Studies on Mobile Agents for Query Retrieval and Web Page Categorization Using Neural Networks

Ali Selamat

June 2003

Doctoral Thesis at Osaka Prefecture University

Abstract

Mobile agent is an emerging technology that is gaining momentum in the field of distributed computing. There are some advantages in using the mobile agent technology compared with a traditional client-server solution. For example, it can reduce a network traffic, it can support a large scale of computations with many computers in a distributed environment, it allows the use of disconnected computing for processing user queries, and it provides more flexibility in the development and maintenance of distributed applications. The goal of this research is based on the application of mobile agent technology in supporting the query retrieval process from the World Wide Web (WWW). Specifically, the methods of dispatching the mobile agents to retrieve the query results from the search engines in WWW have been investigated. We have also considered the ranking and classification methods applied to the query results that have been retrieved by the mobile agents.

The scopes of the research are as follows: First, the effectiveness of mobile agent for query retrieval using the off-line and on-line approaches is investigated. We have found that the query retrieval using the off-line approach by the mobile agent is better compared with the on-line approach. Second, the ranking of query retrieval results that have been retrieved by the mobile agents using the Number of Ordering Score (*NROS*) is investigated. The *Precision* of the query results using the *NROS* is higher than the *Recall* scores. It indicates that from all of the documents returned from the query, a large proportion of the documents is relevant to the user by using the *NROS* approach.

Third, the performance of mobile agents for query retrieval using an extended hierarchical query retrieval (EHQR) approach compared with the hierarchical query retrieval (HQR) approach is investigated. The result shows that the total routing time taken by the mobile agents to retrieve the query results using the EHQR approach is less compared with the

HQR approach. Fourth, the classification of news web pages retrieved by the mobile agents using neural networks based on a background knowledge is evaluated. A new web news categorization approach, namely, a Web Page Classification Method (WPCM) is proposed. The WPCM uses a neural network with inputs obtained by both the principal components and class profile-based features (CPBF). The experimental evaluation demonstrates that the WPCM provides acceptable classification accuracy with the sports news datasets. Finally, we have also overcome the limitation of the principal component analysis-neural networks (PCA-NN method) in supervised data where the characteristic variables that describe smaller classes tend to be lost as a result of the dimensionality reduction by using the WPCM. The classification accuracy on the small classes can be improved although they have been reduced into a small number of principal components.

Acknowledgments

I would like to express my true and sincere thanks and gratitude to my supervisor, Professor Sigeru Omatu, Division of Computer and Systems Sciences, Graduate School of Engineering, Osaka Prefecture University, for his patience, guidance, encouragement, invaluable comments, and advices that made this research possible to complete in time. Also, I would like to thank to Professor Keinosuke Matsumoto, Professor Akira Shiozaki, and Professor Kunio Fukunaga for their comments and recommendations during the preparation of this thesis. Furthermore, I would like to thank Associate Professor Safaai Deris, Faculty of Computing and Information Systems, University of Technology Malaysia (UTM), for giving me a courage to pursue my Ph.D. study under Professor Sigeru Omatu's supervision. Moreover, I would like to thank Mr. Hidekazu Yanagimoto, Assistant Professor Michifumi Yoshioka, Associate Professor Toru Fujinaka, and all members in the Signal Information Processing Laboratory, Division of Computer and Systems Sciences, Osaka Prefecture University for their continuous and constant supports in many aspects of this research.

I would like to thank the Ministry of Education, Culture, Sports, Science and Technology, Japan (MONBUKAGAKUSHO) for supporting my studies while I was in Japan. Also, I would like to thank UTM for giving me permission for my study leave at Osaka Prefecture University, Japan.

I dedicated my work to my wife, Syamsulina Mohamad and my son, Muhamad Firdaus Ali for their patience and encouragement during the many long weekends and late evenings that made the completion of this research and thesis possible. Finally, I would like to thank my brother, Mohd. Hafiz Selamat, for supporting me during my Ph.D. studies at Osaka Prefecture University.

Contents

List of Figures

List of Tables

List of Abbreviations

MaSS	:	Mobile agent search system
LP	:	Local prefetch servers
W	:	The MaSS client agent
X	:	The MaSS server agent
Y	:	The MaSS search agent
Z1,Z2,Z3	:	The local agents at hosts **B, C,** and **D**
A	:	The MaSS server
B	:	A local Yahoo database server
C	:	A local AltaVista database server
D	:	A local HotBot database server
IDF	:	Inverse Document Frequency
n_i	:	The total number of occurrences of term i in the collection
$Max\ n_i$	:	The maximum frequency of any term in the collection
TF_{ij}	:	The term frequency of word i in document Doc_j
i	:	$i = 1, \ldots, N$
j	:	$j = 1, \ldots, n$
$length_j$	:	The total number of unique terms in Doc_j after the stopping and stemming processes
$Erroneous$	:	The number of URLs that have broken and duplicated links
D	:	The total number of duplicated documents that exist in the results set
R	:	The total of similarity score
$Error$	:	The total number of broken links that exist in the results set
ROS	:	Relevant Ordering Score
$Similarity_j$	:	The similarity value of Doc_j
a	:	The system and the expert agree with the assigned category
b	:	The system disagrees with the assigned category but the expert did

