือชื่อนิสิต

สาขาวิชา วิทยาการคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปริญญาวิทยานิพนธ์หลัก.....

ปีการศึกษา 2554

5073862023 : MAJOR COMPUTER SCIENCE

KEYWORDS : SEARCH ENGINE / SEARCH PROTOCOL / ROUTING ALGORITHM /
IMAGE RETRIEVAL / SEARCH MECHANISM

MONLICA WATTANA : FRAMEWORK OF SPECIFIC INFORMATION SEARCH
PROTOCOL. ADVISOR : ASST.PROF. PATTARASINEE BHATTARAKOSOL,
Ph.D., 88 pp.

Since the Internet is a large resource of information that is important to a vast number of people, search engines are implemented for accessing the required resources over the Internet. Every search engine allows users to enter keywords in the search field for searching. Unfortunately, searching by keyword is usually not specific. When users do not know appropriate keywords, the search result is generally a long list of information, much of it irrelevant. Therefore, the protocol named Information Search Protocol (ISP) is proposed. The ISP enables users to enter a partial part of the required URL and keywords; the ISP can narrow down the search. However, the original ISP cannot perform an image search and it has a defect of the routing module that causes network congestion. Therefore, this research proposes the image search mechanism and routing algorithm to apply to the ISP. Since the ISP is modified in all mechanisms, the new version of the ISP is called Specific Information Search Protocol (SISP). The image extraction of the image search mechanism is implemented using Color Region Correlation (CRC). Moreover, this research is also proposing a Patrol Packet (PTP) Algorithm to solve the routing problem. The search result of the SISP is specific. In addition to more effective searches, the PTP can significantly reduce the congestion caused. Therefore, the SISP is a performance search protocol.

Department : Mathematics and Computer Science Student's Signature

Field of Study : Computer Science Advisor's Signature

Academic Year : 2011

ACKNOWLEDGEMENTS

During my years as a Ph.D. student, I have received a lot of tuition, care and friendship from several people, some of which I wish to thank here. I would like to thank the Office of Higher Education Commission, Khon Kean University and Chulalongkorn University for their financial support.

I would like to express my deepest gratitude to my advisor, Assist.Prof.Dr. Pattarasinee Bhattarakosol, to whom with her advice, guidance and care, help me to overcome all the difficulties of the process of research and make this dissertation possible.

My thanks also goes to dissertation committee, Assist.Prof.Dr. Nagul Cooharajanone, Dr. Siripun Sanguansintuku, Assist.Prof.Dr. Ohm Sornil, and Dr. Surapant Meknavin for their advices and guidance about the research activities.

I would like to thank the Department of Computer Science & Engineering at Konkuk University for their facility support during my visiting scholar in 2009-2010, especially, Assoc.Prof.Dr. Sunyoung Han for encouragement, guidance and spending time on my research discussion.

I would also like to thank all lecturers and colleagues at the Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, for their warmest care and moral support. Moreover, I would like to thank the Thai Student in Konkuk University, and all my friends (especially, Miss Wannaporn Thanatkha, Miss Sirorat mongkolsawat, and Miss Phattharapha Triamwetwutikrai) for their helps, care and moral support.

Last but not least, I would like to express my sincere gratitude and deep appreciation to my parents, and my family for constant encouragement, love, and financial supports throughout my life.

CONTENTS

	page
Abstract (Thai)	iv
Abstract (English)	v
Acknowledgements	vi
Contents.....	vii
List of Tables.....	x
List of Figures	xi
Chapter I Introduction	1
1.1 Introduction and Problem Review.....	1
1.2 Definitions	3
1.3 Statement of the Problem	4
1.3.1 Image Search Mechanism.....	5
1.3.2 Routing Module (RM).....	5
1.4 Research Objectives	5
1.5 Scope of the Study	5
1.6 Contributions of the Research	6
1.7 Research Plans	6
1.8 Dissertation Organization	7
Chapter II Background and Literature Review.....	8
2.1 Search Engine	8
2.1.1 Background of the search engine	8
2.2 Overview the Information Search Protocol (ISP).....	9
2.2.1 Embedded Agent for Information Search Protocol (EAISP).....	10
2.2.2 Global Search Engine System (GSES)	12
2.2.2.1 Protocol Interpretation Module (PIM).....	13
2.2.2.2 Search Module (SM).....	13

2.2.2.3 Routing Module (RM).....	13
2.2.2.4 Global Database (GDB)	13
2.2.2.5 Problem of the ISP	14
2.3 Related work of search algorithm	14
2.4 Content-Based Image Retrieval (CBIR)	18
2.4.1 Overview of the Content-Based Image Retrieval (CBIR)	18
2.4.2 Smoothing filter.....	20
2.4.2.1 Average filter	20
2.4.2.2 Median filter	21
2.4.2.3 Gaussian filter.....	22
2.4.3 Color quantization.....	23
2.4.4 Image feature extraction using Color Region Correlation (CRC).....	24
Chapter III Proposed Method	28
3.1 Enhancing Image Feature Extraction.....	28
3.2 Proposed Routing Algorithm.....	34
3.2.1 Patrol Packet (PTP) format.....	34
3.2.2 Patrol Packet (PTP) Algorithm.....	35
3.3 Modified ISP.	37
3.3.1 Format of the Specific Information Search Protocol (SISP).....	38
3.3.2 Embedded Agent for the Specific Information Search Protocol (EASISP)	41
3.3.3 Modified Global Search Engine System (MGSES)	45
3.3.3.1 Classified Module (CM)	45
3.3.3.2 Protocol Interpretation Module (PIM).....	46
3.3.3.3 Search Module (SM).....	47
3.3.3.4 Routing Module (RM).....	51
3.3.3.5 Modified Global Database (MGDB).....	51

3.3.3.6 Example of the PTP Algorithm	53
Chapter IV Implement and Experimental Results	56
4.1 Implementation	56
4.2 Simulation	57
4.2.1 Simulation of proposed routing algorithm.....	58
4.2.2 Simulation of SISP precision	58
4.3 Simulation Results.	58
4.3.1 Number of bytes	58
4.3.1.1 Simulation result of the number of bytes	58
4.3.1.2 Theoretical analysis of the number of bytes	59
4.3.2 Response time.....	62
4.3.2.1 Simulation result of the response time.....	62
4.3.2.2 Theoretical analysis of the response time	63
4.3.3 The performance of the SISP.....	66
4.3.3.1 The information search	66
4.3.3.2 The image search.....	69
Chapter V Discussion and Conclusion	73
5.1 Discussion on Proposed Routing Algorithm	73
5.2 Conclusion on Proposed Routing Algorithm.....	74
5.3 Discussion on the Performance of SISP	75
5.4 Conclusion on the Performance of SISP	76
References.....	78
Biography	88

List of Figures

Figure	page
2.1 The architecture of the ISP	10
2.2 The ISP format	11
2.3 The 3×3 mask of the average filter	20
2.4 Calculating the median value.	21
2.5 An example of the color image quantization	24
2.6 MBRs of region R_1 and R_2 of color c_1	25
3.1 The color quantization of the difference image filter techniques	30
3.2 The process and image example of the enhancing image feature extraction.....	31
3.3 The format of the PTP	34
3.4 The Patrol Packet (PTP) algorithm	36
3.5 The architecture of SISP	38
3.6 The general format of SISP	39
3.7 The SISP format for an information search	39
3.8 The SISP format for an image search.	39
3.9 The algorithm of SISP encapsulation	43
3.10 The algorithm of SISP extraction.....	44
3.11 The modification of GSES.....	45
3.12 The algorithm of Classified Module	46
3.13 The interpretation algorithm.....	47
3.14 The return result algorithm of PIM.....	48
3.15 The algorithm of Search Module.....	50
3.16 The database table of information and image	52
3.17 The database table of routing.....	53
3.18 The example the SISP's flow using the PTP algorithm.....	55
4.1 The SISP browser	57
4.2 Comparison of the number of the sent bytes at n MGSES nodes	59

Figure	page
4.3	The comparison of the number of bytes from the simulation and the number of bytes from the equations 62
4.4	Comparison the response time of three algorithms at n MGSES nodes..... 63
4.5	The flooding algorithm..... 64
4.6	The PTP algorithm 65
4.7	The hierarchical algorithm 66
4.8	Comparison of the number of bytes of SISP and ISP at 15 MGSES nodes 68
4.9	Comparison of the response time of SISP and ISP at 15 MGSES nodes..... 68

List of Tables

Table	page
2.1 Syntactic and Semantics of the ISP format.....	12
3.1 The example of CRC index.....	33
3.2 Syntactic and semantic of the PTP format.....	35
3.3 Syntactic and semantic of the SISP format.....	41
4.1 The description of variables	60
4.2 The information search result of SISP and ISP	67
4.3 The image search result of SISP.....	70
4.3 The image search result of SISP (cont.)	71

CHAPTER I

INTRODUCTION

1.1 Introduction and Problem Review

Over the past decade, many organizations, companies, universities, and governments have implemented web sites to present their information, advertisements or services on the Internet. The Internet is a large source of useful information that is important to users. Additionally, these users can conveniently reach the information or service via the Internet. Presently, the Internet has more information than in the past and the amount of information is growing rapidly. There are many search processes to achieve the required information but it gets more difficult and the returned lists may not be as accurate as desired.

There are various methods to gain access to information from the Internet. One solution is to search from an available search engine. Yahoo, Google, and MSN Search are widely used in order to access information. Since the Internet has huge collections of digital images, image search processes have been developed within the search engines. Many websites have images to provide instant information. Several search engines include a process of an image search facility for users. It is convenient for users to access the information and the image over the Internet. Commonly, the search engine returns a list of web pages. Every search engine allows users to enter key words or required words in the search field. The resulting list obtained from the search process can be short or long, depending on the entered keywords or search methods. Therefore, if the user knows appropriate keywords, the search result can be much specific; otherwise, a long list of documents or irrelevant information (such as information related website advertisements) will be presented. Therefore, it can be time consuming for users to find the required information.

In order to solve this problem, Information Searching Protocol (ISP) [1] was proposed in the year 2006. This protocol enables users to enter a part of the required URL and keywords, e.g. "chula.ac.th" and "mathematics", and the result of search appears as "www.math.sc.chula.ac.th". Thus, the ISP can narrow down the search and reduce the search time for the required information. The ISP and its search mechanism will enhance the efficiency of the search process. Therefore, the ISP is a smart protocol for specific content search over Internet. Unfortunately, the original ISP cannot perform an image search.

One significant factor that affects the performance of the search engine based on the use of the ISP is the search mechanism. In the search mechanism of the original ISP, the query messages are broadcasted to all distributed database servers of the ISP to find a full pathname of the required URL; the result will then be sent back to the client. When the required URL does not exist in the database, a query message is created with new details of destinations and it transfers the query message to other databases that contain links to the current database. The query messages are sent using the flooding technique, where the database server of the ISP will send the query message to all connected links of neighbors except the link of neighbor that sends the query message. Thus, the query message will be sent to the same node several times. As a consequence, congestion happens as the query messages spread. This is a defect of the original ISP.

Following from the problems mentioned above, this research focuses on the implementation of the image search mechanism and the performance of the search mechanism without the congestion problem. The first solution for the current inefficient image search process is adding the image search into the system architecture and mechanism of the ISP. The image search process uses a partial of the image to look for the complete image over the Internet. Thus, the implementation of the image search mechanism enhances the efficiency of the ISP by narrowing down to specific information as needed. Since there are various techniques to find images on the Internet, in this

research the color mapping is applied to the search mechanism of the new version of the ISP. As a consequence, the modified ISP can specify the search both with text and images, and the returned list is shorter.

Another problem in the search mechanism is the congestion where the ISP applies the flooding technique to send queries over the Internet for the required list. In this research, the proposed routing algorithm, named the Patrol Packet (PTP) algorithm, modifies the original flooding mechanism. This altered flooding technique works with a small packet called the Patrol Packet that is sent to check the status of the database of neighbours before sending the modified ISP. Thus, the proposed routing algorithm uses packet messages to find the sending path of the modified ISP without re-bounding the packet.

Since the ISP modifies the format and all mechanisms, the new version of ISP is called Specific Information Searching Protocol (SISP).

1.2 Definitions

DEFINITION 1: A *search protocol* is a protocol of a search engine for searching for and retrieving data. There is a communication standard in the search protocol, such as the format of a query message and an answer message, and a method of exchanging information.

DEFINITION 2: A *routing algorithm* is a search mechanism for finding the required data that is stored on the distributed database. The search mechanism uses a query message that is sent out to the database for looking the required data. When the required data is found, it will be return to the requesting computer.

DEFINITION 3: An *information search* is a search mechanism using the keyword(s).

Consequently, the result of the information search is the list of the URL(s), each URL linking to a web page that contains the keyword(s).

DEFINITION 4: An *image search* is a search mechanism using the color image for finding similar images.

Consequently, the image search result is the list of the URL(s), each URL linking to a web page that contains the required image or a similar image.

DEFINITION 5: A *number of bytes* is the amount of bytes that is sent to all nodes in the distributed system; it is the data flow in the communication channel.

Consequently, if there are too many transferred bytes in the network system, the congestion happens and finally the system is down.

DEFINITION 6: *Response time* is the period of time from sending a query message until receiving the result message.

DEFINITION 7: *Precision* is the fraction of relevant items within the set of retrieved items.

DEFINITION 8: *Relevant image* is an image that is the similar color as the query image and the image is contained in the web page of the require URL.

1.3 Statement of the Problems

Since the Information Searching Protocol (ISP) is a search protocol that supports the multiple search contents for users who know a partial pathname of a required URL and keywords, this protocol can narrow down the search and obtain the search results in a shorter period. Although the ISP is a smart protocol for specific content search over the Internet, its mechanism needs to be improved to increase the performance.

With respect to the ISP mechanism, there are two problems to be considered as follows.

1.3.1 Image Search Mechanism

Since most websites have images to represent their information, the Internet stores a large volume of digital images. Many search engines are embedded with image search mechanisms. The current image search engines allow users to enter keywords or images for the search. The resulting list of the search process often returns too many images and users spend too much time looking for the required image. Therefore, the image search process is added in the architecture system and the mechanism of the ISP to increase the image search performance.

1.3.2 Routing Module (RM)

The previous version of the ISP had a problem in the Routing Module (RM). When the required URL does not exist in the database, the query message will be transferred to other databases that contain links to the current database, with the exception of the link of the neighbor that sends the query message. The routing algorithm uses the flooding techniques for searching the required data on the distributed databases. Thus, the query message is sent to the same node several times. This process is time consuming and causes congestion in the communication channel.

1.4 Research Objectives

The main objectives of this research are listed below:

- To develop a new search protocol for information and images over the Internet.
- To develop a new architecture and search mechanism using the proposed search protocol.
- To develop a routing mechanism of the proposed search protocol.

1.5 Scope of the Study

In this research, the scope of work is constrained as follows.

- To propose a new routing algorithm that supports the routing module of the ISP.
- To implement the ISP that can search information and images on websites.
- An administrator who implements the database of the ISP.
- In the image search, an input image for searching is only a color images.

1.6 Contributions of the Research

This research proposes a new search protocol, the SISP, which supports a high performance search process for all required information with a short response time and a specific returned list as expected. This new protocol enhances the search ability of the Internet users so that the search time can be shortened and the congestion in the communication channel can be avoided.

1.7 Research Plans

In order to achieve these defined objectives, the following tasks will be stated by means of appropriately related work and theoretical techniques.

1. Study concepts of Information Search Protocol (ISP), routing algorithms, search engine systems, search protocols, and content-based image retrieval.
2. Identify the area of the Information Search Protocol (ISP) that is a routing algorithm problem and the image search mechanism.
3. Define a routing algorithm to solve the causes of congestion in the communication channel, i.e. the query message being sent to the same node several times.
4. Define an architecture of a new search engine system for solving the routing algorithm and adding the image search mechanism of the ISP.
5. Define a format of search protocol on the architecture of a new version of ISP according to 4.
6. Implement and simulate the new version of the ISP.

7. Test and evaluate the new version of the ISP and report the results.

1.8 Dissertation Organization

The remainder of the dissertation is organized into four additional chapters as follows.

Chapter 2 introduces the background and related work of the search engine, and the overview of the Information Search Protocol (ISP). Moreover, the related work of the search algorithm and the related work of the content-based image retrieval are presented in this Chapter.

In Chapter 3, the new architecture and mechanism of the ISP that solves the problems of the current ISP is described. This Chapter consists of 3 parts. The first part presents the proposed image feature extraction; the second part explains the proposed routing algorithm; and the third describes the modified ISP that is applied to work with the image search mechanism and the routing algorithm.

Chapter 4 explains the implementation and the simulation of the new version of the ISP. Furthermore, its simulation results and performance analyses are presented in this Chapter in order to show the performance of the proposed routing algorithm and the performance of the new version of the ISP.

Finally, Chapter 5 is the discussion and the conclusions of the proposed routing algorithm and the new version of the ISP.

CHAPTER II

BACKGROUND AND LITERATURE REVIEW

This chapter is divided into 4 parts. The first part discusses search engines. The second part provides an overview of the Information Search Protocol (ISP) and the third part is the related work of the search algorithm. The final part discusses the content-based image retrieval process.

2.1 Search Engine

2.1.1 Background of the search engine

A search engine is a program for looking for web sites in the Internet, which returns a list of matching websites [2]. There are three components of a search engine [3]: Collection, Database, and Search Interface. The first component is Collection. It uses a "Robot" or "Spider" to read the contents in a web page, and if the web page has hyperlinks to other pages, the robot will follow those links to the other pages. Commonly, the robot returns to the site for updating the search engine database every week or month. The data found by the spider is sorted, indexed, and stored in the database. The database is the second component. It contains all the information created from the robot's search. The database copies every web page which the robot visited. The third component is the Search Interface; the interface between the database and the users. Some engines develop a complex strategy with Boolean operators, phrase and proximity searching, and nesting; other engines use a simple keyword for searching. Additionally, there are some mega-tools (meta-search engines) that can search on multiple engine databases using a single interface.

Since the Internet contains a large amount of information, the search engines manage their mechanism for user searching. There are three types of search engines [4]: Human Powered Directories, Crawler Based Search Engines, and Hybrid Search Engines.

Human Powered Directories: The directories of these search engines are prepared by humans. Human editors collect all the listings in directories. They are organized into subject categories and the subjects classify the web pages. Since people look in the listings, the keywords used in web directories are important. The listings of human powered directories are smaller than the listings of most search engines [5]. The two most important directories are the Yahoo Directory and the Open Directory known as Dmoz [6].

Crawler-Based Search Engines: The listings of crawler based search engines are automatically generated. The listings are built by the spiders; not by human collection. The subject categories are not organized; all pages are ranked by a computer algorithm. The search engines are huge and often retrieve a lot of information. In advanced searches, the search engines allow searches within the results of a previous search and enable the user to refine search results [4].

