© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A SCALABLE SEMANTIC PEER-TO-PEER SYSTEM

[1]MOHAMMED AMMARI, [2]DALILA CHIADMI
Mohammadia School of Engineers
University Mohammed V Agdal Rabat Morocco
E-mail: [1]ammari@um5a.ac.ma, [2]chiadmi@emi.ac.ma

ABSTRACT

Peer-to-peer computing is emerging as a new distributed computing paradigm for many novel applications that involve exchange of information among a large number of peers with little centralized coordination. Scalability is without doubt the foremost requirement for a peer-to-peer system. To obtain a high factor of scalability, we partition network search space using a global ontology. Our proposed system takes the form of a semantic layer that can be superimposed on top of any P2P infrastructure. This layer is subdivided as semantic categories through a Hilbert curve which has the merit of good preservation of locality. We choose HyperCup structure to support semantic category in order to increase fault tolerance and overcome bottleneck problem. HyperCup was also selected, because of its efficient broadcast algorithm. Classification within a semantic category will be conducted through multidimensional data analysis. The classification process will be recursively repeated with ontology refinement. We use digital libraries to illustrate our system, and highlight the proposed techniques.

Keywords: *P2P, ontology, HyperCup, Semantic classification, multidimensional data analysis, Correspondence analysis.*

1. INTRODUCTION

Peer-to-peer computing is emerging as a new distributed computing paradigm for many novel applications that involve exchange of information among a large number of peers with little centralized coordination. Grid-computing, massive data storage, instant messaging, videoconferencing, voice over IP, multi-player are some examples. We can classify peer to peer networks as structured or unstructured, depending on the way they are connected and how the data they contain is arranged. In a structured network the connections between nodes are of some regular structure, which allows deterministic and optimal lookup hops (typically O (log N)). In contrast to structured networks, nodes in unstructured networks do not share a regular structure and a unified identifier space. Lookups are thus normally achieved by flooding and using replication in the network.

Scalability is without doubt the foremost requirement for a peer-to-peer system. To obtain a high factor of scalability, we partition network search space using a global ontology. Our proposed system takes the form of a semantic layer that can be superimposed on top of any P2P infrastructure. This layer is subdivided as semantic categories through a Hilbert curve which has the merit of good preservation of locality. We choose HyperCup structure to support semantic category in order to increase fault tolerance and overcome bottleneck problem. HyperCup was also selected, because of its efficient broadcast algorithm. To avoid creating bottlenecks, we will use classification in the various semantic categories. Indeed a division of each semantic category in more specialized sub-categories is essential, which will give increased scalability to our system, and allow it to manage complexity by structuring. The classification within a semantic category will be conducted through multidimensional data analysis. At the end of this classification, the Hilbert curve iteration degree will be increased by a notch in order to support the new emerging classes. This process will be recursively repeated with ontology refinement.

One of the main advantages of our scheme is semantic lookup. For example, suppose that a user wants to find some paper and thesis and technical reports on P2P networks. Instead of sending at least 3 lookup queries, she or he simply annotates the

© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

query by publication (parent concept of paper, thesis and technical report).

We use digital libraries to illustrate our system, and highlight the proposed techniques. In the following section we discuss some related work. In the third section we present our solution before closing on experimental analysis.

2. RELATED WORK

Our work can be classified in different ways. First of all, we introduce an ontology-based DHT. So, all DHT-based P2P overlay networks like CAN [1], Chord [2], Pastry [3], and Tapestry [4] are related to ours. The main difference of our work with previous DHTs is that they are semantic-free. So, they cannot support semantic queries. But, our method is semantic based and can support conceptual queries. Another advantage of our proposal in contrast of the mentioned DHT-based networks, is that the lookup hops number in our network do not depend on the network size and just depends on the Hilbert curve iteration degree. While in the DHT-based networks, cost of lookup increases by growth of the network.

On other side, our work can be classified in the field of metadata semantic indexing. We use a global ontology to organize the search space, so all semantic-based P2P networks can be considered as related works. In the following, we review some of them.

