2010 Fifth International Conference on Internet and Web Applications and Services

A Global knowledge for Information Retrieval in P2P Networks

Anis Ismail, Mohamed Quafafou
LSIS, Domaine Universitaire de Saint-Jérôme
Avenue Escadrille Normandie-Niemen
13397 Marseille cedex 20, France
{anis.ismail, mohamed.quafafou}@univmed.fr

Gilles Nachouki
LINA Laboratory
2 rue de la Houssinière
44322 Nantes cedex 03, France
Gilles.Nachouki@univnantes.Fr

Mohammad Hajjar
Université Libanaise, Institut
Universitaire de Technologie -
Saida – Liban
m_hajjar@ul.edu.lb

Abstract—**In traditional P2P networks, such as Gnutella, peers propagate query messages towards the resource holders by flooding them through the network. However, it is a costly operation since it consumes node and link resources excessively, which are often unnecessarily. There is no reason, for example, for a peer to receive a query message if the peer has no matching resource or is not on the path to a peer holding a matching resource. However, how to quickly discover the right resource in a large-scale P2P network without generating too much network traffic and with minimum possible time remain highly challenging. In this paper, we propose a new peer-to-peer (P2P) search method aiming at exploiting data mining concepts (Decision Tree) to improve search performance for information retrieval in P2P network. We use a PDMS system, which aims to combine a Super-Peer (SP) based network with the capability of managing a data model attached to the peers in the form of relational, xml, or object schemes. Each SP is connected to a Global-Knowledge-Super-Peer (GKSP) that operates with an index (decision tree), to predict the relevant domains (super-peers), to answer a given query. Compared with a super peer-based approach, our proposal architectures show the effect of the data mining with better performance with respect to response time, number of messages, precision and recall.**

Keywords - P2P, data mining, decision tree, routing queries.

I. INTRODUCTION

Querying in traditional P2P (P2P) networks like Gnutella is performed by flooding a query message through the network [1]. In flooding, each node receiving the query forwards the query to all its neighbors. Although it is simple and robust, this approach wastes too much bandwidth because sometimes most of the neighbors who receive the query do not reply. Semantic routing [2] is one of the many schemes developed to solve this problem. It reduces querying over head by trying to find the nodes that might send an answer and forwarding the queries towards those nodes. The distributed nature of P2P network allows all peers to freely exchange resources without suffering from the threats of centralized systems, such as central point of failure, high maintenance costs and low scalability. However, how to find the right information and resource in a large-scale P2P network has remained a critical problem [3] [4].

In the last years, the challenge of P2P computing research has attracted researchers. The work on this area has first gained a highlight when systems such as Napster [30] and Gnutella [31] have emerged as killer-applications for file-sharing between users across the world. On one side, there is the centralized approach of Napster, where peers use a unique central server to search for files and then use the P2P infrastructure to exchange them. On the other side, there is the totally distributed approach of Gnutella, where peers broadcast their file requests to the entire network through connections with neighboring peers.

Both approaches suffer from scalability and robustness problems. Napster presents a central point of failure which compromises the overall stability of the system; whereas in Gnutella, the flooding mechanism used to spread users' requests becomes prohibitive when a large number of nodes exist in the system. These two systems (and their limitations) represent two sides of P2P networks and, since then, they have been used as a basis for improvement in almost all P2P computing research.

Another important aspect of the user experience is how long the user must wait for the results to arrive. This is due to a large part of the mediation process which remains difficult to realize in such a context when the number of servers (SP) increases. Response times tend to be slow in super-peers based networks, because on one hand, the query travels through several super-peers in the network and on the other hand, the super-peers are forced to look for connections (i.e. mappings) in order to route the query. Satisfaction time is simply the time that has elapsed from when the query is first submitted by the user (peer), to when the user receives the overall results.

Ideally, a data mining technique in P2P networks, e.g., P2P classification or P2P clustering, is expected to achieve learning performance that is comparable to that of a regular centralized approach. This, however, is a very difficult task. For instance, P2P classification (also P2P clustering), often faces a number of challenges [5][6], such as scalability (Can the algorithm produce good results within an acceptable duration when there are many peers?).