c	:	The expert disagrees with the assigned category but the system did
Precision	:	A high *Precision* indicates that from all of the documents returned by our query, a large proportion of the documents is relevant to the search
Recall	:	A high *Recall* indicates that from all of the documents in the archive that are relevant to the query, a large number of these documents are returned
F1	:	The average of *Precision* and *Recall*
NROS	:	Number of Relevant Ordering Score
EHQR	:	An extended hierarchical query retrieval (EHQR)
HQR	:	A hierarchical query retrieval approach (HQR)
IQR	:	An itinerary query retrieval approach (IQR)
PQR	:	A parallel query retrieval approach (PQR)
SQR	:	A serial query retrieval approach (SQR)
SHQR	:	A shuttle query retrieval approach (SHQR)
H	:	The maximum number of hosts, i.e. $H = 8, 16, 64$ etc.
h	:	The height of a $tree$
t_p	:	The root of a $tree$
β	:	The branch of a $tree$
T_{HQR}	:	The total time taken by the MaSS search agent to collect the query search results by using the HQR approach
T_{EHQR}	:	The total time taken by the MaSS search agent to collect the query search results by using the EHQR approach
T_{PQR}	:	The total time taken by the MaSS search agent to collect the query search results by using the PQR approach
T_{SQR}	:	The total time taken by the MaSS search agent to collect the query search results by using the SQR approach
T_{SHQR}	:	The total time taken by the MaSS search agent to collect the query search results by using the SHQR approach
$t_{context}$	:	The time taken to create a new MaSS search agent upon a disposal of the previous MaSS search agent when it returns home to the MaSS server after the first trip to the LP servers
W_{send}	:	The time taken to send the query results from the MaSS client agent (**W**) to the MaSS server agent (**X**)
W_{recv}	:	The time taken to receive the query results from the MaSS client agent (**W**) to the MaSS server agent (**X**)
$t(\theta)$	:	The timestamp registered by the MaSS search agent prior to departing from the MaSS server to the LP servers
$t(\phi)$	:	The timestamp computed by the MaSS search agent immediately upon returning to the MaSS server from the LP servers

l	:	The number of trials that have been done in each of the experiments where $l = 1, \ldots, \tau$
τ	:	The maximum number of trials l
PCA-NN	:	Principal component analysis-neural networks
TF-IDF	:	Term frequency-inverse document frequency
WPCM	:	Web page classification method
CPBF	:	Class profile based features
PCA	:	Principal component analysis
Doc_j	:	Refer to each web page document that exists in the news database where $j = 1, \ldots, n$
w_k	:	A word
x_{jk}	:	Terms weight
A	:	A matrix with document-terms weight
df_k	:	A document frequency df_k is the total number of documents in the database that contains the word w_k
idf_k	:	A document frequency idf_k = $log(\frac{n}{df_k})$ where n is the total number of documents in the database
$\overline{x}_k$	:	The mean of m variables in data matrix A
S	:	The matrix S where $S = \{s_{jk}\}$
s_k^2	:	The variance of matrix S
s_{ik}	:	The covariance of matrix S
λ	:	An eigenvalue
$\mathbf{e}$	:	An eigenvector
$\mathbf{E}$	:	An eigenvector matrix where $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \ldots, \mathbf{e}_m]$.
$\mathbf{\Lambda}$	:	A diagonal nonzero eigenvalue matrix
d	:	The desired value of eigenvectors, e.g., 100, 200, 400, etc.
$f_{i1} = \mathbf{e}_1^T \mathbf{x}_i$	:	The set of principal components is represented as $f_{i1} = \mathbf{e}_1^T \mathbf{x}_i$, $f_{i2} = \mathbf{e}_2^T \mathbf{x}_i$, …, $f_{id} = \mathbf{e}_d^T \mathbf{x}_i$, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^T$ and $i = 1, 2, \ldots, n$
M	:	A feature matrix reduced from the $m \times m$ original data size to an $n \times d$ size
n	:	The number of documents in a collection
F_k	:	A frequency of the term k in the entire document collection
R	:	The number of words that have the highest entropy value selected from the CPBF approach
L_{jk}	:	The local weighting of the term k
G_k	:	The global weighting of the term k
p	:	The number of input layers
q	:	The number of hidden layers

f_{d+R}	:	The input values to the neural network are represented by $f_1, f_2, \ldots, f_{d+R}$
η	:	A learning rate
α	:	A momentum rate
t	:	The number of iteration
MSE	:	A mean squared error
T_j	:	The training document
cs_u	:	The desired classification vector corresponding to Doc_j, $u = 1, \ldots, v$
Doc'	:	A new document to be classified
w_k^m	:	The words representing the textual content of the document Doc' and let k denote the term number where $k = 1, .., m$
$P(cs)$	:	The classification score using Bayesian classifier

Chapter 1

Introduction

Mobile agent is an emerging technology that is gaining momentum in the field of distributed computing. There are some advantages in using mobile agent technology compared with traditional client-server solutions. For example, it can reduce a network traffic, it can support a large scale of computations with many computers in a distributed environment, it allows the use of disconnected computing for processing a user queries, and it provides more flexibility in the development and maintenance of distributed applications. The goal of this thesis is based on the application of mobile agent technology in supporting the query retrieval process from the World Wide Web (WWW). Specifically, the methods of dispatching the mobile agents to retrieve the query results from the search engines in WWW have been investigated. We have also considered the ranking and classification methods applied to the query results that have been retrieved by the mobile agents.

This introduction chapter is organized as follows: In Section 1.1, the characteristics of mobile agents are discussed. The evaluation of mobile agent performance for the query retrieval is discussed in Section 1.2. The ranking of query results and the classification of web pages are described in Section 1.3. The goal of the thesis is discussed in Section 1.4. The organization of the thesis is described in Section 1.5.

1.1 Mobile Agents

Mobile agent is an autonomous program that can be transported from one computer to another in a heterogeneous network, such as Internet or an Intranet [27, 66], and can

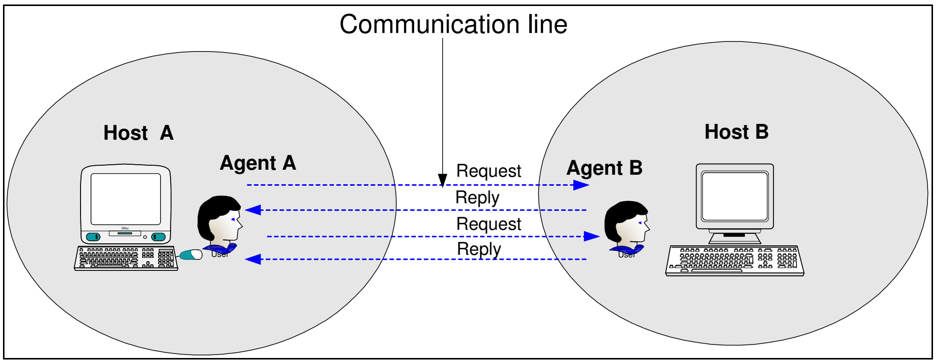

Figure 1.1: A client-server type of communication.