In a work done by Schlosser et al. [5], they changed HyperCup[6] structure by semantic relation between participant nodes. They changed the node joining algorithm, so that nodes that have similar services are placed together in concept clusters. In other words, each main hypercube node is a concept cluster that itself consists of several nodes (computers). The nodes in each cluster are organized in a hypercube or star structure. The dimension of hypercube is chosen with respect to structuring concepts. In the mentioned work, some general concepts are chosen as base concepts and each node instead of pointing to a computer, points to a concept cluster which itself consists of a number of computers. For example, if we have the concepts A, B and C then we will have a three-dimensional hypercube that node 011 shows that the computers in this concept cluster cover concepts B, and C but do not cover concept A. In their method, the number of hypercube dimensions depends on the number of base concepts that are provided in the network. Also, the nodes that are grouped together in one concept cluster are close in meaning of contents that they provide but not in the physical underlying topology location. Real world nodes that are linked together may have a long distance between them in terms of latency and physical traffic. So, the communication cost between them is high and the average transfer speed over the links in the network is lower than the original HyperCup. The main advantages of our work against [5] are that we support lookup operation while they do not support it.

Meanwhile, network Scalability and robustness are common issues on P2P. For example, reduction of communication traffic traversed across ASs (Autonomous Systems) is one of the recent research topics, since it was found that P2P applications occupied a large percentage of the bandwidth between ASs. For traffic reduction across ASs, a hierarchical P2P architecture can be utilized [7] [8]. In the architecture, network peers located in the same AS compose a cluster network, and a super peer is elected for each cluster network. The super peers compose a (top-level) P2P network across ASs. Since the number of peers in the top-level P2P network decreases communication traffic traversing across ASs can be reduced. However, the architecture has harmful effects in terms of reliability since a large amount of traffic is concentrated on the super peers which then become a single point of failure (bottleneck problem).

In a work done by Ammari et al. [9], they proposed a semantic layer that can be superimposed on top of P2P based on distributed hash table. In this article, we propose to extend this solution to all P2P systems. In addition, we propose to change semantic category structure, from super peer to HyperCup in order to increase fault tolerance and overcome bottleneck problem. HyperCup was also selected, because of its efficient broadcast algorithm.

Rostami et al. [10] partition network search space using a consistent static shared upper ontology. They name their approach semantic partition tree (SPT). All resources and queries are annotated using the upper ontology and queries are semantically routed in the overlay network. Also, each node indexes addresses of other nodes that possess contents expressible by the concept it maintains. So, their approach can be conceived as ontology based distributed hash table (DHT). The main advantages of our work against [10] are that we propose the use of multidimensional data analysis to ensure semantic indexing of metadata. Also, our semantic layer is subdivided as semantic

Journal of Theoretical and Applied Information Technology
31st March 2013. Vol. 49 No.3
© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

categories through a Hilbert curve which has the merit of good preservation of locality and semantic affinity.

The paper [11] presents the design and implementation of Bibster, a solution fully implemented, open source, built on top of the JXTA platform. It is a P2P system of bibliographic data exchange among researchers. It uses ontology for data storage, formulation and routing queries, and for presentation of responses. But this system is being optimized due to the attractive alternatives of different underlying modules.

The paper [12] proposes a scalable CGM (Consumer Generated Media) distribution system that delivers newly published content in real time to the appropriate users whose preferences coincide with the content topic. In the proposed system, a semantic P2P network is constructed based on CAN (Content Addressable Network) to distribute content according to the similarity between user preferences and the content. In order to enhance the system, further investigations can be conducted on traffic optimization and seamless network reformation.

Push-type information delivery platforms based on P2P technology were proposed by Ogino [13] and Kacimi [14]. In the system, user context is managed over a distributed hash table (DHT). After publishers retrieve appropriate users from the network, they push information to the users. According to Ogino [13], while various types of user contexts (e.g., sex, age, location, or used terminals) can be supported, the system assumes only the exact matching of the user context. Thus, similarity of user preferences and information cannot be handled as criteria for information delivery.

On the third side, we propose the use of multidimensional data analysis in P2P context. So our paper can be also classified in the field of distributed data mining. Indeed distributed data mining is gaining increasing attention in P2P systems for advanced data driven applications. Data analysis in P2P environments offers a wide spectrum of challenges for the researchers and practitioners. Designing distributed, asynchronous, decentralized algorithms offers many difficult challenges [15] [16]. However, maturing these algorithms and integrating them with real P2P applications offer additional challenges. Most of the P2P data mining algorithms rely upon asymptotic convergence properties. Advanced analysis of such systems such as stability must be quantified. High factor of scalability is the main difference of our work against the previous systems.

