UNIVERSITI PUTRA MALAYSIA

A SOCIAL NETWORK-BASED PEER-TO-PEER MODEL FOR RESOURCE DISCOVERY

AMIR MODARRESI

FSKTM 2009 13

A SOCIAL NETWORK-BASED PEER-TO-PEER MODEL FOR RESOURCE DISCOVERY

AMIR MODARRESI

DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALAYSIA

2009

A SOCIAL NETWORK-BASED PEER-TO-PEER MODEL FOR RESOURCE DISCOVERY

By

AMIR MODARRESI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

October 2009

To My Beloved Mother, Father and little son

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

A SOCIAL NETWORK-BASED PEER-TO-PEER MODEL FOR RESOURCE DISCOVERY

By

AMIR MODARRESI

October 2009

Chairman: **Associate Professor Ali B. Mamat, PhD**

Faculty: **Computer Science and Information Technology**

Peer-to-Peer (P2P) systems are distributed systems consisting of interconnected nodes which provide scalability, fault tolerance, decentralized coordination, self-organization, anonymity, distributed resources and services sharing, lower cost of ownership and better support for creating ad hoc networks. Data sharing, a subset of resource sharing, is one of the attractive topic in P2P systems. Because of autonomy of the nodes, decentralized coordination and volatility of network caused by the autonomy, data sharing is not an easy task in P2P system. Furthermore, there is no guarantee that a node stays in the network for a specific period of time. Hence, the answers to a particular query may be retrieved from different nodes every time. Moreover, the lack of centralized coordinators makes this process harder. These problems in P2P systems lead to a well known problem which is called resource discovery.

Resource discovery are usually fulfilled by two general solutions, namely peer organization and peer selection algorithms. In peer organization solution, an effective logical organization of nodes is designed for easy access to proper nodes; while in peer selection algorithms, an effective query routing process is designed and implemented during query answering process. In current peer organization solutions, lack of generally accepted organization and inaccessibility to all nodes during query answering process are common disadvantages; while in current peer selection algorithms, highly complicated algorithms like complex hash functions, high amount of network traffic and huge indices are common flaws in this category. Interest-based clustering is an example of peer organization while flooding algorithm over random model is an example of peer selection algorithm.

In this thesis, a general model for peer organization based on social network concepts and ontology for precise peer clustering is proposed. The model is a hybrid P2P with several communities organized based on a generally accepted ontology. While the relationship among communities is defined by the ontology, the internal structure of each community obeys social network concept. In this model, all peers with similar interest are gathered in one particular community by considering the proximity of nodes in each community. The advantage of the model is that a particular query with a specific interest is answered by a designated community. This means that, each query is answered by a particular portion of the network, instead of propagating to the entire network.

In order to investigate the performance of the model, a discrete event simulator has been designed and implemented. Several parameters were measured to show the performance of the model. In addition, an algorithm for control flooding, suitable for the model, was designed and tested. The results have shown less number of sent and drop messages which caused less network traffic, higher hit per refer and shorter path in answering queries compared with random and interest-based clustering overlays. Furthermore, the model has shown scalability by increasing the number of sub-communities when the number of nodes in the system increases. Moreover, the proposed routing algorithm has provided less network traffic, higher success and recall rate in comparison with flooding.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk Ijazah Doktor Falsafah

MODEL RAKAN KE RAKAN BERASASKAN RANGKAIAN SOSIAL UNTUK PENEMUAN SUMBER

Oleh

AMIR MODARRESI

October 2009

Pengerusi: **Professor Madya Ali B. Mamat, PhD**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Sistem Rakan-ke-Rakan (P2P) adalah sistem teragih yang mengandungi nod-nod yang saling berkaitan yang menyediakan skalabiliti, toleransi kesilapan, penyelarasan tidak berpusat, pengorganisasian-kendiri, anonimiti, pengagihan sumber dan perkongsian perkhidmatan, pemilikan kos rendah, dan sokongan yang lebih baik untuk penciptaan rangkaian *ad hoc*. Perkongsian data, satu subset bagi perkongsian sumber adalah salah satu topik menarik dalam sistem P2P. Oleh kerana autonomi nod-nod, penyelarasan tidak berpusat dan perubahan rangkaian yang disebabkan oleh autonomi tersebut, perkongsian data bukanlah satu tugas yang mudah dalam sistem P2P. Disamping itu, tidak terdapat jaminan bagi satu nod untuk berada dalam rangkaian bagi satu tempoh masa yang khusus. Dengan itu, jawapan untuk satu pertanyaan tertentu mungkin dicapai dari nod yang berbeza pada setiap masa. Tambahan pula, kekurangan pada penyelarasan berpusat membuatkan proses ini semakin sukar. Masalah-masalah ini dalam sistem-sistem P2P telah menjurus kepada satu masalah yang lebih dikenali sebagai penemuan sumber.

