Journal of Theoretical and Applied Information Technology
20th May 2014. Vol. 63 No.2
© 2005 - 2014 JATIT & LLS. All rights reserved·

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MDPM: AN ALGORITHM FOR MAPPING DISCOVERY IN P2P MEDIATION SYSTEM

[1]SELMA EL YAHYAOUI EL IDRISSI,[2] AHMED ZELLOU,[3] ALI IDRI
[1] Laboratory SIME, ENSIAS. Mohammed V- Souissi University, Morocco
[2] Laboratory SIME, ENSIAS. Mohammed V- Souissi University, Morocco
[3] Laboratory SIME, ENSIAS. Mohammed V- Souissi University, Morocco
E-mail: [1]yahyaoui.selma@gmail.com, [2] zellou@ensias.ma, [3] idri@ensias.ma

ABSTRACT

The information integration systems consist in offering a uniform interface, to provide access to a set of autonomous and distributed information sources. The most important advantage of an information integration system is that allows users to specify what they want, rather than thinking about how to get the responses. The studies in this field have aimed at developing tools allowing a transparent access to data sources disseminated on a network. In particular, there are two major classes of integration systems: the mediation systems based on the paradigm mediator/Wrapper and peer to peer systems (P2P). Recently, other integration systems have emerged as P2P mediation systems. In these systems, the correspondence problem between schemas is a crucial problem. Especially as to integrate different sources requires the identification of the elements which can be dependent between the various diagrams. This is called correspondences or mapping between schemas.

In this paper, we study this problem of mappings discovery and we present an approach of automatic correspondences discovery in a Pure Peer-to-Peer mediation system.

Keywords: *Mediation, P2P System, Mapping, Correspondences.*

1. INTRODUCTION

The peer-to-peer environment is currently a viable solution to allow scale up of the internet. Each peer behaves as both client and server, and provides a part of the whole information of the distributed environment without relying on a central administration. Furthermore, the information sources integration [1] has became a very important field of research because of the explosion sources number and their heterogeneities. The objective is to give the impression of using a homogeneous and centralized system. Two main approaches for the integration systems have been defined based on the localization of the data managed by the system: when the data sources are stored in the integration system we talk about materialized approach or data warehouse [2], On the contrary, when the integrated data are not we talk about virtual approach or mediation system [3].

Recently, P2P mediation systems have emerged. They combine Peer to Peer technology [4] and that of distributed databases and rely on a description of the information sources; we can quote the PIAZZA [5], SomeWhere [6] And PeerDB approach [7]. It comes to mediation systems without global schema where each peer has its own local schema and correspondence between its schema and schemas of the other peers.

The overall structure is as follows. In Section 2, we present the mediation systems and the P2P systems in section 3. In Section 4 we treat the mediation P2P systems which are a combination of two systems discussed in the previous section i.e. P2P and that of mediation. Section 5 presents our proposed approach for discovery mapping in a P2P mediation system. Section 6 provides the detailed validation of our approach. Section 7 summarizes the contribution of this paper.

2. MEDIATION SYSTEM

The mediation approach [8] consists of defining an interface between the agent (human or software) that poses a query and all potentially relevant accessible sources to respond to the query. The objective is to give the impression to interrogate a centralized and homogeneous system while sources interviewed are distributed, autonomous and heterogeneous. This approach presents the interest to be able build an information sources integration system without touching the data remaining in their original sources. The mediator is composed of global schema [9] who defines integrated views based on the data provided by the sources. After receiving the user query by the mediator, it will be rewritten in the form of several sub queries where each sub query is formulated in local schema terms of a single source. It aims to search part of the global query result at a source level. The second component is the adapters that hide the sources linguistic heterogeneity. Each sub-query is translated by the adapter to transform it to the query language used by the source to which it is intended. On the other hand, the adapter translates the response returned by the source to the presentation formalism used by the mediator as shown in Figure 1.

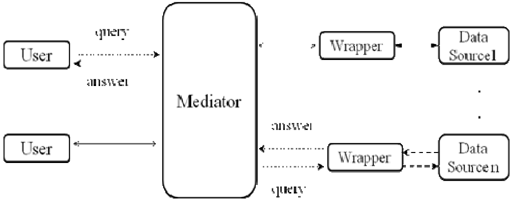

Figure 1: Mediator Architecture.

Although these systems are effective for applications containing a few data sources, they are not very adapted to the new context of integration posed by the web because they are based on a unique global schema. Contrariwise in P2P systems, there is no centralized control or hierarchical organization: each peer is equivalent in functionality and cooperates with other peers in the objective to realize a collective task.

3. P2P SYSTEMS

The peer-to-peer systems have evolved from a simple system for sharing distributed files (with keywords interrogation), as Napster [10] and Gnutella [11] until the advanced systems of data management distributed like Edutella [12] or Piazza [13]; each peer schema is an entry point in the peer-to-peer system. In other words, queries are expressed in the peer local schema. The mappings relating to the local peers are stored at the peer itself.

The super-peer or P2P Hierarchical: In this type of system peers are not equal. Taking into account the capabilities of peers, the peers are more available and powerful are called super Peers are assigned complex tasks such as query routing and the mediation schemas. These Super-Peers constitute a network between them and they assume the responsibilities at the same time query processing and indexing. The major problem with this architecture type is the downfall of network in the failure case which occurs on a server, because there is always a valid connection point to servers.

Structured P2P systems: A structured P2P network is a network where there is no central node; it designates the controlled architecture by a specific structure. The control realized at this level considers in particular the contribution of data for different nodes of the network structures. The data are not assigned to nodes randomly but are accorded to specific locations to facilitate the routing queries submitted by users. These networks are based on an architecture purely decentralized which has mechanisms based on distributed hash tables.

3.1 Purs P2P Systems

In addition, in pure P2P environment the nodes are largely independent, and there is no global control in the global schema form or management of global resources, nor a global schema or data repository. In a P2P Pure system there is no central peer, all Peers in this case have a similar role each peer is connected to a random subset of neighboring peers as shown in Figure 2.