Hybrid Search Engine: A hybrid search engine is different from the traditional text-oriented search engine or the directory-based search engine. This search engine performs by comparing a set of metadata, the primary corpus being the metadata derived from a Web crawler or taxonomic analysis of all contents on the web pages, and the search query of the user [4].

Moreover, most search engines utilize relevancy searching. The search results are ranked, or sorted, according to the criteria. The criteria in the search engines can contain the following: number of terms matched, closeness of terms, location of terms within the document, frequency of terms, document length, and other factors [7].

2.2 Overview of the Information Search Protocol (ISP)

Information Searching Protocol (ISP) [1] is a search protocol that provides support for users who know a partial pathname of a required URL and keywords. An example would be the user entering “www.chula.ac.th” and “math”, the result of search being web pages that have “math” in “www.chula.ac.th”.

The architecture of the ISP consists of 2 main processes. The first process is called the Embedded Agent for Information Search Protocol (EAISP). The second process is performed at the search engine system and is called the Global Search Engine System (GSES). The architecture of the ISP system is presented in Figure 2.1 and the detail of the ISP is illustrated as follows:

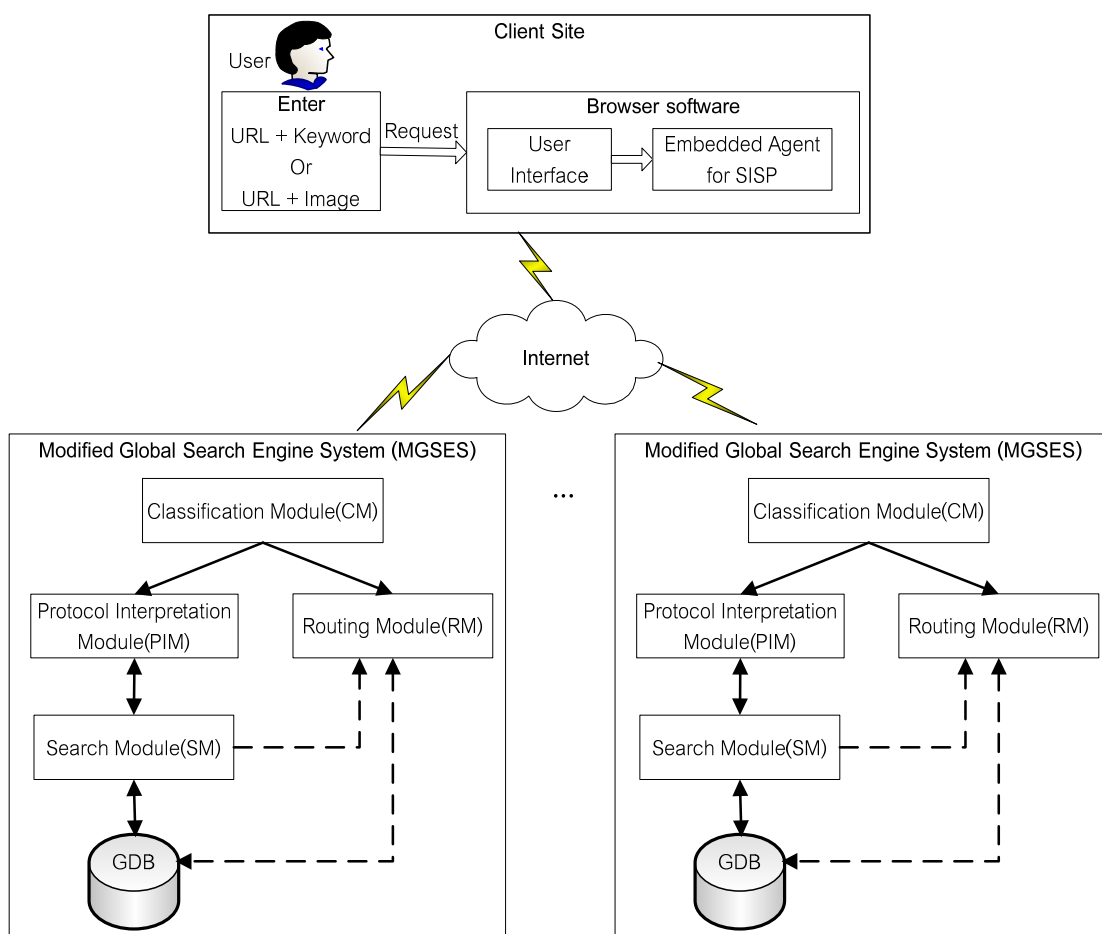


Figure 2.1 The architecture of the ISP

2.2.1 Embedded Agent for Information Search Protocol (EAISP)

This process is located at the client site where it receives a required URL and keywords from a user. The information will be encapsulated into the ISP format before transferring to the second process. The format of the ISP is presented in Figure 2.2.

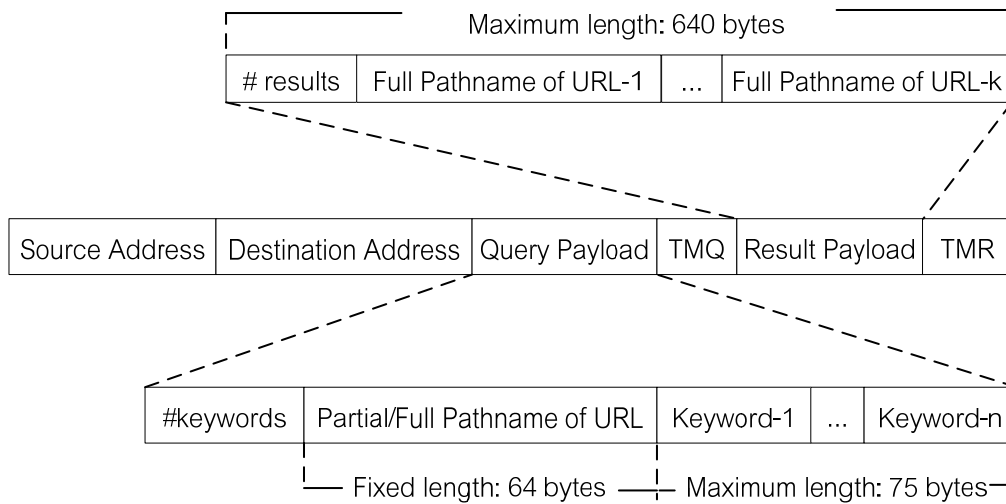


Figure 2.2 The ISP format

In Figure 2.2, the length of the ISP is a dynamic value depending on the number of keywords and the result list. However, the format has two terminate values to indicate the end of the query payload, or the end of the result payload. Each attribute in Figure 2.2 is described in Table 2.1.

Table 2.1 Syntactic and Semantics of the ISP format

Attribute name	Length (byte)	Remark
Source Address	4	The IP address of the user system
Destination Address	4	The IP address of the database search engine
#Keywords	1	Number of search keywords usually is equal to the value of #keywords. Total number of keywords must not exceed 5 words.
Partial/Full Pathname of URL	64	Partial or full pathname of the required URL. Users can enter only one value.
Keyword-i	15	Search keywords
TMQ	1	Terminate key, indicates the end of query text; fixed value in hexadecimal is 'FF'.
#result	1	The total number of URLs listed from the search mechanism; the value of this attribute is k.
Full Pathname of each URL	64	The pathname of the web, may have more than one result; total number of the URLs is not exceed 10 URLs
TMR	1	Terminate key, indicates the end of result text; fixed value in hexadecimal is 'FF'.

When a result is sent back to the client, the EAISP extracts the result payload and sends the full pathname of the URL(s) to the client's user interface of the browser.

2.2.2 Global Search Engine System (GSES).

The GSES consists of three modules: Protocol Interpretation Module (PIM), Search Module (SM), and Routing Module (RM). Moreover, a Global Database (GDB) is installed to store full path of URLs and keywords under that URL. Functions

of each module are described in the following sub topics.

2.2.2.1 Protocol Interpretation Module (PIM)

This module is responsible for extracting the ISP in order to obtain the partial pathname of a URL and keywords of the required content. After obtaining the result of extraction, this result is used to create a search command in an SQL statement. Additionally, when the PIM receives the results from the SM, it will encapsulate the results into the ISP and it send back to the client

2.2.2.2 Search Module (SM)

This module receives the SQL statement from the PIM. The results of the retrieving process can be a full pathname of a URL or a set of full pathnames of URLs. However, there is the possibility that the required URL does not exist in GDB. In this case, the SM will automatically transfer to the Routing Module; otherwise, it will return the results back to the PIM to transfer the result back to the client.

2.2.2.3 Routing Module (RM)

When the required URL does not exist in the GDB, the SM will send a message to the RM. The RM will retrieve IP addresses of other GSEs from the GDB, create the ISPs with new details of the destination, and transfer the ISP to other GSEs that have links to the current GSE. There is a limit for the transfer using a time-to-leave (TTL) technique. The ISP will be transferred until the TTL is equal to zero and the last GSE will send the ISP with a NULL-result back to the client. Therefore, this constraint can eliminate the infinite recursion of the transferring routine.

2.2.2.4 Global Database (GDB)

Each GSE has an installed database named the Global Database (GDB). The data in the GDB consists of fields that indicate the full pathname of URLs, the IP address

of each URL, and defined keywords under each URL. Moreover, the GDB has the address of the GSESs that have links to this GSES.

2.2.2.5 Problem of the ISP

According to the overview of the ISP, the previous version of the ISP has a problem in the Routing Module (RM). When the GSES receives the ISP, the ISP is extracted in order to obtain the partial pathname of a URL and keywords in the Protocol Interpretation Module (PIM), and then the required URL and keywords are sent to the Search Module (SM) for the retrieving process. If the required URL does not exist in the GDB, the SM will send a message to the RM. The RM will retrieve IP addresses of other GSESs from the GDB, create new ISPs that change details of destinations, and send out the ISP to other GSESs that contain links to the current GSES. The routing algorithm of the RM uses flooding techniques for searching the required data on the distributed GSESs. This process is time consuming and causes congestion in the communication channel while searching for the required URL in the GDB of the other GSESs.

2.3 Related work of search algorithm

There are a number of systems that are distributed systems which share resources distributed over the Internet, such as peer-to-peer systems [9], naming systems [10], and distributed databases [11]. Thus, a search resource is important in the distributed system. The objective of a search mechanism is to successfully locate resources while incurring low overhead and delay [12]. Therefore, many researches propose mechanisms to send the query messages to the correct locations that store the requested resource using a minimal number of transferring the query messages. There are various search techniques that have the different trade-offs in their desired characteristics [13-16].

The search techniques of distributed systems are basically classified into two search methods: the informed search, and the blind search [14, 17]. In the informed search method, each node has routing information to refer to the location of the required

data. The other method is the blind search, in which each node does not collect information that supports the search process.

In the informed search, each node collects some routing information that contains parameters for selecting the appropriate neighbor nodes to relay query messages. There are various schemes in the informed search. One of the informed search techniques uses distributed hash tables (DHTs), which are lookup tables of distributed resources (i.e. Chord [18]) in peer-to-peer systems. One of the informed search methods is a semantic model that is used in REMINDIN'[19] and LearningPeerSelection [20]. Another method is a history-of-queries model. The example of history-of-queries is Directed Breadth First Search (DBFS) [21], which proposes to record the neighbor nodes that will quickly return answer messages. The next example is routing indices [22], which proposes to store statistics of document share's neighbors and route the query to a good node. Additionally, Activity Based Search [23] utilizes a method to send the query messages to neighbors that have contributed to previous successful searches. Moreover, several researches [24, 25] improve the informed search using cache technology to store routing information for decreasing response time.

An advantage of the informed search method is the reply time of the result messages, which is less than the blind search method. Moreover, the number of sending query messages is also less than the blind search. However, this search method has a high cost to maintain the routing information. The implementing cost of the informed search is to prepare the parameter of routing information before starting the search method. Another cost is routing information updating. If data on one server is changed, all of its neighbors have to update their routing information. Additionally, the informed search method chooses only some servers that have a higher probability of finding the required data. The query message does not relay to all of the nodes in the system. Thus, if the routing information is not correct, the query message cannot reach the right location and the search might return an incorrect result.

Another basic search method is the blind search. This method uses flooding techniques for sending queries to other servers in the network. For example, Breadth First Search (BFS) is used in peer-to-peer systems such as Gnutella. A node sends a query message to its neighbors, except the neighbor that sent the query message. The starting node initializes a time-to-live (TTL) value in the query message. The TTL is decremented at each hop, with the packet being discarded when the counter becomes zero. The TTL is used for limiting flooding query messages.

The next example is Random Breadth First Search (RBFS) [26], which improves the flooding technique by randomly selecting a portion of its neighbors to send the query messages. RBFS reduces numbers of query messages that are passed on the network. Normalized Flooding [25] uses flooding technique; however, it sends the query messages to some of its neighbors. The number of selected neighbors (δ) is the minimum degree of a node in the network. If a server has neighbor degrees more than δ , it will send the message only to δ nodes in its neighborhood that are randomly selected. Iterative Deeping [13, 17, 27] uses a depth-first search to set the depth in the starting search. The nodes relay the query message using breadth-first search. The search is done when the depth reaches the maximum limit or when the server found the required data. Random walk [13] is another example of the blind search. The search method of random walk is that the node sends one query message to a neighbor that is randomly chosen from all its neighbors. The forwarding message is a randomly chosen neighbor in each step. The query message is called a walker. This method reduces the flooding messages; however, there is a delay in returning the result to user. Therefore, a k-walks algorithm is proposed [28]. The server sends k query messages to an equal number of randomly chosen neighbors, and then each neighbor sends out one message to the next neighbor that is randomly chosen. Another example of random walk is a two-level random walk [29]. There are two policies for the random walk. At the first level, the server selects k_1 random walk with $TTL_1 = l_1$. When the TTL_1 is zero at a particular server, the second policy will be started. The server selects k_2 random walk with $TTL_2 = l_2$.

The next search method is the local flooding with k independent random walks [28]. The idea of this search method is to combine the flooding and random walk. First, the starting node is sent to k neighbors by a local flooding, the value of k neighbor is predefined. If the search finds the required data, the data will be returned. Otherwise, each k nodes starts the independent random walk.

For the random search method, the advantage is to reduce the query messages overhead. However, the random search does not reach all of servers in the system; it selects some nodes that have a higher probability of finding the required data. The query message may not arrive at the right location. Thus, the required information might not be found.

In order to reduce the number of sending query messages, a hierarchical mechanism is considered, such as the naming system. Most naming systems use a hierarchical mechanism to search IP addresses. A well-known naming system is Domain Name System (DNS) [10, 30-32], which maps host names to numerical IP address. The naming structure database is strictly hierarchical of zone. The query messages are sent to the parent and children of the node. Thus, this mechanism can reduce the overhead of sending the query messages and send the query messages to all nodes in the system. However, the hierarchical search mechanism will take a long time if the required data exists in other branches of the hierarchical structure.

Considering the existing search techniques that have been mention above, all search techniques may not send the query message to all nodes on the system, except in the flooding technique and hierarchical mechanism. If the query messages reach all nodes, it can be guaranteed that the results are from the right location. The response time of the hierarchical mechanism may take more time than the response time of flooding techniques if the required data exists in other branches of the hierarchical structure. Thus, the flooding technique is focused on in this research. The advantage of flooding technique is a short response time to return the result messages and the cost of implementation is low. Additionally, this technique can start without preparing the routing

information. However, the flooding technique is a cause of the congestion by distributing query messages.

2.4 Content-Based Image Retrieval (CBIR)

This section provides an overview of Content-Based Image Retrieval (CBIR). Moreover, the background of image processing for image search methods is also presented. These image backgrounds are applied in this research. There are the smoothing filter, the color image quantization, and the image feature extraction using Color Region Correlation (CRC).

2.4.1 Overview of the Content-Based Image Retrieval (CBIR)

Content-based image retrieval (CBIR) [33-40] is a technique used for retrieving similar images from an image database. Content-based image retrieval has been a dynamic field of study for many years [41-42]. The primary features of content-based image retrieval are color, shape, and texture [43]. The image retrieval consists of two main processes; a feature extraction and a feature matching. The feature extraction is most important. The feature should be invariant to image translate, rotate and scale [44].

Example of CBIR algorithm is Query-By-Image Content (QBIC) [45-51]. The QBIC is the first commercial content-based image retrieval system. The system framework and techniques of QBIC have contributed to later image retrieval systems. The color features of QBIC are the average value of (R,G,B), (Y,i,q), (L,a,b), and MTM (Mathematical Transform to Munsell) coordinates, and a k-element color histogram [47]. The texture feature of the QBIC improved the Tamura texture representation [52] (such as combinations of coarseness, contrast, and directionality [46]). In shape feature, the QBIC use shape area, eccentricity, circularity, major axis orientation, and a set of algebraic moment invariants [47, 51].

In the example, there are three features that are used in QBIC: color, texture, and shape. However, the color feature has been extensively studied because it is invariant to scale and translate [53].

In color feature, the color histogram is most commonly applied with color feature representation. The histogram intersection for measuring the similarity of the color histogram is proposed by Swain and Ballard [57]. Ioka [58] and Niblack et al. [59] also introduced the technique to compare the histograms for the similarity between two images. However, most color histograms are very sparse and thus sensitive to noise. Therefore, the cumulated color histogram was proposed by Stricker and Orengo. Their research results presented the advantages of the proposed method over the conventional color histogram method [60].

Moreover, several other color features have been used in image retrieval, such as color moments and color sets. These color features can solve the quantization effects of image in the color histogram. The color moments process was proposed by Stricker and Orengo [60]. The mathematical foundation calculates any color distribution that can be characterized by its moments. Additionally, the information is considered on the low-order moments. The first moment is mean value, and the second and third central moments that are variance value and skewness value, respectively, and were extracted as the color feature representation. Therefore, the weighted Euclidean distance was used for calculating the color similarity of two images.

In addition, an approximation to the color histogram is proposed by Smith and Chang [61, 62]. This method is used for fast searching over large-scale image collections. The first process is to transform the (R, G, B) color space into a perceptually uniform space, for example, HSV, and then quantized the transformed color space into M colors. A color set is defined as a collection of colors from the quantized color space. Since the feature vectors of the color set were binary, a binary search tree was generated. Therefore, the image search is a fast search.

2.4.2 Smoothing filter

The smoothing filter is often used to reduce noise within the image or blur the image. The smoothing image is used in the preprocessing steps, such as removal of a small detail from the image prior to object extraction, or the connecting of a small gap in lines or curves. Moreover, the blurring image can reduce noise on the image [63]. The smoothing filter is usually based on a single value representing the image, such as the average value of the image and the median value. The two values for smoothing an image are called an average filter and a median filter. Additionally, another smoothing filter is a Gaussian filter.

2.4.2.1 Average filter

The average filter or the mean filter is a simple method to implement for smoothing images. Moreover, the average filter is used in Gaussian noise reduction [64-67]. The mean filter is also applied with a derived or texture feature in segmentation process of image [68, 69].