Penemuan sumber biasanya diselesaikan oleh dua penyelesaian umum yang dinamakan algoritma pengorganisasian rakan dan algoritma pemilihan rakan. Dalam penyelesaian pengorganisasian rakan, satu pengorganisasian logikal yang efektif bagi nod-nod direka bentuk untuk memudahkan akses kepada nod yang sesuai manakala dalam algoritma pemilihan rakan, satu proses penghalaan pertanyaan yang efektif direka bentuk dan dilaksanakan semasa proses menjawab pertanyaan. Dalam penyelesaian pengorganisasian rakan yang terkini, kekurangan pengorganisasian penerimaan secara umum dan ketidakboleh capaian kepada semua nod semasa proses menjawab pertanyaan adalah kelemahan biasa manakala dalam algoritma pemilihan rakan, algoritma-algoritma yang sangat kompleks seperti fungsi cincang yang kompleks, jumlah rangkaian trafik yang tinggi dan indeks yang besar adalah kecacatan biasa dalam kategori ini. Kluster berasaskan minat merupakan satu contoh bagi pengorganisasian rakan manakala algoritma pembanjiran ke atas model rawak adalah salah satu contoh bagi algoritma pemilihan rakan.

Dalam tesis ini, satu model umum untuk pengorganisasian rakan berdasarkan konsep rangkaian sosial dan ontologi untuk mengkluster rakan dengan tepat dicadangkan. Model ini adalah satu P2P hibrid dengan beberapa komuniti yang disusun berdasarkan ontologi yang diterima umum. Sementara hubungan antara komuniti-komuniti ditakrifkan oleh ontologi, struktur dalaman setiap komuniti pula adalah mengikut konsep rangkaian sosial. Dalam model ini, semua rakan yang mempunyai minat yang

sama dikumpulkan dalam satu komuniti tertentu dengan mengambil kira nod-nod berhampiran dalam setiap komuniti. Kelebihan model ini ialah satu pertanyaan tertentu beserta satu minat khusus adalah dijawab oleh satu komuniti yang telah direka bentuk. Ini bermakna, setiap pertanyaan dijawab oleh satu bahagian tertentu pada rangkaian tersebut bagi menggantikan penyebaran pada keseluruhan rangkaian.

Bagi tujuan untuk mengkaji prestasi model ini, satu simulator acara berasingan telah direka bentuk dan dilaksanakan. Beberapa parameter telah diukur untuk menunjukkan prestasi model ini. Selain itu, satu algoritma untuk mengawal pembanjiran yang sesuai untuk model ini telah direka bentuk dan diuji. Hasil telah menunjukkan bahawa kurangnya bilangan penghantaran dan penerimaan mesej yang menyebabkan kurangnya rangkaian trafik, *higher hit per refer* dan laluan yang lebih pendek dalam menjawab soalan dibandingkan dengan lapisan secara rawak dan lapisan secara kluster berasaskan minat. Disamping itu, model ini juga telah menunjukkan skalabiliti dengan pertambahan bilangan sub-komuniti apabila bilangan nod dalam sistem bertambah. Selain itu, algoritma penghalaan yang dicadangkan juga telah menyediakan rangkaian trafik yang kurang, kadar panggilan balik dan kejayaan yang lebih tinggi dibandingkan dengan pembanjiran tulen.

ACKNOWLEDGEMENTS

I thank God for all things throughout my voyage of knowledge exploration.

I would like to express my sincere gratitude to my supervisor Associate Professor Dr. Ali Mamat and also my supervisory committee members Associate Professor Dr. Hamidah Ibrahim and Dr. Norwati Mustapha for their guidance and advice throughout this work in making this a success.

My deepest appreciation to my parents for their utmost support and encouragement without which all these would not be possible.

For the others who have directly or indirectly helped me in the completion of my work, I thank you all.

APPROVAL

I certify that an Examination Committee met on 06 / 10 / 2009 to conduct the final examination of **Amir Modarresi** on his Doctor of Philosophy thesis entitled "A Social Network-Based Peer-To-Peer Model for Resource Discovery " in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee were as follows:

Associate Professor Dr. Md. Nasir Sulaiman
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Professor Dr. Mohamed Othman
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Dr. Shamala K. Subramaniam
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Professor Dr. Mahamod Ismail
Faculty of Engineering
Universiti Kebangsaan Malaysia
(External Examiner)

_____
HASANAH MOHD. GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: , 2009

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Ali B. Mamat- Ph.D
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Hamidah Ibrahim-Ph.D
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Dr. Norwati Mustapha- Ph.D
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

_____

HASANAH MOHD GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 10 December 2009

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at University Putra Malaysia or other institution.

