Intelligent Search Technology Combining Semantic Grid and Clustering

Cuncun Wei
Faculty of Engineering, Zhejiang Business Technology Institute, Ningbo 315012, Zhejiang, China
e-mail: weicuncun@163.com

Abstract

It is a critical problem of P2P network about how to efficiently and accurately search resources on P2P network. This thesis mainly starts from improving query efficiency to establish an intelligent search framework. On the basis of Gnutella-flooding search technology, it applies theories of semantic ontology search combined with semantic hash routing table technology and searches accurate answer from the resource library through problem traversal query in the network and node provided in the routing table. Meanwhile, in view of the flaws of the current structuralized and non-structuralized P2P network, it applies a hierarchical clustering method to form a hierarchical semantic web through semantic clustering in the node, domain clustering and global clustering, etc. Experiment shows that this framework could well improve search, query efficiency and fault tolerance in P2P environment, extend and understand user query demands and reach the aim of accurate search.

Keywords*: P2P Network, semantic search, semantic grid, hierarchical clustering, hash routing table*

1. Introduction

The rapid development of information technology provides people with a broad sharing platform, and network retrieval has already become a commonly used channel for people to gain information and people retrieve relevant information through information retrieval tools. This to a certain extent solves the problem of resource classification and retrieval. P2P network has been the research focus in recent years, and the realization of P2P network lies in the efficient search of resources in P2P network. The probability to search resources and the search response time are two key factors affecting search efficiency [1]. Currently, P2P technology has already been widely applied to file sharing in the Internet environment, and these systems could provide data sharing function based on file names. In the large-scale network environment, however, content localization still remains the primary challenge confronted by the in-depth information sharing of P2P system [2]. Moreover, recall and precision of information retrieval are low, making it become increasingly unable to adapt to the retrieval demands.

This thesis mainly starts from the improvement of query efficiency to establish an intelligent search framework. It analyzes the flaws of routing model in the current P2P environment, and establishes a semantic web based on hierarchical semantic clustering. In this network, domain and clustering are applied to exercise a dual management over the network, and the domain and clustering are organized according to a hierarchical structure. Combined with semantic ontology technology, it adopts a node logical structure and node semantic Hash routing table, and realizes precise, accurate, automatic and intelligent network resource search with “ask-answer” query mode. On the platform of semantic grid, it locates retrieval demand in the appropriate data pool to perform retrieval service, so as to further improve precision ratio and recall ratio, and eliminates resource islands.

2. System Framework

This thesis makes transformation on the basis of the original logic structure of key work search in P2P network, and designs the semantic search framework demonstrated in Figure 1.

Received March 23, 2013; Revised May 26, 2013; Accepted June 3, 2013

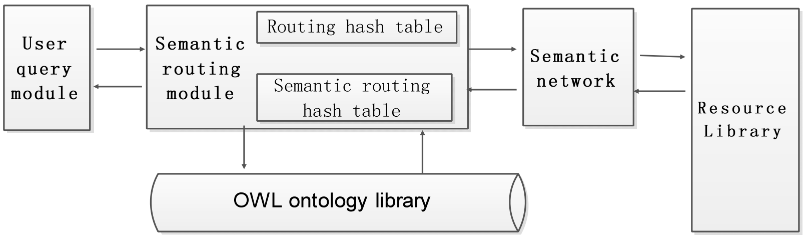

Figure 1. The Semantic Search Framework

It is mainly composed of 5 parts, including user query module, semantic routing module, semantic web, OWL ontology library, and resource library. The framework core is the establishment of semantic web based on semantic clustering and routing Hash table based on semantic grid. What follows focuses on the realization of these two modules.

3. Construction of Semantic Web

3.1. Node Model in P2P Environment

P2P network could be expressed as undirected graph, and the specific definition is as follows:

Definition 1 P2P network undirected graph $G=(V,E)$, in it: V is node set, corrthe networkng to P the edgenode in network; E is edge set, representing the connection between Peer nodes. For any node $u \in V$, $v \in V$, if $(u,v) \in E$, there is absolutely $(v,u) \in E$. For any $u \in V$, its neighbor set is recorded $N_U = \{v \mid (u,v) \in E\}$.

Based on this model, the initial period of our study of P2P network is unorganized and node interconnected, and the construction of semantic web is thus carried out based on the file category saved in the node.