The idea of the average filter is simply to change each pixel value in an image to the average value of its neighbors, including itself [63]. This process has the effect of removing pixel values which are unrepresentative of their surroundings. The average filter is usually thought of as a convolution filter. The mask is the shape and size of the neighborhood to be sampled when computing the mean. The 3×3 mask size is often used, as shown in Figure 2.3. Additionally, if the mask size is bigger, the image will be more smoothing.

$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

Figure 2.3 3×3 mask of the average filter

2.4.2.2 Median filter

The median filter is a nonlinear filter used to eliminate the impulsive noise from an image [70-73]. The method of the median filter is similar to the average filter; however, the median filter is better than the mean filter for preserving useful detail in the image. Furthermore, the median filter is a more robust method than traditional linear filtering because the median filter maintains the sharp edges. Additionally, the image's "salt and pepper" noise can be removed by the median filter [74].

The idea of the median filter is similar to the mean filter; the median filter compares each pixel in the image with its neighbors. Instead of replacing the pixel value with the average of neighboring pixel values, the median filter replaces pixel values with the median of neighboring pixel values. The median value is computed by first sorting all the pixel values from the nearby neighbors into numerical order. Then the considered pixel is replaced by the middle pixel value. If the number of the pixel value is an even number, the average of the two middle pixel values is used. Figure 2.4 presents an example calculation of median filter.

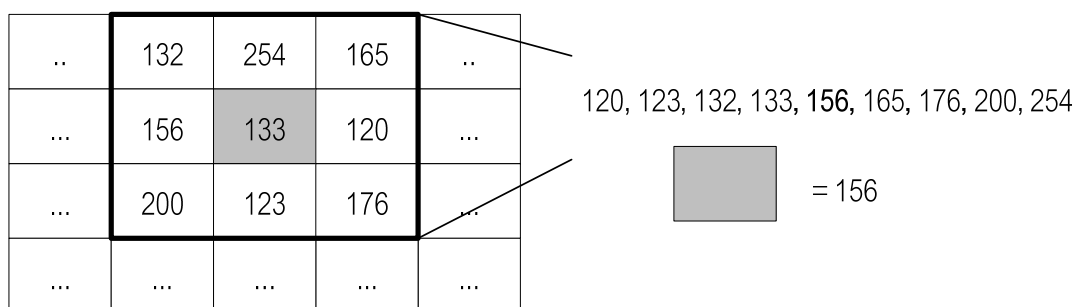


Figure 2.4 Calculating the median value

2.4.2.3 Gaussian filter

A Gaussian filter also known as Gaussian smoothing or Gaussian blur [75]. The Gaussian filter is the blurring of the image using a Gaussian function. The Gaussian filter is also used to remove noise. The Gaussian filter is similar to the average filter; however,

it uses a different mask that is the shape of a Gaussian ('bell-shaped') hump. The mask of the Gaussian filter is explained below.

The Gaussian filter aims to reduce the magnitude of high spatial frequencies in the image proportional to their frequencies. The Gaussian filter decreases magnitude of higher frequencies more. There is more computation time when compared to the average filter. The Gaussian expands to infinity in all directions; however, since it approaches zero exponentially, it can be truncated three or four standard deviation away from its center without noticeably affecting the result.

The equations of a Gaussian function of one dimension and two dimensions are shown in equations 2.1 and 2.2, respectively.

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right) \quad (2.1)$$

$$G(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x^2 + y^2)}{2\sigma^2}\right) \quad (2.2)$$

Where σ is the standard deviation of the Gaussian distribution, x is the distance from the origin in the horizontal axis, and y is the distance from the origin in the vertical axis. When the Gaussian filter is used in two dimensions, the formula generates a surface whose contours are concentric circles with a Gaussian distribution from the center point. The values of the distribution are applied to construct a convolution matrix that is used to the original image. The new value of each pixel is set to a weighted average of that neighborhood pixel. The value of original pixel receives the highest weight and neighboring pixels receive smaller weights as their distance from the original pixel increases. The result is a blur image that preserves boundaries and edges.

2.4.3 Color quantization

Color image quantization is the process of reducing the number of colors in a digital color image [76]. Color image quantization is an important method that can be

applied to many techniques in image processing and computer graphics. Color image quantization can be used in loss compression techniques [77]. A mobile system uses the Color image quantization because the memory of mobile devices is usually small [78]. Moreover, most graphics hardware uses the color image quantization for color lookup tables with a limited number of colors [79]. The Color image quantization can be properly defined as follows [77]:

Let a set of $N_{S'}$ colors where $S' \subset \mathfrak{R}^{N_d}$ and N_d is the dimension of the data space. The color quantization is a map $f_q : S' \rightarrow S''$ where S'' is a set of $N_{S''}$ colors such that $S'' \subset S'$ and $N_{S''} < N_{S'}$. The objective of the map is to minimize the quantization error resulting from replacing a color $c \in S'$ with its quantized value $f_q(c) \rightarrow S''$.

Color image quantization consists of two main steps:

- First is to create a color map that is a small set of colors (normally 8-256 [80]), chosen from the 2^{24} possible combinations of red, green, and blue (RGB) color.
- The second step is to map each color pixel in the color image to one of the colors in the color map.

The example of the color image quantization is illustrated in Figure 2.5.

- The second step is to map each color pixel in the color image to one of the colors in the color map.

The example of the color image quantization is illustrated in Figure 2.5.

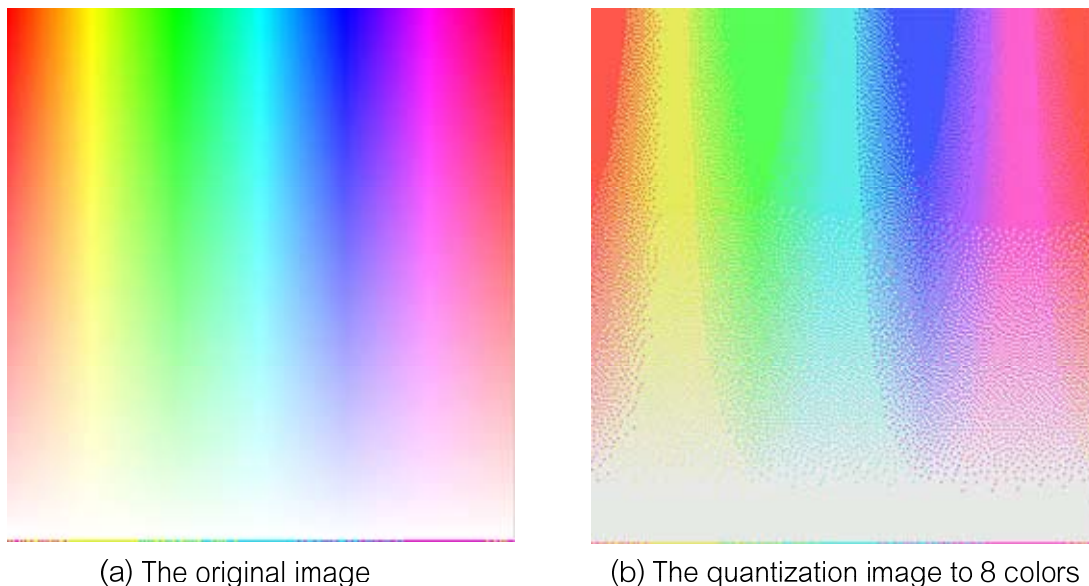


Figure 2.5 An example of the color image quantization

Therefore, the idea of color image quantization is a mapping of the set of colors in the original color image to a much smaller set of colors in the quantized image [81]. Additionally, the color image quantization should reduce the difference between the original images and the quantized images [79].

2.4.4 Image feature extraction using Color Region Correlation (CRC)

This paper focuses on the color feature. The image technique to obtain image indexes, Color Region Correlation (CRC) [82], is applied to work with the proposed protocol. The index of CRC is easy to calculate and invariant to translate, rotate and scale. Moreover, the size of feature index is small and can apply to the new version of the ISP.

The Color Region Correlation (CRC) is the relationship between color-pair regions in an image. The CRC process is used for obtaining the color index. The process of CRC is explained as follows:

Let I be a digital image; the size of I is denoted as $I = m \times n$ pixel. The colors in I are quantized into l colors, c_1, c_2, \dots, c_l . A region $R_{c_i}(I)$ is a group of connected pixels of color c_i in I .

A color-pair relationship is calculated, and then the result is collected in a table of numbers representing the correlation of regions that is called the CRC. Each value of the CRC table, $CRC_{c_i c_j}(I)$, is the average number of regions of color c_j within a minimum bounding rectangle (MBR) of $R_{c_i}(I)$, defined as $MBR(R_{c_i}(I))$. The MBR must fit the region under consideration. The $MBRs$ of R_{c_i} is illustrated in Figure 2.6. Therefore, before $CRC_{c_i c_j}(I)$ can be calculated, the number of $R_{c_j}(I)$ within $MBR(R_{c_i}(I))$, denoted as $nCRC_{c_i c_j}(I)$, must be counted as follows:

$$nCRC_{c_i c_j}(I) = \left| \left\{ R_{c_j}(I) \in MBR(R_{c_i}(I)) \right\} \right| \quad (2.3)$$

And then $CRC_{c_i c_j}(I)$ is computed as follows:

$$CRC_{c_i c_j}(I) = \frac{nCRC_{c_i c_j}(I)}{|R_{c_i}(I)|} \quad (2.4)$$

Where $|R_{c_i}(I)|$ is the number of regions of color(s)

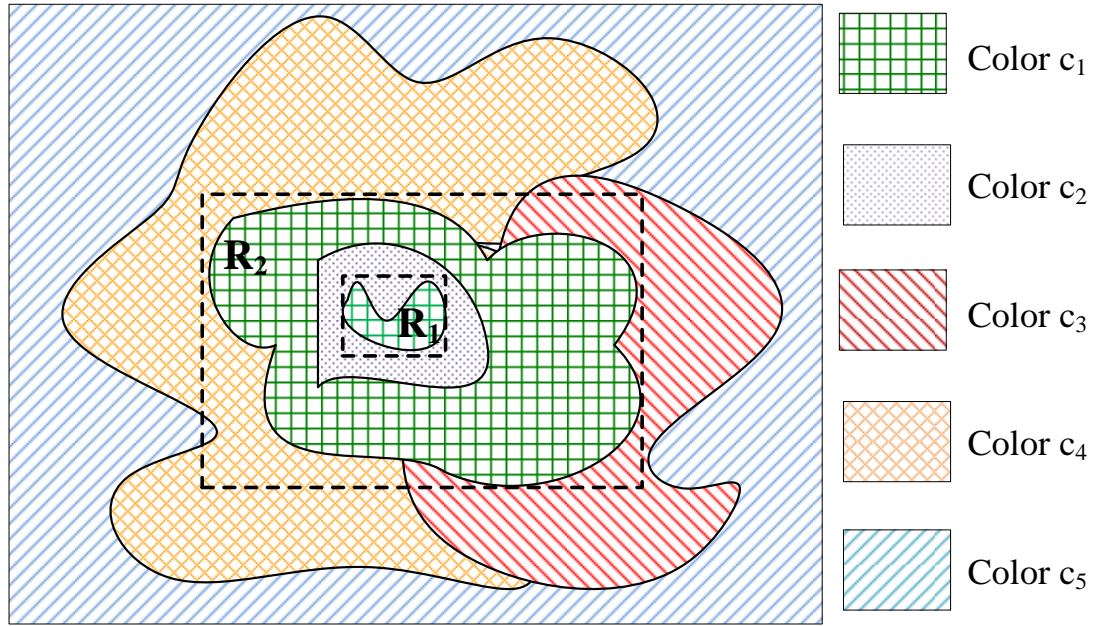


Figure 2.6 MBRs of region R_1 and R_2 of color c_1

Figure 2.6 illustrates MBRs of regions of R_1 and R_2 of color c_1 represented by dashed-line rectangles. The $nCRC_{c_i, c_j}(I)$ of region R_1 can be calculated as follows:

$$nCRC_{c_1 c_1}(I) = 0$$

$$nCRC_{c_1 c_2}(I) = 1$$

$$nCRC_{c_1 c_3}(I) = 0$$

$$nCRC_{c_1 c_4}(I) = 0$$

$$nCRC_{c_1 c_5}(I) = 0$$

Region R_2 can be computed as:

$$nCRC_{c_1 c_1}(I) = 1$$

$$nCRC_{c_1 c_2}(I) = 1$$

$$nCRC_{c_1 c_3}(I) = 1$$

$$nCRC_{c_1 c_4}(I) = 1$$

$$nCRC_{c_1 c_5}(I) = 0$$

Therefore, the summation of the two regions is calculated as follow:

$$nCRC_{c_1 c_1}(I) = 1$$

$$nCRC_{c_1 c_2}(I) = 2$$

$$nCRC_{c_1c_3}(I) = 1$$

$$nCRC_{c_1c_4}(I) = 1$$

$$nCRC_{c_1c_5}(I) = 0$$

The next step is to compute the $CRC_{c_i c_j}(I)$ value:

$$CRC_{c_1c_1}(I) = 1/2 = 0.5$$

$$CRC_{c_1c_2}(I) = 2/2 = 1$$

$$CRC_{c_1c_3}(I) = 1/2 = 0.5$$

$$CRC_{c_1c_4}(I) = 1/2 = 0.5$$

$$CRC_{c_1c_5}(I) = 0/2 = 0$$

However, this calculation only presents the correlations of region of color c_1 and other colors. For the entire CRC table, correlations of other regions of all colors must be calculated. Since CRC calculates every region of each color, it spends considerable time to calculate the entire image.

According to the background and related work in this Chapter, the current search engines return a large number of URLs and images, so the users spend time looking for the required URL. One solution is to use Information Searching Protocol (ISP). This protocol enables users to enter a partial part of the required URL and keywords. Thus, the ISP can narrow down the search and reduce the search time for the required information. Conversely, the routing algorithm of ISP has a defect in that any ISP will be sent to the same search engine system several times, and finally this process causes the traffic overflow. Moreover, the ISP cannot perform the image search. Therefore, this research proposes a new routing algorithm and adds the search image mechanism in the ISP. Considering the search image mechanism, the Color Region Correlation (CRC) is focused on because the index of the CRC is easy to calculate and invariant to translate, rotate, and scale. Moreover, the size of feature index is small. However, the CRC index will spend considerable time calculating if the image has many color regions. Therefore, this research improves the CRC method using the median filter and color quantization technique. The detail of the enhanced CRC method, the proposed routing

algorithm and the new version of the ISP that apply the proposed routing algorithm and the search mechanism are explain in the next Chapter.

CHAPTER III

PROPOSED METHOD

This chapter describes the proposed image search mechanism and the proposed routing algorithm that are applied to increase the performance of the ISP. This chapter is divided into three parts. The first part presents the enhancing image feature extraction. The second part explains the proposed routing algorithm and the third part describes the modified ISP that applies to work with the image search mechanism and the routing algorithm.

3.1 Enhancing Image Feature Extraction

Since the image search aims to find a similar image in a database, image extraction is an important process to extract features of images that are used to compare the similarity between two images. This research focuses on Color Region Correlation (CRC) [82] for retrieving the image because the index of CRC is easily calculated and invariant to translate, rotate, and scale. Moreover, the size of the feature index is small. Therefore, CRC is applied to work with the new version of ISP. The detail of CRC technique is discussed in Chapter II.

The CRC is a set of values representing the correlation of regions by color-pair relationship. The first step of the CRC method is a color quantization of an image in order to reduce the number of colors in the image. In this research, the image is quantized to 8 colors. In the next step, the quantized image is calculated. Each value of CRC is the average number of regions of each color within a minimum bounding rectangle (MBR).

Since this method calculates every region of each color, it takes some time to calculate the entire image. Therefore, this research uses the image filter technique to reduce the number of color regions before quantizing the color image. The image filter techniques of the average filter, the median filter and the gaussian filter are applied to

the image before the quantized color image. Figure 3.1 shows the color quantization effect on the different image filter techniques.



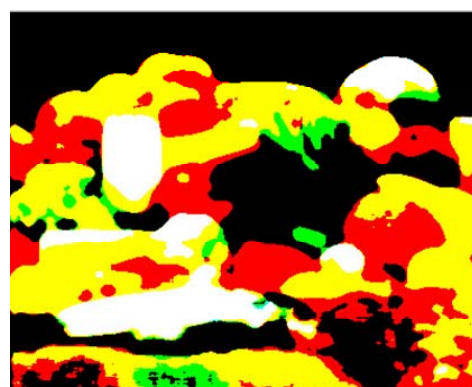
(a) Original image



(b) Quantized color on original image
(3351 color regions)



(c) Quantized color on average filter image
(571 color regions)



(d) Quantized color on median filter image
(293 color regions)



(e) Quantized color on Gaussian filter image
(613 color regions)

Figure 3.1 The color quantization of the different image filter techniques

In Figure 3.1, the number of color regions of the original image quantization is 3,351, the number of color regions of the average filter quantization is 571, the number of color regions of the gaussian filter is 613, and the number of color regions of the median filter is 293. Therefore, the median filter method is the best filter technique because the number of color regions of the median method is minimal. Consequently, the time spent calculating the CRC is shorter using the median filter method. This research adds the median filter method in the CRC process. The process of the enhancing image feature extraction is presented in Figure 3.2.

