PEER-TO-PEER DATA SEARCHING USING RELEVANT PEER SELECTION AND LOCAL AVAILABILITY ESTIMATION

ISKANDAR BIN ISHAK

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

AUGUST 2011

To my beloved:

Wife, *Norkamila binti Roffii*

Lovely Daughter, *Fatimah Zahra*

Father, *Ishak bin Mohd Saris*

Mother, *Sapiah binti Haji Tahir*

Brother, *Osmera bin Ishak*

Sister, *Ispha Anna binti Ishak*

and to all my precious friends

ACKNOWLEDGEMENTS

Alhamdulillah, praise to Allah the Almighty for giving me the strength and guidance to complete this thesis.

First of all, I would to express my appreciation to my supervisor, Prof. Dr. Naomie binti Salim for her precious support, undivided attention and guidance that help me to get through the research. Many thanks to my research mates, Dr. Ammar, Dr. Jehan Zeb, Siriporn, Ladda, and Mohamed Salim who have helped me during my study in UTM and not to forget my ex-schoolmate Dr. Abbas Mohd Zaini, who helped me during my early days in UTM and Johor Bahru.

I would also like to thank to my wonderful parents, Ishak Mohd Saris and Sapiah Haji Tahir, who always supporting me throughout the course. Their sacrifices throughout my study were just priceless. Not to forget to my brother Osmera and my sister Ispha Anna who always there for me when I need them. Special thanks to my in laws in Pahang who has so understood of what I am doing for the past 5 years until I complete my study.

A special thanks to my lovely wife, Norkamila binti Roffii who was always there for me, encourage me when I was down during my study. All the hardship that we have been through have made us even closer and made the life become more meaningful. Last but not least, special appreciation to my lovely daughter, Fatimah Zahra who always cheered me when I am dull. All of you, again...thank you and may Allah blessed us all.

ABSTRACT

Due to the dynamic, robust and resource-intense environment of the peer-to-peer systems, traditional approach of searching has become obsolete as systems becomes bigger and wider. Lack of knowledge, administrative and peers' churning are among the obstacles for researchers to develop an efficient searching mechanism. Proper peer selection is one way of deriving effective searching mechanism. Peers that are considered of relevance, should be chosen to receive the search query in order to increase the search efficiency and obtain better retrieval rates. In this thesis, two approaches have been developed to produce efficient searching mechanism in peer-to-peer system. The first contribution is the flood-based relevant search method where peer relevance determination is based on the query feedback data. Peer relevance is based on whether the peer has answered previous queries and how much the queries answered have content similarity with the ongoing query. A metric graph has been proposed to calculate the relevance of peers based on the values of their query hits and query similarity with the current ongoing query terms. The method is then improved by implementing a dynamic threshold value for relevant peer selection based on the nearest-neighbor concept. The experimental results showed that, the proposed method managed to record high efficiency in search results by having high query hits with low number of messages used. The research also proposed a trust based search approach for unstructured peer-to-peer system. It is developed to reduce the effect of nodes churning on the system. In the trust-based method, three components, namely peer relevance, peer availability and query hops are used. The model uses the boolean concept in decision making for selecting trusted peers. The results obtained showed that the effect of nodes churning in query hits have been reduced from four percent to two percent.