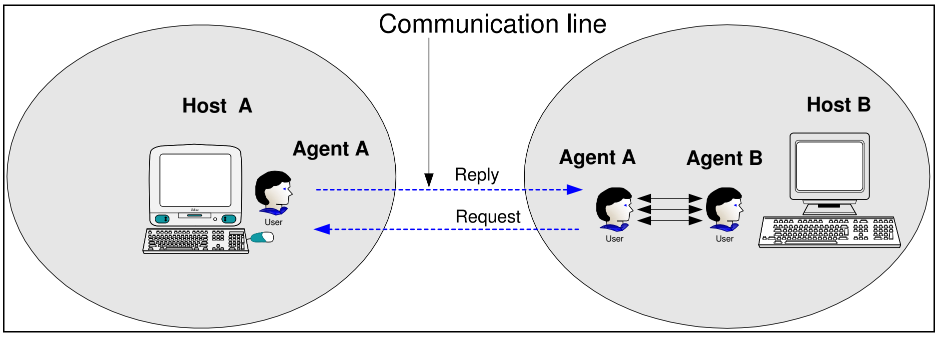

Figure 1.2: A mobile agent type of communication.

continue its execution after it has reached its destination [25],[26],[17]. A mobile agent comprises program code, data, and in some cases the execution state of code. These are packaged into message and sent over a network to a remote location where some codes resume an execution. Mobile agents are particularly suited for applications that require a large amount of data to be transferred between remote locations. It is, therefore, more advantageous to send a mobile agent, whose size is less than the size of the data transfer involved to a remote location to perform multiple local interactions and return the result of the computation. A concept of mobile agent evolves from the concept of a client-server based computing as shown in Fig. 1.1. In the client-server systems, data are passed between the server and the client while the program resides on the client and the server. In contrast with the client-server systems, the mobile agent based systems transfer the program logic along with the data, as shown in Fig. 1.2. This provides mobile agent based systems with features that were unavailable in the traditional client-server systems.

1.1.1 Mobile agents properties

In order to do such tasks that has been explained in a previous section, mobile agents must have the properties such as mobility, path determination, communication, etc. [14]. They are describe as follows:

i. **Mobility**

   Mobility is a property that enables a mobile agent to move from one node in a computer network to another. In a migration process, the system decides when the process should migrate depending on several factors including a load balancing and a resource allocation. However, as the mobile agent is an autonomous entity, it will decide when it is to move, which depends on, whether it has achieved its goal at its current location, etc. Forced migration can also be performed on the mobile agents in special circumstances. For example, a computer might refuse to host a malicious mobile agent that performs a forbidden action.

ii. **Path determination**

   Mobile agent moves from one node of computer network to another with itinerary nodes to visit. The itinerary can either be pre-determined or it can be determined dynamically as the agent gains more information about its environment.

iii. **Communication**

   Unlike stationary agents, which support communication between agents in the environment, mobile agents need to communicate with the local environment on the remote locations that they have visited as well as with other agents that they have encountered at those locations. This requires mobile agents to possess two communication modes. First, a node-oriented communication which is used for interacting with local processes on remote locations they visit. Second, a network-oriented communication such as message passing which is used for interacting with other agents. Most mobile agent systems support synchronous as well as asynchronous modes of communication between agents [17].

1.1.2 Advantages of mobile agents

There are some advantages of using mobile agents for distributed computing applications [66]. They can be described as follows:

i. **They reduce the network load**

   Distributed systems often rely on communication protocols that involve multiple interactions to accomplish a given task. This is especially true when the security measures are enabled. The result is a lot of network traffic. Mobile agents allow us to package the conversation and dispatch it to a destination host, where the interactions can take place locally (see Fig. 1.2). Mobile agents are also useful when it comes to reducing the flow of raw data in the network. When very large volumes of data are stored at remote hosts, these data should be processed in the locality of the data rather than transferred over the network.

ii. **They overcome network latency**

   Critical real-time systems such as robots that are used in the manufacturing processes (i.e., semiconductor factory, chemical plant, etc.) need to respond to changes in their environments in real time. There are some latencies when controlling such systems with different sizes of factory networks. For critical real-time systems, such latencies are not acceptable. Mobile agents offer a solution, since they can be dispatched from a central controller to act locally and directly execute the controller's directions [15].

iii. **They encapsulate protocols**

   When data are exchanged in a distributed system, each host owns the code that implements the protocols needed to properly encode outgoing data and interpret incoming data, respectively. However, as protocols evolve to accommodate new efficiency or security requirements, it is a tedious task to upgrade the protocol codes properly. The result is often that protocols become a legacy problem. Mobile agents, on the other hand, are able to move to remote hosts in order to establish channels based propriety protocols.

iv. **They execute autonomously and asynchronously**

   Often mobile devices have to rely on expensive or fragile network connections. Tasks that require a continuously open connection between a mobile device and a fixed network probably will not economically or technically feasible. To solve this problem, tasks can be embedded into mobile agents, which can be dispatched into the network. After being dispatched, the mobile agents become independent of the creating process and can operate asynchronously and autonomously.

v. **They are robust and fault tolerant**

The ability of mobile agents to react dynamically to unfavorable situations and events makes it easier to build robust and fault-tolerant distributed systems. If a host is being shut down, all agents executing on that machine will be warned and given time to dispatch and continue their operation on another host in the network.

vi. **They are naturally heterogeneous**

Network computing is fundamentally heterogenous, often both from a hardware and a software perspective. The mobile agents are generally computer and transport-layer-independent and are dependent only on their execution environment. Therefore, they provide optimal conditions for different system integration. For example, the integration of computer applications between a Microsoft Windows operating system and a Unix operating system.

