Efficient Semantic-based Content Search in P2P Network*

CHEN Ming[1,2], *LU Yang*[3]

(1. Information Engineering College, Capital Normal University, Beijing 100048, China;
2. Department of Computer Science and Technology, China University of Petroleum, Beijing 102249, China;
3. School of Software, China University of Geosciences, Beijing 100083, China)

Abstract: Content-based full-text search is a challenging problem in Peer-to-Peer (P2P) systems. A distributed P2P overlay network that supports semantic-based content searches, called S-Peer, is proposed. Peers in this overlay are grouped based on the semantics of their data, and self-organized as a semantic overlay network. To reduce overheads incurred by peer joining and leaving in a high-dimensional overlay network, peers are constructed as a one-dimensional semantic space that facilitates efficient routing. The results show the effectiveness, efficiency and scalability of the proposed system.

Key words: Peer-to-Peer; content-based; overlay network
CLC number: TP301 **Document code**: A **Article ID**: 0529-6579 (2009) 01-0118-05

基于语义的对等网络检索机制

陈 明[1,2], 鹿 旸[3]

(1. 首都师范大学信息工程学院, 北京 100048;
2. 中国石油大学计算机科学与技术系, 北京 102249;
3 中国地质大学软件学院, 北京 100083)

摘 要: 提出了一个基于语义的分布式对等覆盖网络S-Peer。网络中的节点基于语义信息聚类, 并自组织成一个语义覆盖网络。S-Peer网络构造了一个一维语义空间, 避免了高维语义空间的维护开销, 提高了资源检索效率。实验结果证明了该算法的有效性和灵活性。

关键词: P2P; 内容; 覆盖网络
中图分类号: TP301

1 Introduction

Peer-to-Peer (P2P) technologies have recently received much attention from academia and industries due to many benefits they offer. In a P2P system, a large number of nodes can potentially be pooled together to share their resources, information and services. However, most existing P2P systems support only title-based searches and are limited in functionality when compared to today's search engines, they lack support for semantic-based content search. The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. In a P2P system, semantic web techniques can be used for expressing the knowledge shared by peers in a well-defined and formal way. In the simple model that we propose, peers use a shared ontology to advertise their expertise in the P2P network. The knowledge about the expertise of other peers forms a semantic overlay network, independent of the underlying network topology.

* 收稿日期: 2008-09-18
基金项目: 国家自然科学基金资助项目 (60072006)
作者简介: 陈明 (1949年生), 男, 教授; E-mail: chenming@cup.edu.cn

If a peer receives a query, it can decide to forward it to peers about which it knows that their expertise is similar to the subject of the query. The advantage of this approach is that queries will not be forwarded to all or a random set of known peers, but only to those that have a good chance of answering it.

In this paper, we address the problem of semantic-based content search in the context of document retrieval. Given a query, which may be a phrase, a statement or even a paragraph, we look for documents that are semantically close to the query. We proposed a distributed P2P overlay network that supports semantic-based content searches, called S-Peer. With such an organization, all information within the network can be organized in a semantic overlay network, and then efficiently indexed.

The rest of this paper is structured as follows. Section 2 describes the related work, with the semantic representation studied in section 3. Section 4 presents the semantic retrieval model. The results of simulations are discussed in section 5. Finally, conclusions of this paper are presented in section 6.

2 Related work

Schwartz[1] describes a method that organizes nodes with similar contents into a group. A search starts with random walk but proceeds more deterministically once it hits in a group with matching contents. Motivated by research in data mining, Cohen et al. [2] use guide-rules to organize nodes into an associative network. Sripanidkulchai et al. [3] extend an existing P2P network by linking a node to other nodes that satisfy previous queries.

Replication has also been explored to improve search efficiency. FastTrack[4] designates high- bandwidth nodes as super-nodes. Each super-node replicates the indices of several other nodes. Cohen et al. [5,6] find that setting the number of object replicas to the square root of the searching rate for an object minimizes the expected search size on successful queries.

Schema-based P2P networks such as Edutella[7] are proposed to combine P2P computing and the Semantic Web. These systems build upon peers that use explicit schemas to describe their contents. They use super-peer based topologies, in which peers are organized for routing queries. However, current schema-based P2P networks still have some shortcomings; queries have to be flooded to every node in the network, making the system difficult to scale. Chirita et al. [8] built a publish/subscribe system on the Edutella P2P infrastructure. This system uses content advertising, subscribing and notifying. However, content advertising may create additional overhead. In our system, a subscription request is first directed to a set of potential producer peers in a semantic cluster. Following that, each producer peer will map the request against its local RDF data.