In this paper, we present the influence of data mining in P2P query routing. Our proposed method focuses on how the query is routed to relevant peers, with minimum query processing at super-peers level in order to improve the answering time of the queries. The important advantage of our approach is how to avoid flooding queries, without large bandwidth consumption. The basic configuration is that peers having similar interests can be grouped together under a super-peer which publishes a theme of interest. In this respect, the introduced notion of expertise plays a crucial role allowing the definition of semantic matching between local data schemes of peers and the domain exposed by a super peer. This baseline configuration is augmented by the introduction of GKSP nodes connecting super peers, and operating with an index (a decision tree) to discover the semantic similarity between a query subject and the underlying peer data schemes. The purpose is to find the relevant super-peers capable of answering a given query. Indeed, the subject of a query indirectly suggests the required expertise to answer the query. This operation should reduce the search load on nodes that have unrelated content, thus avoiding the waste of resources.

978-0-7695-4022-1/10 $26.00 © 2010 IEEE
DOI 10.1109/ICIW.2010.41

The following section presents related works. Section 3 recalls briefly principal concepts of P2P networks and shows the context of our work. Section 4, presents the semantics' routing of queries algorithm in P2P systems called baseline algorithm. Section 5, introduces the Global-Knowledge-Super-Peer (GKSP) network. Section 6 presents the effect of data mining in our baseline algorithm. Section 7 presents the experiments and the evaluations. In Section 8, we present the conclusion

II. RELATAD WORKS

In this section, we firstly discuss the integration of the data mining field in P2P network; then we discuss query routing and the most important approaches of information retrieval in P2P networks.

A. *Data mining in Peer to Peer Network*

Knowledge discovery and data mining from P2P network is a relatively new field with little related literature. Some researchers have developed several different approaches to compute basic operations (e.g. average, sum, max, execution time) on P2P networks. Raahemi et al. [7] present a new approach using data-mining technique. In particular, they used decision tree to classify P2P traffic in IP networks by capturing Internet traffic at a main gateway router, performed preprocessing on the data, selected the most significant attributes, and prepared a training-data set to which the decision-tree algorithm was applied.

Bhaduri et al. [8] propose a P2P decision tree induction algorithm in which every peer learns and maintains the correct decision tree compared to a centralized scenario. This algorithm is completely decentralized, asynchronous, and adapts smoothly to changes in the data and the network. Emekci et al. [9] consider a scenario where multiple data sources are willing to run data mining algorithms over the union of their data; as long as each data source is guaranteed that its information, which does not pertain to another data source, will not be revealed. In particular, they focus on the classification problem and present an efficient algorithm for building a decision tree over an arbitrary number of distributed sources, in a privacy preserving manner, using the ID3 algorithm.

Yingyue et al. [10] present two itinerary planning algorithms, with the goal of maximising the information extracted, while keeping resource usage to a minimum. The ISMAP determines the itinerary before dispatching the mobile agent, while the IDMAP algorithm selects the route on the fly.

B. *Query routing in P2P networks*

Query routing in a P2P network is the process by which a query is routed to a number of relevant communities (SP) instead of being broadcasted to the whole network. In [11], the INGA algorithm is presented. INGA extends the ideas of REMINDIN' [12], where each peer plays the role of a person in a social network. To determine the most appropriate peers, each peer maintains in a lazy manner, a personal semantic shortcut index by analyzing the queries that are initiated by users of the p2p network and that happen to pass through the peer. The main limitation of this routing approach is the unavoidable flooding of the network with messages, when a new peer enters the network or when peers (in lower layers) contain limited information about queries that have already answered in the past. The SQPeer routing strategy [13] uses intentional active schemas (RVL Views) to determine relevant peer bases through the fragmentation of query patterns. However, since each view (active-schema) corresponds to a peer advertisement, it should be broadcasted to the whole p2p network.

Nejdl et al. presented the routing approach based on routing indices in [14]. This approach has been suggested and adapted under various scenarios. It is built upon an RDF-based P2P network. Queries and answers to queries are represented using RDF metadata which we can use together with the RDF metadata describing the content of peers to build explicit routing indices which facilitate more sophisticated routing approaches. Queries can then be distributed relying on these routing indices. On the one hand, these routing indices contain metadata information plus appropriate pointers of other (neighboring) peers indicating the direction where specific metadata (schemas) are used. On the other hand, they do not rely on a single schema but can contain information about arbitrary schemas used in the network. Otherwise, our approach is based on routing distributed indices in order to find the super-peer with minimum query processing, which is the strength of our approach comparing to the approach cited above. Next, we discuss the information retrieval in P2P.

C. *Information Retrieval In P2P Networks*

Information Retrieval (IR) systems keep large amounts of unstructured or weakly structured data, such as text documents or HTML pages, and offer search functionalities to deliver documents relevant to a query. The main challenge of information retrieval in P2P networks is the ability to guide the query to the other peers containing the most relevant answers in a fast and efficient way. Today, researchers, from different areas including database systems, distributed systems, networking and information retrieval, have started to work on efficient, yet semantically powerful search mechanisms in P2P systems.