1.1.3 Mobile agents tasks

As mentioned previously, mobile agents are programs that perform some tasks for the user and are able to migrate from computer to computer. Mobile agents themselves decide when to migrate, where to go and what to do when they get there, and it is also important to note how they do this. Mobile agents are free to act as clients and servers and they can communicate freely with other mobile agents. This communication protocols of mobile agents can be synchronous or asynchronous, depending on the agent implementation language (i.e., java, C++, SmallTalk, etc.) and the services required (i.e., object dispatching, object cloning, etc.) [10]. For example, in order to get a list of data from another agent, a synchronous mode of communication will be used. If a large size of data will be transferred through the network by the agents, an asynchronous mode of communication will be used. Like human beings, the mobile agents need to communicate with the other agents in order to accomplish the assigned tasks. For example, if a mobile agent migrates to a computer it may access the resources of that computer through a local agent that handle the request from other agents.

1.1.4 Mobile agents applications

There are many applications that benefit from the mobile agents paradigm such as an electronic commerce, a personal assistance, a distributed information retrieval, a telecommunication and network services, an information dissemination, a parallel processing, etc. Some of the field of applications are described as follows:

i. **Electronic commerce**

   Mobile agents are well suited for electronic commerce. A commercial transaction may require real-time access to remote resources such as stock quotes and perhaps even agent-to-agent negotiation. Different agents will have different goals and will implement and exercise different strategies to accomplish these goals. The mobile agents have been used to accomplish the tasks that have been assigned to them by their owners [17].

ii. **Distributed information retrieval**

   Information retrieval is a popular example of mobile agent application. Instead of moving large amounts of data to the search engine so that it can create indexes, we can dispatch agents to remote sources where they locally create search indexes that can later be shipped back to the origin. Mobile agents can also perform extended searches that are not constrained by the hours during the creator's computer is in operation [25].

iii. **Parallel processing**

   Given that mobile agents can create a cascade of clones in the network, one potential use of mobile agent technology is to administer parallel processing tasks. If a computation requires so much processor power that it must be distributed among processors, an infrastructure of mobile agent hosts can be a possible approach to allocate the processes [75].

iv. **Information dissemination**

Mobile agents can disseminate information such as news and automatic software updates for vendors. The agents bring new software components as well as the installation procedures directly to the customer's personal computer and autonomously update and manage the software on the computer [74].

1.2 Evaluation of Mobile Agents Performance

Performance of mobile agents in the network applications are based on the number of agents used to execute the tasks that have been requested by a user and the dispatching methods applied on each of the agents to execute the tasks. If there are many agents used to accomplish the tasks, the network bottleneck will occur on the system. Furthermore, if the network routing is not properly configured, a significant time delay will occur on the system. Therefore, a new approach for network management needs to be addressed when applying mobile agent technology to perform the query retrieval tasks.

A hierarchical network management using a mobile agent approach has been discussed by Gavalas et al. [10]. However, the issues related to the increasing number of mobile agents used for managing the hierarchical networks have not been considered. Furthermore, the performances of different types of mobile agents platforms have been investigated by Silva et al. [11], Alouf et al. [12], and Sum et al. [18]. But the architectures of mobile agents for performing the tasks in a network environment have not been fully considered. Moreover, the application of mobile agent for a wireless communication system which is based on a client-server architecture compared with a mobile agent has been described by Villate et al. [13]. Still, the numbers of mobile agents used for the data fetching have not been investigated.

In this thesis, the factors that effect the performance of mobile agents for query retrieval from the WWW such as the number of mobile agents and the dispatching methods applied to the mobile agents to retrieve the query results have been investigated. These factors are related to the path determination and the mobility of mobile agents in the network environment as described in the previous section.

1.3 Ranking and Categorization of Query Retrieval Results

The ranking and categorization of query retrieval results retrieved from the search engines by the mobile agents are two important aspects that affect the users satisfaction when they are searching for information. They are described as follows:

1.3.1 Ranking of query retrieval results

Traditional search engines were never intended to deal with a vast, distributed, heterogeneous collection of documents such as the WWW. The almost complete absence of editorial control over the web documents poses special problems such as coverage, currentness, spamming, dead links, and the manipulating of rankings for commercial advantage [19, 22, 23, 24]. Current search engines in the WWW employ a number of variations of query ranking algorithms compared with the basic term frequency approach. For example, a Google's ranking algorithm [20, 21] that has been used to find the search results relies on their crawlers. These crawlers will capture an amount of information about web pages such as the inverted term index, the font types used for a particular information that are recorded in the web pages, and a distinction is drawn between "plain" and "fancy" hits. Fancy hits involve a match between a query term and a part of a URL, page title, anchor text, and meta-tag. Plain hits involve all other matches against the text of a document.

In this thesis, we propose a new ranking algorithm which is based on the number of broken links and the duplicated links that occur in the query results retrieved by the mobile agent from the WWW. Also, the search results from multiple search engines such as Yahoo, AltaVista, and Google are stored into a single database. Then the collection of search results will be processed by removing the duplicated and broken links that exist in each of the web pages stored in the database, and re-rank each of the query results before they are presented to the users.

1.3.2 Categorization of query retrieval results

While the Internet and e-mail are becoming part of many people's daily routine, many Internet users have been already familiar with the Yahoo [49] directory for searching a particular information, and a Microsoft Outlook's directory which highlights a junk of messages. These are two examples of text classification. The classification of web pages in the Yahoo directory was performed 'manually' where these pages have been assigned to one or more categories by human editors. On the other hand, the users of the Microsoft Outlook can write simple rules to sort the incoming e-mails into folders, or use predefined rules to delete junk e-mails. This is an example of automated text classification. Currently, there are many approaches that have been introduced to categorize the text documents automatically such as a neural network [42, 43, 44], a support vector machine (SVM) [61], a Bayesian classifier [60], etc. In this thesis, the classification of news web pages retrieved by the mobile agents using neural networks based on a background knowledge has been proposed and evaluated.