3. SOLUTION OVERVIEW

Our system takes the form of a semantic layer that can be superimposed on top of any P2P infrastructure. This layer is subdivided as areas through a Hilbert curve which has the merit of a good preservation of locality and semantic affinity. A Hilbert curve is a continuous fractal space-filling curve first described by the German mathematician David Hilbert in 1891. It is demonstrated that the trajectory is dense in the square of departure, therefore equal to the square. The Hilbert curve has been proposed in the context of multidimensional databases because it has a better behavior for preserving locality.

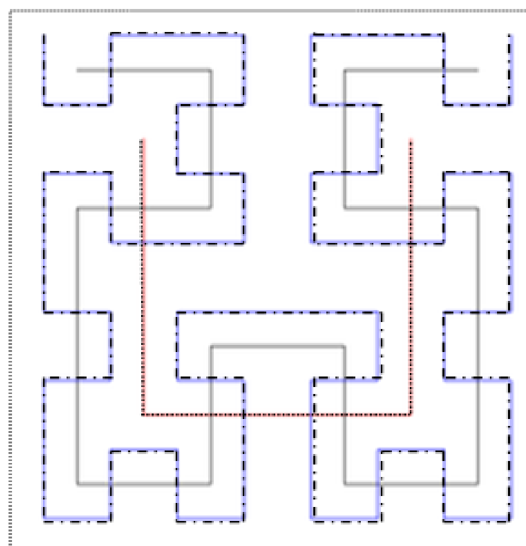

Figure1. First, second and third iteration of the Hilbert curve

We will make a correspondence between the semantic domains extracted from a global ontology and Hilbert areas (see Figure 2). The system is divided in the form of semantic categories. A category is a semantic grouping of records with high internal cohesion and low external coupling. Each semantic category is responsible for managing semantically related records under its jurisdiction. Our scheme is an ontology-based distributed hash table. We mean by this that shared resources in the network are hashed based on their semantics and not just using a semantic-free hash function like SHA-1 [17].

Journal of Theoretical and Applied Information Technology
31st March 2013. Vol. 49 No.3
© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

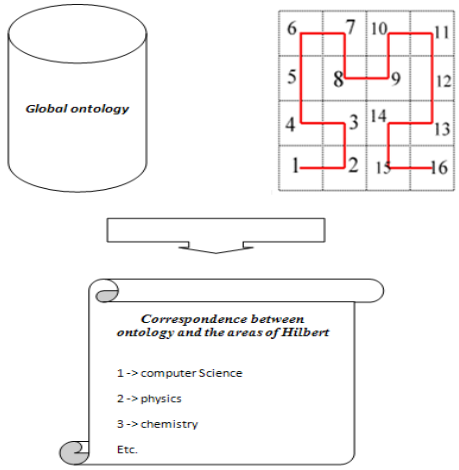

Figure2. Correspondence between the global ontology and areas of Hilbert

As we denoted, each partition of the search space is corresponding to a concept in global ontology. Nodes that reside in a partition, keep addresses of all nodes, scattered in the network, that possess a resource expressible by the concept corresponding to this partition. If there are more than one node in a partition, they form an HyperCup [6] in the partition. HyperCup was selected, because of its efficient broadcast algorithm. It should be noted that when the hypercube is not complete, for example we have five nodes; the hypercube constructing algorithm adds some virtual nodes and creates a complete hypercube. So, in the broadcast algorithm, we always suppose that the hypercube is complete.

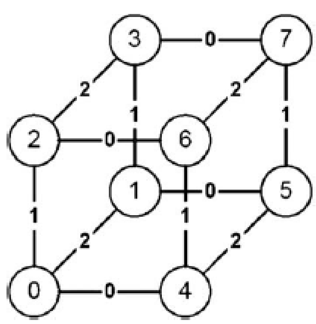

Figure3. Three-dimensional hypercube

HyperCup guarantees that the set of traversed nodes strictly increases during a forwarding process, i.e. nodes receive a message exactly once. It guarantees that exactly N-1 messages are required to reach all nodes in the overlay. Furthermore, the last nodes are reached after $\log_2^N$ forwarding steps.