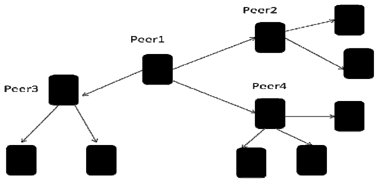

Figure 2: P2P Pur Architecture.

In this architecture each peer manages the sharing without a central peers or super-peer, which allows better reflect the system dependence level. The systems based on the use of super-peer are much more sensitive than pure P2P systems type. For systems that have a completely centralized architecture, it is sufficient to block the central peer to block their operation. For these reasons we chose to work on pure P2P. To interrogate a serval heterogeneous data sources, the information integration systems are based on the definition of a global schema or mapping. Local schemas of information sources are mapped on the global schema of the common interface.

3.2 Mapping Schema in the pure P2P system

Due to the specific characteristics of the pure P2P systems, for example the dynamic and autonomous nature of peer, the approaches that rely on centralized global schemas are no applicable in P2P systems. Thus, the main problem is to maintain the decentralized mapping schema order that query on the Peer schema can be reformulated in a query on the schema of another Peer. The approaches that are used by P2P systems to define and create the mapping between peer schemas can be classified as follows: pairwise schema mapping, mapping based on machine learning techniques, common agreement mapping, and schema mapping using IR techniques. The mapping defines the type of correspondence between the global schema and data sources. Consequently, the mapping specification between schemas determine the difficulty of query rewrite, well as the facility of addition and deleting of sources or to define certain constraints on the sources within the mediation system as shown figure 1.

3.3 Mapping schema in the mediation system

In a mediation system the local's schemas of data sources are mapped in the schema mediation, and mappings are maintained to a mediator. The user sends the query in terms of mediation schema; it will be rewritten in the form of several sub queries where each sub query is formulated in local schema terms of a single source. The wrapper provides translation services between the mediation schema and the local query language.

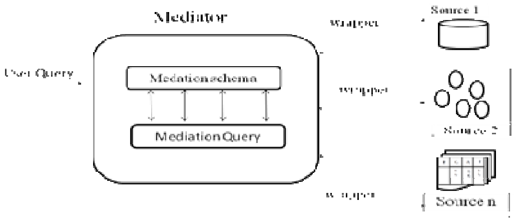

Figure 3: Schema Mapping using a Global Mediation.

In mediation systems, there are two main approaches to defining mappings: GAV (Global-As-View) [14,15] which defines the global schema as views of the local schemas, and LAV [16,17], which describes the local schemas as view of the global schema. In GAV [19,18] approach, the global schema is defined in terms of local schema sources. The first step to implement this approach consists in defining a global schema that covers the data which are interest to user. Thereafter, the global schema is described in terms of the source schema. This means, for each relation the global schema a query specifying how to obtain records of this relationship from relationships sources are written. In this case, rewriting the user query is simple: it is just browse the links between relationships in the global schema and the set of relationships schemas source. However, the failure of this approach is that it is unsuited to the addition of new data sources: Which requires updating the definition of predicates global schema to support new sources.

On one hand, the LAV [20, 21] approach is the opposite of GAV, after the definition of the global schema the content of each source is described as views over the global schema. However, the content of each source are defined as queries extracted from global schema relationships. In this case to response a query using this approach amounts to respond to a query using a set of views. Inter alia, in GAV any change in local schema requires a modification of global schema. In the applications whose local schemas are known at the beginning and they are not changed, GAV is the appropriate approach for this type of applications. It appears from the foregoing that the GAV is not adapted for P2P environments. In LAV, the correspondence between a local schema and global schema can be running locally. Consequently, LAV approach seems best adapted on a large scale of the distributed environments.

There are also other approaches that combine the advantages between the LAV and GAV approach as: GLAV, BAV, BGLAV and HAV. In GLAV approach (Generalized-locale-en-View) [22.23] also called (global Local As View), which combines expressive powers of both LAV and GAV. In a GLAV approach, the independence of a global schema, the maintenance to accommodate new sources, however, instead using a restricted form of first-order logical sentences as in LAV and GAV to define view definitions. Thus, GLAV can derive data using views over source relations, which is beyond the expressive ability of LAV, and it allows conjunctions of global relations, which is beyond the expressive ability of GaV [24]. The BAV approach (Both-As-View) [25] is an approach used to integrate P2P environment; it is also adapted for the integration system whose sources frequently change their schemas. The BAV [26] approach is considered support for evolutionary trends in the global or local level; it allows changes in global schema (public schema), and local schemas sources [27]. The principle of BAV is based on the use of reversible transformations, an important characteristic which establishes the centralized schema integration system (P2P public schema), and the views of local schema with the possibility to extract definitions of local schemas as views on the global schema. This approach has the advantage of specifying mappings between bidirectional schemas. The queries can be changed in both directions between the schemas; it supports the evolution of global and local schemas allowing transformation pathways and schemas to be gradually modify.

Both Global Local As View (BGLAV) is another alternative of GAV and LAV. The approach uses source-to-target mappings based on a conceptual schema as predefined target, and which is specified ontologically and independently of sources. It is easier to maintain the proposed data integration system based on BGLAV than in GAV and LAV, and queries reformulation is reduced to rules deployment. Compared to other data integration approaches, this approach combines the advantages of GAV and LAV, reduces their disadvantages and provides a flexible and scalable alternative for data integration.

BGLAV works in two phases: design and queries processing; in the design phase the system automates the generation of the source-to-target mapping in a synergistic way. Elements mapped by the mapping source-to-target are expressions in addition to the source schema elements that produce virtual target-view elements. This automatically leads to a rewriting of each target element as a union of corresponding virtual target-view elements.

In the treatment phase of the query, the user puts its query in terms of targets relations. The query reformulation is reduced to deploy rules using expressions of the view definition for targets relations in the same way as database systems apply the definitions of view [28].

HAV approach is the combination result of the best of GAV and LAV. There are two layers of mediators in HAV. The bottom layer is composed of multiple mediators called 'specialized mediators', all using the LAV approach for integrating schemas of data sources subset using the same data model. The top layer has a single global mediator that integrates specialized schemas of the bottom layer using the GAV approach.

In the architecture supports HAV, the global schema is not directly linked to data sources. Indeed, a set of partial schemas plays the role of virtual sources: a GAV mapping defines the relation between the global schema and the partial schemas. Each partial schema is defined on a set of data source with the same model using a LAV mapping.

