Association for Information Systems
AIS Electronic Library (AISeL)

MCIS 2008 Proceedings

Mediterranean Conference on Information Systems (MCIS)

10-2008

GLOBAL INDEX CONSTRUCTION FOR DATA INTEGRATION IN LARGE SCALE SYSTEM

Fahima El Hajjej
Faculty of Sciences Tunis, Information Sciences Department, Tunisia, hajjejfahima@gmail.com

Bassem Barkallah
Faculty of Sciences Tunis, Information Sciences Department, Tunisia, Bassem.barkallah@yahoo.fr

Samir Moalla
Faculty of Sciences Tunis, Information Sciences Department, Tunisia, moalla.samir@fst.rnu.tn

GLOBAL INDEX CONSTRUCTION FOR DATA INTEGRATION IN LARGE SCALE SYSTEM

El Hajjej, Fahima, Faculty of Sciences Tunis, Information Sciences Department, hajjejfahima@gmail.com

Barkallah, Bassem, Faculty of Sciences Tunis, Information Sciences Department, Bassem.barkallah@yahoo.fr

Moalla, Samir, Faculty of Sciences Tunis, Information Sciences Department, moalla.samir@fst.rnu.tn

Abstract

Several scientific projects focused on the creation of Peer-to-Peer data management system. The main objective of these systems is to allow data sharing and integration among a large set of distributed, heterogeneous data sources. The emergence of large scale systems provides solutions and brings to surface new challenging unsolved problems, among which, we address the data integration problem.

In order to address this problem, we propose a new data integration approach that allows the semantic integration of heterogeneous and distributed data sources in a Peer-to-Peer environment with high distribution and evolution support. In this paper, we provide an introduction to the approaches; problems and research issues encountered when dealing with data integration.We present our approach and describe the several methods for constructing a global index that is the core of our approach by using semantic similarities. We end our work by an application example.

Keywords: data integration, large scale environment, Peer-to-Peer system, semantic similarity.

1. INTRODUCTION

A large amount of data is available from a large number of distributed, heterogeneous data sources. The data provided may differ in semantics, storage format (unstructured, semi-structured or structured sources) value range, etc…. The heterogeneity is one of the most important problems addressed in the data integration field, the research efforts made in this field aim at providing users access to a set of distributed, heterogeneous and autonomous data sources. Many data integration approaches were proposed by the research community (Chawathe et al 1994) as the multi-databases approach, the federated approach, and the mediated approach. In (Chawathe et al 1994), it was demonstrated that these approaches are unable to satisfy the constraints imposed by the characteristics of the large scale systems (dynamicity, scalability, etc).

Nowadays, large number of organisations has an important number of departments. Each department can use or require different data sources (e.g. relational databases, XML files, object oriented databases and more) to store and access their data. Organisations desperately need take advantage of the data they own and need to capitalize it as an investment that should be profitable, to do so they need a way to manage, interrogate and structure their data that it can be used for decision support or consumer habits analysis. We address the problem of large scale data integration, where the data sources are unknown at design time and are from autonomous organisations. However, traditional data integration approaches (federated approach, multi-databases approach and mediated approach) fail to meet the requirements of a constantly changing environment (user number, preference, description...) and offer limited scaling mechanisms.

The reminder of the paper is organised as follows: Section 2 describes the different approaches for data integration. Section 3 introduces the proposed model and provides the basic definitions. Section 4 describes the algorithmic processes inherent to the global index construction. Section 5, details an application example of the proposed algorithms on a set of local indexes, and Section 6 concludes the paper.

