Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 991–995

ISBN 978-83-60810-22-4

Semantic P2P Search engine

Ilya Rudomilov
Czech Technical University in Prague
Department of Computer Science and Engineering
Email: rudomily@fel.cvut.cz

Prof. Ivan Jelínek
Czech Technical University in Prague
Department of Computer Science and Engineering
Email: jelinek@fel.cvut.cz

***Abstract*—This paper discusses the possibility to use Peer-to-Peer (P2P) scenario for information-retrieval (IR) systems for higher performance and better reliability than classical client-server approach. Our research emphasis has been placed on design intelligent Semantic Peer-to-Peer search engine as multi-agent system (MAS). The main idea of the proposed project is to use semantic model for P2P overlay network, where peers are specified as semantic meta-models by the standardized OWL language from The World Wide Web Consortium. Using semantic model improve the quality of communication between intelligent peers in this P2P network. Undoubtedly, proposed semantic P2P network has all advantages of normal P2P networks and in the first place allow deciding a point with bottle-neck effect (typical problem for client-server applications) by using a set of peers for storing and data processing.**

I. Introduction

RELEVANCE and popularity of Peer-to-Peer networks (P2P) increases every year due to the exponential growth in the number of documents on the Internet and local networks. Already existing and often used Web-search engines with the client-server architecture have problems with storing, processing a large number of documents because of possibilities for centralized solutions (e.g. „bottle-neck effect“). Otherwise, P2P systems provide discributed storing and analyzing data in a set of network members (i. e. "peers").

There is no doubt about the need to find other technologies improve the efficiency of search. One way is to just access a distributed P2P model. This trend is observed not only in commercial areas, but is the subject of set of academic research [3]. High attention to these issues came just at the beginning of the 2000s with the founding The Gnutella network (originally a P2P-file distribution system), which could already be used in a search option [13]. The first of these was developed based on the Gnutella network in 2000 - the search engine InfraSearch [5] and later bought by Sun with name JXTA.

This research is supported by the Grant Agency of the CTU in Prague (grant No. SGS11/129/OHK3/2T/13)and the Visegrad Fund (No. 51100034).

II. P2P background

Peer-to-Peer networks can be classified by used topology, from client-server-like centralized P2P to fully-decentralized P2P networks without central coordination and censorship. For Information-retrieval systems it means indices allocation indices of nodes content, from one centralized index (centralized P2P) to distributed indices among all nodes (decentralized P2P).

A. Centralized P2P

Centralized P2P systems apply advantages of the client-server architecture to P2P networks. There are one or more central servers for nodes coordination (Fig. 1), ensuring of network policy and locate desired documents. Similar to client-server case, node are sending request to server for desired document, but in the centralized P2P network answer contains just addresses of nodes with desired documents for further interaction.

Thereafter the centralized P2P is susceptible to malicious attacks as another centralized systems and single point of failure. Moreover, this category of P2P networks will become a bottleneck for a large number of peers.

Single difference from client-server is in direct data transfer between nodes and thus server did not need to store all files of the network. Follow-up researchers increased competencies of nodes, but defined central server (or few) are necessary for network action.

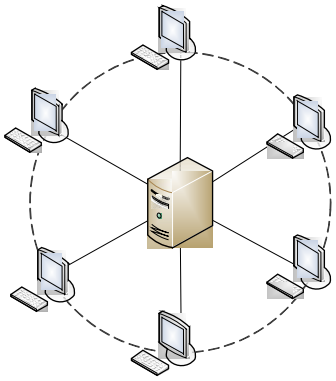

Fig 1. Topology of centralized P2P network

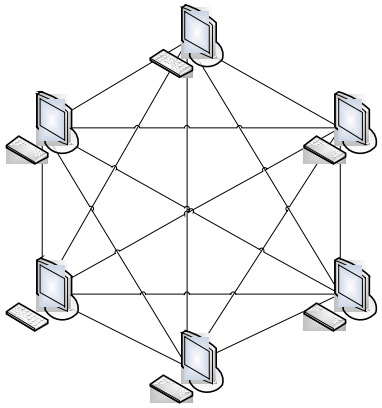

Fig 2. Topology of decentralized P2P network

B. *Decentralized P2P*

A decentralized P2P network is “ideal” P2P network because of one-range topology, equal rights and responsibilities of nodes. There are no central server (Fig. 2) and it assist protection from malicious attacks, censorship, scalable limitations.

The main problem of this type of P2P is coordinating. Nodes in decentralized P2P networks interact directly and coordination are maintained by all nodes. There are two approaches for coordination by using different types of logical network topology, difference between them lies in query transferring between nodes:

1) Unstructured

In an unstructured P2P each node is responsible for its own data, and keeps track of a set of neighbors that it may forward queries to. Nodes have not strict mapping between data and nodes. Classical P2P network Gnutella is related to this category: joining node broadcasts ping-message through whole network and waiting for pong-answer from working nodes. Thereby, flooding is typically method there.