A user query is expressed in terms of the global schema predicates; we get easily a decomposition of this query into sub queries in terms of specialized schemas, by replacing the global schema predicates by their definitions. The answer of queries is easier to achieve because the reformulation of each sub query concern a reduced number of data sources. Adding new sources requires only the definitions of necessary mappings between source schema and specialized schema corresponding [29].

Because of the advantages such as low-effort scalability resulting which involves LAV for the dynamic environments; we privilege LAV than GAV. Being a combination of both approaches GLAV would be another interesting choice that we could use to define mappings in our approach. However, we restrict our considerations to the LAV mapping.

in Table 1 we present a comparative table between the six approaches

Table 1: Comparison Of The Six Approaches.

Approach	Advantages	Disadvantages
GAV	-Rewriting queries is simple.	-is not flexible to accept new sources or eliminate sources into/from the system. Actually, adding or deleting sources might imply modifying the definitions of the global relations.
LAV	- Offers more flexibility to add new sources or delete old ones into/from the integration system, because a new source is just a new view definition.	- Change the global scheme requires a review of integer views. -automating query reformulation has exponential time complexity with respect to query and source schema definitions.
GLAV	-Addition / deletion is simple. -Provides the ability to more expressive mappings.	-The difficulty of query reformulation is the same as in LAV.
BAV	- Supports the - evolution of global schema and local schemas.	-Demanding to be more expensive to examine and treat with BAV that he would not be with the definitions of corresponding views in LAV, GAV or GLAV.
BGLAV	-Each relation in a target schema is predefined and independent of any source schema.	-No ability to semi-automate the specification of source descriptions.
HAV	-Flexible for adding or deleting data sources, -The complexity of query reformulation is reduced.	- The existence of two levels reduces system performance.

4. P2P MEDIATION SYSTEMS

A P2P system operates without central coordination, and therefore provides a dynamic environment that peers can join or leave at any time. Among other advantages we can cite the direct and rapid communications between peers, self-organization, decentralization in the storage and treatment information, and the ability to support a large number of peers, while providing for the fault tolerance. Therefore, to overcome the limitations of mediation system in which the central mediator schema; in addition to being a brake on schema evolution, also complicates information sharing. The coupling between mediation and P2P systems becomes important

The P2P mediation systems are convergence between P2P and mediation systems. Indeed, this type of system is a generalization of a information integration system, because is a dynamic and distributed system whose main goal is the autonomous Sharing of the structured data on a large scale. P2P systems can be classified into systems based on the key, keywords or schemas as the P2P mediation system, and can be seen as an evolution of distributed databases to a larger distribution. The P2P mediation system adopts a fully decentralized approach to information sharing. Each peer maintains its own data source, and therefore, its own schema; it can share all or part with the rest of the system. By distributing the information storage, the treatment and execution of queries between peers, these systems can be scaled up to a large number of nodes, without central control or powerful servers as shown in Figure 4.

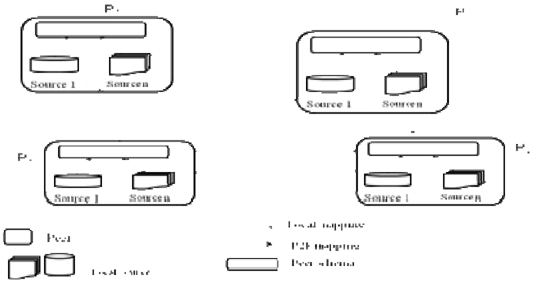

Figure 4: P2P Mediation System.

4.1 Correspondences between schemas

The correspondences define equivalences between the elements of one or more schemas. The objective of establishing correspondences in a P2P system is the setting up of an interrogation environment which masks the heterogeneity and distribution of information sources. The correspondences discovery between schemas is adapted to the case where schemes to integrate are semantically related. In other words, the applicability for cases where the differences between the schemas are mainly structural: The schema elements represent the same information or can be transformed into structurally similar elements in other schemas.

We have chosen five of existing P2P Mediation systems for a comparative study: SenPeer, PIPSINT, PIAZZA, HEPERION, XPeer.

- **SenPeer**

SenPeer is a P2P mediation system based on a Super-Peer architecture type; each peer publishes data using XML data models relational or object. The data mediation in SenPeer requires the construction of a semantic mapping set between the schemas of data. Each peer exports a layer of its data in a model which has a pivot graph structure, semantically enriched with keywords from schemas and intended to guide the discovery of semantic correspondences [30]. Then, these correspondences will be used for query rewrite they allow queries routing and rewriting other schemas. The source schemas are represented in an internal form of network formalism; while protecting the structural relationships between the specific objects in the language schemas.

- **Xpeer**

This system [31] provides an infrastructure of P2P mediation, the shared data are limited to the XML format and the integration language is XQuery. The peers are logically organized in groups on the basis of similarity criteria schemas.

In addition, each peer exports data description to be shared in the tree form. In XPeer Network the peers export a description of the data shared in the form of a tree structure called DataGuide [32]. The Super-Peers are organized to form a tree where each node accommodates a information schema about it sons and the Super-Peers having the same father form a group. The Super-Peers accommodate two information schemas on their sons: the schemas list of their sons the list of schemas is used during the query analysis for the identification of appropriate data sources to respond to user queries.

- **PEPSINT**

This system [33] is based on a P2P architecture type hypride; it allows integrating semantically XML data sources. Thanks to a number of semantic mappings stored in a correspondence table each peer connected to the network is indexed by a Super-Peer. The queries are rewritten by the Super-Peer uses compositions correspondences of the initia peer towards the Super-Peer and the Super-Peer towards the other peers. Every time a new peer joins the network

PEPSINT, the peer is registered and listed in the Super-Peer by establishing of Mapping of its local schema on the global ontology. The mapping is establish by a process of matching schema and stored in the Mapping table of the peer. During the matching schema process the global ontology is prolonged by the integration of local schemas.