Haase et al. [29] propose a model in which peers advertise their expertise in the P2P network. The knowledge about the expertise of other peers forms a semantic topology. Based on the semantic similarity between the subject of a query and the expertise of other peers, a peer can select appropriate peers to forward queries to, instead of broadcasting the query or sending it to a random set of peers.

Papapetrou [15] proposes new approaches to enable distributed IR over P2P with-out limiting the network size or mutilating the IR. The basis of these approaches is an innovative distributed clustering algorithm, which can cluster peers in a P2P network based on their content similarity. This clustering enables significant network savings and also enables new families of distributed IR algorithms. Nottelmann et al. [16] build an IR system over a hierarchical P2P network. The peers there do not maintain a distributed index; instead, some super-peers are assigned the responsibility to keep their peers' summaries, and to forward the queries to the most related of their peers, or to other super-peers. Sharma et al. [17] introduce a system, called IR-Wire, of information retrieval research in the P2P file-sharing domain. This tool maintains many statistics, implements a number of information retrieval ranking functions and contains a data logger and analyzer.

Today, data management in P2P networks provides a promising approach that offers scalability, adaptively to high dynamics, and failure resilience. Although there exist many P2P data management systems in the literature, most of them focus on providing only information retrieval (IR) [18] [19] [20] [21] or filtering (IF) [22] functionality (also referred to as publish/subscribe or alerting), and have no support for a combined service. DHTrie [23] is an exact IR and IF system that stresses retrieval effectiveness, while MAPS [24] provides approximate IR and IF by relaxing recall guarantees to achieve better scalability.

III. Background

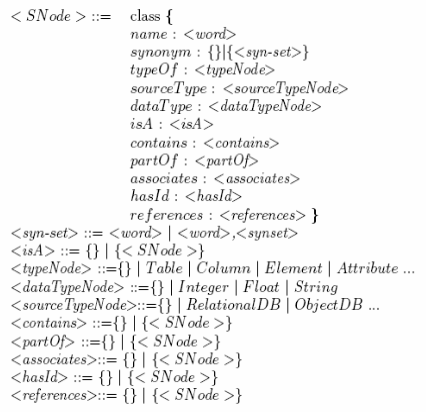
In this paragraph, we introduce the basic notions concerning the architecture of the P2P network and the expertise.

A. *Basic notations*

A Peer is an autonomous entity with the capacity of storage and data processing. In a computer network, a Peer may act as a client or as a server. A P2P is a set of autonomous and self-organized peers (P), connected together through a computer network. The purpose of a P2P network is the sharing of resources (files, databases) distributed on peers by avoiding the appearance of a peer as a central server in this network. We note: P2P = (P, U), P is the set of peers and U represents links (overlay connections) between two peers P_i and P_j, U $\subseteq$ P x P. The super-peer based network that we consider in this paper includes sets of peers and super-peers. We note : P2P = (P $\cup$ SP, K), where P is the set of peers, SP is the set of super-peers and K is the set of overlay links expressed under the format of pairs: (P_i, SP_j) or (SP_j, SP_k) which respectively link a Peer P_i to a Super-Peer SP_j or a Super-Peer SP_j to one or several super-peers SP_k.

A PDMS (Peer Data Management System) combines P2P systems and databases systems. The PDMS, we are considering, is a scale system P2P. Each peer is supposed to hold a database (or an XML document, etc.) with a data schema. Each Super-Peer provides a theme (a semantic domain, a subject, or an idea) representing special interest to a group of peers.

We note R the set of relations reduced in this paper to two relations {Role; IsA} and PDMS={PS $\cup$ SP^T, D, K} where PS represents all the peers of the network with their data schemas S={S1, …., Sp}. A peer is connected to the network with only one data schema. K is the set of overlay links between (super-)peers. Each peer P $\in$ PS is doted of a Data Management System (denoted DMS) able to manage their data. T={T_1,…., T_k} represents the interest themes published by super-peers SP through the network. In our case, each super-peer publishes only one theme and the peers expressed are interested by one theme in T. The themes are not disjoints: two super-peers can publish the same concepts or roles with distinct structures and/or don't use the same vocabulary. D = {D_1, …, D_k} describes the themes in the set of T: D_j describes the theme T_j using sGraph model proposed in [27].



```
< SNode > ::=   class {
                name : <word>
                synonym : {}|{<syn-set>}
                typeOf : <typeNode>
                sourceType : <sourceTypeNode>
                dataType : <dataTypeNode>
                isA : <isA>
                contains : <contains>
                partOf : <partOf>
                associates : <associates>
                hasId : <hasId>
                references : <references> }
<syn-set> ::= <word> | <word>,<synset>
<isA> ::= {} | {< SNode >}
<typeNode> ::={} | Table | Column | Element | Attribute ...
<dataTypeNode> ::={} | Integer | Float | String
<sourceTypeNode>::={} | RelationalDB | ObjectDB ...
<contains> ::={} | {< SNode >}
<partOf> ::={} | {< SNode >}
<associates>::= {} | {< SNode >}
<hasId> ::= {} | {< SNode >}
<references>::= {} | {< SNode >}
```