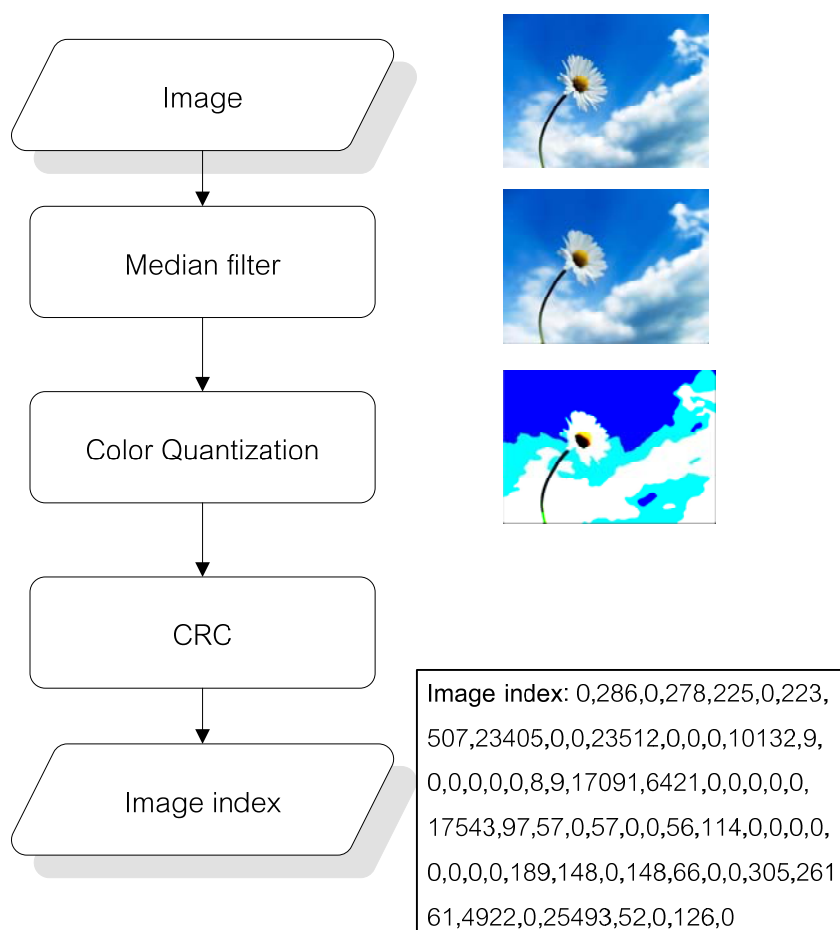


Figure 3.2 The process and image example of the enhancing image feature extraction

As can be seen in Figure 3.2, the enhancing image feature extraction has three processes: the median filter process, the color quantization process, and the CRC process. When the median filter process acquires the input image, it will remove small

details of the image. Then, the median filter image is forwarded to the color quantization process. The color of the median filter image is reduce to eight colors. In the next step, the CRC process obtains the quantized color image, and the result of the CRC process is the CRC index. The size of the CRC index is $8 \times 8 = 64$ values because the CRC is calculated from eight quantized colors. Therefore, the size of the color index is small. The enhanced CRC feature extraction will be applied to work in the embedded process of the new version of the ISP. Moreover, the CRC index of each image will be recorded in the database of the new version of the ISP.

In image search, there is a function that calculates the similarity between the query image and the images in the database. This function is called Image Matching, a function that uses relative distances to calculate the CRC index of two images. The relative distance measure equation is presented in Equation 3.1 and Equation 3.2 as follows.

$$d(Q, P) = \|Q - P\| \quad (3.1)$$

$$= \sum_{i=1}^8 \frac{|q_i - p_i|}{1 + q_i + p_i} \quad (3.2)$$

Where Q is query image.




P is a database image.

q_i is CRC index i of query image

p_i is CRC index i of image database

As an example calculation of the similarity between two images, let IM_q be the query image that is entered by user. $IM1_{db}$ and $IM2_{db}$ are the images that are stored on the database. The CRC index of the 3 images is shown in Table 3.1 and Equation 3.1 and Equation 3.2 are the equations for calculating the similarity value of the images.

Table 3.1 The example of CRC index

Image Name	Image	CRC index
IM_q		0,909,22,1110,353,0,445,1671,57762,0,0, 52296,0,0,0,31773,73,6,0,57,0,0,0,74,23726, 13865,0,0,0,0,0,25130,142,77,0,77,0,0,77, 154,0,0,0,0,0,0,0,335,222,0,222,132,0,0, 450,40718,18446,8,46159,10,0,128,0
$IM1_{db}$		0,2331,5,1667,760,0,953,2130,80861,0,0, 22975,0,0,0,21859,51,29,0,14,1,0,10,30, 24297,14699,1,0,0,0,2,25523,408,217,0, 28,0,0,217,217,0,0,0,0,0,0,0,891,548,0,22, 396,0,0,548,46894,25016,1,43642,20,0,307,0
$IM2_{db}$		0,410994,124670,427570,0,7,0,49166,235092, 0,1,255598,0,854,0,33204,129670,0,0,39313,0, 0,0,0,285082,165424,22968,0,0,203,0,48439,0, 0,0,0,0,0,0,0,1136,2223,0,1142,0,0,0,1297,0,0,0, 0,0,0,0,0,55716,59680,0,58003,0,1504,0,0

$$d(IM_q, IM1_{db}) = \left(\frac{|0-0|}{1+0+0}\right) + \left(\frac{|909-2331|}{1+909+2331}\right) \cdots + \left(\frac{|0-0|}{1+0+0}\right) = 13.3282 \quad (3.3)$$

$$d(IM_q, IM2_{db}) = \left(\frac{|0-0|}{1+0+0}\right) + \left(\frac{|909-410,994|}{1+909+410,994}\right) \cdots + \left(\frac{|0-0|}{1+0+0}\right) = 35.9342 \quad (3.4)$$

According to Equation 3.1 and Equation 3.2, the calculations of the relative distance value show that $IM1_{db}$ and IM_q are more similar than $IM2_{db}$ and IM_q because the relative distance value of the $IM1_{db}$ and the IM_q is less than the relative distance value of the $IM2_{db}$ and IM_q . If the relative distance value is 0, the two images are the same image.

Therefore, if the relative distance value is close to zero, two images are similar. In this research, the relative distance value is 25.6 to define the similarity image; two images are similar in 60% of entire image. If the relative distance value is less than or equal to 25.6, the two images are similar images; otherwise, they are not similar.

3.2 Proposed Routing Algorithm

In order to find the sending path of the ISP without rerouting back to its source, a small packet named a Patrol Packet (PTP) Algorithm is proposed in this research. The PTP supports the routing algorithm for transferring the query message. The PTP format and algorithm are described in the following subtopics.

3.2.1 Patrol Packet (PTP) format

Since the PTP is implemented to assist the routing path of the query message, it contains fields to indicate the sending and receiving status of the query message over the sending path. The format of the PTP is displayed in Figure 3.3 and the meaning of each attribute in Figure 3.3 is described in Table 3.2.

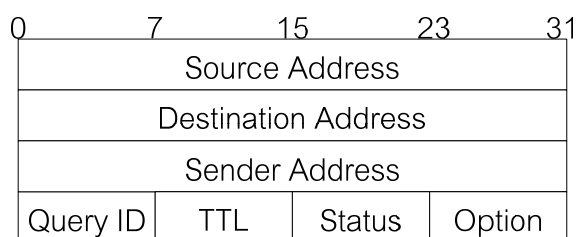


Figure 3.3 The format of the PTP

Table 3.2 Syntactic and semantic of the PTP format

Attribute name	Length (byte)	Remark
Source Address	4	IP address of the client.
Destination Address	4	IP address of the destination GSES.
Sender Address	4	IP address of the source GSES.
Query ID	1	identification number of the search
Status	1	Status of Sending or Receiving; defined as 0: PTP is sent 1: PTP is received with permission to send ISP
Option	1	Null bits

According to Table 3.2, the length of the PTP is 128 bits or 16 bytes, and can be considered as a small packet over the Internet. Thus, the PTP hardly causes any congestion in the communication channel. Moreover, the sending paths identified by the PTP will reduce the number of sending query messages through the network. Therefore, the congestion problem from the ISP and the rebound problem are eliminated.

3.2.2 Patrol Packet (PTP) Algorithm

Patrol Packet (PTP) algorithm is the algorithm that supports the routing algorithm for transferring the query message. Patrol Packet (PTP) algorithm uses the flooding mechanism. Patrol Packet (PTP) algorithm consists of two functions: a sending PTP function and a receiving PTP function. In the sending PTP function, the current node sends the PTP to its neighbors, except the neighbor that sent the query message. The PTP is sent to check the status of the database of the neighbor nodes before sending the query message. In the receiving PTP function, the current node receives the replying

PTP that requests the query message from the neighbor. Thus, this function will send the query message to the neighbor. The PTP algorithm is presented in Figure 3.4.

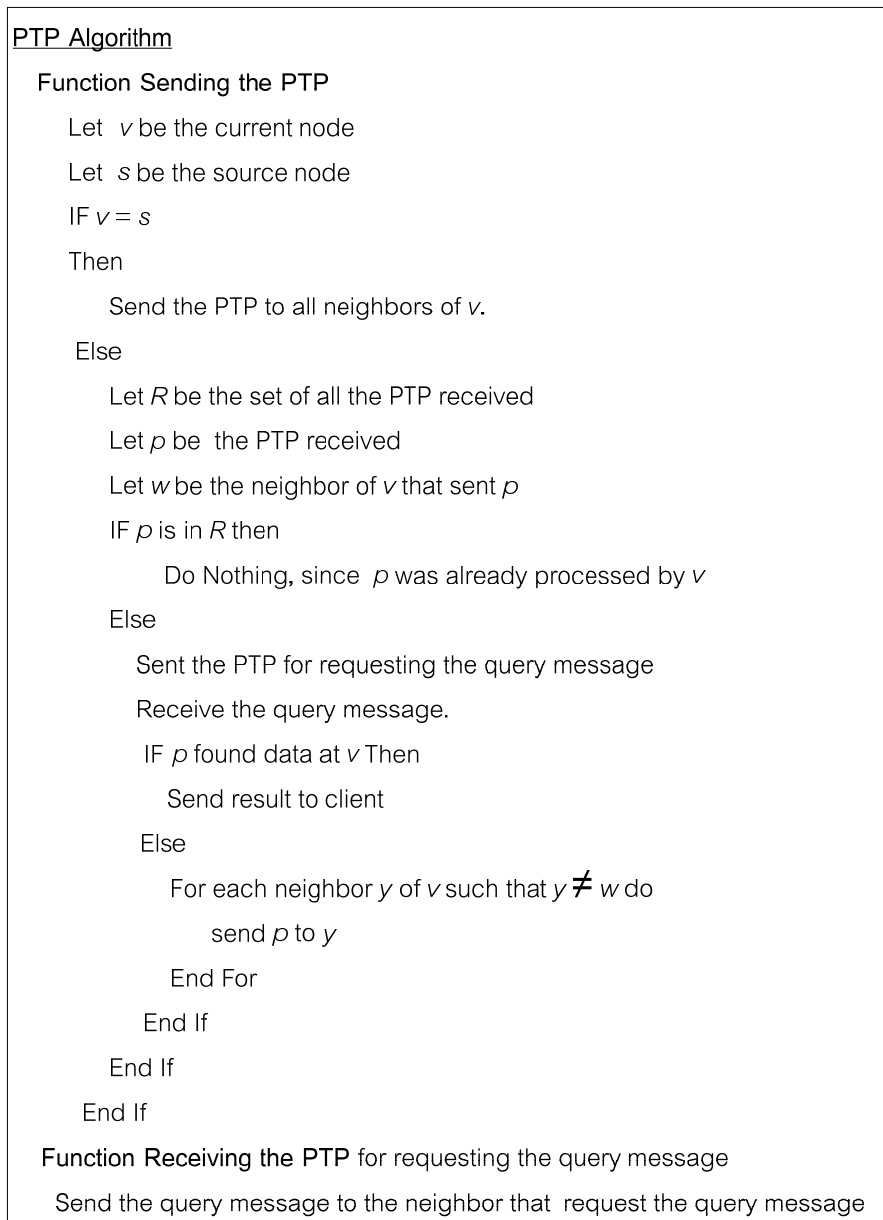


Figure 3.4 The Patrol Packet (PTP) algorithm

Referring to the PTP algorithm in Figure 3.4, it is shown the sending paths are identified by the PTP before sending the query message. So, the PTP algorithm can reduce the number of sending query messages through the network. The congestion problem from the query messages and the rebound problem are eliminated.

3.3 Modified ISP

The proposed routing algorithm and the proposed image retrieval are applied to increase the performance of the ISP. Therefore, the ISP format, the Embedded Agent for Information Search Protocol (EAISP), and the Global Search Engine System (GSES) are modified to work on the proposed routing algorithm and the proposed image retrieval. Furthermore, the Global database (GDB) is also enhanced.

Therefore, the new ISP is called Specific Information Search Protocol (SISP). The modified Embedded Agent for Information Search Protocol (EAISP) is called the Embedded Agent for Specific Information Search Protocol (EASISP). Additionally, the Global Search Engine System (GSES) of the new ISP is called the Modified Global Search Engine System (MGSES). The Global Database has been changed to Modified Global Database (MGDB).

The architecture of SISP is shown in Figure 3.5. The detail of the format and the function of each module are elaborated as follows:

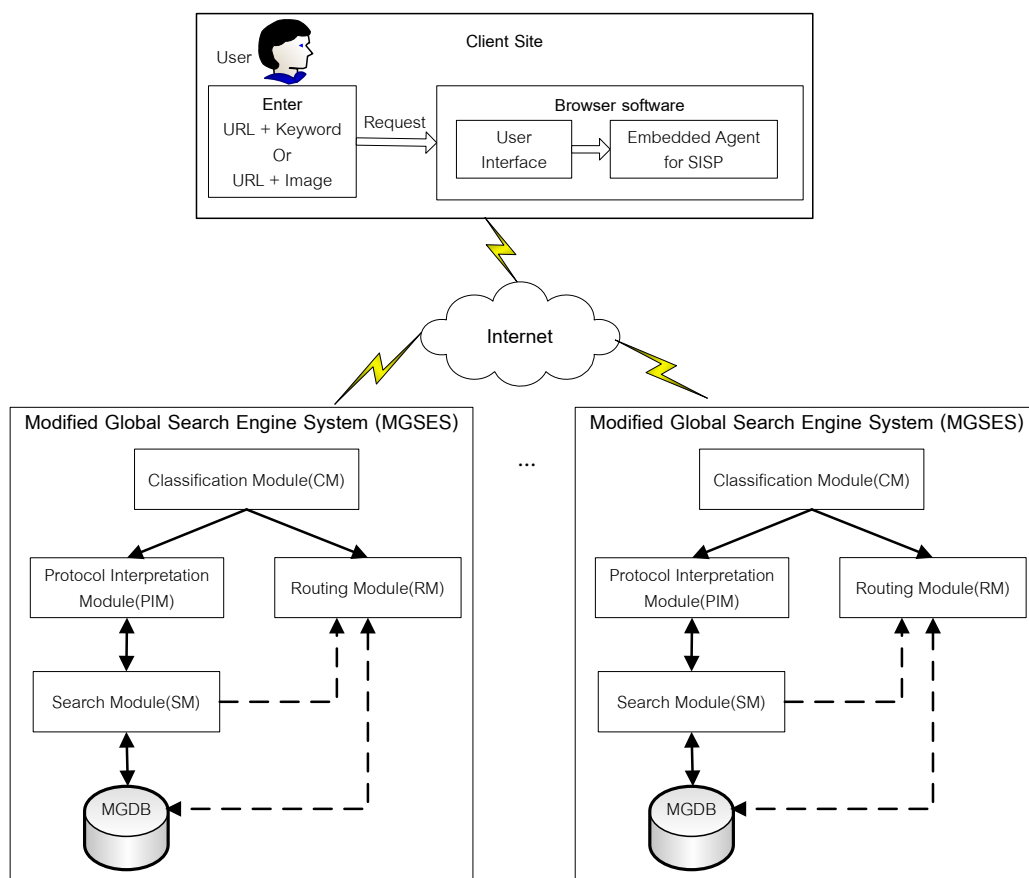


Figure 3.5 The architecture of SISP

3.3.1 Format of the Specific Information Search Protocol (SISP)

Since the SISP was implemented to enhance the search ability of an image search engine using a partial name or full name of a URL and a partial of an image, the format of the protocol must contain the search content that is a combination of a partial name or full name of a URL and the query image. The result of the search process is a list of full pathnames of URLs. Moreover, the content of the URLs in the list will also contain the required image(s) or the similar image(s) for the image search. The general format of SISP is shown in Figure 3.6. Furthermore, the SISP for an information search and the SISP for an images search are presented in Figure 3.7 and Figure 3.8 respectively. The meaning of each attribute is elaborated in Table 3.3.

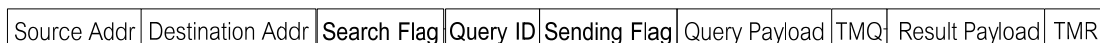


Figure 3.6 The general format of SISP

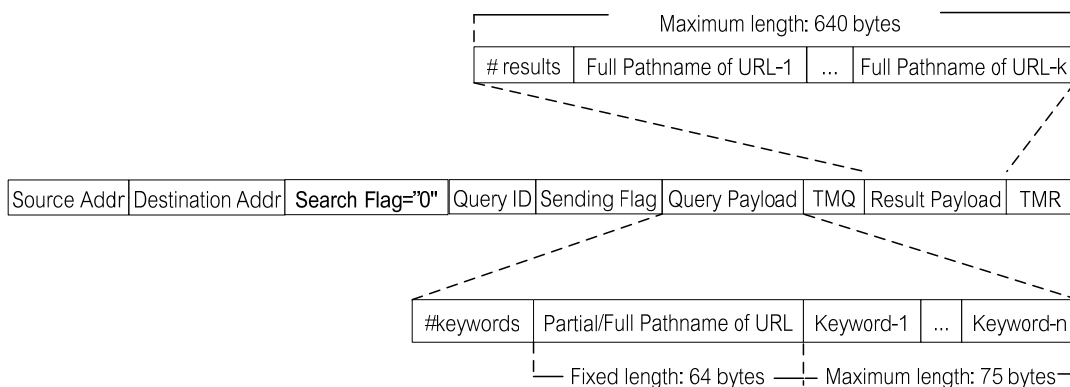


Figure 3.7 The SISP format for an information search

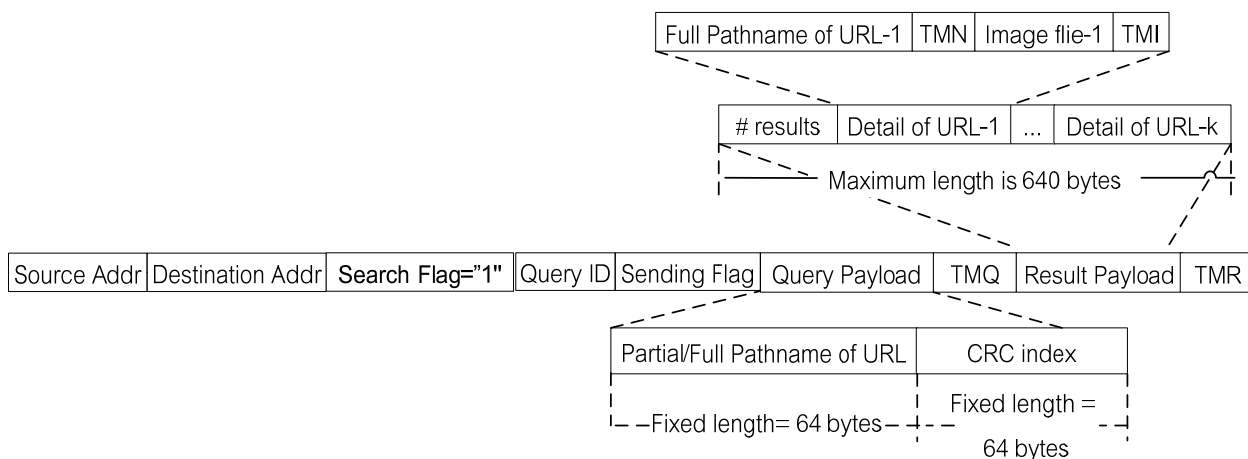


Figure 3.8 The SISP format for an image search

According to the format in Figure 3.6, the Search Flag, Query ID and the Sending Flag fields will be added into the SISP. The Search flag is the value for selecting the information search or the image search. The Search flag of the information search and the image search are "0" and "1" respectively. The Query ID is the identification number of the search and it is used with the Source Address to identify searches in the MGDB. In addition, the Sending Flag is the transfer status of the SISP.