The algorithm works as follows. A node invoking a broadcast sends the broadcast message to all its neighbors, tagging it with the edge label on which the message was sent. Nodes receiving the message restrict the forwarding of the message to those links tagged with higher edge labels. For example, in Fig. 3, assume that node 0 sends a broadcast- at first to all its own neighbors, viz. nodes 4, 2 and 1. Node 4 receives the message on a link tagged as a level 0 link. It forwards the message only to its 1- and 2-neighbors, namely 6 and 5. At the same time, node 2 that has received the message on a level 1 link forwards it to its 2-neighbor, node 3. In the third forwarding step, node 6 relays the message to node7, its 3-neighbor.

All nodes of a semantic category are mirror of each other, in terms of indexed concepts. In other words, all of them maintain same resource information. If the resource information of any node, in a category, changes, it broadcasts changes to other nodes of its category. We should note that nodes of a semantic category may possess different resources, but all of them index the same resources scattered in the network. The big picture of the network is shown (see Figure 4).

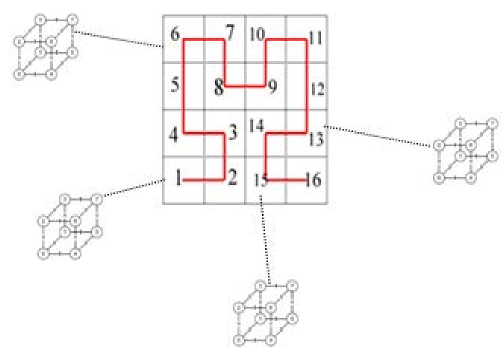

Figure4. Big picture of the proposed scheme

To avoid creating bottlenecks, we will use classification in the various semantic categories. Indeed a division of each semantic category in more specialized sub-categories is essential, which will give increased scalability to our system, and allow it to manage complexity by structuring. The classification within a semantic category will be conducted through multidimensional data analysis. At the end of this operation of classification, the Hilbert curve iteration degree will be increased by a notch in order to support the new emerging classes (see Figure 5).This process will be recursively repeated with ontology refinement.

Journal of Theoretical and Applied Information Technology
31st March 2013. Vol. 49 No.3
© 2005 - 2013 JATIT & LLS. All rights reserved.
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

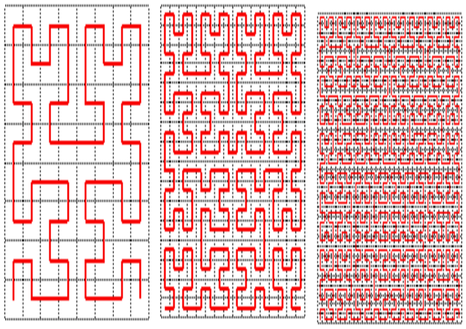

Figure5. Evolution of the Hilbert curve with ontology refinement

We have seen that each semantic category will act as a containment vessel, where the various semantically related documents will be embedded in. Indeed, each semantic category will manage a matrix of contingency ensuring correspondence between I the set of data sources and J the set of terms related to the semantic domain. The contingency matrix is defined by $F = \left(f_{ij}\right)_{(i,j)\in[1,n]*[1,p]}$ where f_{ij} is the frequency of occurrence in the entire population of the couple (i, j) formed by extracted term $T_j, j \in J$ and the data source $S_i, i \in I$. To prevent the genesis of bottlenecks, and overcome the problem of storing a huge and hollow contingency matrix, we will use data sources classification through the multidimensional data analysis. In our case we use factor analysis. Factor analysis is a collection of methods used to examine how underlying constructs influence the responses on a number of measured variables. Factor analyses are performed by examining the pattern of correlations (or covariances) between the observed measures. Measures that are highly correlated (either positively or negatively) are likely influenced by the same factors, while those that are relatively uncorrelated are likely influenced by different factors.

We now have the triplet which allows us to perform a factor analysis:

- a scatter plot $\mathrm{N}(I) = \left\{X_i \in R^n, i \in I\right\}$

Simplex distribution is the set of points X of R^n which has the coordinates x_j defined by: $x_j \geq 0$ satisfying $\sum_{j=1}^{n} x_j = 1$.