Tang et al. [9] applied classical Information Retrieval techniques to P2P systems and built a decentralized P2P information retrieval system called pSearch. The system makes use of a variant of Content-Addressable Networks (CAN) to build the semantic overlay and uses Latent Semantic Indexing (LSI)[10] to map documents into term vectors in the space. Li et al. [11] built a semantic small world network in which peers are clustered based on term vectors computed using LSI. They proposed an adaptive space linearization technique, and constructed the link structures based on small world network theory.

3 Semantic representation of peer

In this part, the way to describe the peer or request as semantic information based on ontology method in S-Peer mechanism. At same time, the similarity computation method of semantic similarity is advanced.

3.1 Ontology

Ontology is the key in semantic web. In the web, ontology gives the meaning of the web. It provides the semantic of word and note. For example, there are Systematized Nomenclature of Medicine (SNOMED)[12] and Unified Medical Language System (UMLS)[13] in field of medicine. Expert can share and remark the information in this field. New Jersey Institute of Technology has designed the ontology of Object-Oriented Healthcare Vocabulary Repository (OOHVR). The project has about 5000 concepts organized in a semantic network and stored in an object-oriented database. It is accessible on the web via any browser. In computer field, the standard classifying model is ACM Topic currently. It is standard classification in computer science domain. The IS-A concept tree of ACM Topic is shown in Figure 1.

The topic hierarchy contains a set of 1287 topics in the computer science domain. However, each researcher is interest in only some of them. In S-Peer, the ontology in different field can be used according to the need of system. The researcher can self-define interesting field. In this paper, the semantic model of topic model in network environment is standardized taking ACM Topic as example.

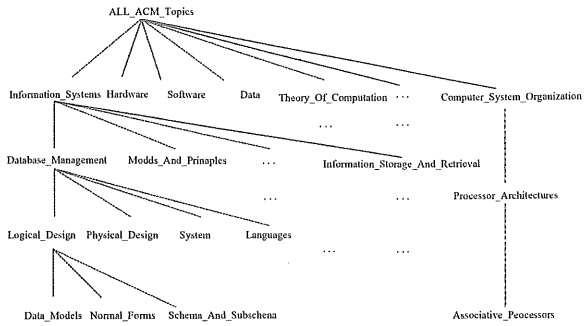

Fig. 1 IS-A concept tree of ACM topic

3.2 Semantic Similarity

Now, the study of semantic similarity can be divided two kinds. One is constructing tree architecture using semantic dictionary to compute semantic similarity. The other one is computing semantic similarity through the way of corpus statistic.

(1) Using the tree hierarchical architecture composed by synonym in semantic dictionary[14], such as WordNet and HowNet, to compute the information entropy or semantic distance between two concepts. Then, the semantic similarity between concepts can be gained.

(2) According to the frequency of the two concepts in the context, the semantic similarity between concepts can be gained through the way of corpus statistic.

The method of corpus statistic is adopted in this paper. According to ACM Topic ontology, every peer classifies the documents in local storage and gives its weight.

Definition 1: Peer semantic representation. $P = \{ < T_I, \lambda_i^P >, i = 1, 2, \cdots, m \}$ denotes the semantic of a peer. In the formula, m is quantity of classification in ACM Topic ontology, T_i is the component of the document on class i, λ_i^P is weight of the peer on class i, it can be gained through computing the frequency of word in the whole peer.

$$\lambda_i^P = \frac{N_i^P}{| P |} \qquad (1)$$

In formula, N_i^P ô is total times number of words of class i on the peer, | P | is the total words number in the peer. So, $P = \{ < T_i, \lambda_i^P >, i = 1, 2, \cdots, m \}$ is aggregate of weighted topic. In practical application, T_i represents a research field of peer, and the weight reflects the concern degree of the peer in Ti field.

Definition 2: Peer similarity. The semantic similarity between peers, or between peer and request, can obtain using the following formula.

$$\mathrm{sim}(P_1, P_2) = \sum_{j=1}^{m} \sum_{i=1}^{m} [\mathrm{sim}(T_i, T_j) \times (\lambda_i^{P1} \times \lambda_j^{P2})] \qquad (2)$$

In formula, sim (T_i, T_j) is the similarity between class T_i and T_j.

The formula 2 can also be used to compute the similarity between peer and request.

In S-Peer, the value range of similarity is in [0, 1]. When two concepts are completely identical, the similarity is 1. Conversely, if the two concepts have not any correlation, the similarity is 0. In other cases, the similarity is from 0 to 1.

4 Semantic Retrieval Model

4.1 Semantic clustering algorithm

In S-Peer, according to the ACM Topic, each class creates a semantic cluster. Nodes with similar semantic information are grouped in the same semantic cluster. To enable search across semantic clusters, an intuitive solution is to construct k-dimensional semantic clusters by connecting each peer to all dimensions of the corresponding clusters such as in [15] and [16]. However, overlay maintenance cost becomes expensive when the number of semantic clusters increases. To reduce overlay maintenance cost, we build a semantic retrieval model which enables the mapping from a k-dimensional semantic space into a one-dimensional semantic space.