1.4 Goal of the Thesis

The goal of the thesis is based on the application of mobile agent technology in supporting the query retrieval process from the WWW. The scopes of the research are as follows: First, the effectiveness of mobile agent for the query retrieval using the off-line and on-line approaches is investigated. Second, the ranking of query retrieval results that have been retrieved by the mobile is investigated. Third, the performances of mobile agents for the query retrieval using a hierarchical approach compared with other approaches are investigated. Fourth, the classification of news web pages retrieved by the mobile agents using neural networks based on a background knowledge is evaluated. Finally, the improvement of classification accuracy on the classes that are represented with a small number of documents will be investigated. The outline of the studies is shown in Fig. 1.3.

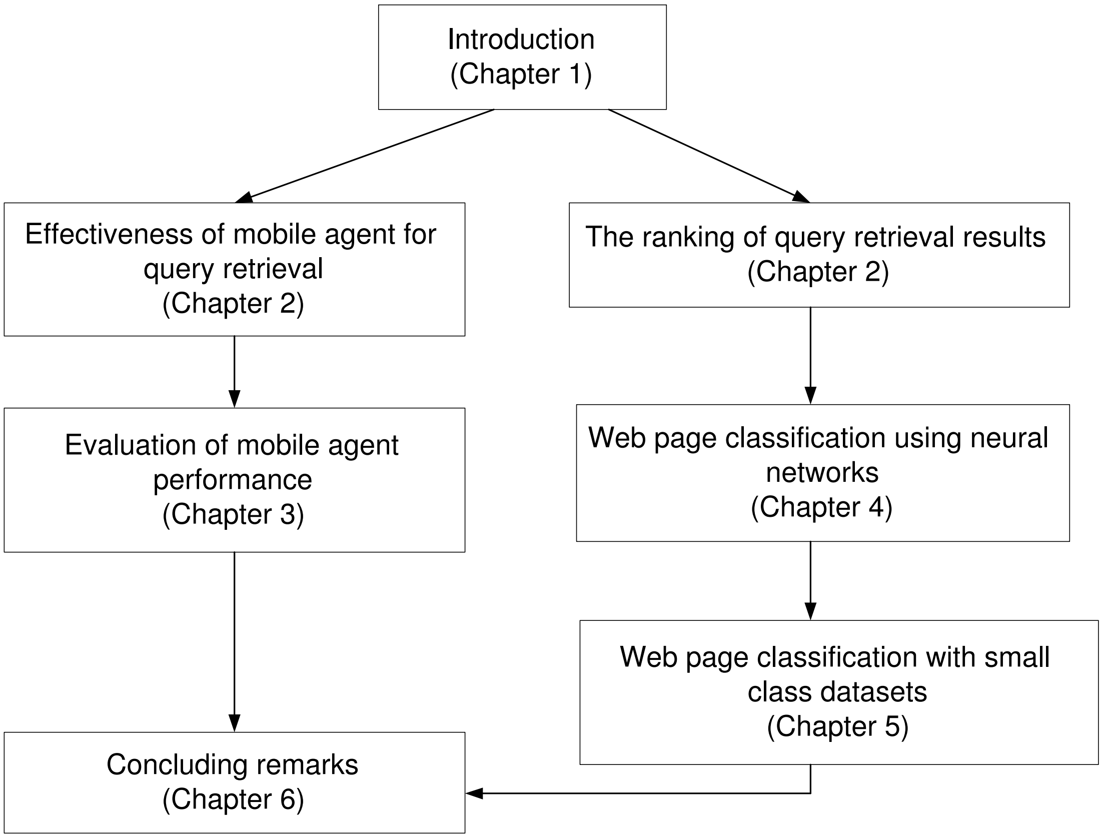

Figure 1.3: The outline of the thesis.

1.5 Organization of the Thesis

The organizations of the thesis are as follows: In Chapter 2, the information searching and retrieving are two important aspects that most of the Internet users do when they are using the Internet. Unfortunately, many available tools that are being used to search a particular information from the Internet such as a search engine could not give the satisfying results. This is due to the problem such as broken and duplicated links. Also the ranking algorithms that are applied by the search engines are different from one to another. This chapter discusses the limitations of current software tools to search in the Internet and proposes a new approach by employing a mobile agent to retrieve the query results in the local search servers. The on-line and off-line query retrievals using mobile agents are evaluated. From experiments, the off-line query retrieval approach has reduced the time spent by a user to retrieve the query results from the Internet compared with the on-line query retrieval approach. The ranking of query results retrieved by the mobile agents are evaluated using the Number of Relevant Ordering Score (*NROS*) method. The effectiveness of the *NROS* has been evaluated using standard information retrieval measures that are *Recall* and *Precision*.

In Chapter 3, the factors that affect the performance of mobile agents in retrieving information from the Internet are the number of agents and the total of routing time taken by the participated agents to complete the assigned tasks have been discussed. Fewer numbers of mobile agents used to execute the tasks will cause lower network traffic and consume less bandwidth, and the total time taken to retrieve the query results has to be kept minimal. In this chapter, we propose a method to improve the time taken to retrieve a query results by the mobile agents in the off-line Mobile Agent Search System (MaSS) by using an extended hierarchical query retrieval (EHQR) approach. The EHQR approach is based on the combination of a parallel and a hierarchical dispatching of mobile agents for query retrieval. Furthermore, in this approach the number of mobile agents has been kept minimal in order to reduce the network loads when retrieving the query results from the local databases. The performance of the EHQR approach has been compared with other query retrieval approaches. Experimental results show that our scheme reduces the total time taken to retrieve the query results compared to other approaches.

In Chapter 4, the news web pages classification using neural networks is discussed. An automatic categorization is the only viable method to deal with the scaling problem of the WWW. In this chapter, we propose a news Web Page Classification Method (WPCM). The WPCM uses a neural network with inputs obtained by both the principal components and class profile-based features (CPBF). Each news web page is represented by the term-weighting scheme. As the number of unique words in the collection set is big, the principal component analysis (PCA) has been used to select the most relevant features for the classification. Then the final output of the PCA is combined with the feature vectors from the class-profile which contains the most regular words in each class before feeding them to the neural networks. We have manually selected the most regular words that exist in each class and weighted them using an entropy weighting scheme. The fixed number of regular words from each class will be used as feature vectors together with the reduced principal components from the PCA. These feature vectors are then used as the input to the neural networks for classification. The WPCM has been compared with the other methods as a benchmark test for the classification accuracy. The experimental evaluation demonstrates that the proposed method provides an acceptable classification accuracy with the sports news datasets.