Figure 1. sGraph node model

Each peer has a schema expressed with a specific data model such as relational, XML, or object data model. To resolve the problem of heterogeneity of models of data sources we chose the common data model sGraph proposed in [27]. A sGraph represents a semantic graph composed of two sets N and E, where N represents the set of nodes of this graph. Each node represents an element of a data source schema and E represents the semantic links between nodes of this graph. Figure 1 gives an example of the model sGraph.

B. *Expertise*

At this step, we only consider data models supported by peers. We distinguish the three following data models, the best known: relational, XML and object. An expertise is defined, in our case, as (a part of) the data schema, expressed with one of the three data models cited above, possessed and published by a Peer in order to share its data with other peers. To facilitate the reconciliation between the data schema of the Peer and the theme described by a Super-Peer, two measures were taken: 1. the expertise of a Peer is expressed with the language of its Super-Peer (i.e. sGraph); 2. The expertise of a Peer is expressed under the format of couple of elements in sGraph, satisfying the following condition: EXP (Pi) = { θ (s_i; s_j) $\in$ SP | (s_i; s_j) $\wedge$ θ $\in$ R}.

IV. SEMANTIC'S ROUTING OF QUERIES - Baseline

A. *Network Configuration*

A new Peer P_j advertises its expertise by sending, to its Super-Peer, a domain advertisement DA_j = (PID; E^j_{XP} ; T_j; ε_{acc}; TTL) containing the Peer ID denoted PID, the suggested expertise E^j_{XP} , the topic area of interest T_j , the minimum semantic similarity value (ε_{acc}) required to establish semantic mapping between the suggested expertise E^j_{XP} and the theme of its Super-Peer. When receiving an expertise E^j_{XP} , a Super-Peer SP_a invokes the semantic matching process to find mappings between its suggested schema and the received expertise. The network is dynamic in the sense that when a super-peer SP leaves the network, the peers of SP attach to the closet super-peer.

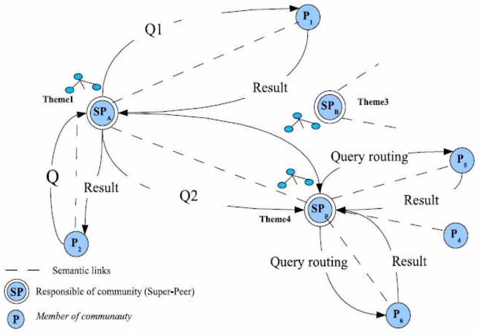


Figure 2. Network configuration and query routing (baseline approach).

B. *4.2 Baseline approach*

A Peer submits its query Q on its local data schema (expertise of peer). This query is sent to its Super-Peer responsible for the domain (see Figure 2). The Super-Peer in its turn suggests, based on the index obtained by the process of mediation (first level), the peers of its domain or the other super-peers that are able to treat this query. Each submitted query received by a Super-Peer, is processed by searching connections (second level of mappings) between the subject of this query and the expertise of peers (of the same domain) or the description of themes of other Super-peers. In its turn, a super-peer from the nearby domain, having received this request, researches among peers (in his domain) that are able to answer this query. This scenario takes time in processing queries. The major problem of this approach is the mediation at the two levels cited above: if we take thousands of peers or super-peers, this approach cannot be scaled due to the mappings at both levels. The following sections describe our approach in order to avoid super-peer, when it's too busy to treat all users' queries, to process the second level of mapping. This approach improves response times of queries and scalability in P2P context by restructuring the network dynamically. To do that, we introduce the concept of Global-Knowledge-Super-Peer (GKSP).