Furthermore, the length of the SISP is a dynamic value depending on the keyword(s) and the number of the results lists. However, there are two terminate values to indicate the end of the query payload or the end of the result payload; and there are two terminate values to indicate the end of the partial/full pathname of URL and the image file in the SISP format for image searching.

Table 3.3 Syntactic and semantic of the SISP format

Attribute name	Length (byte)	Remark
Source Address	4	The IP address of the user system
Destination Address	4	The IP address of the database search engine
Search Flag	1	Flag of search; defined as "0": Information Search "1": Image Search
Query ID	1	Identification number of the search
Sending Flag	1	Status of SISP transfer; defined as 1: the sender is client, thus MGSES does not send PTP and can send the SISP. 0: the sender is MGSES, thus MGSES must send PTP to check status of SISP.
Partial/Full Pathname of URL	64	The partial or full pathname of the required URL. Users can enter only one value.
Keyword-i	15	Search keywords.
CRC index	64	Color region correlation values
TMQ	1	Terminate key; indicates the end of query text, fixed value in hexadecimal is 'FF'.
#result	1	The total number of URLs listed from the search mechanism. The value of this attribute is <i>k</i> .
Full Pathname of each URL	64	The pathname of the web, may have more than one result, total number of the URLs is not exceed 10 URLs
TMN	1	Terminate key; indicates the end of Full Pathname of URL, fixed value in hexadecimal is 'FF'.
TMI	1	Terminate key; indicates the end of image file, fixed value in hexadecimal is 'FF'.
Image file	100	Sample of a result image
TMR	1	Terminate key; indicates the end of result text, fixed value in hexadecimal is 'FF'.

3.3.2 Embedded Agent for the Specific Information Search Protocol (EASISP)

The function of the Embedded Agent for the Specific Information Search Protocol (EASISP) is to create the packet in the new version of the ISP format. Therefore, the Search flag, Query ID, and the Sending Flag fields are added to the SISP. In this module before sending the SISP to the MGSES, the SISP indicates the search types: information and image search. The Query ID is the number of search identification; it is applied with Source Address to identify searching in the MGDB. The Sending Flag is the transfer status of an SISP.

After some part of a required URL and the keyword(s) or partial image are entered through the browser interface, these data will be sent to the EASISP. The EASISP will encapsulate the received URL and keyword(s) or indexes of the query image into the SISP format. Additionally, the Search flag field is set to "0" or "1" for information or image search respectively. Then, this SISP will be transferred to the nearest GSES. Figure 3.9 is the algorithm for encapsulating data into the SISP.

For the image search, there is a function of image extraction that is called CRC function. The function is added in EASISP to extract a feature of the query image and return the CRC index of the query image. There are 64 values of CRC index that may be inserted in the CRC index field of SISP. The CRC index values are used to compare the similarity between the query image and the images that are stored in the MGDB.

```

Algorithm of Encapsulating
Receiving a query from the browser user interface
Put the host name into the Source Address field of SISP
Put the MGSES name into the Destination Address field of SISP
If Keyword field is not NULL Then
    Concatenate the "0" to the SISP // Set the Search flag field
    Concatenate the Query to the SISP
Else
    Call CRC(Query image)
    Concatenate the "1" to the SISP // Set the Search flag field
    Concatenate the URL to the SISP
    Concatenate 'FF' to SISP
    Concatenate CRC index of image
End If
Concatenate 'FF' to SISP
Concatenate NULL to SISP
Concatenate 'FF' to SISP
Send the SISP to the Internet system
End Encapsulating

```

Figure 3.9 The algorithm of SISP encapsulation.

Another algorithm of EASISP is an extraction algorithm. When the result is sent back to the client, EASISP performs the data extraction and sends the full pathname of the URL(s) for information search. Additionally, it sends the image(s) for image search from the result payload to the client's user interface of the browser. The algorithm for extracting information from the received SISP is showed in Figure 3.10.

Algorithm for Extraction

```

Start reading 9th byte at Search flag field,
If Search flag field is "0" Then // Result of Information search
    Do until found 'FF'
        Read next character from SISP
    End Do
    Initial URL_NM to be NULL
    Read a character X
    Do until found 'FF'
        Concatenate X into URL_NM
        Read a character X
    End Do
    Send URL_NM to the browser user interface
Else If Search flag field is "1" Then // Result of Image search
    Do until found 'FF'
        Read next character from IISP
    End Do
    Initial URL_NM to be NULL
    Initial URL_IMG to be NULL
    Read a character X
    Do until found 'FF'
        Concatenate X into URL_NM
        Read a character X
    End Do
    Do until found 'FF'
        Concatenate X into URL_IMG
        Read a character X
    End Do
    Execute image(s) from URL_IMG
    Send URL_NM and image to the browser user interface
Else
    Drop packet
End IF
End Extraction

```

Figure 3.10 The algorithm of SISP extraction.

3.3.3 Modified Global Search Engine System (MGSES)

The GSES is adjusted to work on the search algorithm and the image search retrieval. The new GSES is added to the Classified Module (CM) and modified all modules that are Protocol Interpretation Module (PIM), Search Module (SM), Routing Module (RM), and Global Database (GDB). The Modified GSES is displayed in Figure 3.11 Each module is described as follows.

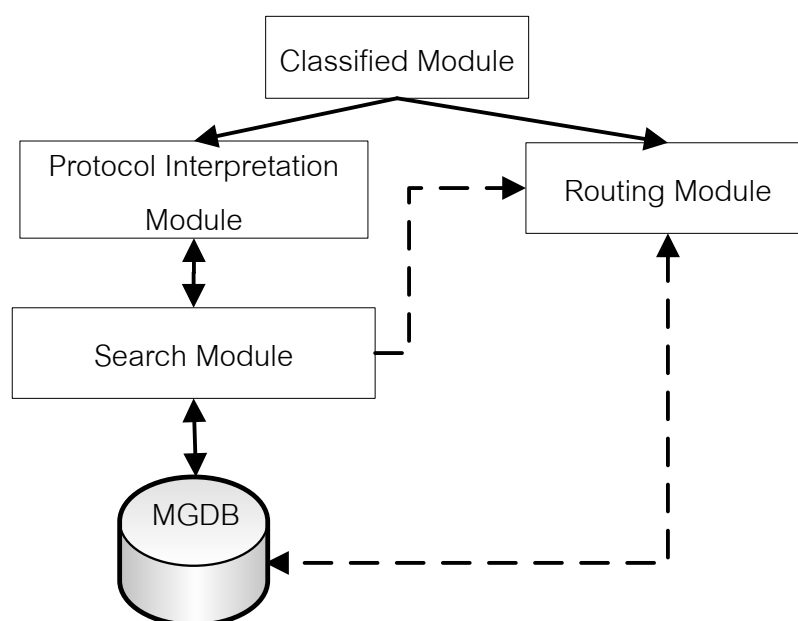


Figure 3.11 The modification of GSES

3.3.3.1 Classified Module (CM)

The CM is responsible for classifying the PTP and the SISP. When the PTP or the SISP is sent to a MGSES, it will be classified by the length of the packet. If the length of the packet is 16 bytes (128 bits), then the packet will be sent to the RM. If the length of the packet is greater than 16 bytes, then the packet will be sent to the SM. The algorithm of CM is described in Figure 3.12.

Algorithm of Classified Module

```
IF length of a packet == 16 bytes Then
    Send the packet to the Routing Module
Else IF length of a packet > 16 bytes Then
    Send the packet to the Search Module
Else
    Drop the packet.
End IF
```

Figure 3.12 The algorithm of Classified Module

3.3.3.2 Protocol Interpretation Module (PIM)

The function of the PIM is to extract the SISP in order to obtain the partial pathname of the URL, and the keywords or the index of image of the required query message. After acquiring the result of the extraction, the result is used to generate a search command in an SQL statement, SQL_P.

For example, in an information search, an entered query is `www.chula.ac.th + "computer science" + "mathematics"`. The PIM will generate an SQL condition that is `URL_NAME = "www.chula.ac.th" AND (KEYWORD = "computer science", OR KEYWORD = "mathematics")`.

For an example of an image search, the PIM uses the URL to generate an SQL condition that is `URL_NAME = www.chula.ac.th` and sends the index of image to the Search Module(SM) for matching the similar image in the MGSES.

The interpretation process is described as an algorithm in Figure 3.13.

```

Algorithm of Interpretation Module
Initial X , URL_PART, URL_PARTL and KEYWORD_PART to be NULL
Read the 9th byte of the Query Payload of SISP
If Search flag field is "0" Then // Query Payload of Information search
    Do until found '+'
        Concatenate X to URL_PARTL
        Read next character into X
    End Do
    Put 'URL_NAME LIKE %' URL_PARTL ' %' into URL_PART
    Read next character into X
    If X != 'FF' Then Put 'KEYWORD LIKE %' into KEYWORD_PART
        Do until found 'FF'
            Concatenate X to KEYWORD_PART
            Read next character into X
            If X = '+' Then
                { Concatenate ' % OR KEYWORD LIKE %' to KEYWORD_PART
                  Read next character into X}
            End Do
            Concatenate ' %' to KEYWORD_PART
            SQL_P = Concatenate ('Select URL_FULLLN from MGDB table' where ' URL_PART 'and' URL_KEYWD)
        Else If Search flag field is "0" Then // Query Payload of Image search
            Initial IND_CRC to be NULL
            Read the 10th byte of the Query Payload of SISP into X
            Do until found 'FF'
                Concatenate X to URL_PARTL
                Read next character into X
            End Do
            Put 'URL_NAME LIKE %' URL_PARTL ' %' into URL_PART
            Read next character into X
            Do until found 'FF'
                Concatenate X to IND_CRC
                Read next character into X
            End Do
            SQL_P = Concatenate ('Select URL_FULLLN from MGDB table' where ', URL_PARTL )
            Send IND_CRC to Search Module.
        Else
            Do nothing
        End If
    End Interpretation

```

Figure 3.13 The interpretation algorithm

Another function of the PIM is to receive the results from the SM. The PIM will encapsulate the search results into the SISP format and send it back to the client. The result of an information search is the full path name of the URL; conversely, the result of an image search is the full path name of the URL and the similar image(s). For an image search, it will send the similar image(s) to the client. Another function of the PIM is described as an algorithm in Figure 3.14.

```

Algorithm for Returning Results
If Search Flag == "0" Then  /// Information Search
  Do until Result(i) = NULL
    Concatenate Result(i) to Result field of SISP
    Concatenate ',' to Result field of SISP
  End Do
  Replace last character of Result field of SISP with blank
  Put 'FF' to TMR field of SISP
Else If Search Flag == "1" Then  /// Image Search
  Initial IMG to be NULL
  Do until Result(i) = NULL
    Concatenate Result(i) to Result field of SISP
    Concatenate ',' to Result field of SISP
    Concatenate Image(i) to IMG
    Concatenate ',' to IMG
  End Do
  Concatenate IMG to Result field of SISP
  Replace last character of Result field of SISP with blank
  Put 'FF' to TMR field of SISP
End If

```

Figure 3.14 The return result algorithm of PIM

3.3.3.3 Search Module (SM)

When the PIM creates the SQL statement, the PIM will send this statement to the SM to perform the retrieving process. For an information search, the results of the retrieving process can be a full pathname of a URL, or a set of full pathnames of URLs.

For an image search, the results of the retrieving process can be a full pathname of a URL, or a set of full pathnames of URLs and an image or a set of images. The search result is dependent on the ability of the query input by users. The algorithm of the SM is shown in Figure 3.15.

Unfortunately, there is an opportunity that the required URL does not exist in MGDB. In that case, the SM will transfer to the Routing Module (RM); otherwise, it will return the results back to the PIM to transfer the search result to the client.

For an image search, there is a function that calculates and compares the similarity image of the query image and the image at the MGDB that is under the return Record of SQL_P. This function is called the Index Matching function and it uses the Euclidean distance equation to calculate the similarity value of CRC index. The details of the Relative distance calculation are presented in Section 3.1.

```
Algorithm of Search Module
If Search Flag == "0" Then /// Information Search
  Run SQL statement
  Let i=0
  If Records is not NULL Then
    Do until end of Records
      Read URL_NM into Result(i)
      Increase i by one
    End Do
    Sent all Result(i) to PIM
  Else
    Transfer to RM
  End IF
Else If Search Flag == "1" Then /// Image Search
  Run SQL_P
  Index_Matching(IND_CRC)
  Let i=0
  If Records is not NULL Then
    Do until end of Records
      Read URL_NM into Result(i)
      Concatenate 'FF' to Result(i)
      Read URL_IMG into Result(i)
      Increase i by one
    End Do
    Sent all Result(i) to PIM

  Else
    Transfer to RM
  End IF
Else
  Do Nothing
End If
End Search
```

Figure 3.15 The algorithm of Search Module

3.3.3.4 Routing Module (RM)

As mentioned in Chapter 2, the routing module of the ISP has a defect. Therefore, the RM is modified using the new routing algorithm that uses a small packet, PTP, to check the status of the SISP on the MGSES.

The RM receives the PTP from the CM and the SISP from the SM when the required URL cannot be found in the MGDB. Thus, the process of the RM is divided into 2 situations, as follows.

Situation 1: The required URL does not exist in the MGDB.

When this situation occurs, the SISP must be sent out to other MGSESs. If the SISP was sent directly from the client, it will automatically be sent to all neighbors of the current MGSES. Otherwise, before sending the SISP to all neighbors, PTP packets must be created and sent out to find the travelling path.

Situation 2: A MGSES receives a PTP packet.

Since a PTP is used to find the sending path, there are two cases to be considered. The first case is the situation that the received PTP is used for a path lookup; status field is '0'. The second case is the situation that the received PTP is the response PTP from other MGSESs; status field is '1'. In the first case, the MGSES receives the PTP for a path lookup. If the MGSES never receives the same content of this PTP; the status field of this PTP is changed to '1'. The MGSES will then reply to this PTP; otherwise, the MGSES does nothing. In the second case, the MGSES receives the PTP that is the replied PTP from other MGSESs. The status field of this PTP is '1', and the MGSES will reply the ISP to the MGSES that sent the PTP.

3.3.3.5 Modified Global Database (MGDB)

Each MGSES has an installed database named the Modified Global Database (MGDB). The MGDB consists of two parts. The first part is the list of the searched URLs,

keywords, and images. The second part is the routing path of the sending SISP as follows.

- **Information and image database**

As mentioned above, the first part is the list of the searched URLs, keywords, and images. The data in this part of the MGDB consists of three tables: URL table (URL_T), the keyword table (KEYWORD_T), and the image table (IMAGE_T). URL_T indicates full pathname of the URL names (URL_NAME) and the identification number of each URL (URL_ID). KEYWORD_T defines URL_ID, keywords under each URL (KEYWORD) and the full pathname of keyword URLs (URL_FULLN). Finally, the data in IMAGE_T consists of fields that are the URL_ID, the full pathname of image URLs (URL_FULLN), the index of CRC (IND_CRC), and image file under each URL. Figure 3.16 shows the database table of information and image.

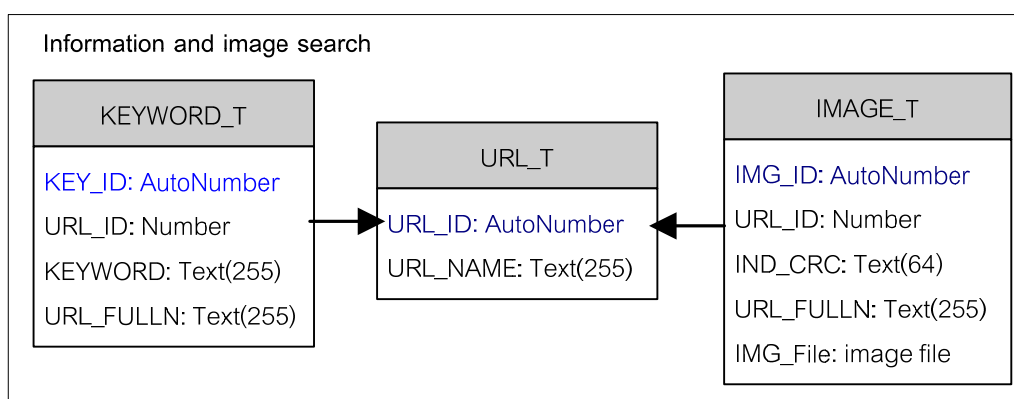


Figure 3.16 The database table of information and image

- **Routing database**

The second part is the routing path of the sending SISP. The data in the second part of the MGDB consists of three tables which are the neighbors table (NEIGHBOR_T), the routing table (ROUTING_T) and the SISP table (SISP_T). NEIGHBOR_T indicates the IP address of the neighbors (NEIGHTBOR_ADDR). ROUTING_T indicates SISP identification (SISP_ID), the neighbor identification (NEIGHBOR_ID), the status of the

PTP (STATUS) and time stamp of arrived small packet. SISP_T indicates the query's identification of the SISP (QUERY_ID), the client address (SOURCE_ADDR), the query of a URL and keyword (QUERY), and the time stamp of the SISP (TIME_STAMP). Figure 3.17 shows the routing database.

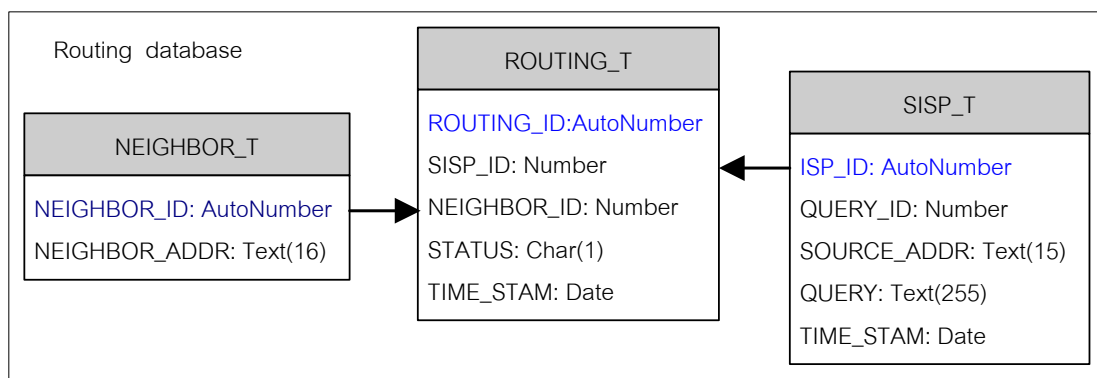


Figure 3.17 The database table of routing

3.3.3.6 Example of the PTP Algorithm

An example of transferring an SISP using the routing of the PTP is shown in Figure 3.18. The details of process are explained as follow.