- **Piazza**

This system is based on a pure Peer-to-Peer architecture; it allows the exchange of relational data XML or RDF [34]. Provides a solution to P2P medition system where the unique logical schema of data integration systems is replaced by a set of mediator schemas that are interlinked to define mappings between the peer schemas. Piazza uses two data integration approaches LAV and GAV for peer mappings. GAV is used to define relations of the mediator's schema over the relations in the sources; and LAV is used to define relations in the sources over the mediated schema. In Piazza system, a reformulation algorithm for query processing is presented that treats both GAV and LAV mappings. However, the Piazza system considers only the schema-level heterogeneity among the peers.

- **HYPERION**

This is a P2P mediation system that treats the problem of mapping data in P2P mediation systems where different peers may use different values to identify the same data. Hyperion relies on mapping tables that list peers of corresponding values for search domains that are used in different peers. Mapping tables provide the establishing for exchanging information between peers. By considering the mapping tables, Hyperion offers a mechanism that rewrite queries between peers; in terms of query answering

4.2 COMPARAISON DES SYSTEMES

To have a comparative vision between the different systems: PEPSINT, XPeer, Hyperion, SenPeer, PIAZZA; we chose a set of comparison criterions that we consider relevant namely:

1. Topology
2. Semantic correspondence: if the peers interested by the data exchange define semantic correspondence or not.
3. Integration with global schema: if integration in these systems is done with a global or without global schema.
4. Approach Mapping: Mapping approaches proposed in the mediation in these integration systems as shown in table 2.

Table21: The 5 Systems Under Different Aspects

Systems	topology	semantic mapping	Global integration schema	Approach mapping
PEPSENT	Super Peer	with	Yes	GAV &LAV
XPeer	Super Peer	without	No	GAV
SenPeer	Super Peer	with	Yes	GAV &LAV
PIAZZA	Pur p2p	with	Yes	GAV &LAV
Hyperion	Unstructu-red	with	No	GLAV

5. MOTIVATION AND PROBLEM STATEMENT

In a mediation P2P system the communications between peers are established via links, called mappings. Thanks to the reasoning mechanisms, a peer can make a special treatment from its own resources. These same mechanisms also operate mappings to propagate to other peers in the network and thus enable realization of the same treatment group by all peer networks. The problem addressed in this paper is to find links between the schemas of different peers in the system.

The following questions must be answered:

1. How to solve the syntactic and semantic heterogeneity of peer?
2. How peer they will route and process their requests?

And we are more interested in answering the following question:

1. How the peer inform the rest of the network of their data?

Solution 1: Is creating global integrated schemas, and to define mappings to connect the global schema and the local source schemas as shown in figure7.

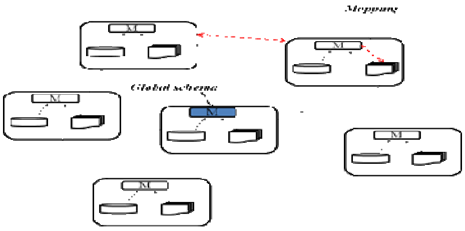

Figure 5: Global Schema In A Network.

The approach of the global schema is not suitable for dynamic distributed environment when two or more peers cooperate, it is wrong determine if the content recovers or they are linked. This requires the specification of mapping and their interrelationship with their neighbors. The Peers Mapping schema in dynamic environments is a difficult problem that has recently attracted much attention. The other problem is the implementation of a distributed architecture that takes into account the autonomy of data sources. The P2P systems provide the infrastructure for the dynamic environment in which autonomous and independent peers can join or leave the network easily and frequently.

Solution 2: in a diversified and large community the discovery of the relevant data and Peers can be expensive. A manner of reducing the cost of discovered in P2P architectures is to group the peers at the communities to limit the scope of research as shown in figure 6. But this method is not effective due to the dynamic nature of the network structure and also of the difficulties posed by the creation of cluster.

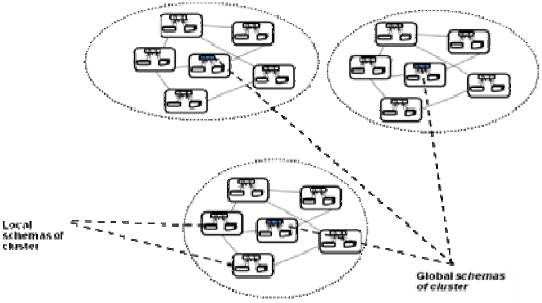

Figure 6: Global Schema In A Cluster.

Knowing that in a P2P mediation system each peer has only its own data, its own local schematic and its own mappings and, due to the fully distributed framework it ignores the schemas, the mappings and the data of other peers. This context, although specific makes the problem of the search brought into of correspondences between schemas very special. The proposed solutions must take into account the absence of the schemas centralized control; we noted that these solutions do not address these challenges.

6. OUR CONTRIBUTION

In our contribution, the network structure follows a pure P2P architecture. In this case it is necessary to add all the constraints related to this architecture type. The problems dealt with in this type of architecture are mainly related to the heterogeneity of schemas. To answer the problems mentioned in the previous section, we propose at first the following solution: is to concentrate on the Pairwise data Source as shown in Figure 10.

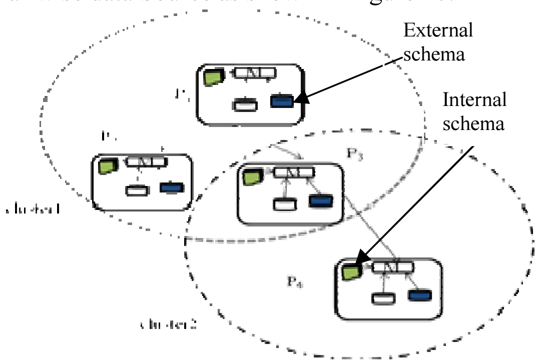

Figure 7: Our Approach Architecture.

This is motivated by the fact that all nodes in a P2P network are considered as clients and servers at the same time, and does not require centralized infrastructure to manage the data transformation. Clustering consists in group similar Peers, in general, if two peers are physically close, they can be considered similar in the P2P system: they will be placed in the same cluster. To facilitate the heterogeneities, we introduced two additional layers in our hierarchy of Peers: intern-Peer schema, as each of the Peers in the same cluster has an internal schema which acts as a schema mediation and external-schema-Peer a copy of "Peer information", and finally copies the information received from its neighbors. The Peers are identified by the hash of a particular attribute as identity. Then, we define a cluster as a set of Peers sharing the same information category; we also define a Peer as an entry point for the cluster.