V. Global-Knowledge-Super-Peer

Our proposed System is a P2P system based on an organization of peers around super-peers according to their proposed themes. The super-peers are connected to a Global-Knowledge-Super-Peer (GKSP). GKSP is the engine that predicts the super-peers having peers, which may have relevant data to answer queries with minimum query processing using decision tree. By consequence, this improves the answering time of the queries issued from peers (see Figure 2). The super-peer architecture allows the heterogeneity of peers by assigning more responsibility to peers able to assume them. Therefore, certain Peers, called Global-Knowledge-Super-Peers, have an additional computing power and greater bandwidth, resources, and performing administrative tasks. They are responsible for routing queries to relevant super-peers, allowing not only to reduce efforts of compilation of queries but also to prevent the spread of queries in the network. In each domain, there is a super-peer connected to a Global-Knowledge-Super-Peer where we have an index to identify the most relevant Super-Peers to provide good results of queries. Otherwise, if the Global-Knowledge-Super-Peer didn't find the relevant super-peers form its index for a given query, it returns the query to its parents to work with the baseline in order to find the answer to this query. In order to see the effect of the data mining (decision tree) in the baseline network, we suggest to run our simulations in two configurations. One with the baseline connection between the super-peers (Hybrid Global-Knowledge-super-peers network (HGKSPN)), and else with non connection with the super-peers (Global Knowledge Network (GKN)).

VI. Effect of Data Mining in Baseline Apperoach

Our algorithm of semantic query routing is composed of three stages:

– In the first step (the step of baseline approach), the semantic routing algorithm exploits the expertise of (super-)peers and the two levels of mappings in order to forward a query Q to only relevant super-peers. Each super-peer in its turn forwards this query to relevant Peers of its domain. The followings sub-steps are necessary in order to process the query:

1. Extract the subject of this query (Sub(Q)(Q of peer P_2);
2. Select, by this super-peer (SP_A), the most relevant peers (P_1) for the query and the other super-peers (SP_P)(by matching the subject of the query to the set of expertise Exp(p_2) of peers or to the themes of super-peers). The selection is based on a function CAP that measures the capacity of a peer or a super-peer on answering a given query;

$$Cap(P,Q) = \frac{1}{\mathrm{Sub(Q)}} \sum_{S \in Sub(Q)} \underset{e \in Exp(P)}{Max} S_S(s,e) \qquad (3)$$

3. Once the set of relevant (super-)Peers has been identified, the Super-Peer sends the query to those promising peers or super-peers close to them by using their ID, IP addresses and the underlying physical network. The advantage of this step is that it permits us, for the second step, to collect information about the queries received by super-peers and the relevant super (-peers) selected in order to process it.

– The second step exploits the Hybrid Global-Knowledge-Super-Peers Network (HGKSPN) with the baseline approach. This step is very useful when the performance of the system is low. This step runs in four stages:

1. The Super-Peer (SP_6) sends the query (Q of P_1) directly to the Global-Knowledge-Super-Peer (GKSP);
2. The Global-Knowledge-super-peer (GKSP) identifies (without mapping) the relevant super-peers (SP_6, SP_8 and SP_5) of this query by consulting its index IND (obtained by applying decision tree algorithms);
3. Each selected super-peer (SP_6, SP_8 and SP_5) sends the query to relevant peers (P_1, P_3 and P_4);
4. If there is no result in the index of GKSP, then the GKSP returns the query to the super-peer (SP_6) to be treated with first step;
5. The final result of selected peers (P_1, P_3, P_4 and P_5) is returned (Index way + baseline way).

– The third step exploits the Global-Knowledge-Super-Peers (GKSP) network without the baseline approach (without any connection between super-peers). This step is very useful, where we can see the use of the data mining in the P2P context and its effects in the performance of our proposed system. This step runs in three stages:

1. The super-peer (SP_6) sends the query directly to the Global-Knowledge-Super-Peer (GKSP);

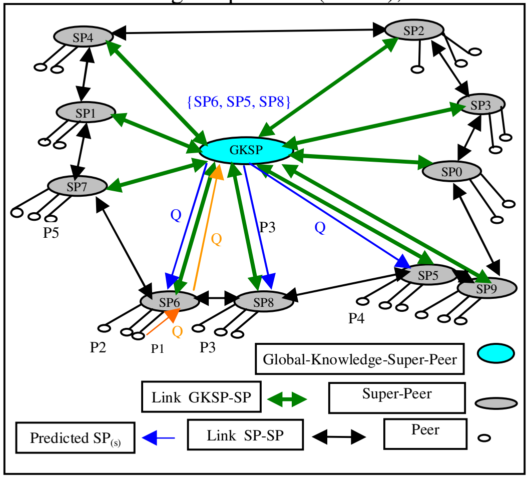


Figure 3. Hybrid Global Knowledge (super-)peers network (HGKSPN).