Firstly, the SISP from client is sent to the nearest MGSES, the MGSES A. The SM searches for the required URL and updates the SISP_T in the MGDB. If the MGDB has the required URL, the GESE A sends the result query back to the client. If the required URL is not found, the detail of the SISP will be changed from Flag '1' (for the first time of sending the SISP) to '0'. After that, the MGSES A sends the SISP to the MGSES A's neighbours, which are the MGSES B and the MGSES E. In this case the MGSES A does not send the PTP to check MGSES A's neighbours because the MGSES A is the first MGSES which receives the SISP, thus the MGSES A can distribute the SISP to its neighbours.

Afterwards the MGSES B and the MGSES E receive the SISP and search for the required URL and the SISP_T in their MGDB will be updated. In the example, the MGSES B will not be able to obtain the required URL in its MGDB and the RM of MGSES

B sends the PTP to the MGSES neighbours, MGSES C, to check the status. The MGSES C does not have the SISP of client and Query_ID in its SISP_T; thus, the MGSES C allows the MGSES B to send the SISP to the MGSES C by replying the PTP that changes the status to "1". When the MGSES C receives the SISP from MGSES B, the MGSES C will update the SISP_T and start searching for the required URL in its MGDB, which is not locally available. So, the RM creates the PTP and sends to the MGSES C's neighbour, MGSES D.

MGSES E, when it receives the SISP from the MGSES A, will update the SISP_T and start searching for the required URL in its MGDB, which does not contain the required URL. Thus, the RM creates the PTP and sends it to the MGSES E's neighbour, MGSES C and MGSES D. The MGSES C and the MGSES D receive the PTP and search for the SISP of the client and the Query_ID and the Routing_T in their MGDBs will be updated. In example, the MGSES C has already received the SISP, so it does nothing. The MGSES D does not have the ISP of the client and Query_ID in its SISP_T; thus, the MGSES D allows the MGSES E to send the SISP to the MGSES D by replying PTP that changes the status to "1". The MGSES D will update the SISP_T and search the required URL in its MGDB.

When the MGSES C sends the PTP to the MGSES C's neighbour that is the MGSES D. Then, the SISP_T of the MGSES D will be checked with the PTP while the ROUTING_T will be updated. Since, the MGSES D has received the SISP, it will not allow the MGSES C to send the ISP again so the MGSES D does nothing.

In the example, the MGSES D has the required URL. The result of the searching will be inserted into the SISP sent from the MGSES D and is sent back to the client.

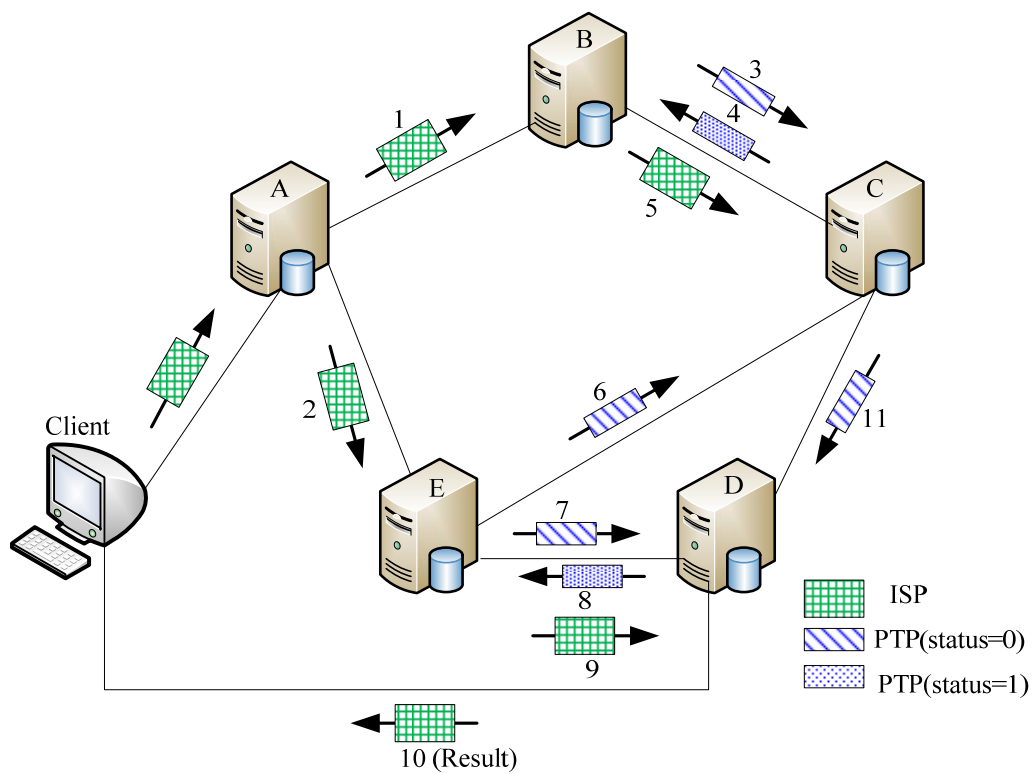


Figure 3.18 The example the SISPs' flow using the PTP algorithm

CHAPTER IV

IMPLEMENTATION AND EXPERIMENTAL RESULTS

In order to prove the efficiency of the SISP, a simulation was implemented to validate and evaluate the performance of the SISP. The performance indicators are the response time, and the number of sent bytes in the communication channel. Moreover, the performance of the SISP is compared with the performance of the ISP. Therefore, the implementation, the simulation, simulation result, and the performance analysis to demonstrate the performance of the proposed scheme are elaborated as follow.

4.1 Implementation

The simulation was implemented on an HP 2 Quad Cores with XEON Processors and 16 GB main memory running Ubuntu 10.04 Desktop. The SISP is simulated on a virtual machine environment that creates the client and the MGSEs. Moreover, there are several software packages installed on the MGSEs. The database management system employs MySQL Server version 5.1, and the phpMyAdmin runs on an Apache2 Web server is used to deal with MySQL Server. The system of the SISP was implemented as a java program. The PTP algorithm transfers using UDP and the SISP uses TCP for the transmission. Moreover, the browser of the SISP is also implemented using the java program; it prepares three fields for entering the partial URL, keywords and image part directory. Furthermore, there are three buttons; two buttons for selecting method for the search (information search or the image search) and another one for browsing the image. The SISP browser is shown in Figure 4.1.

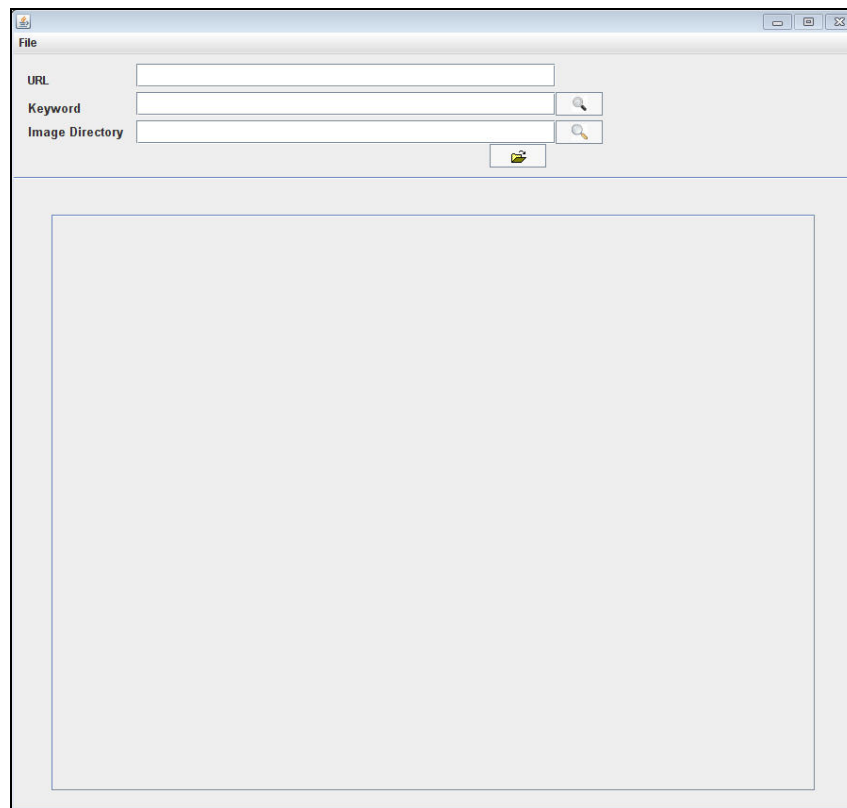


Figure 4.1 The SISP browser

Moreover, each MGSES stores the URL, keywords and images. There are 1,435 records for the information search and 400 records for the image search. All records are distributed on all of the MGSES. The URL, keywords, and images of MGSES were collected from the Google search engine by typing the complete required URL and keywords on the search field. For example; `suksitnives site:chula.ac.th`, or entering in an advanced search menu by typing the required keywords in the “all these word” field and the required URL “site or domain” field on the advanced search page.

4.2 Simulation

The simulation was divided in two parts: the simulation of the proposed routing algorithm and the simulation of the performance of the SISP. The details of two simulations are presented in the following subtopics.

4.2.1 The simulation of the proposed routing algorithm

In order to prove the efficiency of the proposed routing algorithm, the number of sent bytes and the response time are considered. The simulation was performed by implementing three routing algorithms; the flooding algorithm, the hierarchical algorithm, and the proposed algorithm. The flooding algorithm and the hierarchical algorithm are two standard methods to reach the resource location. Moreover, all of the MGSES nodes were connected with a mesh topology for simulating the flooding algorithm and the proposed algorithm. The hierarchical structure was implemented for the hierarchical algorithm. The required URL is stored in one of the MGSES nodes that will send the result message to the client. Three of the routing algorithms were tested using ten sets of required URLs and keywords that are encapsulated in the SISP format at the client site and sent out to the MGSES. The MGSES creates n nodes for $n = 3, 5, 7, 9, 11, 13$ and 15 nodes on each simulation.

4.2.2 The simulation of the SISP performance

In order to prove the performance of the SISP mechanism, the search result of the information and image search of the SISP is compared with the ISP. The MGSES creates 15 nodes for the simulation. For information search, two of the search engines are tested using 10 sets of the partial name of required URL and keywords. For image search, the ISP cannot search an image. Therefore, the SISP is tested using 10 sets of the partial name of the required URL and image.

4.3 Simulation Results

4.3.1 Number of bytes

4.3.1.1 Simulation result of the number of bytes

The number of bytes is a parameter to measure the performance of the search method. The simulation used ten sets of required URL and keywords that are

encapsulated in the SISP format at the client site and sent out to the MGSES node and the number of receiving byte is recorded at every node.

Figure 4.2 shows the comparison of the average number of sent bytes among three algorithms at n MGSES nodes. The result shows that the number of sent bytes of the flooding algorithm is higher than the number of sent bytes of the proposed routing algorithm and the number of sent bytes of the hierarchical algorithm is the least. When the number of MGSES nodes is increased, the number of bytes of the flooding algorithm is much higher than the number of bytes of the PTP algorithm.

According to the experimental result, the proposed routing algorithm can reduce the number of bytes of the flooding algorithm by more than 50%. The number of sent bytes of the hierarchical algorithm is slightly increased when the number of MGSES nodes is raised.

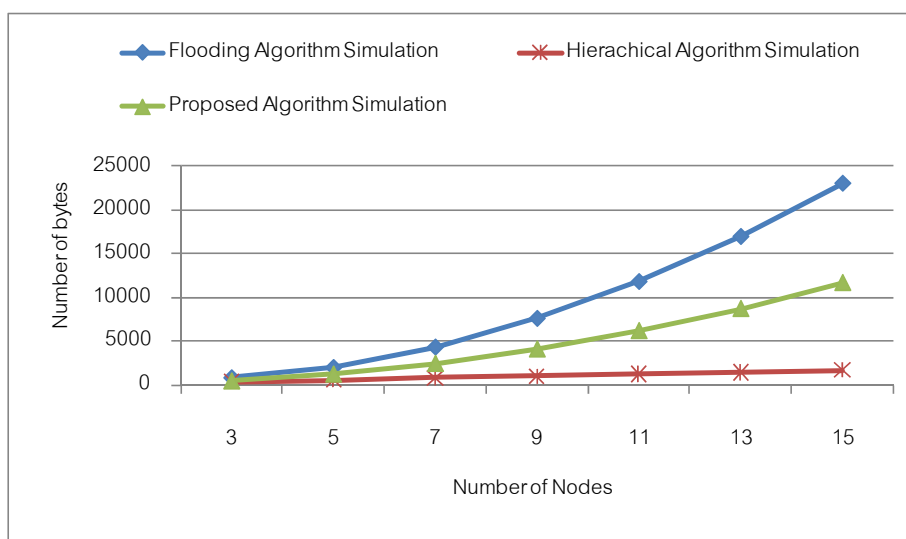


Figure 4.2 Comparison of the number of the sent bytes at n MGSES nodes

4.3.1.2 Theoretical analysis of the number of bytes

Since the PTP algorithm is used to search the routing path of the SISP, the size of the PTP must be small to avoid congestion problems. The size of the PTP is 16 bytes and, the size of the SISP is a dynamic value depending on the length of

keyword. From the SISP format, the length of the ISP is more than 89 bytes and the length of the SISP is larger than 91 bytes because two bytes of the Query ID and the Flag field are added. The number of bytes of the PTP is much smaller than the number of bytes of the SISP. Therefore, the PTP can easily and quickly check the SISP sending and receiving status in other MGSEs and also to report to the current MGSE. The current MGSE will consider other MGSEs to send the SISP.

The number of the sending SISP is the number of times that the SISP is sent to search the required URL in the network from the first MGSE to other MGSEs until the required URL is found. The description of variables is shown in Table 4.1.

Table 4.1 The description of variables

Variable	Description
$NumberISP$	The total number of sending ISP
$NumberSISP$	The total number of sending SISP
$NumberPTP$	The total number of sending PTP
$NumberByte$	The total number of bytes of all sending
$node_i$	The MGSEs node number i
N	The number of MGSEs
$link(node_i)$	The number of links to neighbors of $node_i$

The calculation of the number of sending ISPs is Equation 4.1 and the total number of bytes of all sending ISPs is shown in Equation 4.2:

$$NumberISP = \sum_{i=1}^n link(node_i) - n \quad (4.1)$$

$$NumberByte \geq \left(\sum_{i=1}^n link(node_i) - n \right) \times 89 \quad (4.2)$$

The PTP reduces the number of sending of modified ISPs, which is shown in Equation 4.3.

$$NumberSISP = n-1 \quad (4.3)$$

The number of sending PTPs is the number of times that the PTP is sent to search for the routing path in the network from the first MGSES to other MGSESs until the required URL is found.

The number of sending PTPs can be calculated as in Equation 4.4:

$$NumberPTP = \sum_{i=1}^n link(node_i) - (n - 1) \quad (4.4)$$

Therefore, the total number of bytes of all sending using the PTP and the SISP is shown in Equation 4.5:

$$NumberByte \geq \left(\sum_{i=1}^n link(node_i) - (n - 1) \right) \times 16 + (n - 1) \times 91 \quad (4.5)$$

In the hierarchical algorithm, the number of sending SISP is the number of MGSES nodes, n . Thus, the number of bytes can be calculated as in Equation 4.6:

$$NumberByte \geq n \times 89 \quad (4.6)$$

The simulation uses mesh topology in the flooding algorithm and the proposed algorithm. Thus, the summation of links of each node is $n(n-1)$. Therefore, the total of number of bytes of the previous algorithm is shown in Equation 4.7:

$$\begin{aligned} NumberByte &= (n(n - 1) - (n - 1)) \times 89 \\ &= (n^2 - 2n + 1) \times 89 \\ &= 89n^2 - 178n + 89 \end{aligned} \quad (4.7)$$

The total number of bytes of the PTP algorithm is shown in Equation 4.8:

$$\begin{aligned} NumberByte &\geq \{(n(n - 1)) - (n - 1)\} \times 16 + ((n - 1) \times 91) \\ &\geq ((n - 1) \times 91) + (n^2 - 2n + 1) \times 16 \\ &\geq 16n^2 + 59n - 75 \end{aligned} \quad (4.8)$$

Equations 4.6, 4.7 and 4.8 can represent the value of n using $n = 3, 5, 7, 9, 11, 13$ and 15 . Figure 4.3 shows the calculation results and the comparison of the number of bytes of the simulation. It can be seen that these results are similar to each other.

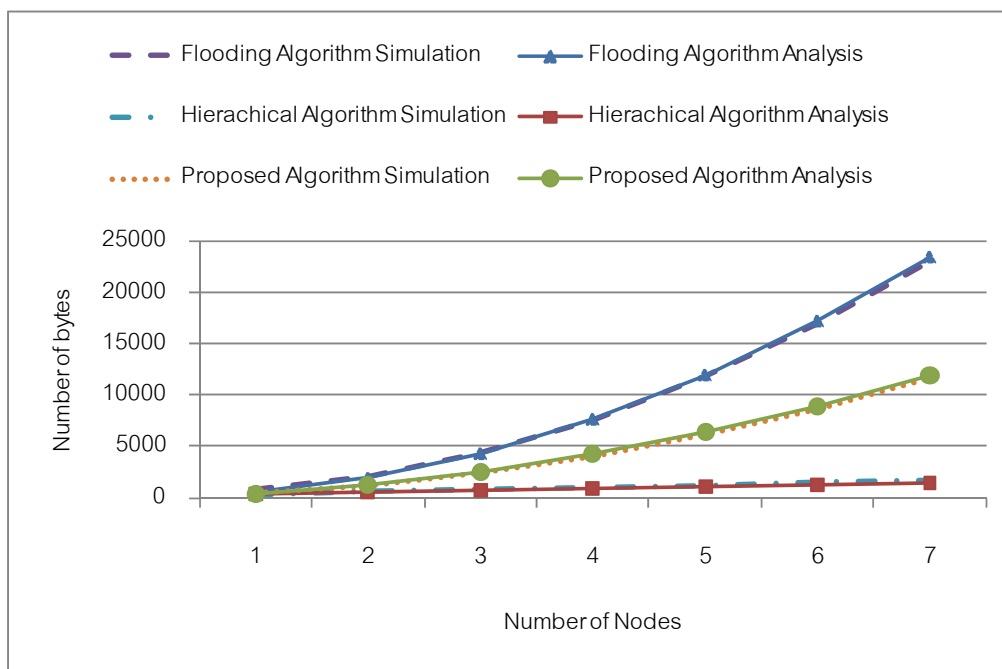


Figure. 4.3 The comparison of the number of bytes from the simulation and the number of bytes from the equations.

