Conference or Workshop Item:
Pham, Tran Vu, Lau, Lydia M.S. and Dew, Peter M. (2006) An adaptive approach to P2P resource discovery in distributed scientific research communities. In: Sixth International Workshop on Global and Peer-to-Peer Computing (GP2P) in conjunction with IEEE/ACM International Symposium on Cluster Computing and the Grid 2006, 16-19 May 2006, Singapore.

An Adaptive Approach to P2P Resource Discovery in Distributed Scientific Research Communities

Tran Vu Pham
School of Computing
University of Leeds
Leeds, LS2 9JT, UK
tranp@comp.leeds.ac.uk

Lydia MS Lau
School of Computing
University of Leeds
Leeds, LS2 9JT, UK
llau@comp.leeds.ac.uk

Peter M Dew
Informatics Institute
School of Computing
University of Leeds
Leeds, LS2 9JT, UK
dew@comp.leeds.ac.uk

Abstract

Resource discovery in a distributed environment is always a challenging issue. It is even more difficult to provide an efficient query routing mechanism while still able to support complex query processing in a decentralised P2P environment. This paper presents an adaptive approach to P2P resource discovery. It separates the routing of queries from query matching mechanism so that an effective combination could be explored. Three properties of scientific research communities provide the grounding for the method: the existence of common interest groups, the willingness to share resources of common interests and the transitive relationship in the sharing behaviour. By exploiting these properties, search queries can be efficiently forwarded to those who are more likely to have the answers to improve the quality of search results and to reduce the network traffic. Experimental results have provided some evidence to confirm the efficiency of this adaptive approach.

1. Introduction

Resource discovery in a P2P environment remains a challenging issue, despite many P2P applications have been introduced and commercialised, such as Napster and Kazaa [8]. The challenge is in the search for an efficient method of routing complex queries within the decentralised environment. The current most popular resource discovery techniques can only partially meet the requirements. The basic flooding technique used in Gnutella-like systems [5] is able to support various kind of queries but not scalable as the population of peers of the network grows [1]. On the other hand, the indexing method using distributed hash table is scalable in term of query routing, but can only support exact or partial keyword matching [15, 16, 20]. It fails to process complex queries, such as those require matching the semantic of document contents.

Resource discovery methods which based on the users' interests have recently emerged to improve the routing of complex queries [6, 17, 18, 19]. Instead of sending queries blindly to every peer in the network, these methods try to forward the queries to peers that are most likely to have the answers. The number of fruitless attempts can then be reduced. However, these approaches often require a complex network topology [17, 18] and/or clustering of peers into groups of common interest [6]. With clustering methods, a peer is assigned to only one particular group of common interest. Queries about other interests will not be efficiently routed.

However, it is common that scientists have more than one area of interest, especially in multidisciplinary research. In these research communities, scientists require knowledge, expertise and resources from a number of related areas.

In this paper, an adaptive approach is introduced. It exploits the existence of different common interest groups in distributed research communities. Instead of clustering peers into different groups, it allows individual peers to specify their own views of the virtual world based on their interests. Each peer can then adaptively learn which active peers in the community share similar interests. This will help the routing of queries more precisely to the peers that are most likely to have the answers.

This adaptive approach was originally designed for the Collaborative e-Science Architecture (CeSA) [13, 14], which combined P2P environments with computational and data grids using a service oriented architecture. However, it is equally applicable for general purpose P2P resource discovery.

In the next section, rationale behind the design is explained. The adaptive approach and the experiments will then be discussed in detail in subsequent sections. Similar approaches will be reviewed and compared. The conclusion at the end confirms the potential of the proposed approach and its application. There will also be discussion about potential extensions to this work.

2. Requirements from Scientific Research Communities

The requirements for resource discovery have been drawn from our previous work on the CeSA [13, 14]. The CeSA is an architecture which aims to support collaborations within distributed scientific communities. The CeSA consists of two environments (grid and P2P) loosely connected by Grid services [4][1]. The grid environment consists of 'heavy duty' resource providers. The P2P provides an environment for more interactive collaborations amongst members of the community. This is where an adaptive discovery method of resources is required.

When designing an efficient algorithm for resource discovery, two important factors that need to be studied are: the kind of resources to be discovered and the characteristics of user communities.