2. The Global-Knowledge-super-peer (GKSP) identifies (without make mapping) the relevant super-peers (SP$_6$, SP$_8$ and SP$_5$) for this query by consulting its index IND (obtained by applying decision tree algorithms);
3. Each selected Super-Peer (SP$_6$, SP8 and SP$_5$) sends the query to relevant peers (P$_1$, P$_3$ and P$_4$);
4. The final result of selected peers (P$_1$, P$_3$ and P$_4$) is returned (Index way only).

VII. EXPERIMENTAL AND EVALUATIONS

Decision trees represent a supervised approach of classification. The Weka classifier package has its own version of C4.5 known as J48. J48 is actually a latest version of C4.5. It was the last public version of this family of algorithms before the commercial implementation C5.0 [28] had been released. It builds a decision tree model by analyzing training data, and uses this model to classify user data. The decision tree is able to classify approximately 92% of the data correctly.

We describe the performance evaluation of our routing algorithm with a SimJava-based simulator [26]. We start the simulation by creating super-peers with their corresponding expertise and domain. Then, peers choose their domain to belong to a super-peer that gives the peer its expertise. Each super-peer creates a semantic mapping with other super-peers by sending part of its expertise to it. A component of query indicates the class of the examples in the decision tree structure. The instances are classified by sorting them down the tree from the first component of the query to other component of the query. All experiments were run on a machine Core 2 Duo 1.83GHZ with 4 GB RAM, 250 GB Hard disk and Windows Vista operating system. Evaluating the performance of P2P network is an important part to understand how useful it can be in the real world. As with all P2P applications, the first question is whether P2P is scalable. Our systems were evaluated with different set of parameters i.e. number of Peers, number of Super-peer etc. Evaluation results were quite encouraging. There are many dimensions in which scalability can be evaluated: one important metric is the answer time of a given query takes. We run simulations on P2P network in three different sizes. Peers are bounded to super-peers according to their interest. Each peer sends N queries to its SP, on its schema with its query language, which sends the query to a GKSP node in order to precise which Super-peer(s) can answer the given query, in the HGKSPN architecture. In our simulation, we took N=1.

- At first one, we modified the number of Peers (300, 600,..., 3000 Peers) and Super-peers (10, 12 ,14, 16, 20,..., 34) in both Architectures to measure the execution time.

The graphs shown in Figures 4, 5, 6 and 7 are the results of our simulations. They demonstrate the performance of using the GKSP with a decision tree to route Queries to relevant P2P Super-peers. In the first observation, the difference in the execution times between 300 and 900 peers in the HGKSPN architecture is small (See Figure 4). The execution time was measured as the repository size increased. Measurements, shown in Figure 4, show that the time, increased in the HGKSPN architecture, is less than the baseline architecture at 3000 Peers. The response time decreases about 23 % in the HGKSPN architecture compared to the baseline architecture. It is due to the processing of queries at SP level in baseline architecture. Otherwise, in the HGKSPN and GKN networks, the processing of queries is at GKSP level using decision tree. We observe that the time in GKN is also less then HGKSPN; this depends on the processing of queries at GKSP level only in GKN. Contrarily, in HGKSPN, we use the baseline method to process the queries not answered by the GKSP. So in the GKN, we win some time but we lose some precision concerning the queries responses. Measurements, in Figure 5, show the increasing number of messages in the GKN architecture is more then the baseline architecture. It is due to the presence of a higher layer “GKSP” increasing the number of messages.

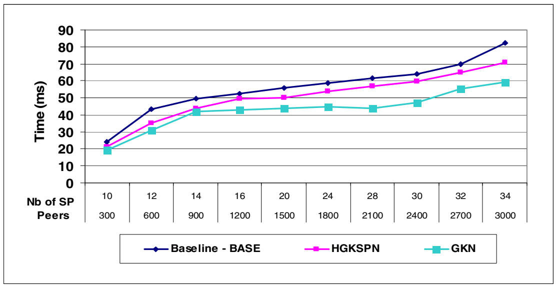


Figure 4. Responce time

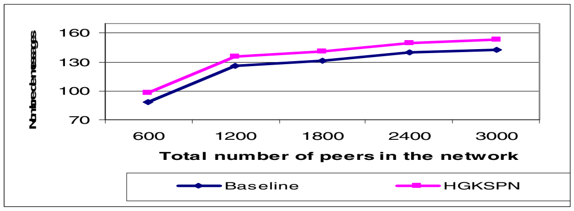


Figure 5. Number of Messages

- The most popular measure of information retrieval in our systems is the precision and recall.