4.3.2 Response time

4.3.2.1 Simulation result of the response time

Another parameter of the search performance is the response time. Figure 4.4 shows the response time of three algorithms. When the number of MGSES nodes is increased, the response time of the hierarchical algorithm is higher than the response time of both the flooding algorithm and the proposed algorithm. The response time of the flooding algorithm and the proposed algorithm are slightly increased when the number of the MGSES nodes is higher. The response time of the proposed algorithm is slightly higher than response time of the flooding algorithm because the proposed algorithm checks the status of the MGSES before sending the query messages.

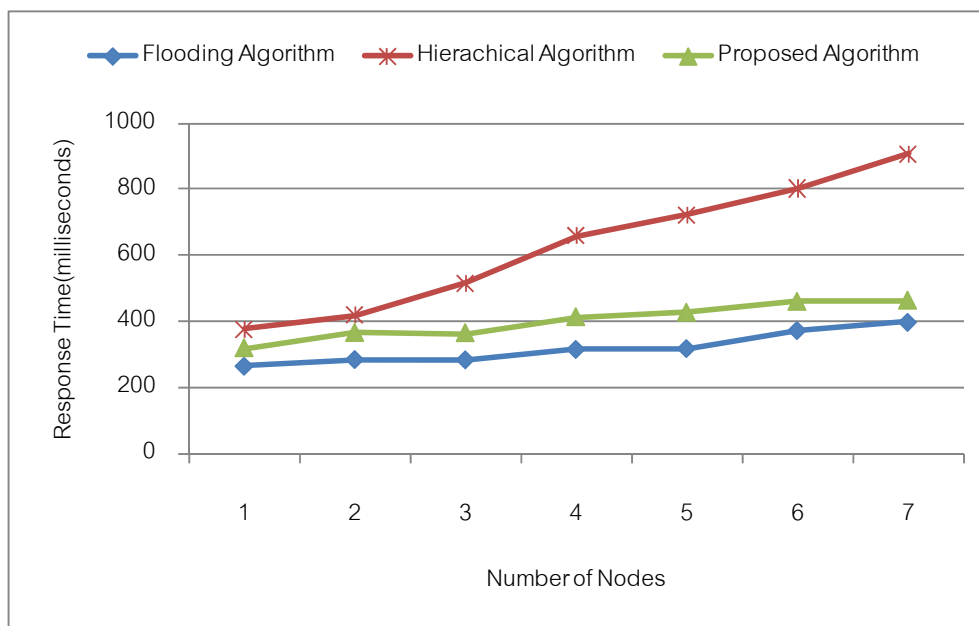


Figure 4.4 Comparison on the response time of three algorithms using n MGSES nodes

4.3.2.2 Theoretical analysis of the response time

The time complexity of each algorithm is calculated for the response time of the search algorithm. The algorithm of the flooding algorithm, the PTP algorithm, and the hierarchical algorithm are shown in Figure 4.5, Figure 4.6, and Figure 4.7, respectively.

According to the flooding algorithm in Figure 4.5, the time complexity of this algorithms is $O(n^2)$ where n is the number of nodes. For PTP algorithm in Figure 4.6, the time complexity of this algorithm is $O(n^2+n)$, because the PTP algorithm uses the flooding technique to check the status for sending query messages. The query messages will then be sent to the node. However, the current network is mesh topology. Thus, the response time of the flooding algorithm and the PTP algorithm are $O(1)$ because the client sends the query messages to the nearest neighbor and the nearest neighbor will send to all neighbors in the mesh topology. The response time is a constant value. Therefore, the response time of the flooding algorithm and the PTP algorithm are $O(1)$. According to the hierarchical algorithm in Figure 4.7, the time complexity of this algorithms is $O(\log n)$.

Thus, the response time of the flooding algorithm and the PTP algorithm is better than the response time of the hierarchical algorithm.

Flooding Algorithm

```

Let  $v$  be the current node
Let  $s$  be the source node
IF  $v = s$  Then
    Send the query message to all neighbors of  $v$ .
Else
    Let  $R$  be the set of all the query message received
    Let  $q$  be the query message received
    Let  $w$  be the neighbor of  $v$  that sent  $p$ 
    IF  $q$  found data at  $v$  Then
        Send result to client
    Else
        IF  $q$  is in  $R$  Then
            Do Nothing, since  $q$  was already processed by  $v$ 
        Else
            For each neighbor  $y$  of  $v$  such that  $y \neq w$  do
                send  $q$  to  $y$ 
            End For
        End If
    End If
End If

```

Figure 4.5 The flooding algorithm

PTP Algorithm**Function Sending the PTP**

Let v be the current node

Let s be the source node

IF $v = s$

Then

Send the PTP to all neighbors of v .

Else

Let R be the set of all the PTP received

Let p be the PTP received

Let w be the neighbor of v that sent p

IF p is in R then

Do Nothing, since p was already processed by v

Else

Send the PTP for requesting the query message

Receive the query message.

IF p found data at v Then

Send result to client

Else

For each neighbor y of v such that $y \neq w$ do

send p to y

End

End

End

End

Function Receiving the PTP for requesting the query message

Send the query message to the neighbor that request the query message

Figure 4.6 The PTP algorithm

```

Hierarchical Algorithm
Let  $v$  be the current node
Let  $q$  be the query message
Hierarchical_search( $v, q$ )
IF  $q$  found data at  $v$  Then
    Send result to client
Else
    For each children  $y$  of  $v$  do
        Send  $q$  to  $y$ 
        Hierarchical_search( $y, q$ )
    End
End

```

Figure 4.7 The hierarchical algorithm

4.3.3 The performance of the SISP

In order to evaluate the performance of the SISP, the experiment is divided into two parts: the information search, and the image search. The information search is compared with the ISP using the number of search results, the number of bytes, and time response time. The image search provides the number of search results, the number of bytes, and the response time.

4.3.3.1 The information search

The result of the information search of the SISP is the number of search results, the number of bytes, and the response time. The number of search results, the number of bytes, and the response time are compared with the number of search results, the number of bytes, and the response time of the ISP by entering the partial name of URL and keyword(s). The result of the information search of the SISP and the ISP are shown in Table 4.2. In addition, Figure 4.8 and Figure 4.9 show the comparisons of the number of bytes and response time of SISP and ISP respectively.

Table 4.2 The information search result of SISP and ISP

No	Partial of URL	Keyword	SISP			ISP		
			Number of search results	Number of bytes	Response time	Number of search results	Number of bytes	Response time
1	chula.ac	computer	7	114	195	7	112	181
2	apple	iphone	8	112	180	8	110	175
3	sfcinemacity	transformers	7	19625	605	7	39,068	542
4	scb.co	exchange rate	4	17139	582	4	23,161	530
5	tourismthailand	phuket	6	13263	573	6	23,322	541
6	4shared	Bruno Mars	8	13006	531	8	22,570	521
7	longdo.com	framework	8	17961	661	8	23,373	641
8	youtube	lucky	9	15939	591	9	23,108	541
9	amazon	books	5	13629	555	5	21,582	546
10	cnn	economy	11	11367	564	11	21,976	541
11	cnet	notebook	14	15871	546	14	22,173	501
12	microsoft	windows7	14	16081	373	14	22,173	341
13	bbc	libya	13	13397	590	13	21,582	552
14	ebay	ipad	7	13377	580	7	21,385	562
15	ebay	antique	10	14191	531	10	21,976	503
16	bizrate	canon	4	13177	254	4	21,582	210
17	wmg	linkin park	4	11759	554	4	22,764	411
18	esl	conversation	12	13575	580	12	22,961	521
19	ubuntu	tutorial	8	15817	531	8	28,445	460
20	oracle	Jdeveloper	8	14133	576	8	22,567	551

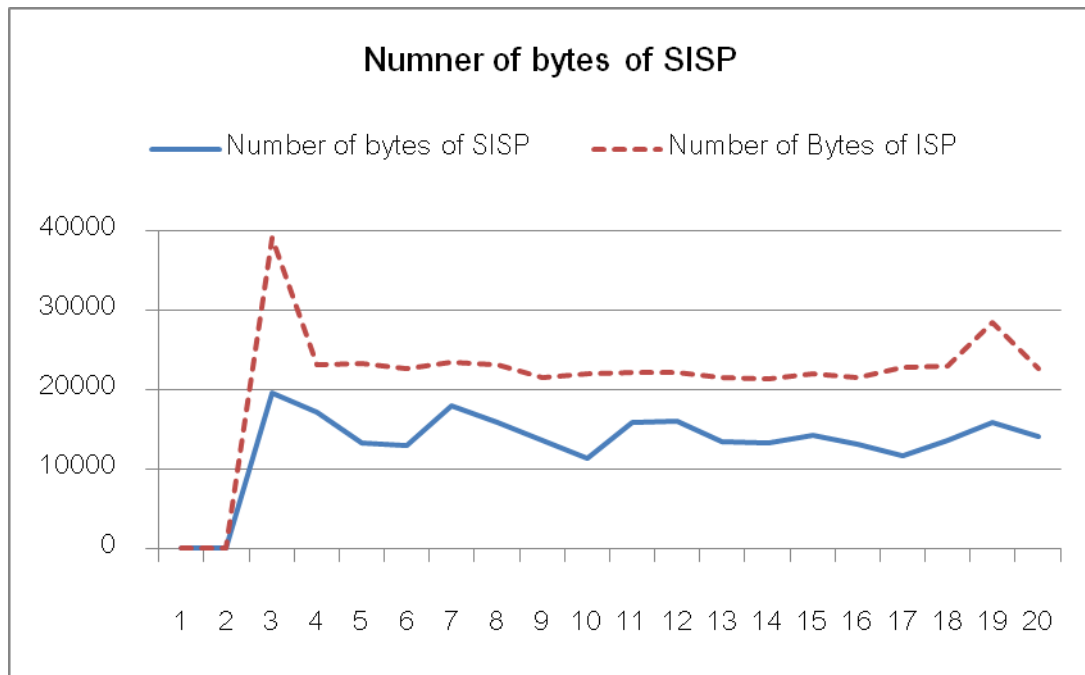


Figure 4.8 Comparison of the number of bytes of SISP and ISP at 15 MGSES nodes

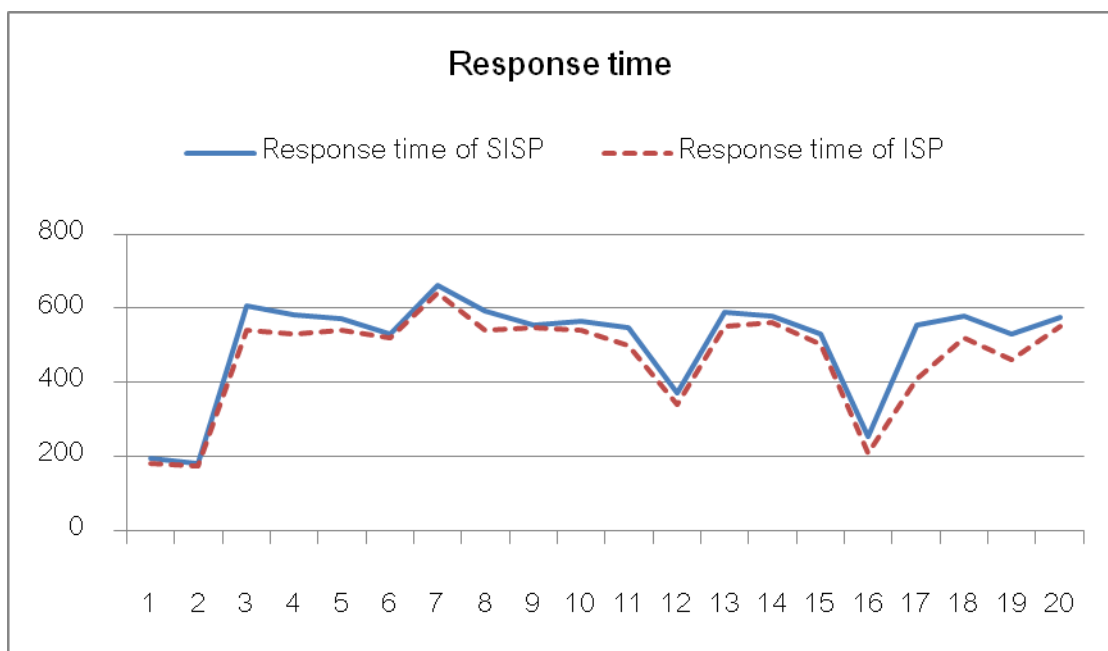


Figure 4.9 Comparison of the response time of SISP and ISP at 15 MGSES nodes

Referring to Table 4.2, all of the search result numbers of the SISP and the ISP are equal. In Figure 4.8, the number of bytes of the ISP is greater than the number of bytes of the SISP. The SISP can reduce the number of bytes to send the query messages. However, in Figure 4.9, the response time of the SISP is slightly higher than the response time of the ISP because the SISP has the process of PTP to check the status of the MGSES before sending the query messages. In Figure 4.8, the number of bytes of data set numbers 1 and 2 are very small because the required URLs are on the nearest MGSES of the client. Therefore, when the query message is sent to the nearest MGSES, the MGSES can reply with the results to the client without sending the query message to other MGSEs. Furthermore, in Figure 4.9, the response time of data set numbers 1 and 2 are very low because the required URLs are on the nearest MGSES of the client. Therefore, the nearest MGSES can reply the results to client immediately.


4.3.3.2 The image search

The search result of the image search of the SISP is the number of search results, the number of bytes, and response time. Since the ISP does not perform the image search, the ISP cannot compare with the SISP. The partial name of the URL and image are entered for testing the image search. The result of image search of the SISP is presented in Table 4.3.

Table 4.3 The image search result of SISP

No	Partial name of URL	Image	Number of search results	Number of bytes	Response time
1	jesad.com		2	281	311
2	21stcenturyfood.com		3	303	480
3	wallpaper77.com		1	14268	1822
4	koiaichaku.com		3	15231	4166
5	all4myspace.com		3	14847	9481
6	worldtravellers.com		2	15087	1292
7	statravel.com		14	14041	5959
8	liiopet.com		4	15012	2484
9	wallpapersweb.com		1	14562	3135
10	wikipedia.org		5	15450	4298

Table 4.3 The image search result of SISP (cont.)

No	Partial name of URL	Image	Number of search results	Number of bytes	Response time
11	nasa.gov		12	13587	4285
12	uiowa.edu		2	14007	3255
13	photofurl.com		8	14291	3555
14	myspace.com		2	14825	2744
15	chula.ac.th		3	14865	2664
16	myspace.com		1	13877	721
17	mycatspace.com		4	15117	2674
18	multiply.com		4	15118	2408
19	bigcatrescue.org		7	14637	2942
20	things-to-do-in-thailand.com		3	15057	1642

Referring to Table 4.3, the number of bytes and the response time of the data set numbers 1 and 2 are smaller than the others because the required URLs of the data set numbers 1 and 2 are on the nearest MGSES of client. Therefore, when the query message is sent to the nearest MGSES, the MGSES can reply with the results to the client immediately, without sending the query message to other MGSESs. Additionally, the number of bytes and the response time of the data set numbers 1 and 2 are small.

CHAPTER V

DISCUSSION AND CONCLUSION

5.1 Discussion on Proposed Routing Algorithm

Due to the fact that there are a large amount of shared information resources distributed over the Internet, search protocols with efficient search algorithms are significant factors for users' successful searches. Unfortunately, the available search mechanisms cannot return the expected results as needed by users. Moreover, some mechanisms may cause network congestion and high delay when searching. Thus, the development of a search mechanism called Patrol Packet Algorithm (PTP) is proposed for the Information Search Protocol (ISP).

The proposed search technique, PTP, has modified the flooding technique so that the query messages are sent to all nodes with a small response time and a guaranteed returned list from required locations. Although this technique was altered from the original flooding technique by the use of a small packet to check the status of receiving query message, the visited nodes and number of distributed bytes are less than in the original method. Another factor-of-performance indicator is the response time, and this test has shown that the response time of the flooding algorithm is slightly faster than the response time of the PTP algorithm. This is because there is a check process for the status of receiving query message, before sending the query message.

The other performance comparison was performed between the PTP and the hierarchical technique that is the standard searching technique. The result indicates that the number of bytes of the PTP is more than the hierarchical technique because of the pattern the hierarchical technique uses to send the queries; it sends the query message from the parent to the children. Therefore, the number of sending queries is equal to the number of nodes, which is the same as the PTP algorithm. However, the PTP has an overhead from sending the PTP packet for node checking. Thus, the performance of the hierarchical technique is better than PTP when considering the number of bytes flowing

in the communication channel. Since the physical network of the Internet is the mesh topology, the response time from the PTP algorithm is better than the hierarchical algorithm.

5.2 Conclusion on Proposed Routing Algorithm

The required information is distributed over various web sites. Thus, the search method is important for the proper search information. However, the existing search engine techniques generally obtain a large number of web site lists, and the users have to take time to find the specific information. Therefore, the ISP is proposed for a specific search by entering a part of a URL and some keywords to find the full path name of the required URLs. The search results have shown a shorter list of more specific required URLs. However, the defect of the previous ISP mechanism is the routing system.

Therefore, the PTP that is proposed and implemented in this paper is intended to decrease the ISP duplication and rebounding to the same MGSES. The simulation results of the PTP algorithm and the routing of the previous ISP show that the proposed routing algorithm reduces the number of bytes in a communication channel by more than 50%. Furthermore, the numerical analysis has shown that the number of bytes of the proposed algorithm is less than the previous routing algorithm. Thus, the PTP algorithm can avoid the congestion caused by transferring the ISP in the system.

According to the results of performance testing of the PTP, this algorithm is suitable for the search requirements over the distributed system where a large of query messages spread over the communication channel. Moreover, the PTP algorithm can apply to other systems that send the query messages to all nodes for finding the distributed resources such as peer-to-peer systems and distributed databases.

5.3 Discussion on the Performance of SISP

Since the Internet contains enormous amounts of information and images, there are various methods to gain access to the information and images over the Internet. However, search processing to reach the required information and images is more difficult. In order to access specific information and images, the users search from available search engines, such as Google, Yahoo, and MSN Search. Most search engines allow users to enter keywords or required words in the search field. The result list of searching acquired from the search process can be long or short depending on the entered keywords. Therefore, if the user enters suitable keywords, the search result will be specific; otherwise, a long list of irrelevant information and images will appear. Therefore, it is a time consuming for users to find the required information and images.

In order to narrow down the search result, Information Searching Protocol (ISP) was proposed. This protocol enables users to enter a partial part of the required URL and keywords. Thus, the ISP can narrow down the search and reduce the search time for the required information. Conversely, the ISP cannot perform the image search. Therefore, this research enhances the ISP mechanism for searching for information and images.

This research proposes an SISP that can search information and images. In order to add the search image mechanism to SISP, the Color Region Correlation (CRC) is used for the image feature extraction process because the CRC index is simple to calculate and invariant to translate, rotate, and scale. Moreover, the size of feature index is small. This research also improves the CRC method using the median filter to reduce the CRC calculation of the color region number.