3.2. Node Clustering

Node clustering is to classify files saved in the node. During this process, characteristic vector represents structured clustering, which could greatly reduce the amount of data to be processed. On the basis of semantic similarity measurement, it clusters information resources in the network, and cluster the semantically similar information resources, so as to form a semantic clustering tree [3]. Clustering characteristic vector is defined as follows:

Definition 2 Semantic clustering characteristic is a triple $CF = \{SM, N, \theta\}$, and N is the number of clustering object, θ is semantic similarity threshold of objects inside sub-clustering, SM is semantic description $SM = \{Name, Attribute, H^c, R\}$ of sub-clustering, and SM equals to a composite information unit entity, including a name set $Name$ of SM, entity attributes set $Attribute$, entity semantic relationship set H^c and R. $Name$ saves several terms to generalize this sub-clustering, $Name \in T$.

The specific clustering method could apply hierarchical clustering method based on characteristic vector.

3.3. Select Domain Super Node

Assume W_i is a series of node in space i, with $|W_i|$ nodes altogether. All nodes have their respective roles as either super node or ordinary node. They control inter-domain clustering together [4]. Ideally, domain super node must be in the center of this domain. In consideration of load balance, the number of nodes inside the domain must be similar.

The space of the same size must be constructed as far as possible, or the number of nodes in each space is close to S_w. Assume that IP of the node P_i is IP_{pi}. Sometimes, we may accidentally select some cheating behaviors or error-prone nodes. In order to avoid the selection of this inefficient node as the super node, parameter T is introduced. T is defined as the minute of that time. If $\left(IP_{pi}+T\right)MODS_{W}=0$, P_i is selected as domain super node. This could bring consistent separation to the selected domain super node in the whole network.

3.4. Generation of Domain

When node P_i is selected as the domain super node, it will issue detection message to the node directly connected to it. When node P_i receives this detection message, it will make the following responses:

(1) If P_i is unoccupied: ① change its state into occupied; ② send its identity to domain super node, describe what resources it contains and indicate whether it is willing to act as domain super node or clustering super node; ③ sendthe neighboring nodemation to neighboring node of P_i, rather than the node from which it is sent.

(2) If P_i is occupied: P_i sends to the node from which it receives detection information an occupied message, including information about its own domain super node. Similarly, it sends its own domain super node the information about P_i. In this way, domain super node of both sides could get to know information related to its own neighbor. After the process above is completed, the following result could be gained:

① Every domain super node includes a series of nodes in its own domain, and has their resource message;

② Every ordinary node knows its own domain super node;

③ Every domain super node knows the domain super node in its neighboring domain.

This algorithm guarantees that all the nodes of the network are interconnected. This is of vital importance to search in P2P network, or it will be impossible to gain resources saved in these nodes.

3.5. Domain Clustering

After one domain and its domain super node are determined, one needs to start to search characteristic vector of node, and to have inter inner domain clustering on the basis of these characteristic vectors.

(1) Domain super node i sends detection information of characteristic vector to all the nodes inside the domain.

(2) When node P_i receives this detection information of characteristic vector, it sends its characteristic vector F_i to domain super node.

(3) After receiving these characteristic vectors, domain super node starts to implement clustering inside the domain and generates a series of clustering based on a series of characteristic vectors. Characteristic vector F_i presents the specific characteristic of clustering C_i.

(4) Domain super node selects one clustering super node R_i for each clusteselectionand this process of select could be carried out according to the resource message it provides, such as CPU and connectivity, so as to select a node with better performance.

(5) Lastly, what is saved in domain super node is a series of clustering description $\left(CD\right)$, and every description indicates a clustering C_i. $\left(CD\right)$ includes clustering mark C_i, characteristic vector F_i, clustering member $\left\{P\right\}$, of this clustering and clustering super node R of this clustering, i.e. $CD_i=\left(C_{i,}F_{i,}\left(P\right),R\right)$.

(6) Domain super node sends every clustering super node a clustering description CD_s CD that contains all clustering in the domain. The clustering super node sends node in every clustering a message including $\left(C_{i,}F_{i,}R\right)$.

During this process, a node may contain many types of files, and it needs to take into account the optimal clustering of these nodes inside this domain[5]. In the practical application, the user mainly takes into account the query time and efficiency, etc. The method suggested here is: only when this node has a certain amount of files belonging to this clustering can this node be included in this clustering.