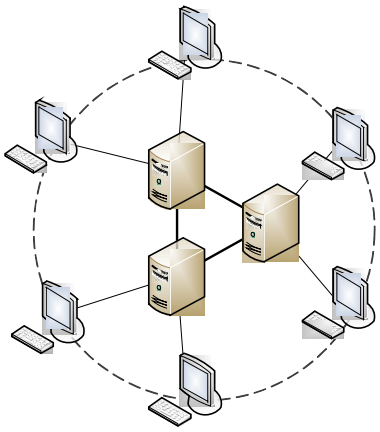

Fig 3. Topology of hybrid P2P network

2) Structured

A structured P2P supports nodes mapping by using predefined strategies (in the first place distributed hash tables, DHT). Nodes can interact with another with more or less strict information about them thanks. As a result, a query can be routed to the node who can have desired documents with the high probability. Majority of structured P2P adopt the key-based routing: CAN, Chord, and Pastry.

C. *Hybrid P2P*

The main advantage of centralized P2P systems is that they are providing a quick and reliable resource locating; the main problem is server limitations. On the other hand, decentralized P2P require more time in resource locating. Different researchers propose to combine techniques of both networks in hybrid P2P concept, which is semi-decentralized and uses a fashion of super-nodes instead of one central server (Fig. 3). There are no difficulties with nodes locating, but problems with super-nodes maintaining for building one-range top level of network.

One of the most famous hybrid P2P network is Edutella (2004), German project for reliable exchange of educational materials. The two-level network is maintained by a set of super-nodes, which are composing HyperCuP level. All documents (document, papers, and videos) in the network are described via defined RDF schema and stored on appropriated super-nodes. A query routing is provided by exchanging RDF indices between super-nodes.

III. Multi-agent approach for P2P

The one of the main modern trends of architecture for P2P search engines is Multi-agent systems (MAS) [9], which consists of a number of intelligent agents, each of which operates independently for the benefit of the whole system. This method allows you to create a fully decentralized or semi-decentralized P2P network, in which, the respective agents work independently or with privileged local coordinating agents [1].

Some modern P2P search engines are developed as a "meta-search engines" [7] for parsing and joining search results from popular commercial engines using [4] and these meta-results are often classified according to own rules [2]. However these systems are based on commercial search engines, and for improved their performance and therefore can not operate independently. These systems may improve search results, but do not work autonomously.

Another actual way to solve is decentralized search engine, among which we can mention the experimental YaCy search engine. YaCy was developed at the University of Karlsruhe in 2003 [16]. YaCy is a P2P search engine without main server with indices and where each Linux-server with an installed YaCy separate downloads, indexes the Web and processes user queries to search for documents throw other servers in the YaCy network. YaCy uses distributed hash tables (DHT) for defined, simple and effective allocation

documents between nodes: nodes calculate (key; value) pairs for all documents in the network and then use these pairs for looking for and getting required file by participating in required DHT. DHT is a classical communication mechanism for P2P networks since P2P networks with popular file-sharing protocols like BitTorrent, Gnutella, Napster, etc [10]. However YaCy uses 4 predefined servers with node lists, therefore YaCy is not fully-decentralized P2P.

IV. Semantic P2P networks

Information-retrieval systems is one of the main problems of P2P networking through problems with generation and distribution indices among nodes. Although P2P is suitable for content-sharing systems thanks to using DHT from some attributes (filename, author and so on), sharing of documents is connected to more complicated and bigger indiced.

YaCy search engine respects the concept of a decentralized P2P system, but its performance is very small [12] of the principal reasons the time cost of the communication between peers by using DHT. This may resolve the problem uses the idea of semantic P2P (SP2P) [8], in which peers are described as meta-models in the Semantic Web according to standard OWL. SP2P systems eliminate problems associated with the use of common ontologies (e.g. maintenance, scalability). However, we have to explicit to provide explicit semantic mappings (i.e. definitions of semantic relationship) between nodes [8].

The use of Semantic Web techniques in Peer-to-Peer traced back to the project SWAP (Semantic Web and Peer-to-Peer, 2002), coordinated by the University of Karlsruhe. During the project, the researchers analysed the potential of Semantic Web Technologies for Peer-to-Peer, prepared by method descriptions and software prototypes. Researcher on the project Steffen Staab updated and published research results in 2006 [11].