$$\text{Precision} = \frac{\text{\# of relevant responses retrieved}}{\text{total \# of retrieved Responses}} \qquad (4)$$

$$\text{Recall} \quad = \frac{\text{\# of relevant responses retrieved}}{\text{total \# of relevant Responses}} \qquad (5)$$

Measurements, in Figure 6, show the precision of the GKN compared to the Baseline. We can observe that there is no big difference between these architectures (between 600 peers and 3000 peers), which means the stability of our architectures while increasing the number of peers and super-peers. This experiment is designed to measure the accuracy of data (since precision is almost not affected completely by the network size), which is the recall (See Figure 7). The recall increases with the size of the network and reaches a percentage of almost 90 % in the GKN architecture and about 91 % in the baseline architecture. Finally, our prototype raises some interesting performance issues in responding time of queries.

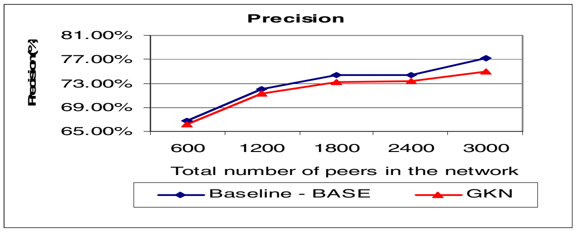


Figure 6. Precicion rate

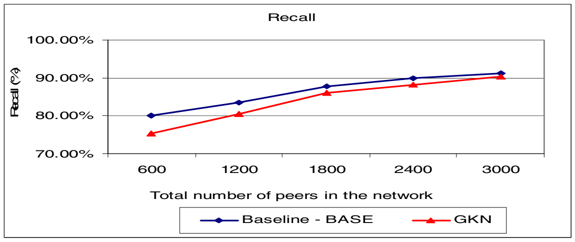


Figure 7. Recall rate

VIII. Conclusion

In general, efficient search in P2P networks could be improved by applying data mining in query routing. Classification based on decision trees is one of the important problems in data mining. In this paper, we proposed an architecture using data mining (decision tree) for query routing in P2P networks. Our experimental results show that the integration of the data mining in the P2P network improves the answering time of a given query and therefore gave us a high performance without consuming the bandwidth. The advantage of this architecture is the robustness in Queries routing in P2P Network, respecting very importing issues such as answering time of queries.

Acknowledgment

This work has been done as a part of the project "recherché intelligente d'information multimedia multilingue arabe" by the franco-libanais comity Cedre.

References

[1] S. Androutsellis-Theotokis. A survey of P2P file sharing technologies. Electronic Trading Research Unit (ELTRUN), Athens University of Economics and Business; 2002.

[2] S. Joseph, NeuroGrid: semantically routing queries in P2P networks. In: Revised papers from the networking workshops on web engineering and peer- to-peer computing. London, UK: Springer; 2002, pp. 202–14.

[3] J. K. Kim, H.K. Kim, and Y. H. Cho, A user-oriented contents recommendation system in P2P architecture. Expert Systems with Applications, 34(1), 2008, pp. 300–312.

[4] C. Niu, J. Wang, R. Shen, L. Shen, and H. Luo, Cooperativeness prediction in P2P networks. Expert Systems with Applications. 35(3), 2008, pp. 1267-1274.

[5] P. Küngas and M. Matskin, Semantic web service composition through a p2p-based multi-agent environment,“ Springer-Verlag, Lecture Notes on Computer Science, vol. 4118, 2006, pp. 106-119.

[6] S. Datta, K. Bhaduri, C. Giannella, R. Wolff, and H. Kargupta, Distributed data mining in P2P networks. IEEE Internet Computing, Special issue on Distributed Data Mining 10(4), 2006, pp. 18-26.

[7] B. Raahemi, A. Hayajneh, and P. Rabinovitch, P2P IP Traffic Classification Using Decision Tree and IP Layer Attributes, International Journal of Business Data Communications and Networking, Vol. 3, Issue 4, 2007, pp. 60-74.

[8] K. Bhaduri, R. Wolff, C. Giannella, and H. Kargupta, Distributed Decision Tree Induction in P2P Systems. Statistical Analysis and Data Mining Journal, Vol. 1 , Issue 2, 2008, pp. 85-103.