In Chapter 5, the limitation of the principal component analysis-neural networks (PCA-NN method) in supervised data where the characteristic variables that describe smaller classes tend to be lost as a result of the dimensionality reduction can be addressed by using the WPCM has been discussed. The classification accuracy on the small classes can be improved although they have been reduced into a small number of principal components. The WPCM has been compared with the other classification approaches as the benchmark tests for the classification accuracy of the classes that are represented with a small number of documents. The *F1* measure which is a kind of average of *Precision* and *Recall* has been used to evaluate the classification accuracy of the datasets. From the experiment, we have found that the *F1* measure on the sports news datasets using the WPCM provides acceptable classification accuracy of the classes that are represented with a small number of documents compared with other approaches.

In Chapter 6, the conclusions of the thesis are described.

Bibliography

[1] A. Selamat, S. Omatu, H. Yanagimoto, T. Fujinaka, M. Yoshioka, "Effectiveness of Mobile Agent for Query Retrieval," *IEEJ Transactions on Electronics, Information and Systems Society,* Vol. 122-C, No. 8, pp. 1367-1373, 2002.

[2] A. Selamat, S. Omatu, H. Yanagimoto, T. Fujinaka, M. Yoshioka, "Web News Categorization Using Neural Networks," *IEEJ Transactions on Electronics, Information and Systems Society,* Vol. 123-C, No. 5, pp. 1020-1026, 2003.

[3] A. Selamat, S. Omatu, "Feature Selection and Categorization of Web Pages Using Neural Networks," *Int'l. Journal of Information Sciences,* Elsevier Science Inc., *(in press).*

[4] A. Selamat, S. Omatu, H. Yanagimoto, T. Fujinaka, M. Yoshioka, "Extended Hierarchical Mobile Agent for Query Retrieval," *IEEJ Transactions on Electronics, Information and Systems Society, (under submission).*

[5] A. Selamat, S. Omatu, "A Neural Network Approach to Web Page Categorization Using Background Knowledge," *IEEE Transactions on Knowledge and Data Engineering, (under submission).*

[6] A. Selamat, S. Omatu, H. Yanagimoto, "Web Page Classification Using Neural Networks Based on Augmented PCA," *Int'l Symposium on Artificial Life and Robotics,* AROB'2003, Oita, Japan, 2003.

[7] A. Selamat, S. Omatu, "Web Page Classification Using Neural Networks," *Int'l Conferences on Neural Networks,* IJCNN'2003, Portland, USA, 2003.

[8] A. Selamat, S. Omatu, H. Yanagimoto, "Mobile Agent Search System for Information Retrieval from the Internet: MaSS," *Int'l Conference on Knowledge-Based Intelligent Information Engineering,* KES'2001, IOS Press, Netherland, pp. 914-918, 2001.

[9] A. Selamat, S. Omatu,"Route Selection by Mobile Agent for Query Retrieval," *IEEE Int'l Symposium on Computational Intelligence in Robotics and Automation,* CIRA'2003, Kobe, Japan, 2003.

[10] D. Gavalas, D. Greenwood, M. Ghanbari, M. O'Mahony, "Hierarchical Network Management: A Scalable and Dynamic Mobile Agent Based Approach," *Computer Networks,* Elsevier Science, Vol. 38, pp. 693-711, 2002.

[11] L.M. Silva, G. Soares, P. Martins, V. Batista, L. Santos, "Comparing the Performance of Mobile Agent Systems: A Study of Benchmarking," *Computer Communications,* Elsevier Science, Vol. 23, pp. 769-778, 2000.

[12] S. Alouf, F. Huet, P. Nain, "Forwarders vs. Centralized Server: An Evaluation of Two Approaches for Locating Mobile Agents," *An International Journal of Performance Evaluation,* Elsevier Science, Vol. 49, pp. 299-319, 2002.

[13] Y. Villate, A. Illarramendi, and E. Pitoura, "Keep Your Data Safe and Available While Roaming," *Journal of Mobile Networks and Applications,* Kuwler Academic Publishers, Vol. 7, pp. 315-328, 2002.

[14] S. Papastavrou, G. Samaras, E. Pitoura, "Mobile Agents for World Wide Web Distributed Database Access," *IEEE Trans. on Knowledge and Data Engineering,* Vol. 12, No. 5, 2000.

[15] A. Loitta, G. Knight, G. Pavlou, "On Efficiency and Scalability of Decentralized Monitoring Using Mobile Agents," *Proceedings of DSOM '99, Zurich, Switzerland, October 11-13, 1999,* LNCS 1700, Springer, 1999.

[16] C. Bohoris, A. Loitta , G. Pavlou, "Software Agent Constrained Mobility for Network Performance Monitoring," *Proceedings of the 6th IFIP Conference on Intelligence in Networks (SmartNet 2000), Vienna, Austria,* ed. H.R. van As, Kluwer, pp. 367-387, 2000.

[17] P. Dasgupta, N. Narasimhan, L.E. Moser, P.M. Melliar-Smith, "MAgNET: Mobile Agents for Networked Electronic Trading," *IEEE Transactions on Knowledge and Data Engineering,* Vol. 11, No. 4, 1999.

[18] J. Sum, H. Shen, G. Young, J. Wu, C.S. Leung, "Analysis on Extended Ant Routing Algorithms for Network Routing and Management," *Journal of Supercomputing,* Vol. 24, No. 3, pp. 327-340, 2003.

[19] P. Jackson, I. Moulinier, *Natural Language Processing for Online Applications, Text Retrieval, Extraction and Categorization,* John Benjamins Publishing Company, Amsterdam, The Netherlands, 2002.