Each data source S_i is associated with the point X_i in the n-dimensional space, which has the following coordinates

$$x_{ij} = \frac{f_{ij}}{f_{i.}}$$

$f_{i.}$ Data source frequency in the entire population. We prove that:

$$\sum_{j=1}^{p} x_{ij} = \sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}} = \frac{\sum_{j=1}^{p} f_{ij}}{f_{i.}} = 1$$

So each data source is located on the simplex of distribution R^n

- A mass distribution $X_i \rightarrow f_{i.}$

At each point X_i we affect the mass

$$f_{i.} = \sum_{j=1}^{P} f_{ij}$$

Therefore, we find information size concerning the data source. This information has been lost in the transition from the source S_i to the point Xi

- Distance χ^2 to measure the similarity between points.

Indeed, to calculate the distance between $X_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)^t$ and $X_{i'} = \left(x_{i'1}, x_{i'2}, \ldots, x_{i'p}\right)^t$ we reduce by $G = \left(f_{.1}, f_{.2}, \ldots, f_{.p}\right)$

$$d^2\left(X_i, X_{i'}\right) = \sum_{j\in J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2$$

with $f_{.j}$ represents the frequency of the term T_j in the entire population.

The distance χ_2 was chosen because it satisfies the distributional equivalence principle (Benzécri, 1973: vol. I, p. 23). This principle can be stated in a simplified form as follows: if two columns (resp., two rows) have the same relative values, then

Journal of Theoretical and Applied Information Technology
31st March 2013. Vol. 49 No.3
© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

merging them does not affect the distances between rows (resp., columns).

The integration of new data sources is made as follows: The information element to insert spreads gradually to the way of an epidemic in a human group. This element can spread to different semantic categories through an epidemic protocol [18]. Any new data source to insert will be considered as a fictitious source $S_{fictive}$ in each semantic category. The data source to insert can belong to one or more semantic categories. For example a record in the field of biochemistry may belong to semantic categories "Chemistry" and "Biology". This will allow our system to manage the overlap between the different areas.

The search for relevant data sources within a semantic category will be made according to the algorithm below:

- Receiving a request in the form of XQuery.
- Extraction of terms from the predicate.
- Semantic enrichment of terms extracted from the predicate, with semantically related terms from an ontology constructed from the data schemas of semantic category.
- Construction of the sub matrix of contingency $F = \left(f_{ij}\right)_{(i,j)\in[1,n]*[1,p]}$
- Discharge of sources S_i checking :

$$\forall j \in J \quad \frac{f_{ij}}{f_{i.}} < \upsilon_{pertinence}$$

Where $\upsilon_{pertinence}$ represents the relevance threshold parameter set by the system administrator.

- Ranking of selected data source, with relevance order.

4. EXPERIMENTAL ANALYSIS

We use digital libraries to illustrate our proposals, and highlight classification techniques based on Correspondence analysis (Benzécri, 1973; Greenacre, 1984, 1993a; Lebart, Morineau and Warwick, 1984). Indeed the Correspondence analysis is one of a family of methods based on the singular value decomposition, and has become a standard method for graphically displaying tables of nonnegative data. The method is particularly popular in the social and environmental sciences for analyzing frequency data (see, for example, Greenacre and Blasius (1994) and ter Braak (1985) respectively). As emphasised by Benzécri, who originally developed correspondence analysis as a method for exploring large tables of counts in linguistics.

We choose "Computer Science" as a semantic category. We consider that we have 10 data sources, and we restrict the relevant candidate terms to 14, J = {ROM, BUS, ASP, PHP, C, Oracle, MySQL, Drivers, CPU, C + +, Linux, RAM, Compilation, JAVA}. The frequency of keyword occurrence in a data source is evaluated according to Likert scale from 1 to 9 (see table1). The simulation is performed using the software XLSTAT 2011.

Table1. Contingency matrix ensuring the correspondence between sources and terms.