ABSTRAK

Sifat yang dinamik, lasak dan keperluan sumber yang tinggi dalam persekitaran sistem rakan-ke-rakan menyebabkan pendekatan pencarian secara tradisional tidak lagi sesuai apabila sistem-sistem tersebut semakin berkembang dan meluas. Kekurangan pengetahuan, pentadbiran dan penggodakan rakan adalah di antara kekangan bagi penyelidik dalam menghasilkan mekanisme pencarian yang efisien. Pemilihan rakan yang tepat adalah salah satu cara dalam menghasilan mekanisme pencarian yang efisien. Hanya rakan-rakan yang relevan sahaja harus dipilih untuk menerima kata carian bagi meningkatkan keberkesanan pencarian dan memperoleh kadar dapatan semula yang lebih baik. Dalam tesis ini, dua kaedah telah disumbangkan dalam menghasilkan mekanisme pencarian yang berkesan dalam sistem rakan-ke-rakan. Sumbangan pertama ialah kaedah carian relevan yang berasaskan pembanjiran di mana penentuan kerelevanan rakan mengikut data maklumbalas carian. Rakan yang relevan adalah berdasarkan sama ada sesebuah rakan itu pernah menjawab pertanyaan-pertanyaan yang lalu dan kesamaan antara kandungan rakan dan kata carian yang sedang diproses. Sebuah graf matriks telah dicadangkan untuk mengira nilai kerelevanan sesebuah rakan berdasarkan jumlah padanan carian dan kesaksamaan kata carian dengan carian yang sedang diproses. Kaedah tersebut kemudian diperbaiki dengan menggunakan nilai ambang yang dinamik untuk pemilihan rakan yang relevan berdasarkan konsep jiran-terdekat. Keputusan eksperimen menunjukkan bahawa kaedah yang dicadangkan telah merekodkan keberkesanan yang tinggi dalam keputusan carian melalui penghasilan padanan carian menggunakan jumlah mesej yang rendah. Penyelidikan ini juga mencadangkan carian berdasarkan kepercayaan untuk sistem rakan-ke-rakan tanpa struktur. Dalam kaedah ini, tiga komponen telah digunakan iaitu kerelevanan rakan, keberadaan rakan dan lompatan carian. Model ini menggunakan konsep Boolean dalam penentuan keputusan untuk memilih rakan yang paling dipercayai. Keputusan eksperimen telah menunjukkan bahawa kesan dari penggodakan rakan ke atas nilai padanan carian telah berkurang dari empat peratus ke dua peratus.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

APS	-	Adaptive Probabilistic Search
BFS	-	Breadth-First Search
CAN	-	Content Addressable Network
DHT	-	Distributed Hash Tables
ID	-	Identity Document
IR	-	Inverse Ranking
ISM	-	Intelligent Search Mechanism
LRU	-	Least-Recently Used
MQH	-	Most-Query Hits
P2P	-	Peer-to-peer
RAM	-	Random Access Memory
RNN	-	Relevance Nearest Neighbour
RP	-	Reduction Percentage
TTL	-	Time-to-Live
WWW	-	World Wide Web

CHAPTER 1

INTRODUCTION

1.1 Motivation

Peer-to-peer data sharing systems are distributed systems that provide users with the facility to share data over the Internet. The system contains interconnected peers, which are decentralized and non-autonomous. The peers are distributed and scattered across the Internet. Data sharing activity which includes data searching is the core feature in the system to support better data discovery. In recent years, peer-to-peer networks have become one of the medium for Internet users to share resources. In a peer-to-peer network, a peer interchangeably acts as client and a server of the system. It holds the key to realizing the potential of enabling Internet users to share services between each other. Peer-to-peer offers an attractive solution for easy resource sharing through its distributed characteristics, easy-to-use application that can easily be obtained and used by the Internet users.

However, due to the lack of coordination, autonomy and knowledge, searching is not an easy task. Furthermore, peer availability does not guarantee that a peer might stay in the system for a period of time. Therefore it is an utmost important to choose relevant peers to route query message in order to reduce the number of messages used and answering time for better searching in unstructured peer-to-peer network without the loss of the common unstructured peer-to-peer identity and characteristics.

This chapter will discuss the research background, problem statements and the research questions related to the main research. This chapter also includes the research objectives, research scope and significances of study and expected contributions.

1.2 Problem Statement

In a peer-to-peer environment, searching is the most important feature that a peer-to-peer system need to have. Based on literatures, most of the peer-to-peer system are in unstructured form (Li and Liao 2005), (Stutzbach *et al.* 2005), (Schmid and Wattenhofer 2007). Having no fixed structure, unstructured peer-to-peer poses a great challenge in developing good and efficient search methods. Lack of knowledge, administration, and system dynamics are threats that requires the need for efficient search mechanism in order to perform well in such distributed system. Among the problems exist in searching over the unstructured peer-to-peer are:

i. **Peer Selection Problem in Unstructured Peer-to-peer network searching**: Peer selection problems have been the main problem in peer-to-peer system.