[20] G. Meghabghab, "Google's Web Page Ranking Applied to Different Topological Web Graph Structures," *Journal of the American Society for Information Science and Technology,* Vol. 52, No. 9, pp. 736-747, 2001.

[21] A. Arasu, J. Cho, H. Gracia-Molina, A. Paepcke, and S. Raghavan, "Searching the Web," *ACM Transactions on Internet Technology (TOIT),* Vol.1, No. 1, pp. 2-43, 2001.

[22] Q. Zhanzi, M. Hemmje, E.J. Neuhold, "Using Link Types in Web Page Ranking and Filtering," *Proceedings of the Second International Conference on Web Information Systems Engineering,* 2001, Vol. 1, pp. 311-320, 2001.

[23] R. Baeza-Yates, C. Castillo, "Relating Web Characteristics With Link Based Web Page Ranking," *Proc. on String Processing and Information Retrieval, SPIRE '2001,* pp. 21-32, 2001.

[24] W. Chang, G. Sheikholeslami, J. Wang, A. Zhang, "Data Resource Selection in Distributed Visual Information Systems," *IEEE Transactions on Knowledge and Data Engineering,* Vol. 10, No. 6, pp. 926-946, 1998.

[25] R. Gray, D. Kotz, S. Nog, D. Rus and G. Cybenko, "Mobile Agents: The Next Generation in Distributed Computing," *Proc. of the Second Aizu Int'l. Symposium on Parallel Algorithms and Architectures Synthesis,* Japan, pages 8-24, 1997.

[26] H. S. Nwana and D. T. Ndumu, "A Perspective on Software Agents Research," *Knowledge Engineering Review,* Vol. 14, No. 2, pp. 125-142, 1999.

[27] D. Johansen, N. P. Sudmann and R. V. Renesse, "Performance Issues in TACOMA," *3rd. Workshop on Mobile Object Systems, 11th Europeean Conference on Object-Oriented Programming, Jyvaskyla,* Finland, 1997.

[28] O. E. Zamir, "Clustering Web Documents: A Phrase-Method for grouping Search engine Results," *PhD Thesis*, University of Washington, 1999.

[29] MetaCrawler available at http://www.metacrawler.com, 1999.

[30] M. E. Muller, "An Intelligent Multi-Agent Architecture for Information Retrieval from the Internet," *Technical Report,* Institute for Semantic Information Processing, University of Osnabruck, Germany, 1999.

[31] C. Wai Ho and A. Goh, "Jamaica: A World Wide Web Profiler," *Internet Research: Electronic Networking Applications and Policy,* Vol. 9, No. 2, pp. 129-139, 1999.

[32] TACOMA availabe at http://www.cs.uit.no, 1997.

[33] P. Maes, Y. Lashkari, and M. Metral, "Collaborative Interface Agents," *In: Readings in Agents, edited by Michael N. Huhns & Munidar P. Singh,* Morgan Kaufmann Publishers, Inc., 1997.

[34] Compaq Web Language, available at http://www.research.compaq.com/SRC/WebL/, 1997.

[35] H.M. Lee, C.M. Chen, C.W. Hwang, "A Neural Network Document Classifier with Linguistic Feature Selection," *Proc. of 13th Int'l Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, New Orleans, Louisiana, USA, IEA/AIE 2000, pp. 555-560, 2000.

[36] Z.Q. Liu and Y. Zhang, "A Competitive Neural Network Approach to Web-Page Categorization," *Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* Vol. 9, No. 6, pp. 731-741, 2001.

[37] D. A. Karras and B. G. Mertzios, "A Robust Meaning Extraction Methodology Using Supervised Neural Networks," *Proc. of Australian Joint Conf. on Artificial Intelligence,* pp. 498-510, 2002.

[38] A. Rauber, D. Merkl, and M. Dittenbach, "The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data," *IEEE Trans. on Neural Networks,* Vol. 13, No. 6, pp. 1331-1341, 2002.

[39] J.H. Chang, J.W. Lee, Y. Kim, and B.T. Zhang, "Topic Extraction from Text Documents Using Multiple-Cause Networks," *Lecture Notes in Artificial Intelligence vol. 2417 (Proceedings of 7th Pacific Rim International Conference on Artificial Intelligence 2002),* pp. 434-443, 2002.

[40] G. L. Gentili, M. Marinilli, A. Micarelli, F. Sciarrone, "Text Categorization In An Intelligent Agent for Filtering Information On The Web," *Int'l Journal of Pattern Recognition and Artificial Intelligence,* Vol. 15, No. 3, pp. 527-549, 2001.

[41] T. Kohonen, S. Kaski, K. Lagus, J. Salojävi, J. Honkela, V. Paatero, and A. Saarela, "Self Organization Of A Massive Text Document Collection," *in: Oja, E. and Kaski, S., editors, Kohonen Maps, Elsevier, Amsterdam,* pp. 171-182, 1999.

[42] S. Wermeter, "Neural Network Agents For Learning Semantic Text Classification," *Information Retrieval,* Vol. 3, No. 2, pp. 87-103, 2000.

[43] E.M. Ruiz, P. Srinivasan: "Hierarchical Text Categorization Using Neural Networks," *Information Retrieval,* Vol. 5, No. 1, pp. 87-118, 2002.

[44] S.Y. Lam and D.L. Lee, "Feature Reduction for Neural Network Based on Text Categorization," *Proc. of the 6th International Conference on Database Systems for Advanced Applications,* Hsinchu, Taiwan, 1999.

[45] F. Sebastini, "Machine Learning in Automated Text Categorization," *ACM Comput. Surveys,* Vol. 34, No. 1, pp. 1-47, 2002.

[46] G. Karypis and E.H. Sam, "Fast Dimensionality Reduction Algorithm With Applications to Document Retrieval & Categorization," *Proc. of the 2000 ACM CIKM Int'l Conf. on Information and Knowledge Management,* McLean, VA, USA, pp. 12-19, 2000.

[47] S. Zelikovitz and H. Hirsh, "Improving Short Text Classification Using Unlabeled Background Knowledge," *Proc. of the Seventeenth Int'l Conf. on Machine Learning,* Stanford University, USA, 2000.