3.6. Global Clustering

The process of inter-domain global clustering is as follows:

(1) After generating a domain, every domain super node saves some information of its neighboring domain, including the information of the neighboring domain super node. The network is then essentially a network composed of several domains, just like several nodes in the initial state

(2) In the hierarchical domain network, i level of domain is composed of a series of domains of $i-1$ level, and this number is generally $|SW|$ (SW refers to a group of domain, while $|SW|$ refers to the number of domains in this group of domain) on average. In other words, $i-1$ level of domain super nodes has 1/$|SW|$ ratio of opportunity to become i level of domain super node. Here, while determining domain super node, the method applied is the same as that used in selecting the domain super node in forming the initial domain. It asks for special attention here that i level of domain super node must be selected from $i-1$ level of domain super node.

(3) The same method is applied to i level of domain super node while generating i level of domain. In this way, i level of domain super node could get to know the information of neighboring domain on its same level.

(4) The same as collecting characteristic vector in the basic clustering, domain super node on level i collects $N_C|SW|$ clustering descriptions of the previous level. Similarly, it implements clustering algorithm to generate advanced clustering. In order to restrict the complexity of the next step, the number of clustering shall not exceed a certain amount, and shall select one clustering super node for every clustering.

(5) Clustering description passes among node representatives, and finally all the clustering representatives gain all the clustering information of the lower level and upper level in the hierarchical structure.

Through the algorithm above, there is only one domain super node left, without any neighbor. It makes the end of inter-domain global clustering.

The final structure realization:

① Every node saves relevant information of domain super node.

② Domain super node of Level 1 knows how many nodes are included in its domain, and has information of domain super node of Level 2 of the higher domain it is subordinate to.

③ Domain super node on level i ($i>1$) knows domain super node of level $i-1$ subordinate to its lower domain and the information of domain super node of information of $i+1$ level.

④ Every domain super node has clustering super node information in its own jurisdiction domain.

Finally, hierarchical network structure is formed. When a certain node in the network submits query demand, first of all, one should analyze what category this demand belongs to, and then submit it to the corresponding clustering super node. The clustering super node makes query in this clustering and submits this query to the clustering super node of the upper clustering [6], so as to reach the purpose of making query in other domains to gain more information. The query is mainly carried out among clustering, and domain only plays a role of understanding the whole network and assisting the generation of clustering, which is not involved in the practical application.

4. Routing Module Based on Semantic Network

Semantic grid adopts metadata to describe information in the grid, which well defines information and service and better enable the coordinated work between computer and people. Most importantly, it describes all resources, including service, in a manner that could be processed by the machine, aiming to realize the intelligent interoperability of semantic and machine [7].

Gnutella protocol is a P2P protocol used to release retrieval, mainly used in completely decentralized resource search [8]. Flooding is the search technology based on this protocol. This technology does not need to calculate topology and relevant routing to maintain the network, and only requires sending the information node to all packages in a broadcasting manner. It enjoys simple maintenance and highly efficient performance, accompanied by flaws like the existence of massive redundant connection, additional network traffic flow, a large consumption of network bandwidth, and query method based on “key words” search technology, which could not solve the problem of appearance of irrelevant “result” in user query [9].

This thesis makes a simple transformation to the Flooding method of the Gnutella protocol in P2P network system. Based on local semantic ontology technology of node, it improves disadvantages existing on Gnutella-flooding method, so as to reach the purpose of the user accurate query.

Semantic routing module shown in Chart 1 maintains key word query of node sharing resource library in the original P2P manner, and substitutes the original key word processing module that receives a query with semantic ontology module. The specific OWL ontology library in this thesis adopts a currently programmed semantic ontology semantic ontology--- Resource Description Framework (RDF) [11]. It also adds semantic Hash resources table, and revises Hash routing table. Although Hash routing table still includes IP, Port and specific sharing resources of P2P key word query node, it newly adds a currently mature semantic ontology (library) and semantic Hash sharing resources table, which is expressed with natural semantics in a “question---answer” manner [10]. Moreover, every answer and the former question in the question search path have explicit inclusion relationships.