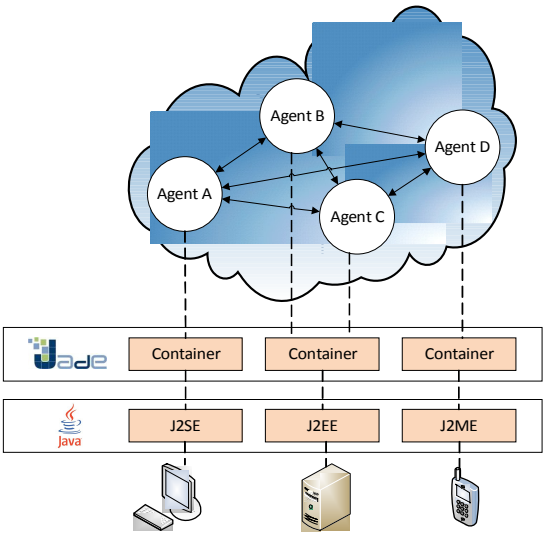

Fig 4. Topology of hybrid P2P network

V. FIPA and JADE

Open-source JADE framework (Java Agent Development framework) [14] is a free framework for developing Java-based intelligent multi-agent systems and in addition, according to standards from the FIPA (Foundation for Intelligent Physical Agents) [15], a major non-commercial group in the multi-intelligent systems. The FIPA's membership includes Toshiba Corp., Siemens, Boeing Company, RWTH Aachen University, etc. The widely adopted FIPA standards are the Agent Management and Agent Communication Language (FIPA-ACL) specifications, which already in use as an industry standard.

It is a modern and popular environment, which can be used without restriction or need major interventions and other collaborators in the research. One can certainly believe that the principles of the proposed system will be used not only as a research subject, but in practical applications.

Using JADE for implementation MAS-based P2P applications is a common practice [9] and has obvious advantages:

- Interoperability: JADE is according to FIPA specifications;
- Portability: Java allows to use different platforms and JADE-based implementation can run on J2EE, J2SE, J2ME environment;
- Easy of use: JADE is a set of APIs, which has GUI for a nodes management.

VI. Our approach

The idea of our project involves the design, testing and implementation of semantic P2P search engine.

The first phases of the project is devoted to a theoretical model of the system, which will have a decentralized architecture and therefore will not use any central node and a set of documents and indices will be placed on any intelligent agent, the amount which will create a multi-agent system (MAS). The theoretical model is based on the ontological reasoning: fully-decentralised P2P network. We build on the progress achieved in this field [13], which expands on the idea of using the semantic model of P2P architectures [8]. Agents will have the same rights and functionality.

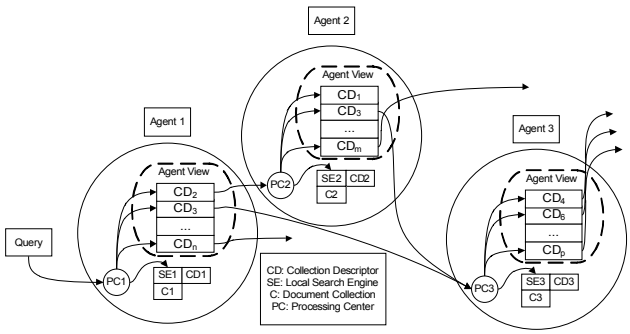

Fig 5. Structure of proposed SP2P network

A. Node structure

Each intelligent agent will include (Fig. 5):

• A set of documents ("Document Collection") with available information. Document Collection is used by Local Search Engine in searching progress.

• Semantic structure of the interaction of node's Document Collection on neighbour nodes („Collection Descriptor"). Collection Descriptor provides information about neighbours (e.g. IP-adresses) and their content. This component works similar to signature of node: nodes distribute their Collection Descriptors in the network with any reconnection, similar to common P2P-practice (e.g. in Gnutella network) of sending notification ping in time of reconnection to the network. Search engine for searching information around documents on the node („Local Search Engine"). Component looks for relevant documents around Document Collection storage to incoming request.

• Processing center for incoming requests and sending results of searching („Processing Center"). All requests in the network are passing throw nodes Processing Centers, which manage sending local requests to the Local Search Engine and coordinate request resending to another nodes because of information about neighbour from their Collection Descriptors.

• Information about other P2P network nodes („Agent-view structure") is a set of received Collection Descriptors with common methods to parse and store.

B. Topology

We suppose to use modified Chord (i.e. DHT-based) model of structured decentralized P2P with possibility to distribute semantic indices among nodes. Chord represents a one-dimensional circular fashion of 2^m nodes (Fig. 6), where each node have a ID (i.e. IP-address) and which is arranging from 0 to $2^m - 1$. Each node has a predecessor (counter-clockwise node) and successor (the next node in a clockwise direction).