	ROM	BUS	ASP	php	C	Oracle	MySQL	Drivers	CPU	C++	Linux	RAM	Compil	Java
S1	6	6	3	2	4	4	3	4	5	4	4	5	4	3
S2	4	4	4	4	5	5	4	4	4	4	4	4	4	4
S3	5	5	4	3	4	4	2	5	6	3	4	5	4	4
S4	6	4	2	2	4	6	4	4	5	3	3	6	5	4
S5	4	4	4	4	4	5	4	4	3	4	5	4	4	4
S6	5	5	3	3	3	5	3	4	5	4	5	6	4	3
S7	5	5	4	3	3	4	3	6	5	2	4	5	5	3
S8	4	3	4	4	5	6	4	3	4	4	3	4	4	4
S9	6	5	2	2	3	4	4	4	6	3	4	6	4	3
S10	4	3	4	4	4	5	4	4	3	4	5	4	5	4

The interest of the Correspondence Analysis is to provide a graphical representation of scatter plots I and J (dual cloud) in a space with lower dimension. The aim is to highlight the hidden correspondences that the numerical calculation does not immediately reveal. Traditionally, the representation is done in two dimensions by plotting the first factorial plane. The quality of the graphical representation can be evaluated by the eigenvalues histogram. If the sum of the first two eigenvalues represents a large portion of the total inertia, the quality of graphics is considered good. In our case, the quality is good since the first two eigenvalues account for 73.83% of the total variance (see Figure 6).

Journal of Theoretical and Applied Information Technology
31st March 2013. Vol. 49 No.3
© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

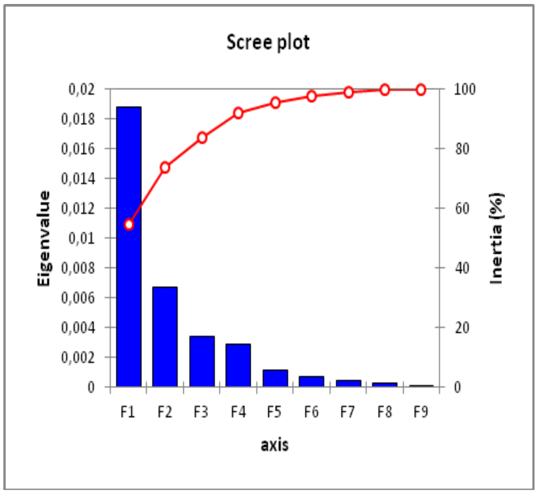

Figure6. Histogram of eigenvalues

The main advantage of correspondence analysis is simultaneous graphical representation of the terms of two variables. When the quality is good (73.83% in our case) the graphical representation can easily interpret the data. The proximity between two points-rows (data sources) (such as S10 and S5) reflects a similar keywords profile (see Figure 7). The proximity between two points-columns (keywords) (such as C, C + +, Java ...) reflects a similar profile of data sources. The simultaneous representation of points-rows and points-columns is used to identify the variables responsible for some proximities. This can be exploited to search efficiently. Thus the relevant data sources will be close - as defined in the distance χ^2 - of the keywords used in a search.

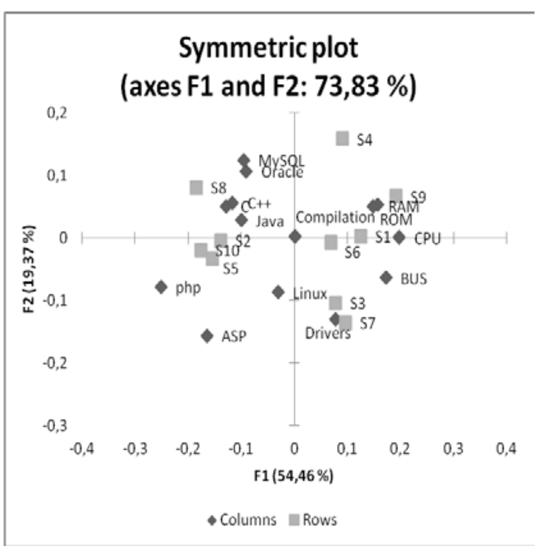

Figure 7. Simultaneous graphical representation of data sources and keywords

We can note that the sources S8, S10, S5 and S2 rather cover the area of Software (databases, programming languages,etc.) while the data sources S4, S9, S1, S6, S3, S7 tend to cover the area of hardware (CPU, bus, RAM, ROM, etc.). That said the semantic category “Computer Science” will be divided into two semantic categories, more specialized and smaller (“Software” and “Hardware”). So, data sources classification allow us to overcome the problem of storing a huge and hollow contingency matrix, and permits us to manage complexity by structuring.