2. RELATED WORK

Data integration is defined as a set of services allowing transparent and uniform access to a set of distributed, heterogeneous and autonomous data sources. In the literature, we can distinguish three major data integration approaches (Barkallah et al 2007):

- *Multi databases approach:* This approach does not require a global schema, and the data interrogation and access is performed directly on the sources through a multi databases language as MSQL, where the user knows all the details concerning the source he queries (structure, data types, value range, location…).
- *Federated approach* (Susanne et al 1999): This approach requires a global schema which integrates all the data sources schemas, that central element need to be defined at design time and all the process is based upon it. It insures heterogeneity and transparency, but can not handle the dynamicity of the data sources.
- *Mediated approach:* the mediated approach has been used to integrate data from distributed heterogeneous sources, where a mediator abstracts the user from problems caused by different locations, query languages and protocols of the different sources. A key issue in the mediated approach is the mapping expression that can be represented using LAV, GAV, GLAV and BAV.

Research on integration systems has been converging toward mediated architecture. This approach is hardly applicable on large scale system since it requires a central mediated schema which limits the evolutions capabilities of the local schemas and complicates the distribution of data (Hacid et al 2005). However, no generic solution exists to the data integration problem (Valduriez et al 2004).

We describe the two major data sources description formalisms provided in the literature: schema and ontology, and introduce the main characteristics of our approach.

Schema integration is the process of combining data source schemas into a coherent global schema in order to reduce data redundancy in heterogeneous data source systems (Joseph et al 2005). It is often hard to combine different data source schemas because of the different data models or structural differences in the data representation and storage. At data integration, several issues must be addressed. We focus on the problem of heterogeneity, more specifically on semantic heterogeneity – that is, problems related to semantically equivalent concepts or semantically related/unrelated concepts. In order to address the problem of semantic heterogeneity previously described, we apply to ontologies as a semantic support for data integration.

Ontology integration involves the use of ontology(s) to effectively combine data from multiple heterogeneous sources. The effectiveness of ontology based data integration is closely tied to the consistency and expressivity of the ontology used in the integration process. Ontology gives the name and the descriptions of the entities of specific domains using predicates that represent relationship between these entities (Agustina et al 2005). Therefore, ontology might be used for data integration as data source descriptor because of its potential to describe the semantic of data sources (Agustina et al 2003). Ontology representation for data source cannot capture real word semantics, but only logical relations between predicates, so this solution can be tuned to express real word semantics.

Our work describes a new approach to large scale semantic integration of heterogeneous data sources. In this approach we will use the relationship among the entities of different sources. Each data source presents its data trough a set of concept. A concept is identified by a name and description. In addition, we are working on including similarity functions to give a more precise comparison among the terms of different local data sources. We focus on the semantics of the words used to describe the concepts and not only the relation between concepts that are usually expressed using ontologies.

In the following Sections of this paper, we are attempting to detail our approach of large scale data integration.

3. MODEL DESCRIPTION

In this section, we present the setting where our approach can be deployed, the basic concepts used to define the proposed model.

3.1. Environment description

The environment on which we operate is a cluster characterized by a great stability in term of nodes number and a high degree of homogeneity (network, OS, data semantic, etc). A sub-cluster is composed of a set of neighbour's nodes. Each of the cluster nodes provides a description of its data through concepts (data interface).

A node N_i includes a set of data instances D_j described by a set of concepts C_k. A concept is defined by a name, and a description of its components. The suggested model is based on the use of two index structures. The first presented structure is the Local Index (LI) which persists in each node in order to describe the set of its concepts. The second presented structure is a Global Index (GI) composed of all the public concepts available at a cluster and specifies the semantic link between them.

The two index structures are described as follows:

- Each entry of the Local Index (LI) structure is represented as follows:

(CN)	**(CD)**
Concept Name	**{Concept Description}**

  - The first element (CN) represents the name of concept C_i.
  - The second element (CD) is the description of concept C_i.
- A GI entry is described as follows:

(PC)	**(SCL)**
<Primary Concept >	**Similar Concepts List**

  - The couple $<C_i, N_i>$ defines the primary element (PC), where C_i represents the concept describing the data included in node N_i.
  - The second element (SCL) represents the set of similar concepts to the concept C_i which is described in PC of the same entry. Each element is defined as follows: $<C_j, N_j, Val\_sim (C_i, C_j) >$, where C_j is the concept included in node N_j and similar to C_j having a value of $Val\_sim(C_i, C_j)$[1].