With regard to the resources, the types of resources that need to be discovered in scientific communities are usually information about the availability of Grid resources, experimental datasets, research papers or working documents. Each of these categories requires a different method of query matching. For example, to discover information about Grid resources, i.e. a Grid Service, a query might look for information about service providers, input and output parameters. To search for a research dataset, the information required could be authors, time of experiment and so on. Complex query matching techniques are required in these cases. Therefore, a good method for resource discovery needs to support different types of query matching techniques.

Scientific research communities are dynamic and distributed. These communities consist of members from all over the world. Scientists might develop new interests anytime. In addition, scientists today are often undertaking multidisciplinary research (e.g. in e-Science). Therefore, they might participate in different interest groups at the same time. These characteristics affect the behaviours of resource sharing amongst scientists within the communities. A good resource discovery method needs to support and can also exploit these characteristics to improve its efficiency.

In conclusion, an efficient resource discovery method for scientific P2P environment needs to:

- be scalable, a requirement that any efficient discovery technique needs to address
- support any types of queries and query matching techniques
- deal with the dynamics in P2P environments
- support different interests of scientists

[1]It is equally applicable to WS-Resource Framework, which is replacing Grid services [3]

3. The Adaptive Approach to P2P Resource Discovery

A resource discovery process often involves two major steps: 'routing of query' and 'matching of query'. The goal of the adaptive method is to provide an efficient mechanism for routing search queries in a pure P2P environment. Instead of sending a query blindly to its neighbouring peers in the network, a peer should try to send the search query to peers that are most likely to have answers. This would improve quality of results and also reduce unnecessary network traffic. The adaptive approach implements this idea by a learning mechanism to help a peer in learning from past query results about interests of other peers in the environment and in detecting (and adapt to) any changes in their interests.

As this approach separates the routing from the query matching, the routing component can be used with any types of queries and any query matching techniques.

3.1. Underlying Properties

The adaptive method for resource discovery in P2P environment exploits three properties which emerged from a previous case study on the characteristics of scientific research communities [13]. Property 1 provides a conceptual model for the grouping of peers. Property 2 is the basis to develop a learning mechanism and Property 3 underpins the routing algorithm for queries.

Property 1: There is the existence of groups of common interests within a research community.

In a large scientific community, collaborations usually take place amongst groups of scientists who are working on similar or the same topics. This is similar to the small world concept [7, 9, 10, 11]. However, a scientist may work on a number of related or overlapping research topics. Hence, he/she can participate in different groups.

The following 2 properties are drawn from Property 1.

Property 2: Scientists who have a common interest often need and share a common set of resources for that particular interest.

Property 3: Transitive relationships about 'interest in resources' exist in scientific research communities.

3.2. The Operations

The adaptive resource discovery method consists of three operations: (i) describing peer interests using ontology, (ii) recording peers with similar interests using past query results and, (iii) routing search queries. These operations are discussed in detail in following subsections.

Describing Peer Interests. Initially, each peer in the network is provided with an initial set of classification ontology. This set of ontology is globally recognized within the community and covers a wide range of interests that exist in the target user community. It is similar to eBay or Yahoo directory but the interests are from a scientific domain. Using this set of ontology, a user can describe his/her peer interests.

As an example, Fig. 1 shows a fraction of the global ontology that might be used to describe the e-Science domain.

```
e-Science\
    Biology\
        BioChemistry\
        BioInformatics\
    Chemistry\
        Atmospheric Chemistry\
        Combustion Chemistry\
        Environmental Chemistry\
    Computing\
        Computer Visualisation\
        Informatics\
        Grid Computing\
```

Figure 1. A fraction of an intial global ontology for e-Science community

The ontology in Fig. 1 starts with 'e-Science' as the general domain. In the 'e-Science' domain, there are three sub domains: 'Biology', 'Chemistry' and 'Computing'. Similarly, the 'Biology' domain can further classified as 'BioChemistry' and 'BioInformatics' and so on. The ontology provided in Fig. 1 is only a very simple ontology for illustration purpose. In reality, the initial ontology should contain much more detail.