A cluster is formed by the Peers in the same subnet and sharing common information to ensure continuity of service in a network. Clusters are independently controlled and dynamically reconfigured in motion peers. This network architecture's main advantage is the robust against topological changes caused by the movement of peers, peer failure, and the peer connects / disconnects.

Example:

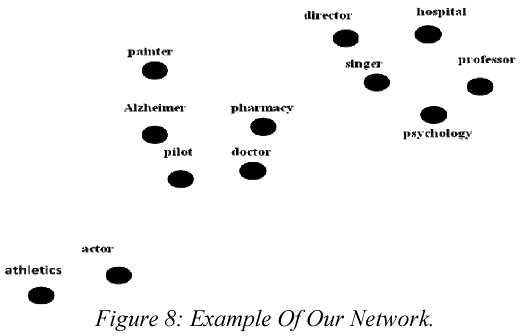

Figure 8: Example Of Our Network.

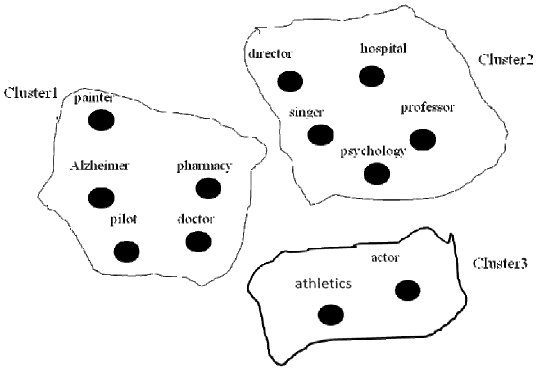

Figure 9: The Physical Topology.

Our network is organized according to a physical partition, so each cluster contains the peers that are neighbours and nearest. Consider NC is the number of cluster, and Ci is the i cluster with a Unique ID for each i cluster and X = {C1, C2, ..., CN} is the set of all the clusters, each peer also has a unique ID . In our approach there are two links types: local links connect Peers in a cluster; and the global links Peers connect different clusters. Two Peers are that bind by a local link are local neighbors, and if they connect with a global link then are global neighbors. If a peer who has a global link with a neighboring cluster is the intermediate Peer. Then if new Peer is connected to the network it obtains local neighbors by a broadcaster message, after it may have an IP address and retrieve the mask cluster and the peers in the cluster respond him with their IDs and their IP addresses. When the new peer gets enough local neighbors he searches global neighbors through its local neighbors. If the new peer is not local neighbor forms a new cluster; we assume that initially each peer knows a list of peers belonging to other clusters through intermediate Peer which has this list this list is enriched during its operation.

In our approach each Peer contains a neighbor table at the local level, a table neighbor at the global level; and a correspondence table (which contains the address of Peers with which it must share information) and a neighboring table clusters for Peers intermediate (Intermediate table). To search the available Peers in the network, the Peer propagates a query message through its neighbors included an H value of TTL (Time to Live). Each Peer retransmits the request to its neighbors depending on the H value.

- If H> 0, it retransmits the query to all its neighbors, included the local neighbors and global neighbors.
- If H = 0, it does not retransmits query only to its local neighbors.

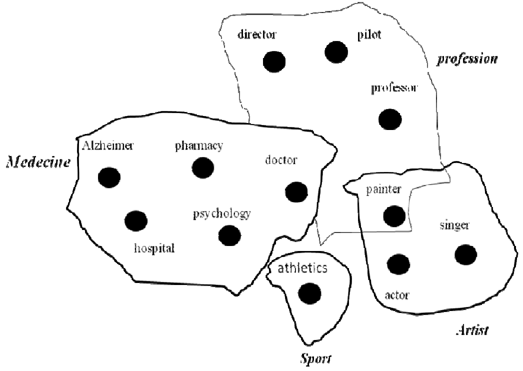

Figure 10: Regrouping Of Network In Cluster Based On Domains.

In the latter case we have grouped our network according to an architecture based on the concept or by domain. According to the example of figure 3 there are four clusters let us know that each cluster contains peer of the same domain; for painter it can be in the cluster "profession" as it can belongs to "artist "and it is the same case for Doctor. On the other hand the athletics can not to belong to any clusters which exist therefore it will build its own cluster.

6.1 Discovery Correspondences Principle

The resources discovery is affected as follows: each Peer has neighbors to which it publishes a copy of the information that receives. When a peer wants to access information, it first consults its internal schema in order to verify if they have the information or not. In the contrary case, it sends the query to one of its neighbors who has information whose identifier is similar to the information required; this routing is performed until achieve the required resource. The access to information is by traversing in opposite direction the way taken by the discovery query. In each browsed Peer the information is copied in the external schema of these peers. This approach allows improving the quality of routing because if the same query reproduced Intermediate peers can respond without routing the query until Peer which contains the information initially requested. In addition the network topology tends to be organized because in throughout the application, the Peers specialize on the similar Key information.

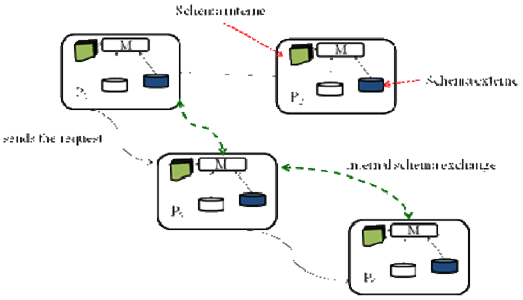

Figure 11: Correspondences Discovery.

6.2 Correspondence Levels In Our Approach

In our approach Peers define a correspondence between their local schemas and any other schema which seems to be interesting for them. The system tries to extract correspondences between the schemas that have no predefined mapping relying on the correspondence transitivity that was defined. However, there will be three levels of matches: As first level is the initialization step which is performed on each peer. The set of local data is summarized using an internal schema. In each Peer there is Mapping between Si (schema source) and its internal schema and of mapping between the sources schemas. The second level is when all Peers send a copy of their data information of which they have towards their cluster neighbors (correspondences between Peers in the same cluster). And finally the third level when the intermediate peer (the peer belongs to two clusters at the same time) send a copy its external schema Peer to the other cluster. (Correspondence between two different clusters Peers as shown figure 13).