4.1. Semantic Hash Routing Table

Semantic search network makes a massive use of semantic Hash tables, because the query process of “question-answer” in semantic network resources is realized by utilizing semantic ontology included in user node of semantic Hash table to settle a specific problem. It is a (question, communication mode, node) mapping correspondence table, together with the design of the judgment of performance evaluation of the answer. Node includes IP, Pot and the specific sharing resource description, etc. In other words, every node cites a detailed answer of address and information in ontology library, capable of solving a problem accurately. The management of the Hash routing table includes the addition and cancellation of the table.

(1) Increase of the routing table. When the logic node receives Ping order with node ability description, it sends a ping instruction to answer, and queries whether its communication method asks for encryption and encryption method, followed by checking whether it really has declared node ability. When it passes the examination, it is added in the routing table, and immediately sends the notice of ability enhancement to all other neighboring nodes.

(2) Cancellation of routing table. Basically, all that has time strategy or long-term setting aside, or long-term failure of contact, or when it is certain that the ability is no longer needed could be cancelled. However, in order to maintain the stability of node ability, the node Hash routing table should not change frequently.

4.2. Improvement Based on Gnutella-flooding Protocol Search Method

Because of machine and network performance of Gnutella-flooding protocol in P2P network, it could not maintain a large capacity of routing table, or there will be much flooding information, and result in massive search redundancy [12]. Meanwhile, it does not have the function to judge the relationship between search answer and question. Semantic ontology search adopts Gnutella-flooding protocol query, introduces semantic ontology library through changing its routing table into a Hash routing table [13], and forms a large and effective routing table for maintenance as much as possible, so as to maintain its sufficiently strong query ability, without the problem of flooding information. As every node is interconnected and selects routing in the tremendous routing table, there may be one message that really sends query message, and the corresponding answer could be selected in the routing table. Node semantic ontology sends node ability query, or Query, to the next node through Hash routing table, i.e. Query i+1, which realizes the traversal query problem in network. The traversal process of Query is also the process of refinement of problems. The user will gradually gain an accurate answer, so as to solve a series of problems related to the field from the shallow to the deep.

5. Algorithm Analysis

During the process of constructing semantic web, there is the analysis of the algorithm time complexity of constructing hierarchical clustering. Assume the height of tree as h (the height of tree is determined by the size of memory and page h □ n). The fission factor of the tree is F , and the insertion of one point asks for a number of nodes to be visited as $l+\log Fh$. The cost to insert all objects to generate a tree is $P(n*F(1+\log Fh)$, but generally tree is not generated once, but through several times of reconstruction. Assume that there are at most $l(l$ □ $n)$ leave nodes in need of reconstruction, and the cost of rebuilding tree is $P(n*F(1+\log Fh)$. Therefore, the average time complexity is $P(n*l*F(l+\log Fh))\approx P(n)$.

The experiment verification tells that search framework based on semantic web is greatly improved compared to the non-organized flooding method in efficiency, and the specific experiment result is shown in Figure 2 and Table 1.

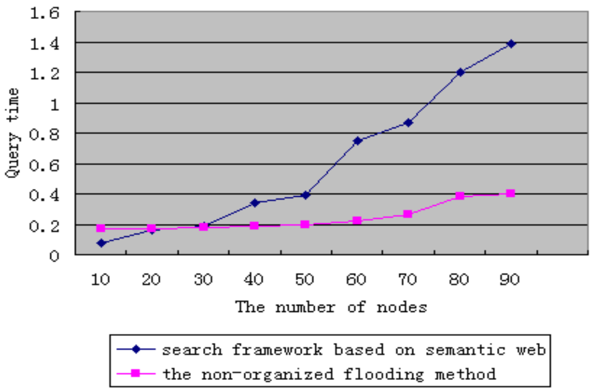

Figure 2. Two Methods of Query Time

Table 1. The Experimental Results of Data

query time\The number of nodes	10	20	30	40	50	60	70	80	90
search framework based on semantic web	0.08	0.16	0.19	0.34	0.39	0.75	0.87	1.2	1.39
the non-organized flooding method	0.17	0.17	0.18	0.19	0.2	0.22	0.26	0.38	0.4

It could be seen from this contrast chart that when there are few nodes, efficiency of flooding method is better than the construction of semantic web, because flooding method reduces the construction and maintenance time of the whole network topology. However, as network nodes continuously increase, flooding method query time is also increasing rapidly. By contrast, in a semantic web environment, the speed of increase of query time is obviously slower. Therefore, in the distributed and dispersed environment with a large scale of nodes, the construction of semantic web contributes more to the practical application.