Using the global ontology provided, individual scientists start to describe their interests. If a scientist has only a general interest in 'Biology', the classification can just be 'e-Science\Biology'. However, if the scientist has more specific interests within 'Biology', for instance 'BioInformatics', the classification associated with the peer should be 'e-Science\Biology\BioInformatics'. 'Informatics' within 'Computing' might also be of interest. Hence, 'e-Science\Computing\Informatics' is added to the peer's description of interest. The final description of the peer interest for our example will be a subset of the initial ontology, as illustrated in Fig. 2.

```
e-Science\
    Biology\
        BioInformatics\
    Chemistry\
    Computing\
        Informatics\
```

Figure 2. Description of a peer's interests

Recording Peers with Similar Interests. This is a learning operation which takes place through out the life time of a peer. The learning process helps the peer to update its knowledge and to adapt itself to the environment. The learning process is based on Property 2 described above.

For each peer, a Query History Tree (QHT) is constructed. The backbone of the QHT is the classification ontology used to describe the peer's interests. Each node (branch or leaf) of the QHT represents a classification of a peer's interest. Attached to each node of the tree is a 'peer list' which records the peers that have previously answered queries on the interest represented by the node. Each entry to a 'peer list' must contain enough information to identify a peer in the network. Depending on the P2P application used, an entry could be a peer ID or a pair of IP address and port number. Initially, these lists are empty and will be updated during the life time of the peer. Fig. 3 gives an example of the QHT for the peer used in Fig. 2.

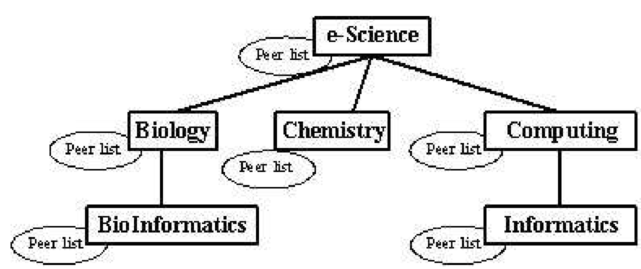

Figure 3. A query history tree of a peer

When a query is issued by a peer, a classification tag defining the area of interests will be attached. This classification tag is a tree path from the root of the peer's

QHT to the node that represents the interest. Following the previous example, if a query is looking for resources about 'BioInformatics', a classification tag 'e-Science\Biology\BioInformatics' will be assigned to it. When receiving query results, the peers with valid responses will be added into the peer list attached to the node.

The decision on whether a response is valid or not depends on the implementation strategy. If full automation is chosen, any peer answering will be added to the list. However, this method might not produce a very accurate list, as it commonly happens that a response to a search query is not necessarily a relevant answer to the query. If accuracy is preferred, the validation will be left to the user. With this approach, only peers with valid answers will be added to the peer list of the node. For complex queries, the second approach is preferable, in order to improve the quality of the peer lists.

A peer list attached to each node of a QHT can also be a priority list. The use of prioritisation strategy decides how a peer should adapt to changes in the environment. Depending on the characteristics of the community, priority can be given to peers that have previously provided the largest number of valid answers or the most recent valid answer. If 'most recent valid answer' is prioritised, the peer will respond quicker to changes in the environment. Hence, it is more appropriate for a fluid or newly set up group. However, in a more static environment, the use of 'largest number of valid answers' will provide more reliable query results.

Routing of Queries. The routing mechanism aims to utilise the existence of common interest groups for more effective routing. By using QHTs constructed in the previous step together with the transitive relationship of 'interest in resources' amongst the peers (Property 3), a 'network' of peers with a common interest can be traced (Property 1). This overlay network can be used for routing the query to the peers most likely to provide an answer. The following will explain the routing in detail.

Each search query is associated with a Time-To-Live (TTL), a fan-out value (f) and a classification tag:

- TTL: the maximum number of hops that the query can travel within the network. This value is defined by the application.
- Fan-out value (f): the number of peers to which a peer will forward a query message. This value is defined by the application.
- Classification tag: the construction of classification tag is explained in the previous subsection. It is used for routing the query and for recording query results.

The routing process is carried out when a peer issues a search query or when it receives a search query from another peer. The knowledge contained in the QHT of the peer is used to guide the routing of query messages to next appropriate peers within the environment.

When issuing a query message at a peer:

- The user specifies the topic that the query message is looking for (e.g. 'BioInformatics').
- A classification tag will be constructed and attached to the query (e.g. 'e-Science\Biology\BioInformatics').
- The peer will then look up at the node of its QHT that is pointed to by the classification tag and pick up from the peer list of the node first f peers to forward the query to.

On receiving a search query, a peer will:

- Attempt to answer the query by searching in its local storage for relevant resources.
- If there is an answer, reply directly to the requesting peer.
- If the query has not reached its TTL, then use the classification tag attached to the query to look up in its QHT for the first f peers (similar to the previous case), and forward the query to the selected peers.