Each concept C_i, existing in the cluster, is stored under the element **PC.** The similar concepts to C_i are described in the corresponding **SCL**. Unless these similar concepts verify a minimum degree of similarity, they shouldn't be taken into account. A concept C_i with no similar concepts is called single, and the corresponding **SCL** will be empty.

3.2. Similarity measure

On our proposed approach relies on semantic similarity held between concepts, we present the metrics used for the similarity measures. The global index construction is based on the grouping (gathering) of similar concepts. We use one of the following semantic similarity measures: Res, Jcn and Lin (Zargayouna et al 2004). These similarity measures are based on the information content. The more

[1] It is calculated according to one of the methods of measurement used (Resnik, Lin, Jiang and Contrath)

information two concepts share, the more similar they are, and the information shared by two concepts is indicated by the information content of the concepts that subsume them (Resnik 1995).

4. GLOBAL INDEX CONSTRUCTION ALGORITHMS

In this section, we describe the algorithmic processes suggested to achieve the global index construction. To reduce the time allocated to the global index construction, we divide the cluster into several sub-clusters. The construction on each sub-cluster global index will be performed simultaneously; process used for the construction will be relayed in parallel at each sub-cluster level. In this case, the global index construction is performed through two phases:

- Sub-cluster global index construction,
- Cluster global index construction.

4.1. **Sub-cluster global index construction**

We provide two methods for building the global index at a sub-cluster level:

- The first method called sequential architecture (ordered),
- The second method called hierarchical architecture (Layered).

4.1.1. Sequential architecture

The first method consists of passing down from node to node the computed indexes that will be used for the construction of the global index. This method requires the execution of two algorithms:

- Alg1: Build the first Global Index GI_1 by combining (computing the similarity degrees held between) the concepts described in the two local indexes taken as first entry.
- Alg2: Construction of an intermediate version of the Global Index GI_i (i>1). It consists of updating the GI_1 produced at the first step by computing the similarities between the actual node local index concepts and GI_1.

The construction of this index illustrated by Figure1 displaying the following notations:

- LI_i: Local Index located in node *i.*
- GI_1: First Global Index.
- GI_i: Global Index version *i*.

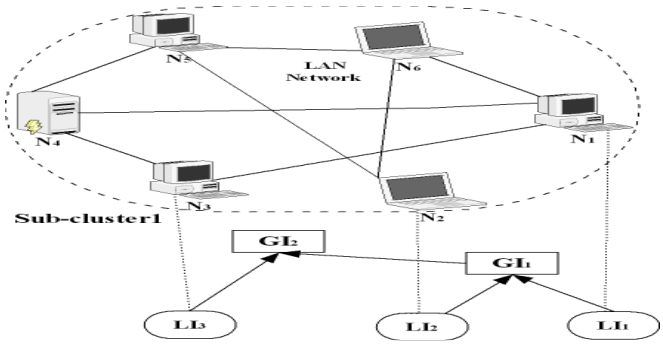

Figure1. Sequential architecture for sub-cluster global index construction

4.1.2. Hierarchical architecture

The second method uses a hierarchical architecture (Layered). The index construction in this case requires the execution of two or three algorithms according to the number of sub-cluster nodes:

- Alg1: Build the first Global Index GI_1 by combining (computing the similarity degrees held between) the concepts described in the two local indexes taken as first entry.
- Alg3: Construction of an intermediate version of the Global Index GI_i by combining the concepts described in two intermediate global indexes produced at the previous level.

- Alg2: Alg2 is used when the number of sub-cluster nodes is odd. It allows the construction of the final Global Index version of Sub-Cluster GI_{SC} by combining the partial index produced at the last level (top of the hierarchy and the remaining local index).