[48] S.T. Dumais, "Improving The Retrieval of Information from External Sources," *Behavior Research Methods, Instruments and Computers,* Vol. 23, No. 2, pp. 229-236, 1991.

[49] The Yahoo Web Pages, available from http://www.yahoo.com, 2001.

[50] The CNN Web News Pages, available from http://www.cnn.com, 2001.

[51] The Japan Times Web Pages, available from http://www.japantimes.co.jp, 2001.

[52] H. Mase and H. Tsuji, "Experiments on Automatic Web Page Categorization for Information Retrieval System," *Journal of Info. Processing Society of Japan,* Vol. 42, No. 2, pp. 334-347, 2001.

[53] R. A. Calvo, M. Partridge, and M. Jabri, "A Comparative Study of Principal Components Analysis Techniques," *Proc. Ninth Australian Conf. on Neural Networks,* Brisbane, QLD., pp. 276-281, 1998.

[54] R. A. Johnson and W. D. Wichern, *Applied Multivariate Statical Analysis,* Fifth-edition, Prentice Hall, USA, 2002.

[55] A. Selamat, S. Omatu, H. Yanagimoto, "Web News Classification Using Neural Networks Based on PCA," *Proc. of SICE 2002*, Osaka International Convention Center, Osaka, Japan, pp. 2388-2393, 2002.

[56] S. Jones, Karen, and P. Willet, *Readings in Information Retrieval,* San Francisco, USA, 1997.

[57] Salton & McGill, *Introduction to Modern Information Retrieval,* New York, McGraw-Hill, USA, 1983.

[58] J.P. Yang, V. Honavar, and , L. Miller, "Mobile Intelligent Agents for Document Classification and Retrieval: A Machine Learning Approach," *Proc. of the European Symposium on Cybernetics and Systems Research,* Vienna, Austria, pp. 707-712, 1998.

[59] R.R. Korfhage, *Information Storage and Retrieval,* John Wiley and Sons Inc., USA, 1997.

[60] T. Joachims, "Probabilistic Analysis of The Rocchio Algorithm with TFIDF for Text Categorization," *Proc. of Int'l Conf. on Machine Learning,* Nashville, TN, USA, pp. 143-151, 1997.

[61] T. Joachims, "A Statistical Learning Model of Text Classification with Support Vector Machines," *Proc. of the Conf. on Research and Development in Information Retrieval (SIGIR),* ACM, 2001.

[62] A.K. McCallum, "Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering," available from http://www.cs.cmu.edu/~mccallum/bow, 1996.

[63] D.D. Lewis, "Evaluating and Optimizing Autonomous Text Classification Systems," *E. A. Fox, P. Ingwersen, and R. Fidel (Ed.), SIGIR '95: Proc. of the 18th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, New York, pp. 246-254, 1995.

[64] A.S. Weigend, E. D. Wiener, and J. O. Pedersen, "Exploiting Hierarchy in Text Categorization," *Information Retrieval,* Vol. 1, No. 3, pp. 193-216, 1999.

[65] R. A. Calvo and H.A. Ceccatto, "Intelligent Document Classification," *Intelligent Data Analysis,* Vol. 4, No. 5, 2000.

[66] D. Lange and M. Oshima, “Programming and Developing Java Mobile Agents With Aglet Mobile Agent,” Addison Wesley, 1998.

[67] J.W. Baek, J.H. Yeo, G. Kim, H. Y. Yeom, “Cost Effective Mobile Agent Planning for Distributed Information Retrieval”, *International Conference on Distributed Computing Systems (ICDCS),* pp. 65-72, April 2001.

[68] A. Selamat and S. Omatu, “Extended Hierarchical Mobile Agent for Query Retrieval”, *Proc. of Kansai-section Joint Convention of Institutes of Electrical Engineering,* Kinki University, Osaka, Japan, Nov. 9-10, 2002.

[69] T. Kawamura, S. Joseph, A. Ohsuga, S. Honiden, “Designing Multi-Agent Systems based on Pairwise Agent Interactions”, *IEICE Trans. Inf. & Syst.,* Vol. E 84-D, No. 8, pp. 968-980, 2001.

[70] L. Ismail, D. Hagimont, “A Performance Evaluation of the Mobile Agent Paradigm”, *Fourteenth Conference on Object-Oriented Programming, Systems, Languages and Applications (OOPSLA),* Denver, USA, pp. 306-313, November 1999.

[71] R. Jha and S. Iyer, “Performance Evaluation of Mobile Agents for E-commerce Applications”, *Lecture Notes in Computer Science series,* Springer-Verlag, Vol. 2228, pp. 331-340, 2001.

[72] P. Cogan, J. Gomoluch, and M. Schroeder, “Quantitative and Qualitative Comparison Of Distributed Information Processing Using Mobile Agents Realised in RMI and Voyager”, *International Journal of Software Engineering and Knowledge Engineering (IJSEKE),* Vol. 11, No. 5, pp. 583-606, 2001.

[73] M. G. Rubinstein, O. C. M. B. Duarte, and G. Pujolle, “Evaluating the Performance of a Network Management Application Based on Mobile Agents”, *IFIP Networking 2002, Pisa, Italy,* published in LNCS 2345, Enrico Gregori, Marco Conti, Andrew Campbell, Guy Omidyar, and Moshe Zukerman (Eds.), Springer-Verlag, pp. 515-526, 2002.

[74] W. Theilmann, K. Rothermel,“Optimizing the Dissemination of Mobile Agents for Distributed Information Filtering”, *IEEE Concurrency,* Vol. 8, No. 2, pp. 53-61, 2000.

[75] M. Dikaiakos, M. Kyriakou, G. Samaras, “Performance Evaluation of Mobile-agent Middleware: A Hierarchical Approach,” *In Proceedings of the 5th IEEE International Conference on Mobile Agents,* J.P. Picco (ed.), Lecture Notes in Computer Science series, Springer-Verlag, Vol. 2240, pp. 244-259, 2001.