A few possibilities might happen when routing search queries:

(i) If the peer list being pointed to by the classification tag is empty (for example, in the initial state, when peer lists in the tree have not been populated) or the number of peers in the peer list is fewer than f, then the following steps will be done:

- Traverse up the tree and pick up the peers in peer lists held by parent nodes until the requested number is met, with the priority given to closest parents.
- If the request is still not met, forward the queries the peers selected by the previous step and some randomly selected neighbouring peers to have enough f peers or as closest as possible.

(ii) If the classification tag carried by the query does not match any node of the QHT of the local peer, then partial mapping between the classification tag and the local peers QHT will be used. This can happen when the current local peer and the peer that issues the query have different interests or different description of interests. The partial mapping will start from the root of the tree and the root of the classification tag. Only the matching part of the classification tag with the QHT will be used by the local peer as if it was the classification tag. The procedure used to select peers to forward the query to will be exactly the same as in the previous case. If no match is found, the query will be forwarded to random neighbours.

(iii) To avoid forwarding a query to a peer more than once, a loop detection technique is used. This technique requires each query to keep a record of peers it has visited. Before forwarding a query to other peers, a peer will check the path that the query has taken so far, and will only forward the query to peers not in the record. However, with the current routing mechanism, at every hop on its way, a query message is cloned into f copies before being forwarded to the next set of peers. It is possible for the 'same' query to arrive at a peer via more than one route taken by different 'cloned queries'. This kind of duplication cannot be eliminated by the loop detection technique.

4. Experiments

In order to test the adaptive approach, two experiments were run with the following objectives:

- The first experiment was to evaluate the efficiency of the adaptive resource discovery method by comparing its performance with the basic blind (random) flooding method.
- The second experiment was to analyse the relationship between resource distribution of the network and the efficiency of the adaptive method.

These experiments were conducted in a simulated conditions.

4.1. Experiment 1 - Evaluating the Adaptive Approach

Two simulations were set up. One was for the blind flooding method, and the other was for the adaptive method.

In order to compare the proposed adaptive approach with the blind flooding method, the same fan-out value f (3) and TTL (6) were used in each of the simulations. The same network configuration and pattern of resource distribution were also used in both simulations.

Network configuration: The simulated network was set up with 10,000 peers. The network topology was randomly generated so that every peer would be connected to at least 3 and maximum 6 neighbours.

Resource distribution: Each piece of resource was randomly enumerated as an integer in range 0 to 4,999 (inclusive) and was assigned to one of 500 categories, based on its value. Each of these categories represents a topic of interest. Four consecutive categories were assigned to each peer to represent the interests of the associated scientist. In the simulation of the adaptive method, the categories (areas of interests) form the ontology.

Each peer was assigned with randomly five pieces of resources, ranged within its assigned categories.

Measurement: The following two measurements were taken in each of the simulations:

- *Query hit rate*: calculated by 'the number of queries that have answers' over 'the total number of queries issued' in a specified period.
- *Network traffic*: measured by 'the total number of query messages' passed around in the network during a specified period.

A more efficient method should produce higher hit rates with lower network traffic.

Process: A total of 400,000 queries were generated by all peers in the network for each of the simulations. Queries produced by a peer were restricted within its assigned categories. After every 5000 queries, a hit rate was calculated and recorded. Network traffic after every 5000 queries was also recorded. In this experiment, for both methods, when a peer found an answer for a query in its local storage, it would stop forwarding the query.

Results: The graph in Fig. 4 shows the hit rate comparison between the blind flooding method (light colour line) and the proposed adaptive method (dark line). As seen from the graph, the hit rate of the blind flooding method, calculated after each 5000 queries, fluctuated below 30 percent, while, the hit rate of the adaptive method grew gradually when the number of queries increased. After about 325,000 queries, the hit rate of the adaptive method reached 90 percent. It became stable at 93 percent, after 360,000 queries were issued.

The hit rate of the adaptive method improved dramatically when the number of queries increased. This is because it took into account the characteristics of resource distribution within the environment. At the beginning, the hit rate of this method was roughly the same as the blind flooding method. However, when the learning progressed, peers accumulated more knowledge about the environment. Therefore, search queries were forwarded to more appropriate destinations. The hit rate levelled when it had learned quite enough about its environment. In the blind flooding method, search queries were always forwarded randomly to other peers, hence the hit rate was almost the same, no matter how many queries had been issued.