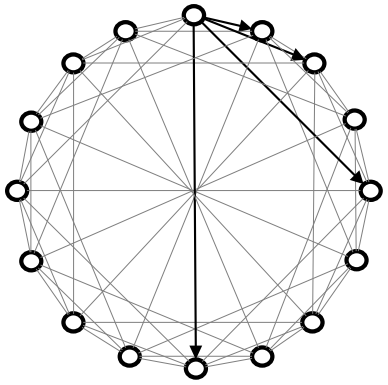

Fig 6. Instance of 16-node Chord network

C. Semantic mapping

Nodes should communicate with each other using the semantic network map to be created from an "agent-view structure" of agents, allocated on nodes. Semantic P2P is such a P2P, which combines the advantages of structured (in which the nodes are defined) and unstructured (in which network topology is not defined) P2P, and assume that because it does not have the disadvantages of both types, i.e. nodes will have addresses for the fastest routing topology but not defined and the nodes can be disconnected without any problem with the network reliability.

Similar OWL-based ontology for simple agents was developed by Michal Laclavik and his colleagues as AgentOWL project in 2006-2009 [6]. In addition, their project was implemented by JADE framework and we can use their experience in our network.

VII. Future work

On the next phase we will research opportunities to integrate generated semantic indices of nodes into Chord protocol. Finally, we will implement an experimental system using JADE framework, which is suitable for this project because of possibility to use XML (and RDF) for communication between agents and OWL for mapping nodes. We propose to conduct research efficiency of this experimental Semantic P2P network in comparison with existed P2P search engines (e.g. YaCy search engine).

References

[1] W. Galuba and S. Girdzijauskas: Peer to Peer Overlay Networks: Structure, Routing and Maintenance. *Encyclopedia of Database Systems, Part 16,* Spriner, 2009, p. 2056-2061.

[2] H. Gylfason, O. Khan, and G. Schoenebeck: Chora: Expert-Based P2P Web Search. *Agents and Peer-to-Peer Computing. Lecture Notes in Computer Science, Volume 4461/2008*, Springer, 2008, p. 74-85.

[3] T. Kathiravelu: Approaches to P2P Internet Application Development. *Proceedings of the Joint International Conference on Autonomic and Autonomous Systems and International Conference on Networking and Services (ICAS/ICNS 2005)*, 2005.

[4] I. Kovalev, M. Rusakov, and M. Tsarev: Search Crossplatform Multiagent System. *Automatic Documentation and Mathematical Linguistics, 2010, Vol. 44, No. 1*, Allerton Press, p. 53–55.

[5] Iraklis A. Klampanos, J. Barnes, and J. Jose: *Evaluating Peer-to-Peer Networking for Information Retrieval within the Context of Meta-searching*, 2006, p. 530.

[6] M. Laclavik, M. Babik, Z. Balogh, and L. Hluchy: AgentOWL: Semantic Knowledge Model and Agent Architecture. *Computing and Informatics. Vol. 25, no. 5 (2006)*, p. 419-437. ISSN 1335-9150, Chapters 1, 4, 5.

[7] J. Lehikoinen, I. Salminen, A. Aaltonen, P. Huuskonen, and J. Kaario: *Personal and Ubiquitous Computing, Volume 10, Number 6*, Longon: Springer, 2006, p. 357-367.

[8] A. Mawlood-Yunis, M. Weiss, and N. Santoro: From P2P to reliable semantic P2P systems. *Peer-to-Peer Networking and Applications, 2010, Volume 3, Number 4*, Springer, p. 363-381.

[9] A. Poggi and M. Tomaiuolo: Integrating Peer-to-Peer and Multi-agent Technologies for the Realization of Content Sharing Applications. *Studies in Computational Intelligence, 2011, Volume 324, Information Retrieval and Mining in Distributed Environments*, Springer, p. 93-107

[10] C. Roncancio, M. del Pilar Villamil, C. Labbé, and P. Serrano-Alvarado: Data Sharing in DHT Based P2P Systems. *Lecture Notes in Computer Science, 2009, Volume 5740, Transactions on Large-Scale Data- and Knowledge-Centered Systems I*, Springer, p. 327-352
[11] S. Staab and H. Stuckenschmidt: *Semantic Web and Peer-to-Peer,* Springer, 2006.
[12] G. Weikum: Peer-to-Peer Web Search. *Encyclopedia of Database Systems*, Saarbrücken: Springer Science+Business Media, 2009, p. 2082-2085.
[13] H. Zhang and V. Lesser: *Toward Peer-to-Peer Based Semantic Search Engines: An Organizational Approach*, Saarbrücken: VDM Verlag, 2008. ISBN 3639084799.
[14] JADE Software Framework (2009), http://jade.tilab.com/
[15] FIPA Specifications (2000), http://www.fipa.org
[16] YaCy Project (2006), http://www.yacyweb.de