5. CONCLUSION

This manuscript studies the semantic query application in P2P systems. A scalable network partitioning solution is proposed to deal with node lookup issue. Our proposed system takes the form of a semantic layer that can be superimposed on top of any P2P infrastructure. The aim is to confer the semantic aspect to P2P in order to respond to semantic queries. We used multidimensional data analysis to ensure semantic indexing of metadata. We opted for a Hilbert curve to support the emergence of new more specialized classes. We chosed HyperCup structure to support semantic category in order to increase fault tolerance and overcome bottleneck problem. HyperCup was also selected, because of its efficient broadcast algorithm.

REFRENCES:

[1] S. Ratnasamy, M. Handley, R. Karp, and S. Shenker (2001) A scalable content-addressable network. *In Proc. ACM SIGCOMM*.

[2] I. Stoica, R. Morris, D. Karger, M. Kaashoek, and H. Balakrishnan (2001) Chord: A scalable peer-to-peer lookup service for internet applications. *In Proc. ACM SIGCOMM*.

[3] A. Rowstron, and P. Druschel (2001) Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. *In Proceedings of Middleware*.

[4] B.Y. Zhao, J. Kubiatowicz, and A.D. Joseph (2001) Tapestry: An infrastructure for fault tolerant wide-area location and routing. *Technical Report UCB/CSD-01-1141*, University of California, Berkeley.

© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[5] M. Schlosser, M. Sintek, S. Decker, W. Nejdl (2002) Hypercup – hypercubes,ontologies and efficient search on p2p networks, *in: International Workshop on Agents and Peer-to-Peer Computing*, Bologna, Italy.

[6] M. Schlosser, M. Sintek, S. Decker, W. Nejdl (2001) Hypercup shaping up peer-to-peer networks, *Technical Report, Stanford University*.

[7] K. Shin et al. (2009) "A Proximity-Based Self-Organizing Hierarchical Overlay Framework for Distributed Hash Tables." *IEICE Trans. on Fundamentals*, pp. 1651-1662, B Vol. E90-B, No. 7.

[8] I. Martinez-Yelmo et al. (2008) "Routing Performance in a Hierarchical DHT-based Overlay Network", *Proceeding of 16th Euromicro Conference on Parallel, Distributed and Network-Based Processing*.

[9] M. Ammari, D. Chiadmi, L. Benhlima (2011) A semantic layer for a Peer-to-Peer based on a distributed hash table. *The International Conference on Informatics Engineering and Information Science*, Kuala lumpur Malaysia.

[10] H. Rostami, J. Habibi, E. Livani (2009) Semantic partitioning of peer-to-peer search space, *Computer Communications*.

[11] J. Broekstra, P. Haase, F.V. Harmelen, M. Menken, P. Mika, B. Schnizler, R. Siebes (2005) Bibster-a semantics-based bibliographic peer-to-peer systems, *in: The Second Workshop on Semantics in Peer-to-Peer and Grid Computing*.

[12] T. Warabino and Y. Kishi, (2011) Scalable CGM distribution system based on a semantic peer-to-peer network. *Asia-Pacific Conference on Communications*, pp. 399-404

[13] N. Ogino et al. (2009) "Performance Evaluation of Distributed Context Management System Based on Peer-to-Peer Technology." *IEICE Trans. on Fundamentals*, pp711-728, B Vol. J92-B, No.4.

[14] M. Kacimi et al. (2008) "Content-Based Information Routing and Retrieval in Cluster-based P2P Overlay networks." *Proc. of 2008 IEEE International Conference on Signal Image Technology and Internet Based Systems*.

[15] R. Wolff, K. Bhaduri, H. Kargupta (2005) Local 12 thresholding based data mining in peer-to-peer systems.*Technical Report TR-CS-05-11*, Department of Computer Science, UMBC.

[16] S. Datta, C. Giannella, H. Kargupta (2005) K-means clustering over peer-to-peer networks. *In Proceedings of the 8th International Workshop on High Performance and Distributed Mining (HPDM'05). In conjunction with the SIAM International Conference on Data Mining*.

[17] D. Eastlake, P. Jones (2001) Us secure hash algorithm1 (sha1). Available from: *<http://tools.ietf.org/html/rfc3174>*.

[18] M. Jelasity, A. Montresor, and O. Babaoglu. (2009) T-Man: Gossip-based fast overlay topology construction. *Computer Networks*, 53(13):2321–2339.