Similarly, the graph on Fig. 5 shows the number of messages passed in the network after every 5000 queries for both cases. As expected, the number of messages needed for every 5000 queries by the blind flooding remained roughly the same (just over 1,800,000). In the case of using adaptive method, the number of messages required for every

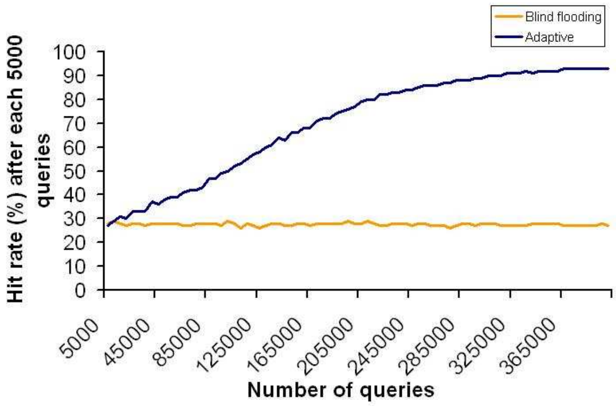

Figure 4. Hit rate comparison between the blind flooding method and the adaptive method

5000 queries decreased when the total number of queries increased. This is easy to explain. As the hit rate increases, fewer number of query messages would be passed on from peer to peer.

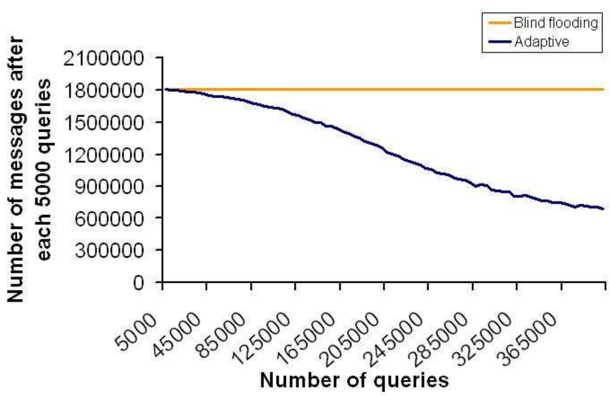

Figure 5. Messages passed in the network when using the flooding method and the adaptive method

In conclusion, this experiment has shown that the adaptive method is more efficient than the blind flooding method. After a certain number of queries are issued, the learning process will help peers to adapt to its environment. As the result, the query hit rate will increase.

4.2. Experiment 2 - Effect of Resource Distribution

In order to analyse the effect of resource distribution on the proposed adaptive approach, a number of simulations using this method were run using different patterns of distribution.

Network configuration: This experiment used the same network configuration as in the previous experiment.

Resource distribution: Resources and categories were enumerated as the same way as in the previous experiment. The only difference was the number of resource categories assigned to individual peers. In this experiment, several runs of the adaptive method were performed. In each run, peers were assigned with a different number of resource categories (2, 6 and 10).

As a contrast, random distribution of resources was also experimented. Simulations were run on the following two implementations of random distribution using the adaptive method:

- The whole resource domain was classified into 500 categories as the previous simulations. However, as peers could have any resource within the resource range, each peer was assigned with all of these 500 categories.
- Resources in the network were treated as in one category. All peers could have any number of resources within resource range of 0 to 4,999.

Measurement: For each simulation, hit rate for every 5000 queries was calculated for comparison.

Process: The measurement process for each simulation was done exactly in the same way as in the previous experiment.

Results: The graph in Fig. 6 shows different hit rates returned by simulations using different patterns of resource distribution.

As shown on the graph, as the number of categories assigned to each peer increased, it took longer for the hit rate to rise. This result concurred that the learning outcome will be better when each peer has resources in fewer categories. When each peer has limited amount of resources (as in this experiment), and if the resources on each peer are spread over so many of categories (interests), it will be harder to learn accurately a peers interests. Therefore the learning outcome will be less accurate. It will take longer for the network to produce optimal query results.