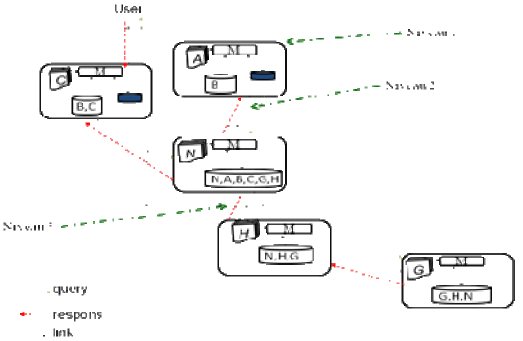

Figure 12: Query And Response Path.

6.3 Queries Treatment

System login: when a user wishes to have information, it will be necessary to send a query to a peer as (P1). The local Peers discovery (which belongs to the same group): When a user sends its query to a peer, this peer sends a broadcast message on the network which other Peers of the cluster are active. Only these Peers will respond the broadcast message. We are then connected to the cluster network and the peer receiving a request must respond the query indicating its identity for example. After the intermediate peers repeat the same operation so on to maintain consistent information. Pees Discovered with required information: In practice when seeking information that is on the cluster network that we broadcast request. It is sent initially to neighboring Peer, then by degrees propagated in the other cluster until finding Peer who holds the information searched. The response consisted the identity of this homologue follows the same path in reverse.

6.4 Distributed Clustering Algorithm

In our proposed algorithm each peer has an identifier *peerId* the algorithm executed by each peer network in a distributed manner. The partitioning of a network in Cluster requires the exchange of specific information; for that a discovery packet particular has been created. The fields in this package called *ConstructClusterPacket* exchanged during the execution of the partitioning algorithm are:

- TTL: Time To Live the package lifetime.
- PeerType: type Peer sender of the packet.
- PSID: global identifier ID of the packet Peer sender.
- Subject: package subject or the message type sent it can be: an invitation a peer with another- a message to ask the internal schema of another peer or to have the type of the peer which can be simple a peer or an intermediate peer.
- PDtId: identifier ID Peer recipient.
- ClusterId: Cluster identifier where the transmitter Peer package belongs.
- The Subject field of a ConstructClusterPacket package defines its subject it can have three different values:

Invitation: This is a invitation package to a certain cluster, the packet is initially sent by the peer-inviting it is rebroadcast by Peers which joins this cluster according to their situation. The first packet broadcast by the Peer-inviting has a TTL equal to N (Peer Number in network).

SI (Internal Schema): This is a package of refusal invitation this package is sent when Peer, already assigned to a certain cluster or when it receives an invitation package to another cluster which is of another domain. It also serves to define the intermediate Peers and complete the

IntermediateTable. Peer can belong to two domains at the same time as in the case of Doctor and Painter in Figure 13, and complete the *IntermediateTable* table.

Intermediate: This package is used typically by intermediate Peer to inform its neighbors of its type. This package is also used to define Peers intermediaries and to complete the *IntermediateTable* table.

- New_Peer: This type of packet is exchanged in the appearance of new Peer in the network. It is used to all attribute this new Peer to certain cluster.

```
  INPUT : N
   PART I :
    IF (P is an inviting Peer) THEN
  Send a packet P' (P.PeerId, ANY, P.ClusterID Invitation,
Normal, N-1)
    ENDIF
  PART II :
       IF (P receives a packet P') THEN
   PART II.1 :
       IF (P'.Subject = Invitation) THEN
       IF (P.ClusterId is undefined) THEN
    in the Cluster [P.ClusterID := P'.ClusterID]
        IF (P'.TTL = 0) THEN
     P.  PeerType := Intermediate Send a P''(P.PeerId, ANY,
P.ClusterID,
     Intermediate, P.PeerType, 1)
       ELSE
  Send a packet P''(P.PeerId, ANY, P.ClusterID, Invitation,
P.PeerType, P'.TTL - 1)
          ENDIF
      ELSE
     IF (P.ClusterID ≠ P'.ClusterID) THEN
      P.PeerType := intermediate Send a packet P''(P.PeerId,
ANY, P.ClusterID,SI P.PeerType,1)
P. IntermediateTable.add(P'.ClusterID, P'.PSId)
        ENDIF
    ENDIF
 ENDIF
PART II.2 :
 IF (P'.Subject = Intermediate Or P'.Subject = SI) THEN
 IF (P.ClusterID ≠ P'.ClusterID) THEN
 P.PeerType:=Intermediate P. IntermediateTable.add(P'.ClusterID,
P'.PSId)
 IF (IntermediateTable is modified) THEN
Send a packet P''(P.PeerId, ANY, P.ClusterID, Intermediate,
P.PeerType, 1)
    ENDIF
   ENDIF
ENDIF
PART III :
IF (P'.Subject = New_Peer) THEN Send a packet P''(P.PeerId,
P'.PSId, P.PeerID, Invitation, P.PeerType, 1)
   ENDIF
ENDIF
 OUTPUT : IntermediateTable, ClusterID, PeerType at PeerP
```

6.5 Discussion

1. Each peer can determine its own cluster. The cluster ID of each peer is equal to its ID or is equal to the cluster ID of its neighbors.

2. Each peer must have its clusterid once it becomes part of a cluster. This cluster ID will be broadcast at this time and will not be changed before the algorithm stops. Thus, each peer can determine its cluster.

3. In the dynamic network: the peers can change their site and they can be eliminated or added. A topological change occurs when a peer connects or disconnects from / to all or part of its neighbors, which modifies the structure of cluster and network thereafter.

4. Each entry of the *IntermediateTable* table contains following information:

- NeighclusterId: identification of neighbouring cluster of peer.
- IntermediatePeerIds: the list of Peers belong to neighbor cluster which are in direct communication with the intermediate Peer having the IntermediateTable table; ces Peers are also the Peers of the type Intermediate in their cluster.
- TimeOut: the live time of the entry in the table, this information is used to maintain the table IntermediateTable.