The global index construction using this architecture is illustrated by Figure 2 displaying the following notations:

- LI_i: Local Index located in node *i*.
- GI_i: Global Index produced at the *i* level.

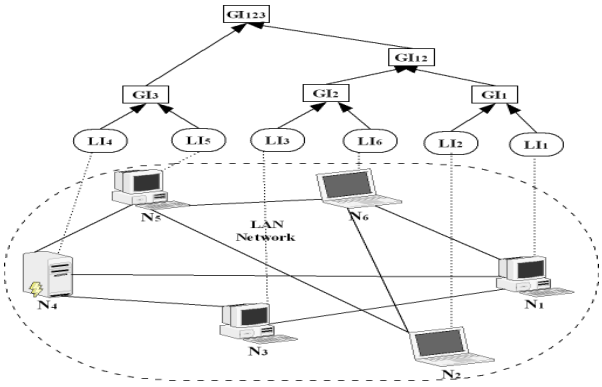

Figure2. Hierarchical architecture for sub-cluster global index construction

4.2. Cluster global index construction

Coming after the construction of sub-cluster global indexes, Alg3 is executed to build the cluster global index. To do so, we can use one of the following architectures: sequential architecture or hierarchical architecture. The global index construction according to the sequential architecture is illustrated by Figure 3. In this figure, we adapt the following notations:

- GI_{SCi}: Sub-Cluster Global Index *i*
- GI_{SCij}: Global Index assembling the Sub-Cluster global indexes *i* and *j*

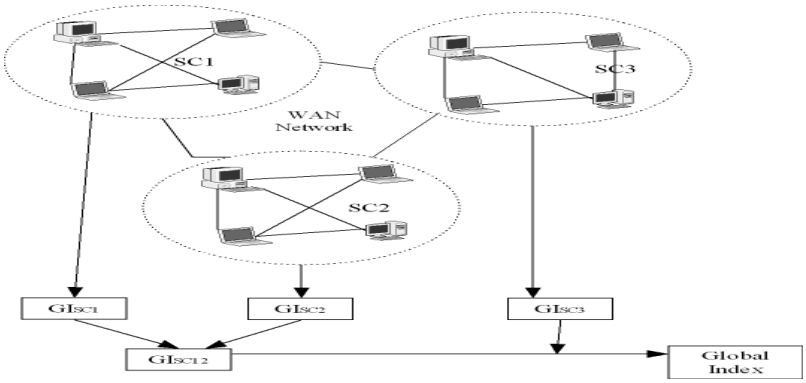

Figure3. Cluster global index construction

4.3. Construction algorithm

In this section we describe the three algorithms used for the cluster global index construction. Each algorithm has a number of main stages. Each stage embodies a set of tasks that must be achieved. We will briefly explain each stage.

- **Alg1:** Building the first Global Index GI_1 by combining the concepts described in the first two local indexes. The different stages of Alg1 algorithm are ordered as follows:
  - *Stage1*: taking out the set of concepts from LIi to be listed unto (registered under) entry (PC) of the global index under construction,

- *Stage2*: compute the semantic similarity between the concept PC of each entry and the other concepts of LIj,
- *Stage3*: verify the condition of similarity established by the user,
- *Stage4*: If the condition is verified, add the resulting concept to the similar concepts list of the considered concept,
- *Stage5*: return to stage1 with LIj.