The two methods of applying the adaptive approach to a random distribution of network resources produced two contradicting results. By classifying all resources to one category, the query hit rate kept decreasing when the number of queries increased. Whereas in the other case, the hit rate produced increased overtime, despite slowly. This is

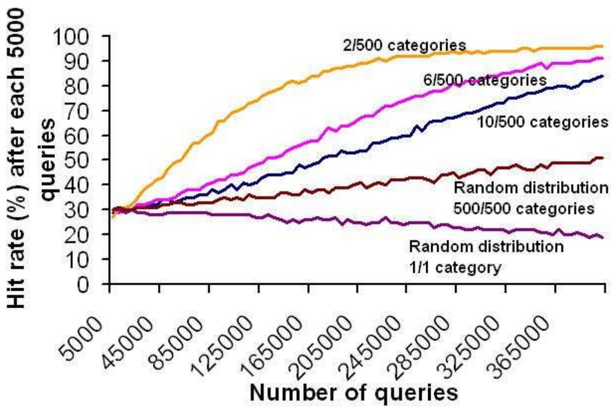

Figure 6. Query hit rates of simulations on different resource distribution configurations

because when treating the whole resource domain as one category, the use of 'peer-list' for learning not only did not help, but also encouraged a 'group-think' scenario which means a smaller group of peers seem to satisfy each other's query, hence having very little opportunity to explore peers outside the group. As a result, the coverage of query messages (for a query) was reduced. This scope was even smaller than the coverage of a query routed by the blind flooding method.

In summary, this experiment had two important outcomes. Firstly, it has shown that resource distribution of the network has effect on performance of the adaptive method. Secondly, the adaptive method can also be used in a random distribution network if resources are categorised, despite the fact that previous researches have revealed that this kind of distribution does not commonly exist in real life [7, 9, 10, 11].

5. Related Work

At the beginning of this paper, it was mentioned that some approaches to P2P resource discovery that based on user interest pattern are emerging. In this section, an analysis on some of these approaches was conducted.

The HyperCuP system used ontology to organise peers into groups of similar interests using a hypercube topology network [17, 18]. Search queries were forwarded to interest groups to produce a better hit rate and reduce redundant query messages. This approach required complex construction of the structured hypercube topology network. When joining the network, a peer declared its interest so that the network could put the peer into the cluster of its interest. As P2P is a dynamic environment, a peer might change its interest over time. Constantly updating the network would result in high cost. Furthermore, it would be more complicated if peers had more than one interest.

The small world pattern introduced by Newman [9, 10, 11] was used to develop an information dissemination algorithm [6]. The basic principle of this algorithm is to build clusters of peers of similar interests. A search query (or a piece of information to be disseminated) is targeted to relevant clusters, where the query is most likely to be answered, to reduce network traffic and to increase query hit rate. There are some issues need to be resolved with this approach. Firstly, what size the cluster should be to achieve optimal performance? If the cluster size is very large, the quality of results will be decreased. However, if the cluster size is small, some relevant information will not be retrieved. Secondly, in the case when a peer has more than one interest, queries of interests other than the common interest of cluster that the peer is assigned to will be less efficiently distributed.

Along this line, Cohen et al. proposed an algorithm to build associative overlays based on guild rules to route search queries [2]. Possession rule, which grouped together peers that shared a common data item, was proposed as a guide rule. This approach required a traced index of peers that participated on the rule. The use of document names for possession rules made it unable to deal with the semantic similarity of document contents. Alternatively, Sripanidkulchai et al. introduced an architecture in which a peer's view of the semantic overlay was a list of peers that had previously had answers to its queries [19]. The future queries would be forwarded directly to peers in the list, as shortcuts. This method is simple to implement. However, similar to the cluster approach, the size of the list has a strong effect on the search results. If the users have many different interests, the hit rate will decrease.

In comparison with the above approaches, the proposed adaptive method provides users with more flexibility. Although also exploiting the small world pattern, or interest-based locality, it does not require complex construction of the network. Only a general classification ontology for the domain is required at the beginning. It does also not limit peer users to any particular cluster. The users can participate (implicitly) in any group by declaring their interests. The algorithm will adaptively locate the group. In case there are groups with too big or small size (that might affect quality of search results) the users can use the query history tree to further classify the big groups or to merge the small groups to a larger one.

6. Conclusion and Future Work

The paper has described in detail an adaptive approach to resource discovery in a P2P environment. This adaptive approach takes into account the characteristics of scientific research communities in order to provide an efficient way of routing search queries. As the routing is separated from query matching, this adaptive approach can be used with any types of queries and query matching techniques.