6.6 Scenarios

1. Initially, Clusters are formed by Peers-inviting (at the beginning each peer has its clean cluster). These Peers-inviting broadcast an invitation package once to their neighbors (Part I). This invitation package has a TTL equals N (Mean number of peer in the network); this section of code is executed by Peers-inviting. With the aim of specify the type of each Peer PeerType and complete the IntermediateTable (storing necessary information in the construction phase of the routing table between clusters). Peers also exchange packages whose subjects are SI or Intermediate.

2. When a peer receives an invitation package (Part II.1) it treats it according to its situation:

- If it is already affected to a certain cluster.
- If it is not affected it joined same cluster of Peer transmitting packet if they are of the same domain.

3. If the TTL of the message is zero, then the peer is a peer Intermediate it broadcasts to its neighbors a packet of Intermediate type of TTL equal to 1. Otherwise, it rebroadcasts the message after its TTL decremented of 1.

4. When Peer receives a packet of the SI type or Intermediate type of Peer belonging to another cluster (Part II.2), it changes its type in Intermediate and saves the Peerid and the Peer clustered transmitting packet in its IntermediateTable table if they do not exist.

5. If the IntermediateTable table were modified the Peer broadcasts a packet Intermediate type of TTL equal to 1. In this way, the exchange of the packets Intermediate stops when there are no

more new information to insert in the IntermediateTable table.

6. When a peer receives a packet type New_Peer (Part III), this means there is a new peer that appears in its neighborhood and in the network. For this the peer responds by an invitation packet (Invitation type) of TTL equal to 1. If they are of the same domain he joins its cluster.

7. At the end of the execution of this algorithm, each peer must be assigned to a certain cluster and have defined its type. The Peers Intermediaries (of Intermediate type) have IntermediateTable used later in the construction phase of the table routing between the clusters.

7. VALIDATION OF OUR APPROACH

We have presented the scenario mapping discovery for our approach (based on broadcast and exchange external schemas between peers). In this section, we present an implementation of the scenario that has been adopted in our approach which consists to realize, indeed the mappings between schemas peers. We first describe the environment in which we performed our implementation. The scenario and the implementation of this implementation are then detail.

Environment Implementation: Client / Client architecture with java sockets. To validate our proposition, we used the following tools:

- Java to develop the application.
- XML: for the knowledge management and the data archiving.
- XSD summarizing information which has a peer in our network.

We summarize the mapping discovery process that we have done. Indeed, this process comprises the steps following:

1. Cluster Management it consists to:
   - At the beginning each peer has its own cluster
   - An ID is automatically associated with this cluster and they are stored in a XML file containing all the clusters with their peers.
   - A peer sends a message broadcaste to its neighbors in the network.
   - A XML File is generated containing information on the clusters.
2. Management peers it consists to:
   - When a peer connects it sends its external schema to its neighbors.
   - If it shares the same domain with another peer, an ID result of a not-invertible hash function (SHA1) is associated with this peer.
   - Everything is stored in an XML file reserved for each peer and automatically generated. Below an example of the XML file generated after the addition of each peer.
3. Network topology

The peeers P1 P2 P3 P4 are peers of the same cluster (medicine) P5 and P9 are intermediate peers.

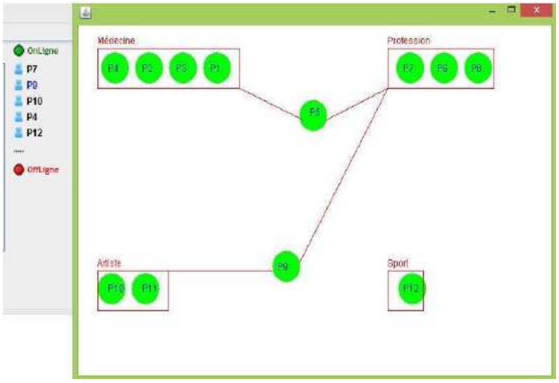

Figure 13: Network Topology.

8. CONCLUSION

Several mediation systems problems are related to scaling up. Any integration system must provide the solutions to the following problems: (1) How to provide an integrated global view of data represented through different conceptualizations? (2) How to identify and specify the mapping between semantically related data? (3) How to update different data bases is such an integrated global view? It is role is to provide the user with a view a transparent and interrogation uniform of information without the user not caring about the source of informed nor of their origin format. In this work we have treated the discovered correspondence problem in the “P2P mediation” integration systems; we proposed an algorithm for clustering by domain on a pure P2P network architecture. We were bring to establish a study on the different features offered by this algorithm in a P2P network, to properly design an architecture which is based on the project and to achieve the objectives set. We have the GAV approach was not adapted to the integration of a new source when the number of sources became too high.

Then initially it is supposed that the LAV is the most adapted for our approach. It remains to study the robustness of the scaling and the optimality of the clustering algorithm, to improve the two bases of availability and of performance, also storing information of apartment just other peer where they will be stored, therefore we will create a new schema for each peer that will call it "schema dictionary". Without forgetting to improve the validity of our P2P network as all the other quality criteria of a P2P network the aspect security and the application ergonomics.

REFRENCES:

[1] A. Y. LEVY," *Combining artificial intelligence and databases for data integration, Artificial Intelligence Today",* Springer Berlin Heidelberg, 1999**p**. 249-268,.

[2] J. Widom. «Research problems in data warehousing». *In Conference on Information and Knowledge Management, (1995).* **P.**25-30

[3] G. WIEDERHOLD, Mediators in the architecture of future information systems, *IEEE* Computer, 1992. p.38-49,.

[4] M. Lenzerini. «*Principles of p2p data integration*» In Zohra Bellahsene and Peter McBrien, editors, DIWeb, (2004).

[5] A. Halevy, Z. Ives, I. Tatarinov, and Peter Mork. "*Piazza: data management infrastructure for semantic web applications".* WWW '03 Proceedings of the 12th international conference on World Wide Web. P.556-567

[6] P. Adjiman, P. Chatalic, F. Goasdoué, M-C. Rousset, and L. Simon. «*Somewhere in the semantic web*». International workshop on principles and practice of semantic web reasoning, (2005).P; 1—16.