[9] F. Emekci, O.D. Sahin, D. Agrawal, and A. El Abbadi, Privacy preserving decision tree learning over multiple parties, Data & Knowledge Engineering, Vol. 63, Issue 2, 2007, pp. 348-361.

[10] Y. Xu and H. Qi, Dynamic mobile agent migration in Wireless Sensor Networks, International Journal of Ad Hoc and Ubiquitous Computing - Vol. 2, No.1/2, 2007, pp. 73 – 82.

[11] C. Tempich, A. Löser, and J. Heinzmann, Community Based Ranking in P2P Networks. In Proceedings of the 4th International Conference on Ontologies, Databases and Applications of Semantics, 2005, pp. 1261-1278.

[12] C. Tempich, S. Staab, and A. Wranik, REMINDIN': Semantic Query Routing in P2P Networks Based on Social Metaphors. In Proceedings of the 13th International World Wide Web Conference, 2004, pp. 640-649.

[13] G. Kokkinidis and V. Christophides, Semantic Query Routing and Processing in P2P Database Systems: The ICS-FORTH SQPeer Middleware. In Proceedings of the 3rd Hellenic Data Management Symposium, 2004, pp. 486-495.

[14] W. Nejdl, M. Wolpers, W. Siberski, A. Löser, I. Bruckhorst, M. Schlosser, and C. Schmitz, Super-Peer-Based Routing Strategies for RDF-Based P2P Systems. In Proceedings of the 2nd International Workshop On Databases, Information Systems and P2P Computing, Toronto, Canada, September 2004.

[15] O. Papapetrou, Full-text indexing and Information Retrieval in P2P Systems, ACM International Conference Proceeding Series, EDBT, Nantes, France; Vol. 326, 2008, pp. 49-57.

[16] H. Nottelmann and N. Fuhr, A decision-theoretic model for decentralised query routing in hierarchical P2P networks. In ECIR, 2007, pp. 148-159.

[17] S. Sharma, L. T. Nguyen, and D. Jia, IR-Wire: A Research Tool for P2P Information Retrieval. In Proc. ACM Wkshp. Open Source Inf. Retr., 2006.

[18] M. Bender, S. Michel, P. Triantaffllou, G. Weikum, and C. Zimmer, MINERVA: Collaborative P2P Search. In VLDB, 2005, pp. 1263-1266.

[19] J. Lu and J. Callan, Federated Search of Text-based Digital Libraries in Hierarchical P2P Networks. n Advances in Information Retrieval, 27th European Conference on IR Research In ECIR, 2005, pp. 52-66.

[20] J. Stribling, I. Councill, J. Li, M. Kaashoek, D. Karger, R. Morris and S. Shenker, Overcite: A Cooperative Digital Research Library. In IPTPS, 2005.

[21] M. Bender, S. Michel, P. Trianta_llou, G. Weikum, and C. Zimmer, Improving Collection Selection with Overlap-Awareness. In SIGIR, 2005, pp. 67 - 74 .

[22] I. Aekaterinidis and P. Trianta_llou, PastryStrings: A Comprehensive Content- Based Publish/Subscribe DHT Network. In ICDCS, 2006.

[23] C. Tryfonopoulos, S. Idreos, and M. Koubarakis, Publish/Subscribe Functionality in IR Environments using Structured Overlay Networks. In SIGIR, 2005, pp. 322-329.

[24] C. Zimmer, C. Tryfonopoulos, K. Berberich, M. Koubarakis, and G. Weikum, Approximate Information Filtering in P2P Networks. In WISE, 2008, pp. 6-19.

[25] C. Zimmer, C. Tryfonopoulos, and G. Weikum, MinervaDL: An Architecture for Information Retrieval and Filtering in Distributed Digital Libraries. In ECDL, 2007, pp. 148-160.

[26] F. Howell and R. McNab, "simjava: a discrete event simulation package for Java with applications in computer systems modelling", in proc. First International Conference on Web-based Modelling and Simulation, San Diego CA, Society for Computer Simulation, 1998.

[27] D. Faye, G. Nachouki, and P. Valduriez. Semantic query routing in senpeer, a p2p data management system. NBIS, 2007, pp. 365-374.

[28] J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kauffman, 1993.

[29] Peter Haase, Ronny Siebes, Frank van Harmelen: Peer Selection in P2P Networks with Semantic Topologies. ICSNW, 2004, pp. 108-125.

[30] Napster homepage. http://www.napster.com.

[31] Gnutella homepage. http://www.gnutella.com.