Many approaches have been proposed to solve this problem that uses query feedback methods (Cohen and Shenker 2002), (Lv *et al.* 2002), (Zeinalipour-Yazti 2004), (Zeinalipour-Yazti *et al.* 2005). However, resource discovery is still a challenge in peer-to-peer network, especially in an unstructured peer-to-peer system that is very dynamic and robust. A fair weight should be given to both query hits and content similarity in which a peer's relevance value can be determined and thus assist in peer selection. On the other hand, the concept of trust has been considered as a method in decision making in distributed-computing. It is anticipated that applying the concept of trust-based method can provide better peer selection thus achieving better searching performance.

ii. **Network Congestion Problem**: Excessive usage of message in the network and wasted or unsent messages to the desired peer is another major problem in peer-to-peer searching (Cohen and Shenker 2002), (Crespo and Molina 2002), (Lv *et al.* 2002), (Yang and Garcia-Molina 2002), (Zeinalipour-Yazti 2004), (Zeinalipour-Yazti *et al.* 2005). This problem is the source of network congestion problem since wasted messages filled the network and thus reduces the performance of searching over the system. This problem can be solved by reducing the message usage thus avoiding network congestion.

iii. **Churn Rate Problem**: Churn rate or the rate of peers leaving the system causes a negative impact towards the system's searching performance (Marti and Molina 2006). Peer that can be a candidate to answer a query can suddenly be offline, thus sending query message to this peer will in result zero feedback, and make the searching to be less-efficient. This problem can be solved through estimation of peers' availability. Since the research is done on unstructured peer-to-peer network, knowledge about other peers is solely based on the query feedback data. Therefore, a good estimation of peers availability and unavailability can reduce the effect of churning.

The questions of this research that needed to be answered are:

1. How to determine the relevance of nodes in unstructured peer-to-peer networks, using query feedback data?

2. How to select relevant nodes that have high probability of responding to the queries sent.
3. How to estimate other peers' availability so that query messages can be sent to them?

1.3 Research Objectives

The ultimate goal of this work is to produce a searching mechanism in unstructured peer-to-peer networks based on the similarity of the peers' content and the popularity of the peers. The objectives of this research are:

i. To proposed a new method for selecting relevant peers for routing search query in unstructured peer-to-peer network. The selection method bases on a very simple and simple method of probability and vector space model. The respective specific objective is:

   a) To develop a query feedback based scheme for selecting relevant peers when routing a search query across unstructured peer-to-peer network and subsequently, optimizing the usage of the available resources (query feedback data) for efficient searching in the network.

ii. To develop an efficient search mechanism in unstructured peer-to-peer networks that can reduce the number of messages used for searching across the network. Subsequently avoiding congestion in the network and maintaining low search cost.

iii. To develop localized nodes availability estimation method based on query feedback. Node availability will tell whether the nodes are available for sending queries due to the random join and departure of nodes in the peer-to-peer system. The estimation method will be incorporated to the proposed searching scheme and thus can reduce the negative effects of peers' unavailability on query retrievals.

1.4 Research Scope

Peer-to-peer system is a large research area that covers computer science subjects such as computer network, query routing, node selection and information retrieval. The main focus of this research is on the searching technique across unstructured peer-to-peer network. The physical parameter of the network is not considered in this study. In order to have a realistic peer-to-peer simulations, each peer will have its own delay time representing both its joining and lifetime on the network and the delay that shows the peers leaving the network.

It is based on the simulation of data with a degree of replication, which means, each files can be found on more than one peer. In the simulation of the research, the search method is mainly based on flooding technique that will not involve structural changes on the physical network as well as overlay network.

1.5 Research Significance

This research develops a new method in determining relevant peer selection in searching across unstructured peer-to-peer network based on past query feedback. A new algorithm of flooding across the network has also been developed that ensures effective searching by reducing the number of messages used and increases query hits.

The performance of the peer-to-peer searching can be improved in terms of:

i. Efficient searching method – submitting the query to the peers that are similar and related to the query and hence reducing the number of messages and number of hops through, which will help to increase the efficiency of the peer-to-peer system.

ii. Better retrieval results– submitting the query to the peers that are related to the query content will also increase the probability of finding good and superior result of the query.

iii. Reducing the effect of peer churning – estimating the peers' availability before the query message is sent to the peers can reduce the effects of peer churning.