[7] W. S. Ng, B. C. Ooi, K-L. Tan, and A. Zhou. «*Peerdb: A p2p-based system for distributed data sharing*», (2003). 19th International Conference on 5-8 March 2003.p 633-644

[8] G. Wiederhold ,"*Mediators in the architecture of future information systems*". Computer (Volume:25 , Issue: 3) . March 1992. P 38-49.

[9] B. Alexe L.Chiticariu R.J.Miller W.Tan. "*MUSE: Mapping Understanding and deSign by Example*". Proceedings of International conference on Data Engineering (ICDE) 7-12 April 2008. P 10-19.

[10] Napster. *http://www.napster.com.*

[11] Gnutella. *http://gnutella.wego.com.*

[12] W. Nedjl, B.Wolf, C. Qu, S. Decker,M. Sintek, A. Naeve,M. Nilsson, M. Palmer, and T. Risch. "*Edutella: a p2p networking infrastructure based on rdf*". WWW '02 Proceedings of the 11th international conference on World Wide Web .P.604-615. ACM New York, NY, USA ©2002

[13] A. Halevy, Z. G. Ives, P. MorK, I. Tatarinov, «*Piazza: Data Management Infrastructure for Semantic Web Applications*», Proceedings of the twelfth international conference on World Wide Web Budapest, (2003).P. 556-567,

[14] G. Molina, and al., "*The TSIMMIS project: Integration of heterogenous information sources*". The 16th Meeting of the Information Processing Society of Japan, (1994)

[15] I. Koffina, G. Serfiotis, V. Christophides, "*Foundations for Information Integration: A State of the Art*". Sixth Framework Programme, 2005.

[16] Papakonstantinou, Y., and al.,"*Object fusion in mediator system*". In proceedings of International Conference on Very Large Dtabases (VLDB), Bombay, India, September 1996. P. 413-424

[17] S. Adali, and al., "*Query Caching and Optimization in Distributed Mediator Systems*". Proceedings of the ACM SIGMOD International Conference on Management of Data, Montreal, Canada, ACM, New York, June 1996. P. 137-146.

[18] E. Mesrobian, R. Muntz, E. Shek, S. Nittel, M. LaRouche and M. Kriguer, "*OASIS: an open architecture scientific information system, "in Proceedings of RIDE*", New Orleans 1996, IEEE Press. **P**. 107 – 116.

[19] G.-L. G. Gomez. "*Construction automatisée de l'ontologie de systèmes médiateurs. Application à des systèmes intégrant des services standards accessibles via le Web* ''. PhD thesis, Université Paris XI Orsay, 2005.

[20] A. Y. Levy, A. Rajaraman, and J. J. Ordille. "*The World Wide Web as a collection of views: Query processing in the information manifold*". Proceedings of the International Workshop on Materialized Views: Techniques and Applications (VIEW'1996), June 1996. pages 43–55.

[21] M. Friedman, and D. Weld, "*Efficient execution of information gathering plans*". In Proceedings of the International Joint Conference on Artificial Intelligence, Nagoya, Japan. 1997. P. 785-791.

[22] Friedman, M., and al., "*Navigational plans for data integration*". In Proceedings of the National Conference on Artificial Intelligence.1999. P. 67—73.

[23] A. Calì. "*Reasoning in data integration system: why LAV and GAV are siblings*". In Proc. of the 14th International Symposium on Methodologies for Intelligent Systems (ISMIS), 2003. P. 562-571.

[24] O. Kapitskaia, A. Tomasic, and P. Valduriez, "*Dealing with discrepancies in wrapper functionality*", Tech. Rep.RR-3138, INRIA, 1997. P. 327—349.

[25] M. Boyd, and al. "*AutoMed: A BAV Data Integration System for Heterogeneous Data Sources*". In Advanced Information Systems Engineering 16th International Conference, CAiSE, Riga, Latvia, June 7-11, Proceedings. Springer-Verlag, 2004. P. 82–97.

[26] P. McBrien, A. Poulovassilis. "*Data integration by bi-directional schema transformation rules*". In: 19th International Conference on Data Engineering, ICDE'03, March 5 - March 8, 2003

[27] A. Zellou, "*A solution for Xquery queries rewriting in LAV approach in a mediation system*", Caree 2008.

[28] L; Xu, W. Embley, "*Combining the Best of Global-as-View and Local as-View for Data Integration*", In Information Systems Technologies and Its Applications - ISTA. 123–136, 2004.

[29] F. L. Lahmer, "*Une approche Hybride d'intégration de sources de données hétérogènes dans les datawarehouses*", Université Mentouri de Constantine, Algérie, 11 décembre 2011.

[30] D. Faye, "*Médiation de données sémantique dans SenPeer, un système pair-à-pair de gestion de données* ", PHD Thesis, Nantes, France, October 2007.

[31] C. SARTIANI. P.MANGHI, G. GHELLI, G.CONFORTI . "*XPeer: A Self-organizing XMLP2P Database System*", Proceedings of the First EDBT Workshop on P2P and Databases (P2P&DB 2004), Crete, Greece July, 2004.P. 456-465.

[32] R.Goldman, J.Widom: "*DataGuides: Enabling query formulation and optimization in semistructured databases*". In: VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25- 29, 1997, Athens, Greece, Morgan Kaufmann (1997) 436–445.

[33] I. F. CRUZ, H. XIAO, F. HSU, "*Peer-to-Peer Semantic Integration ofXML and RDF Data Sources*", Third International Workshop on

Agents and Peer-to-Peer Computing (AP2PC 2004), July, 2004.P. 108-119.

[34] I. F. CRUZ, H. XIAO, F. HSU, "*Peer-to-Peer Semantic Integration of XML and RDF Data Sources*", Third International Workshop on Agents and Peer-to-Peer Computing (AP2PC 2004), July, 2004. .P. 108-119.

[35] M. Arenas, V. Kantrev Kementsietsidis A,Kiringa I, Miller R.J, Mylopoulos J,"*The hyperion Projet ct :from Data Integration to Data Coordination*", Sigmod. Record 32(3), p.53-58, 2003