Definition 3: Similar peer. The threshold value of peer P_1 is σ, $P_1 = \{ < T_I, \lambda_i^{P_1} >, i = 1, 2, \cdots, m \}$, $P_2 = \{ < T_i, \lambda_i^{P_2} >, i = 1, 2, \cdots, m \}$. If $Sim(P_1, P_2) > \sigma$, the peer P_1 and P_2 can be thought as similar peer.

Definition 4: Semantic correlation class. Given threshold value σ, $P_1 = \{ < T_1, \lambda_i^{P_1} >, i = 1, 2, \cdots, m \}$. If $\lambda_i^P \geqslant \sigma$, T_i can be called as semantic correlation class of peer.

Definition 5: Super semantic correlation class. For all semantic correlation classes, the one which has biggest weight can be called super semantic correlation class .

After peers obtain their , they join the corresponding semantic cluster. Peers who have similar semantic information will establish relationship as neighbor. Usually, peers in the same cluster have higher similar probability. Thus, in the routing process, query will be prior routed to the peer in the same cluster.

4.2 Semantic overlay network

When a new peer joins the network, it will send an advertisement including its IP address and semantic

information. When node P receive node Q' s advertisement, it compute the similarity firstly, and judge whether node Q is its similar node or not. If node Q is similar node of node P, node P will append Q to its lookup table. Each node maintains a local lookup table shown in Table 1.

Table 1　Lookup table

No.	Semantic description	latency
*	$\{<T_i, \lambda_i^P>, i, 1, 2, \cdots, m\}$	60
*	$\{<T_i, \lambda_i^P>, i, 1, 2, \cdots, m\}$	217
	.....	
T_1	$\{<T_i, \lambda_i^P>, i, 1, 2, \cdots, m\}$	16
T_3	$\{<T_i, \lambda_i^P>, i, 1, 2, \cdots, m\}$	39

The asterisk in the first column denotes the peer in the same cluster with the local peer. The second column of lookup table shows the peer' s semantic description, and the last column shows the latency from here to the corresponding peer. Queries will be prior routed to the peer which has lower latency. Similar peers establish relationship as neighbor, thus peers self-organize a semantic overlay network. If a peer receives a query, it can decide to forward it to peers about which it knows that their expertise is similar to the subject of the query.

5　Evaluation

We use simulation to evaluate the effectiveness of S-Peer. The simulation results demonstrate S-Peer' s advantages in overhead, recall, and search efficiency.

5.1　Overhead

In this experiment, we evaluated search overhead by comparing search costs among S-Peer and Gnutella. All nodes are randomly distributed among the system. 10^4 requests are randomly distributed and targeted among all nodes.

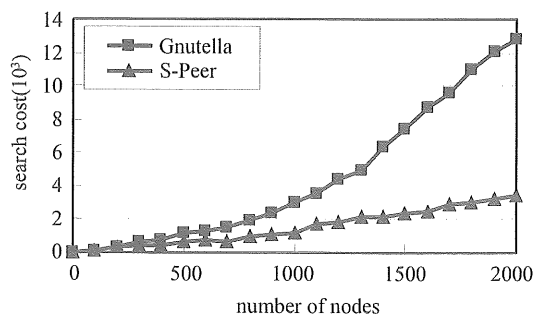

Fig. 2　Search cost for increasing numbers of nodes

As shown in Figure 2, the search cost of Gnutella increases rapidly when the network size grows. In contrast, S-Peer significantly reduce the search cost.

5.2　Recall

To evaluate a P2P system, we use precision and recall measures known from classical Information Retrieval. These measures are defined as follows:

$$\text{Re call} = \frac{\text{Docs}_{\text{relevant}}\text{Docs}_{\text{returned}}}{\text{Docs}_{\text{relevant}}} \quad (3)$$

$\text{Docs}_{\text{relevant}}$ denotes the set of relevant documents in the network, meaning that the terms in the query match their meta-data description, and $\text{Docs}_{\text{returned}}$ being the set of returned documents. The recall indicates how many of the relevant documents are returned.

In this experiment, each node randomly issue requests. Figure 3 plots the recall for increasing numbers of nodes.

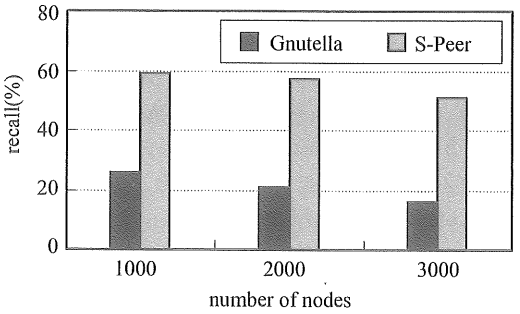

Fig. 3　Recall for increasing numbers of nodes

In the figure, S-Peer lead higher recall than Gnutella. Since clustering in the network focuses queries to a small set of peers, and reduces the number of randomly discovered peers. Also, a shared ontology for semantic similarity improves the recall rate of the system compared with an approach that relies on exact matches, such as a simple keyword based approach.