1.6 Assumption and Limitations

This research does not attempt to develop a new architecture of peer-to-peer. In fact, it aims to develop a new searching technique for peer-to-peer environment on unstructured peer-to-peer networks. Apart from that, only a random, scale-free network is considered in

this thesis. Random model is chosen because it is a simple type of network to be designed and built. The proposed work does not focus on the lower level of the network, but only focuses on the application level of a peer-to-peer system. In which it is similar with what had been done in other peer-to-peer applications such as Gnutella and Napster (both will be described later in Chapter 2) in terms of peer selection for the purpose of peer-to-peer searching.

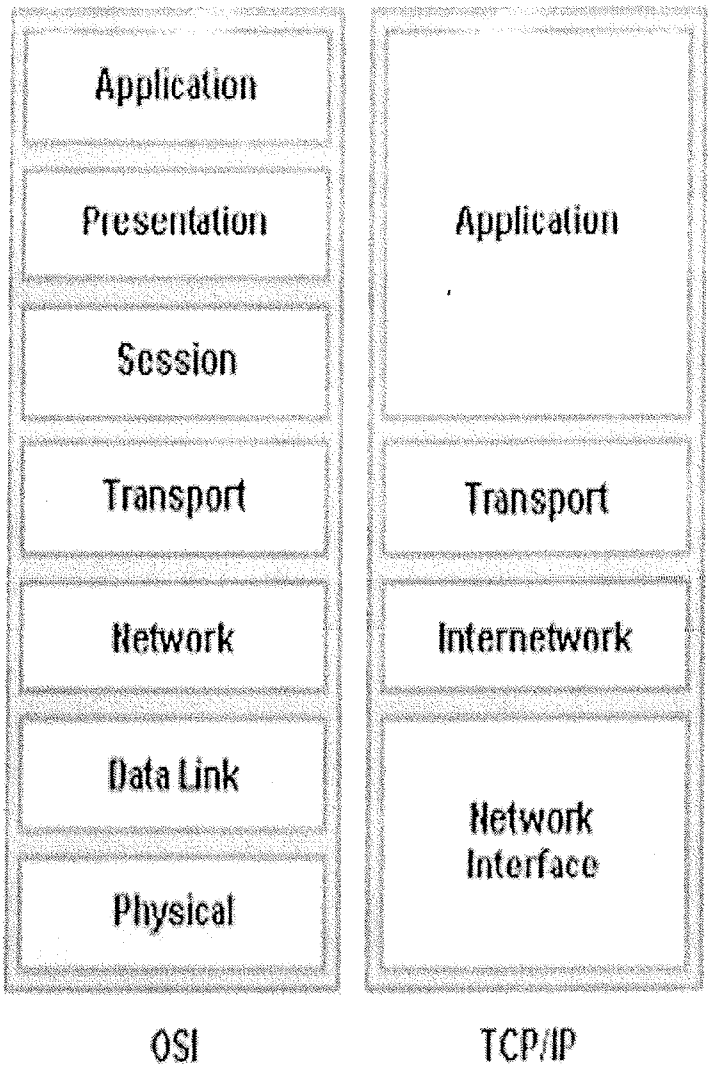

Figure 1.1: OSI and TCP/IP Layer

1.7 Organization of the thesis

The remainder of the thesis is organized as follows. Chapter 2 gives a general introduction to the background of peer-to-peer networks and Chapter 3 explained the research methodology applied in this thesis. Chapter 4 describes in details the relevance based searching method while Chapter 5 describes the churn-friendly trust based searching method. Finally Chapter 6 concludes this thesis and points out some possible future research directions.

REFERENCES

(2006) WordNet

Retrieved 19 September 2008, from http://www.wordnet.princeton.edu

(2003) NeuroGrid

Retrieved 19 September 2008, from http://sourceforge.net/projects/neurogrid

(2006a) Cygwin

Retrieved 19 September 2008, from http://www.cygwin.com

(2006b) NeuroGrid

Retrieved 19 September 2008, from http://sourceforge.net/projects/neurogrid

(2008a) Gnutella

Retrieved 19 September 2008, from http://rfc-gnutella.sourceforge.net/

(2008b) Kazaa

Retrieved 19 September 2008, from http://www.kazaa.com

(2008c) Limewire

Retrieved 19 September 2008, from http://www.limewire.com