6. Conclusion

Based on the network structure of P2P environment, in order to improve query efficiency, this thesis establishes an intelligent search framework combining semantic grid and clustering. In view of the flaws of this framework in structuralized and non-structuralized network, it proposes the construction of a hierarchical semantic clustering net of dual management. This network is highly adaptable to distribute and dispersed P2P network, together with convenient maintenance and high efficiency. Meanwhile, on the basis of network Gnutella-flooding search technology, it applies theories of semantic ontology search method, describes the resource framework through semantic ontology libraries. Combined with semantic Hash routing table technology, through the traversal query of problem in the network, it searches accurate answer from sharing resource library of nodes provided by routing table, realizes combination of ontology and Petri network, achieves the effective organization of a great deal of information on the internet, extends users' retrieval demands in the retrieval system to understand the real retrieval intention of user, and achieves the aim of accurate search.

Acknowledgements

This work was supported by Zhejiang province department of education scientific research achievements of the project in 2011 year (Y201121970) and The 2012 Ningbo returnees innovative entrepreneurial projects (2012B72012); Zhejiang province excellent young teachers of universities and colleges funding projects in 2010 year (The Zhejiang Department of Education (2010) No. 175, 304.); Ningbo educational planning issues of 2012(YGH120).

References

[1] Habib Rostami, Jafar Habibi, Emad Livani. Semantic routing of search queries in P2P networks. *Journal of Parallel and Distributed Computing*. 2008; 68(12): 1590-1602.

[2] Nikolaos D Doulamis, Pantelis N Karamolegkos, Anastasios Doulamis, Ioannis Nikolakopoulos. Exploiting semantic proximities for content search over p2p networks. *Computer Communications.* 2009; 32(5): 814-827.

[3] Adrian Kuhn, Stéphane Ducasse, Tudor Gîrba. Semantic clustering: Identifying topics in source code. *Information and Software Technology.* 2007; 49(3): 230-240.

[4] Gloria Bordogna, Gabriella Pasi. A quality driven Hierarchical Data Divisive Soft Clustering for information retrieval. *Knowledge-Based Systems.* 2012; 26(1): 9-19.

[5] Bettina Fazzinga, Giorgio Gianforme, Georg Gottlob, Thomas Lukasiewicz. Semantic Web search based on ontological conjunctive queries. *Web Semantics: Science, Services and Agents on the World Wide Web.* 2011; 9(4): 453-473.

[6] John WT Lee, Daniel S Yeung, Eric CC Tsang. Hierarchical clustering based on ordinal consistency. *Pattern Recognition.* 2005; 38(11): 1913-1925.

[7] Andrew Flahive, David Taniar, Wenny Rahayu, Bernady O. Apduhan. Ontology tailoring in the Semantic Grid. *Computer Standards & Interfaces.* 2009; 31(5): 870-885.

[8] Evangelos Pournaras, Georgios Exarchakos, Nick Antonopoulos. Load-driven neighbourhood reconfiguration of Gnutella overlay. *Computer Communications.* 2008; 31(13): 3030-3039.

[9] Grzegorz Chmaj, Krzysztof Walkowiak. A P2P computing system for overlay networks. *Future Generation Computer Systems.* 2013; 29(1): 242-249.

[10] Panos Kalnis, Wee Siong Ng, Beng Chin Ooi, Kian-Lee Tan. Answering similarity queries in peer-to-peer networks. *Information Systems.* 2006; 31(1): 57-72.

[11] K Munir, M Odeh, R McClatchey. Ontology-driven relational query formulation using the semantic and assertional capabilities of OWL-DL. *Knowledge-Based Systems.* 2012; 35(1): 144-159.

[12] Elena Meshkova, Janne Riihijärvi, Marina Petrova, Petri Mähönen. A survey on resource discovery mechanisms, peer-to-peer and service discovery frameworks. *Computer Networks.* 2008; 52(11): 2097-2128.

[13] Habib Rostami, Jafar Habibi, Emad Livani. Semantic routing of search queries in P2P networks. *Journal of Parallel and Distributed Computing.* 2008; 68(12): 1590-1602.