5.3　Search efficiency

In this experiment , we compared average path

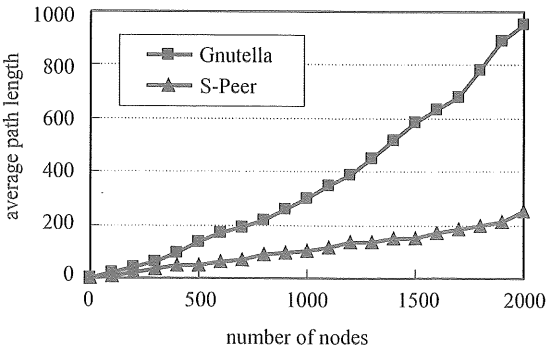

Fig. 4　Average path length for increasing numbers of nodes

length among S-Peer and Gnutella. Each node randomly issue requests.

As shown in Figure 4, the average path lengths for S-Peer increase slowly with the network size as compared to Gnutella.

6 Conclusion

In this paper we studied how to improve the efficiency of a P2P system by clustering nodes, and a distributed P2P overlay network that supports semantic-based content searches, called S-Peer, is proposed. The results from a range of experiments show that S-Peer works effectively and has a good tradeoff between search efficiency and search cost.

References:

[1] SCHWARTZ M. A scalable, Non-hierarchical resource discovery mechanism based on probabilistic protocols [R]. Technical Report CU-CS-474 - 90, University of Colorado, 1990.

[2] COHEN E, FIAT A, KAPLAN H. Associative search in peer to peer networks: Harnessing Latent Semantics [C]// IEEE INFOCOM'03, April 2003.

[3] SRIPANIDKULCHAI K, MAGGS B, ZHANG H. Enabling efficient content location and retrieval in peer-to-peer systems by exploiting locality in interests[J]. ACM SIGCOMM Computer Communication Review, 2002, 32(1): 1.

[4] FAST Track. http://www.fasttrack.nu.

[5] COHEN E, SHENKER S. Replication strategies in unstructured peer-to-peer networks[C]//ACM SIGCOMM'02, 2002.

[6] LV Q, CAO P, COHEN E, et al. Search and replication in unstructured peer-to-peer networks [C]//ICS'02, June 2002.

[7] Nejdl W, Wolpers M, Siberski W, et al. Super-peer-based routing and clustering strategies for RDF-based peer-to-peer networks [C]//Proceedings of 12th World Wide Web Conference, May 2003.

[8] CHIRITA P A, IDREOS S, KOUBARAKIS M, et al. Publish/subscribe for RDF-based P2P networks [C]// Proceedings of the 1st European Semantic Web Symposium. Greece, May 2004.

[9] TANG C Q, XU Z C, DWARKADAS S. Peer-to-peer information retrieval using self-organizing semantic overlay networks [C]// Proceedings of ACM SIGCOMM. Karlsruhe, Germany, August 2003.

[10] DEERWESTER S C, DUMAIS S T, LANDAUER T K, et al. Indexing by latent semantic analysis[J]. Journal of the American Society of Information Science, 1990, 41(6): 391 - 407.

[11] LI M, LEE W C, SIVASUBRAMANIAM A, et al. A small world overlay network for semantic based search in P2P[C]//Proceedings of the 2nd workshop on semantics in peer-to-peer and grid computing in conjunction with the world wide web conference. May 2004.

[12] PRICE C, SPACKMAN K. SNOMED Clinical Terms [J]. British Journal of Healthcare Computing and Information Management, 2000, 17(2): 27 - 31.

[13] HUMPHREYS B L, LINDBERG D A B. The UMLS Project: Making the conceptual connection between users and the information they need[J]. Bulletin of the Medical Library Association, 1993, 81(2): 170 - 177

[14] BUDANITSKY A, HIRST G. Evaluating word net based measures of lexical semantic relatedness[J]. Computational Linguistics, 2006, 32(1): 13 - 47.

[15] CRESPO A, GARCIA-MOLINA H. Semantic overlay networks for P2P systems[R]. Stanford University, January 2003.

[16] GU T, TAN E, PUNG H K, et al. ContextPeers: scalable peer-to-peer search for context information[C]//Proceedings of international workshop on innovations in web infrastructure, in conjunction with the 14th world wide web conference (WWW 2005), Japan, May 2005.