- **Alg2:** Building the intermediate version of the Global Index GI_i (i>1). While in entry, this algorithm takes the GI_i index and another Local Index LI_k of the same sub-cluster. The stages of this algorithm are ordered as following:
  - *Stage1*: compute the semantic similarity between the PC concept of each GI_i entry and the GI_k concepts.
  - *Stage2*: once the value of similarity is made up for, GI_k is added to the SCL of the entry.
  - *Stage3*: The algorithm checks up (passes over; comes through) one to one the local index concepts and adds them to the GI_i entries, and finally we assign to each concept the GI_i concepts similar to.
- **Alg3:** Building the intermediate version of the Global Index GI_i by combining the concepts described in two intermediate global indexes produced at the previous level.Upon entry, this algorithm takes out in pair the global indexes (GI_{SCi}, GI_{SCj}). The stages of this algorithm are as follows:
  - *Stage1*: coming through the sub-cluster global index GI_{SCi} and bringing about to similar concepts list (SCL) of each entry of this index, while checking up the condition of similarity established (founded; imposed; conceived) by the user, the sub-cluster global index concepts GI_{scj},
  - *Stage2*: assigning new entries to GI_{SCi} index. These entries, allotted the semantic links with the AG_{SGi} concepts, represent the GI_{SGj} index concepts.

5. APPLICATION EXEMPLE

We provide a real case application example of the proposed algorithms in order to create a cluster global index. The cluster on which the example is run is composed of two sub-clusters SC_1 and SC_2. The SC_1 sub-cluster is composed of three nodes N_1, N_2 and N_3 and the SC_2 sub-cluster is composed of five nodes N_4, N_5, N_6, N_7 and N_8. The SC_1 global index sub-cluster will be constructed according to the sequential architecture while the SC_2 global index construction will be carried out following the hierarchical architecture.

The putting into effect of the differing stages of the global index construction results in the following:

5.1. Global index construction relating to SC_1

The sub-cluster SC_1 is made up of three nodes whose local indexes are presented as follows:

NC	CD
Car	mark, color
Automobile	Mark, supplier
Home	Structure, address

Table 1. LI_1 located in the N_1 node

CN	CD
Bus	Manufacturing, types
Driver	Name
House	shape, layout
Auto	type, design

Table 2. LI_2 located in the N_2 node

CN	CD
Hospital	type, departments, funding
Huilding	Planning, design

Table 3. LI_3 located in the N_3 node

Application of Alg1

We apply the Alg1 algorithm to two Sub-Cluster SC_1 local indexes LI_1 (Table 1) and LI_2 (Table 2).

CP	LCS
< car, N_1 >	<bus, N_2, 0,69> ; <auto, N_2, 1>
< automobile, N_1>	<bus, N_2, 0,65> ; <auto, N_2, 1> ; <driver, N_2, 0,51>
< home, N_1 >	<house, N_2, 0,96>
< bus, N_2 >	<car, N_1, 0,69> ; <automobile, N_1, 0,65>
< driver, N_2 >	< automobile, N_1, 0,51 >
< house, N_2 >	< home, N_1, 0,96 >
< auto, N_2 >	<automobile, N_1, 1> ;< car, N_1, 1>

Table 4. First Global Index GI_1

In Table 4, the concept « car » of N_1 is similar to the concept « bus » of N_2 having a value of 0,69. This value is defined by one of the similarity measurements previsiouly mentioned in sub-section 3.2. The same concept « car » also is similar to concept « auto » of the same node having a value of 1.

Application of Alg2

We are now taking into consideration the third local index LI_3 of node N_3 of the same sub-cluster SG_1. The second Alg2 algorithm takes while in entry local index LI_3 as well as first global index GI_1 (Table 4). It results in global index GI_2.

CP	LCS
< car, N_1 >	<bus, N_2, 0,69> ; <auto, N_2, 1>
< automobile, N_1>	<bus, N_2, 0,65> ; <auto, N_2, 1> ; <driver, N_2, 0,51>
< home, N_1>	<house, N_2, 0,96> ; < Building, N_3, 0,60 >
< bus, N_2>	<car, N_1, 0,69> ; <automobile, N_1, 0,65>
< driver, N_2>	<automobile, N_1, 0,51>
< house, N_2 >	<home, N_1, 0,96> ; < building, N_3, 0,60 >
< auto, N_2>	<automobile, N_1, 1> ; <car, N_1, 1>
< hospital, N_3>	
< building, N_3>	< house, N_2, 0,60 > ; < home, N_1, 1 >

Table 5. Global index GI_2

GI_2 includes (is comprehensive of) all the concepts found at Sub-Cluster *SC_1 level* besides similar concepts.