The performance evaluation of the SISP is the number of search results, the number of bytes, and the response time to perform the information search compared with the ISP. The number of bytes of SISP is equal to the number of bytes of ISP by entering the partial name of URL and keyword(s). Therefore, even though the search

performance of SISP is same the ISP, SISP can narrow down the result to the require URL. Moreover, the number of bytes of SISP is less than the number of bytes of ISP. The SISP can decrease the number of bytes in the communication channel by more than 50%. This is because the PTP is used look for sending path of SISP. However, the response time of SISP is slightly slower than the response time of ISP. This is because there is a PTP process for checking the status of receiving SISP, before sending the SISP.

In image search, since the ISP does not perform the image search, the ISP cannot compare with the SISP. The SISP can search using the partial pathname of a URL and the query image. Therefore, the SISP is more efficient than ISP.

Even though the SISP is an efficient search mechanism allowing users to search from a partial path of a URL, if a user enters the wrong partial pathname of URL, or misspells of a partial pathname of a URL, the return result will be NULL.

5.4 Conclusion on the Performance of SISP

As information and images are widely distributed over the Internet, there are many websites containing information and images. Unfortunately, searching for specific information and images over the Internet using the available search engine techniques usually result in a large list of URLs. These results are time consuming for users to find the specific need in a short period of time.

Therefore, this research proposes the SISP, which is implemented to narrow down the search results of information and image searches. This protocol is adjusted from the ISP that supports the search using a pathname of URL and keywords to focus the search result to a specific content. Therefore, the SISP has added the image search mechanism. The image extraction of the SISP uses Color Region Correlation (CRC). The CRC index is simple to calculate and invariant to translate, rotate, and scale. Furthermore, the feature index is small in size. However, the CRC method is improved to reduce the CRC calculation of the color region number using the median filter.

According to the search results of the SISP, the SISP can better search information and images; it is more useful than the ISP. Additionally, SISP can reduce the number of bytes in the communication channel and the accuracy of the search result is still maintained. Thus, the performance of the SISP is obviously higher than the performance of the ISP.

References

- [1] Bhattarakosol, P., and Preechaveerakul, L. Information Searching Protocol: A Smart Protocol for Specific Content Search over Internet. Proceedings of SPIE's International Symposium on Optics East 2006 (IT405). (October 2006).
- [2] Lien, D., and Yan P. Measuring the efficiency of search engines: an application of data envelopment analysis. Applied Economics. 31, 12 (1999): 1581.
- [3] Web Searching: Search Engine Components. Dartmouth Biomedical Libraries. [Online]. Available from: <http://www.dartmouth.edu/~library/biomed/guides/web-search/components.html> [2012, April 16].
- [4] Bhatia, M.P.S. and Gupta, D. Discussion on Web Crawlers of Search Engine. Proceedings of 2nd National Conference on Challenges & Opportunities in Information Technology (COIT-2008). (March 2008): 227-230.
- [5] Chung, C. and Clarke, C. Topic-oriented collaborative crawling. CIKM. (2002): 34-42.
- [6] Chen L., and Scullion C. An Empirical Study and Evaluation on the Cost-Effectiveness of Pay-Per-Click e-Marketing. International Conference on Management of e-Commerce and e-Government (ICMeCG09). (2009)
- [7] Jurafsky, D., Martin, J. H. Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd ed. Pearson Prentice Hall. 2000.
- [8] Tümer, M. A. S. and Bitirim, Y. An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Halka. International Conference on Internet Monitoring and Protection (ICIMP). (2009). pp. 51-55.

- [9] Winter, J. Routing of structured queries in large-scale distributed systems. LSDS-IR 2008, 11-18.
- [10] Albitz, P. and Lui, C. DNS and BIND. 3rd ed. O'Reilly & Associates Inc, 1998.
- [11] Doshi, P. and Raisinghani, V. Review of dynamic query optimization strategies in distributed database. Electronics Computer Technology (ICECT), 2011 3rd International Conference on. 6 (2011): 145 – 149.
- [12] Bisnik, N., and Abouzeid, A. Modeling and Analysis of Random Walk Search Algorithms in P2P Networks. Proc. of International Workshop on Hot Topics in Peer-to-Peer Systems. (July 2005): 95 - 103.
- [13] Lv. C, Cao. P, Cohen. E, Li. K, and Shenker. S. Search and replication in unstructured peer-to-peer networks. Proc.16th Int. Conf. Supercomputing. (2002): 84-95.
- [14] Tsoumakos, D., and Roussopoulos, N. Adaptive Probabilistic Search for Peer-to-Peer Networks. Proc.3rd Int. Conf. P2P Computing. (2003): 102–109.
- [15] Li, X., and Wu, J.. Searching techniques in peer-to-peer networks. Handbook of Theoretical and Algorithmic Aspects of Sensor, Ad Hoc Wireless, and Peer-to-Peer Networks. (2005).
- [16] Lin, T. and Wang, H. Search Performance Analysis in Peer-to-Peer Networks. Proc. of the Third International Conference on Peer-to-Peer Computing (P2P'03). (2003): 204- 205.
- [17] Tsoumakos, D. and Roussopoulos, N. Analysis and Comparison of P2P Search Methods. Proc.1st Int. Conf. Scalable Information Systems (INFOSCALE 2006). 25 (2006).

- [18] Stoica, I., Morris, R., Karger, D., Kaashoek, F. and Balakrishnan, H. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. ACM SIGCOMM Conference. (2001).
- [19] Tempich, C. Staab, S. and Wranik, A. "REMINDIN"* Semantic Query Routing in Peer-to-Peer Networks Based on Social Metaphors. Proceedings of the 13th International World Wide Web Conference. (2004): 640-649.
- [20] Yeferny, T. and Arour, K. LearningPeerSelection: A Query Routing Approach for Information Retrieval in P2P systems. International Conference on Internet and Web Applications and Services (ICIW 2010), (May 2010): 235–241.
- [21] Kalogeraki, V., Gunopulos, D. ,and Zeinalipour-Yazti, D. A local searchmechanism for peer-to-peer networks. Proc. of the Eleventh International Conference on Information and knowledge Management. (2002). pp. 300 - 307
- [22] Crespo, A. and Garcia-Molina, H. Routing Indices for Peer-to-Peer Systems. Proc. Of 22nd Intl. Conference on Distributed Computing System, 2002s. 23- 32.
- [23] Yuan, F., Liu, J.and, Yin, C. A Scalable Search Algorithm for Unstructured Peer-to-Peer Networks. Proc. of Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/ Distributed Computing - SNPD 2007. (2007): 199 – 204.
- [24] Seshadri, S. and Brian, F. Cooper: Routing Queries through a Peer-to-Peer InfoBeacons Network Using Information Retrieval Techniques. IEEE Trans. Parallel Distrib. Syst. 18, 12 (2007): 1754-1765.
- [25] Rozlina, M. and Buckingham, C. D. Preprocessing for Improved Query Routing in Super-peer P2P Systems. 2008 IEEE Region 5 Technical, Professional, and Student Conference 2008, (2008).

- [26] Kalogeraki, V., Gunopulos, D. and Zeinalipour-Yazti, D. A local search mechanism for peer-to-peer networks. Proc. of the Eleventh International Conference on Information and Knowledge Management. (2002): 300-307.
- [27] Yang, B. and Garcia-Molina, H. Improving search in peer-to-peer systems. Proc. of the 22nd IEEE International Conference on Distributed Computing (IEEE ICDCS'02). 2002.
- [28] Dorrigiv, R., L'opez-Ortiz, Al., and Pratat, P. Search Algorithms for Unstructured Peer-to-Peer Networks. Proc. of 32nd IEEE Conference on Local Computer Networks (LCN '07). (2007): 343-352.
- [29] Jawhar, I. and Wu, J. A two-level random walk search protocol for peer-to-peer networks. Proc. of the 8th World Multi-Conference on Systemics, Cybernetics and Informatics. (2004).
- [30] Mockapetris, P. V. Development of the Domain Name System. Proceedings of ACM SUGCOMM'88. (1988): 123 - 133.
- [31] Mockapetris, P. V. Domain names-concepts and facilities. RFC 1034. (Nov. 1987).
- [32] Mockapetris, P. V. Domain names-implementation and Specification. RFC 1035. (Nov. 1987).
- [33] Datta, R., Joshi, D., Li, J. and Wang, J. Image Retrieval: Ideas, Influences, and Trends of the New Age. Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval. (Nov. 2005).
- [34] Carson, C., Belongie, S., Greenspan, H. and Malik, J. Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying. IEEE Trans. On PAMI. 24, 8 (2002): 1026-1038.

- [35] Chen, Y. and Wang, J. Z. A Region-Based Fuzzy Feature Matching Approach to Content- Based Image Retrieval. IEEE Trans. on PAMI. 24, 9 (2002): 1252-1267.
- [36] Natsev, A., Rastogi, R. and Shim, K. WALRUS: A Similarity Retrieval Algorithm for Image Databases. Proc. ACM SIGMOD Int. Conf. Management of Data. (1999): 395–406.
- [37] Li, J., Wang, J.Z. and Wiederhold, G. IRM: Integrated Region Matching for Image Retrieval. Proc. of the 8th ACM Int. Conf. on Multimedia, (Oct. 2000): 147-156.
- [38] Mezaris, V., Kompatsiaris, I. and Strintzis, M. G. Region-based Image Retrieval Using an Object Ontology and Relevance Feedback. Eurasip Journal on Applied Signal Processing. 2004, 6 (2004): 886-901.
- [39] Ma, W.Y. and Manjunath, B.S. “NETRA: A Toolbox for Navigating Large Image Databases. Proc. IEEE Int. Conf. on Image Processing. 1, (Oct. 1997): 568–571.
- [40] Niblack W., et al. The QBIC Project: Querying Images by Content Using Color, Texture, and Shape. Proc. SPIE. 1908, (Feb. 1993): 173–187,
- [41] Lew, (Ed.) M. S. Principles of Visual Information Retrieval. Springer. 2001.
- [42] Smeulders, A. W. M., Worring, Santini, M. S., Gupta, A. and Jain, R. Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. Pattern Analysis and Machine Intelligence. 22, 12 (2000): 1349-1380.
- [43] Wang, Z., Zheng, Q., and Sun, J. An Effective Content-based Web Image Searching Engine Algorithm. Proceedings of the 2010 IEEE ICMIT. (2010)
- [44] Swain, M., and Ballard, D. Color indexing. International Journal of Computer Vision 7. 1 (1991).
- [45] Daneels, D., Campenhout, D., Niblack, W., Equitz, W., Barber, R., Bellon, E., and Fierens, F. Interactive outlining: An improved approach using active

- contours. Proc. SPIE Storage and Retrieval for Image and Video Databases. (1993).
- [46] Equitz, W., and Niblack, W. Retrieving Images from a Database using Texture—Alogrithms from the QBIC System, Technical Report RJ 9805, Computer Science, IBM Research Report. (May 1994).
- [47] Faloutsos, C., Flickner, M., Niblack, W., Petkovic, D., Equitz, W., and Barber, R. Efficient and Effective Queryingby Image Content, Technical Report, IBM Research Report. (1993).
- [48.] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafine, J., Lee, D., Petkovic, D., Steele, D. ,and Yanker, P. Query by image and video content: The QBIC system. IEEE Computer. (1995).
- [49] Lee, D., Barber, R., Niblack, W., Flickner, M., Hafner, J., and Petkovic, D. Indexing for complex queries on a query-by-content image database. Proc. IEEE Int. Conf. on Image Proc., (1994).
- [50.] Niblack, W., Barber, R., and et al. The QBIC project: Querying images by content using color, texture and shape. Proc. SPIE Storage and Retrieval for Image and Video Databases. (Feb 1994).
- [51] Scassellati, B., Alexopoulos, S., and Flickner, M. Retrieving images by 2d shape:Acomparison of computation methods with human perceptual judgments. Proc. SPIE Storage and Retrieval for Image and Video Databases. (1994).
- [52] Tamura, H., Mori, S., and Yamawaki, T. Texture features corresponding to visual perception. IEEE Trans. On Sys., Man. and Cyb. SMC-8(6), (1978).

- [53] Lei, Z., Fuzong, L., and Bo, Z. A CBIR Method Based On Color-Spatial Feature. Technical Report, Department of Computer Science and Technology, Tsinghua University.
- [54] McCamy, C. S., Marcus, H., and Davidson, J. G. A color-rendition chart. Journal of Applied Photographic Engineering. 2, 3 (1976).
- [55] Miyahara, M. Mathematical transform of (r,g,b) color data to munsell (h,s,v) color data. SPIE Visual Commun.Image Process. 1001, 1988.
- [56] Wang, J., Yang, W.-J., and Acharya, R. Color clustering techniques for color-content-based image retrieval from image databases. Proc. IEEE Conf. on Multimedia Computing and Systems. (1997).
- [57] Swain, M. and Ballard, D. Color indexing. International Journal of Computer Vision. 7, 1 (1991).
- [58] Ioka, M. A Method of Defining the Similarity of Images on the Basis of Color Information. Technical Report RT-0030, IBM Research, Tokyo Research Laboratory. (Nov. 1989).
- [59] Niblack, W., Barber, R., and et al. The QBIC project: Querying images by content using color, texture and shape. Proc. SPIE Storage and Retrieval for Image and Video Databases. (Feb. 1994).
- [60]. Stricker, M. and Orengo, M. Similarity of color images. Proc. SPIE Storage and Retrieval for Image and Video Databases. (1995).
- [61] Smith, J. R., and Chang, S.-F. Single color extraction and image query. Proc. IEEE Int. Conf. on Image Proc. (1995).

- [62] Smith, J. R., and Chang, S.-F. Tools and techniques for color image retrieval. IS & T/SPIE Proceedings, Vol. 2670, Storage & Retrieval for Image and Video Databases IV. (1995).
- [63] Gonzalez, R.C., and Woods, R.E. Digital Image Processing. 2nd ed. Pearson Education, 2002.
- [64] Pratt, W.K. Digital Image Processing. 3rd ed., Wiley, New York, 2001.
- [65] Gonzalez, R.C., Woods, R.E., and Eddins, S.L. Digital Image Processing Using MATLAB. Pearson Education, 2004.
- [66] Jain, A.K. Fundamentals of Digital Image Processing. Prentice-Hall, Englewood Cliffs, NJ, 1989.
- [67] Rosenfeld, A., and Kak., A.C. Digital Picture Processing. Academic Press, New York, 1982.
- [68] Pal, S.K., Ghosh, A., and Shankar., B. Uma. Segmentation of remotely sensed images with fuzzy thresholding, and quantitative evaluation. International J.Remote Sensing. 21, 11 (2000): 2269–2300.
- [69] Jain, A., and Zongker, D. Feature Selection: Evaluation, Application, and Small Sample Performance. IEEE Trans. Pattern Anal. Mach. Intell. 19, 2 (1997): 153–158.
- [70] Yin, L., Yang, R., Gabbouj, M., and Neuvo., Y. Weighted Median Filters: A Tutorial. IEEE Trans. on Circuits and Systems. 43, 3. (March 1996): 157-192.
- [71] Haralick, R.M., and Shapiro, L.G. Computer and Robot Vision. Addison-Wesley, 1 (1992).
- [72] Baxes, G.A. Digital Image Processing. Principles & Applications. Wiley & Sons. (1994).

- [73] Lim, Jae S. Two-Dimensional Signal and Image Processing. (1990): 469-476.
- [74] P. Bojarczak, and S. Osowski. Denoising of Images – a Comparison of Different Filtering Approaches. WSEAS Transactions on Computers. 3 (July 2004): 738, 744.
- [75] Gaussian blur. Wikipedia. [Online]. Available from: http://en.wikipedia.org/wiki/Gaussian_blur. [2012, April 10]
- [76] Braquelaire, J., and Brun, L. Comparison and optimization of methods of color image quantization. IEEE Transactions on Image Processing. 6, 7 (1997): 1048-1052.
- [77] Velho, L., Gomes, J., and Sobreiro M. Color image quantization by pairwise clustering. Proceedings of the 10th Brazilian Symposium on Computer Graphics and Image Processing. (1997): 203-207.
- [78] Rui, X., Chang, C., and Srikanthan T. On the initialization and training methods for Kohonen self-organizing feature maps in color image quantization. Proceedings of the First IEEE International Workshop on Electronic Design. Test and Applications. (2002).
- [79] Freisleben, B., and Schrader A. An evolutionary approach to color image quantization. Proceedings of IEEE International Conference on Evolutionary Computation. (1997): 459-464.
- [80] Scheunders, P. A Genetic C-means Clustering Algorithm Applied to Color Image quantization. Pattern Recognition. 30, 6 (1997): 859-866.
- [81] Xiang, Z., and Joy, G. Color Image Quantization by Agglomerative Clustering. IEEE Computer Graphics and Applications. 14, 3 (1994): 44-48.

- [82] C. Paiboolsirikul, and N. Covanisaruch. Feasibility Study for Image Indexing by ColorRegion Correlation. Proceedings ICEP. (2002).

Biography

Name: Miss Monlica Wattana.

Date of Birth: 25th May, 1982.

Educations:

- Ph.D., Program Computer Science, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Thailand, (June 2007 - May 2012).
- Visiting scholar, Department of Computer Science and Engineering, Konkuk University, Korea, (September 2009 - August 2010)
- M.Sc. Program Computer Science, Faculty of Science, Khon Kaen University, Thailand, (June 2004 - November 2006).
- B.Sc. Program Computer Science, Faculty of Science, Khon Kaen University, Thailand, (June 2000 - March 2004).

Publication papers:

- M. Wattana, P. Bhattarakosol and S. Han, Information Searching Protocol for Color Image (ISPCI). Proceedings of 2010 International Conference on Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE, Chengdu, China, 2010.
- M. Wattana and P. Bhattarakosol, Patrol Packet (PTP) For Routing Algorithm of Information Searching Protocol (ISP). Proceedings of 2009 International Conference on Future Networks (ICFN 2009). Bangkok, Thailand, 2009.
- M. Wattana and P. Bhattarakosol, Information Searching Protocol for E-Commerce (ISPEC), Proceedings of 2009 International Conference on Wireless Information Networks & Information System (WINBIS 09), Kathmandu, Nepal, 2009.
- M. Wattana and P. Bhattarakosol, Patrol Packet Algorithm: a smart routing algorithm for the naming system, Proceedings of 2008 International Conference on Networks, Applications, Protocols, and Services (NetApps2008), Universiti Utara Malaysia, Malaysia, 2008.

Scholarship:

- THE 90th ANNIVERSARY OF CHULALONGKORN UNIVERSITY FUND (Ratchadaphiseksomphot Endowment Fund), (May 2011).
- The University Development Committee (UDC) Scholarship Program of the Office of the Higher Education Commission, Thailand. (June 2004 – May 2011).