_____

AMIR MODARRESI

Date: March 20, 2009

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Meaning
ACM	Association for Computing Machinery
AI	Artificial Intelligence
API	Application Programming Interface
APS	Adaptive Probabilistic Search
BFS	Breadth First Search
CAG	Content Aware Group
CAN	Content Addressable Network
CIDR	Classless Inter-Domain Routing
DAML	DARPA Agent Markup Language
DARPA	Defense Advanced Research Projects Agency
DHT	Distributed Hash Table
DRLP	Distributed Resource Location Protocol
GUESS	Gnutella UDP Extension for Scalable Searches
IP	Internet Protocol
LOM	Learning Object Metadata
LSI	Latent Semantic Indexing
OIL	Ontology Inference Layer
OWL	Web Ontology Language
P2P	Peer-to-Peer
RDF	Resource Description Framework

RDFS	Resource Description Framework Schema
RSS	Rich Site Summary
SeRQL	Sesame RDF Query Language
SHA	Secure Hash Algorithm
SPARQL	SPARQL Protocol And RDF Query Language
SVD	Singular Value Decomposition
TTL	Time To Live
UDP	User Datagram Protocol
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VSM	Vector Space Model
W3C	The World Wide Web Consortium
XML	Extensible Markup Language

CHAPTER 1

INTRODUCTION

1.1 Motivation

Search systems in libraries are a well-known form of search systems. Since search systems have been used for a long time, there is a rich knowledge behind them. In conventional way, searching such as those done in libraries is centralized. In such systems, there is usually a big computer or cluster of computers which answers users' queries. Such systems maintain a manually index system like libraries catalog, or automated crawling indexer in search engines like Google or Excite.

The main advantage of centralized search engines is that they do not create unnecessary network traffic. Queries are sent directly to servers and answers are returned to initiator in a similar way. When the search engine is fast, answers reach to users in less than a second and most of the time provided answers satisfy users. However, there are many disadvantages of centralized search engines. Organizations can control such systems very easily. The policy of providing data and censorship are some examples of such controls. Even if such controls do not exist, search engine companies can dictate their own policies. For example, companies can change the order of information which is sent to users. Nowadays this is a business. By charging extra fee from information providers, any order of sending information to users is possible; therefore, what is shown on the screen does not represent importance or publicity of the information.

Another disadvantage is the possibility of using the privacy of users. Since computers usually maintain the same unique IP-addresses, search engines can track their queries and interests. Search engines can send advertisements based on users' interest along with requested data. This way of advertising is more effective than sending the same content to all users.

From technical point of view, there are other disadvantages involved. Centralized searching systems usually need big storages and processing power. These requirements affect the scalability of systems and increase the cost of ownership. A centralized coordinator is also needed to control these requirements.

These disadvantages are caused to look for an alternative, namely decentralized search systems. In this kind of systems, all computers usually have the same functionality and a node does not have any control on the other nodes. This generally refers to *Peer-to-Peer (P2P)* networks. In a short period, different kinds of P2P were introduced from fully decentralized to partially centralized systems. In *purely decentralized* systems like FreeNet and Gnutella, there is no central coordinator for organizing the network or usage of resources and communication lines. Computers are connected to each other without any particular organization or hierarchy and send a search query to all connected computers with the hope of finding answers in reasonable steps or hops. Since this structure is flat, the routes of messages are unpredictable. This characteristics protects the system from censorship or at least make it very difficult. Furthermore, each computer is autonomous; in other words, each computer determines when and in what

extend it makes its resources available to other computers. The privacy of users is also protected better than centralized systems; because just some parts of the user's request may be sent to other nodes. Although pure decentralized systems solve censorship and privacy problems, they introduce another problem which is network traffic. This is due to the fact that, there is no organization in the network; and global view of the network is not identifiable; therefore query initiator may not have any idea about proper resources for answers. Consequently, the location of the information providers can be hidden easily.

Semi-decentralized systems were more successful than pure centralized ones. Popularity of file sharing systems like Napster, BitTorrent and KaZaa, and the instant messaging systems like MSN-Messenger proves this claim. Semi-decentralized systems, as its name indicates, have a hierarchical structure or a centralized component. Although data is transferred directly between provider and consumer, there is a centralized matchmaking system which usually identifies information providers. These semi-decentralized systems still suffer from privacy and censorship problems but not as much as centralized systems.

Generally speaking, P2P systems - without considering any categorization - provide more scalability, greater fault tolerance, decentralized coordination, self-organization, anonymity, distributed resources and services sharing, lower cost of ownership and better support for creating ad hoc networks.