5.2. Global index construction related to SC_2

We use to construct the global index a hierarchical architecture. The local indexes of Sub-Cluster SC_2 are presented as follows.

CN	CD
University	Type, departments
Plain	Type

Table 6. LI_4 located in N_4

CN	CD
Ship	Architecture, measuring, types

Table 7. LI_5 located in N_5

CN	CD
Show	Categories
Building	Design

Table 8. LI_6 located in the N_6

CN	CD
Tablelands	geography, history, type
Hull	Form

Table 9. *LI_7 located in N_7*

CN	CD
Sea	nomenclature,
Medicine	practice, education, branches

Table 10. *LI_8 located in N_8*

Application of Alg1

Restrictedly in the beginning of construction is Alg1 applied to pass over to level 1. We apply the Alg1 algorithm to the two local indexes LI_4 and LI_5 to come out with GI_{45} index. We at the same time track on the same process applying it to the two local indexes LI_6 and LI_7.

CP	LCS
< university, N_4 >	
< plain, N_4 >	< ship, N_5, 0,51 >
< ship, N_5 >	< plain, N_4, 0,51 >

Table 11. *Global index GI_{45}*

CP	LCS
< Show, N_6 >	
<building, N_6>	<tablelands, N_7, 0,50>
<tablelands, N_7>	<building, N_6, 0,50>
<hull, N_7>	

Table 12. *Global index GI_{67}*

Application of Alg3

In coming to level 2, Alg3 algorithm is done with. While in entry it takes two already built indexes GI_{45} and GI_{67} resulting in GI_{4567}.

CP	LCS
< university, N_4 >	
< plain, N_4 >	< ship, N_5, 0,51 >
< ship, N_5 >	< plain, N_4, 0,51 >, <hull, N_7, 0,67>
< Show, N_6 >	
<building, N_6>	<tablelands, N_7, 0,50>
<tablelands, N_7>	<building, N_6, 0,50>
<hull, N_7>	< ship, N_5, 0,61 >

Table 13. *Global index GI_{4567}*

Application of Alg2

It is noticed that at the Sub-Cluster level (SC_2) we have before us an odd number of nodes. The N_8 node had not been taking part in global index construction with its Local Index LI_8. Henceforth, the Alg2 algorithm displays the global index of all the sub-cluster, taking in its entries the indexes GI_{4567} and LI_8. The application of Alg2 results in the following table:

CP	LCS
< university, N_4 >	<medicine, N_8, 0,64>
< plain, N_4 >	< ship, N_5, 0,51 >
< ship, N_5 >	< plain, N_4, 0,51 >, <hull, N_7, 0,67>, <sea, N_8, 0,50>
< Show, N_6 >	
<building, N_6>	<tablelands, N_7, 0,50>
<tablelands, N_7>	<building, N_6, 0,50>
<hull, N_7>	< ship, N_5, 0,61 >
<sea, N_8>	< ship, N_5, 0,50 >
<medicine, N_8>	< university, N_4, 0,64 >

Table 14. *Global index GI_{45678}*

5.3. **Cluster Global index construction**

Application of Alg3

The construction of GI_{SC2} is carried out at same time with that of GI_{SC1}**.** The Alg3 algorithm takes up the two Global Indexes (GI_{SC1}, GI_{SC2}) which relate to the Sub-Clusters SC_1 and SC_2. GI_{SC1} and GI_{SC2} being brought about, the algorithm Alg3 has produced the cluster global index (Table *15*).