The experiments showed that this approach can significantly improve query hit rate in comparison with blind flooding method. It can also offer users with more flexibility in adjusting their preferences.

Random topology networks were used to compare the efficiency of the blind flooding and the adaptive methods and to analyse the effect of resource distribution on the proposed method. Future experiments can be conducted on realistic configurations of network topology and resource distribution to get a better insight into the behaviour the adaptive method in a real environment.

The use of ontology also poses a new challenge: the management of ontology in a distributed and decentralised environment. An earlier paper [12] proposed an evolutionary approach which allows individuals to manage (extend, adapt and share) their own ontologies after being given an initial one. As the community evolves and with increasing sharing, the most common ontology may be promoted to be used in the whole group.

References

[1] Y. Chawathe, S. Ratnasamy, L. Breslau, N. Lanham, and S. Shenker. Making gnutella-like p2p systems scalable. In *Proceedings of the ACM SIGCOMM 2003*, pages 407–418, 2003.

[2] E. Cohen, A. Fiat, and H. Kaplan. Associative search in peer to peer networks: Harnessing latent semantics. In *Proceedings of the IEEE Infocom'03*, San Francisco, 2003.

[3] K. Czajkowski, D. Ferguson, I. Foster, J. Frey, S. Graham, T. Maguire, D. Snelling, and S. Tuecke. *From OGSI to WS-Resource Framework: Refactoring and Evolution*, May 2004.

[4] I. Foster, C. Kesselman, J. Nick, and S. Tuecke. *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration*. Open Grid Service Infrastructure WG, Global Grid Forum, June 2002.

[5] Gnutella. http://www.gnutella.com, 2006.

[6] A. Iamnitchi and I. Foster. Interest-aware information dissemination in small-world communities. In *The 14th IEEE International Symposium on High Performance Distributed Computing*, pages 167–175, North Carolina, USA, 2005.

[7] A. Iamnitchi, M. Ripeanu, and I. Foster. Small-world file-sharing communities. In *Infocom*, Hong Kong, 2004.

[8] Kazaa. http://www.kazaa.com/, 2006.

[9] M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Physcal Review E*, 64, 2001.

[10] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighed networks, and centrality. *Physcal Review E*, 64, 2001.

[11] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci.*, 98(2):404–409, 2001.

[12] T. V. Pham, L. M. Lau, and P. M. Dew. The integration of Grid and Peer-to-peer to support scientific collaboration. In D. Michaelides and L. Moreau, editors, *Proceedings of GGF11 Semantic Grid Applications Workshop*, pages 71–77, Honolulu, USA, Jun. 2004.

[13] T. V. Pham, L. M. Lau, P. M. Dew, and M. J. Pilling. Collaborative e-Science architecture for reaction kinetics research community. In *Proceedings of the Challenges of Large Applications in Distributed Environments (CLADE) workshop*, pages 13–22, North Carolina, USA, July 2005. IEEE Computer Society Press.

[14] T. V. Pham, L. M. Lau, P. M. Dew, and M. J. Pilling. A collaborative e-Science architecture towards a virtual research environment. In S. J. Cox and D. W. Walker, editors, *Proceedings of the 4th e-Science All Hands Meeting (AHM'05)*, Nottingham, UK, September 2005.

[15] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A scalable content-addressable network. In *Proceedings of the ACM SIGCOMM 2001*, pages 161–172, 2001.

[16] A. Rowstron and P. Drusche. Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems. In *Proc. of the 18th IFIP/ACM International Conference on Distributed Systems Platforms (Middleware 2001)*, Heidelberg, Germany, November 2001.

[17] M. Schlosser, M. S. S. Decker, and W. Nejdl. A scalable and ontology-based P2P infrastructure for Semantic Web Services. In *Proceedings of the Second IEEE International Conference on Peer-to-Peer Computing*, pages 104–111, Sep. 2002.

[18] M. Schlosser, M. Sintek, S. Decker, and W. Nejdl. Hypercup - hypercubes, ontologies and efficient search on p2p networks. In *Intl. Workshop on Agents and Peer-to-Peer Computing*, Bologna, Italy, 2002.

[19] K. Sripanidkulchai, B. Maggs, and H. Zhang. Efficient content location using interest-based locality in peer-to-peer systems. In *Infocom*, San Francisco, USA, 2003.

[20] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *Proceedings of the ACM SIGCOMM 2001*, pages 149–160, 2001.