CP	LCS
< car, N_1 >	<bus, N_2, 0,69> ; <auto, N_2, 1> ; < ship, N_5, 0,51 >
< automobile, N_1>	<bus, N_2, 0,65> ; <auto, N_2, 1> ; <driver, N_2, 0,51> ; < ship, N_5, 0,59 >
< home, N_1>	<house, N_2, 0,96> ;< Building, N_3, 0,60 >
< bus, N_2>	<car, N_1, 0,69> ; <automobile, N_1, 0,65> ; < ship, N_5, 0,51 >
< driver, N_2>	<automobile, N_1, 0,51>
< house, N_2 >	<home, N_1, 0,96> ; < building, N_3, 0,60 >
< auto, N_2>	<automobile, N_1, 1> ; <car, N_1, 1> ; < ship, N_5, 0,56 >
< hospital, N_3>	< university, N_4, 0,54 > ; <medicine, N_8, 0,64>
< building, N_3>	< house, N_2, 0,60 > ; < home, N_1, 1 >; <building, N_6, 1>
< university, N_4 >	<medicine, N_8, 0,64> ; < hospital, N_3, 0,54>
< plain, N_4 >	< ship, N_5, 0,51 >
< ship, N_5 >	< plain, N_4, 0,51 >; <car, N_1, 0,51>; <hull, N_7, 0,67>; <sea, N_8, 0,50>; <bus, N_2, 0,51> ; <auto, N_2, 0,56> ;
< Show, N_6 >	
<building, N_6>	<tablelands, N_7, 0,50> ; < building, N_3, 1>
<tablelands, N_7>	<building, N_6, 0,50>
<hull, N_7>	< ship, N_5, 0,61 >
<sea, N_8>	< ship, N_5, 0,50 >
<medicine, N_8>	< university, N_4, 0,64 > ; < hospital, N_3, 0 ,64>

Table 15. Cluster global index

6. CONCLUSION

In this article, we presented a new approach to a large scale data integration system. Within this framework, we worked out a model for the global index construction which makes it possible to combine all the existing concepts on the network as well as the semantic links established between them. On the basis of this model, we presented the algorithms of global index construction and we illustrated them for an example.

As a prospect in this work, we propose to find a method for the dividing up of the global index, placing it in a grid.

7. REFERENCES

Agustina, B. and Alejandra, C. and Nieves R, B. (2003). An ontology approach to data integration. Journal of Computer Science and Technology.

Agustina, B. and Alejandra, C. and Nieves R., B. (2005). Ontology-Based Data Integration. Encyclopedia of Database Technologies and Applications. 450-456

Barkallah, B. and Moalla, S. (2007). Data integration Models for large scale systems: a comparative study. Proceedings of ICSSEA. Paris, France.

Chawathe, S. and Garcia-Molina, H. and Hammer, J. and Ireland, K. and Papakonstantinou, Y. and Ullman, J. and Widom, J. (1994). Integration of heterogeneous information sources. Proceedings of IPSI conference. Tokyo, Japan.

Hacid, M. and Reynaud, R.C. (2005). L'intégration de sources de données. Revue Information Interaction, Intelligence (I3) Une Revue en Sciences du Traitement de l'information.
Joseph A, G. (2005). Data, Schema, Ontology and Logic Integration. Logic Journal of the IGPL. Volume 13, Number 6. Oxford.
Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. International Joint Conference for Artificial Intelligence (IJCAI-95). 448-453.
Susanne, B. and Ralf-Detlef, K. and Ulf,L. and Herbert, W. (1999). Federated information systems: concepts, terminology and architectures.
Ted, P. and Siddharth, P. and Jason, M. (2004). WordNet ::Similarity - Measuring the Relatedness of Concepts. Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2004). Boston.
Valduriez, P. and Pacitti, E. (2004). Data Management in Large-scale P2P Systems. 6th Int Conf on High Performance Computing in Computational Sciences. Valence, Espagne.
Zargayouna, H. and Salotti, S. (2004). Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. In Proceedings of IC.