Conceptual Synopses of Semantics in Social Networks Sharing Structured Data

Verena Kantere[1], Maria-Eirini Politou[2], and Timos Sellis[2]

[1] Ecole Polytechnique Fédérale de Lausanne
verena.kantere@epfl.ch
[2] School of Electr. and Comp. Engineering,
National Technical University of Athens
{politou,timos}@dbnet.ece.ntua.gr

Abstract. We are interested in the problem of data sharing in overlay networks that have a social structure, i.e. participants are linked and exchange data with others w.r.t. the similarity of their data semantics. In this paper we propose a methodology to produce conceptual synopses for the semantics that are encapsulated in the schemas of relational data that are shared in a social network. These synopses are constructed solely based on semantics that can be deduced from schemas themselves with some optional additional conceptual clarifications. The produced synopses represent in a concentrated way the current semantics. Existing or new participants can refer to these synopses in order to determine their interest in the network. We present a methodology that employs the conceptual synopsis for the construction of a mediating schema. These can be used as global interfaces for sharing of information in the social network. Furthermore, we extend our methodology in order to compress the conceptual synopsis such that infrequent concepts are eliminated and the respective inferred global schema encapsulates the most popular semantics of the social network.

1 Introduction

Social networks are structures that map semantic relations of the members to overlay links. In such a network, linkage usually follows the unstructured model, i.e. members are connected to the most similar others and are aware of a small part of the network.

We are interested in social networks that share structured data, i.e. data that adhere to a schema; Our focus is the relational model, since it is the most commonly-used one in practice to represent the structured data. Linked, or else, *acquainted* members of such networks create and maintain sets of mappings between the schemas that their data conform to. These mappings are necessary in order for the acquaintees to understand each other, not only in terms of semantics, but also in terms of data structure; these mappings offer a way to organize and compromise their intra data-sharing [1, 5].

In a broad network which hosts a set of social groups, prospective members need guidance in order to select the groups they desire to participate. Thus, they would benefit from information that is related to member ids or names, but, more essentially, to the content that is shared. Members of such networks need additional assistance in order to match their data to the data of members with similar interest. Actually, they would

R. Meersman and Z. Tari (Eds.): OTM 2008, Part II, LNCS 5332, pp. 1367–1384, 2008.

benefit even more from summarizations of semantics, if they could infer from them a mapping scheme for their own data to other shared data.

A global conceptual synopsis and, furthermore, a respective mediating schema gives the opportunity to the participants of the social network to get answers that adhere better to the semantics of their queries for data, since query loss of information due to successive query rewriting is avoided [7, 14]. Beyond this, the conceptual synopsis enables joining members to obtain an overall idea, make an "educated guess" about the semantics of the data shared in this social network. Moreover, the respective mediating schema can be used for the creation of direct mappings that facilitate the data exchange with the total of participants.

A practical problem related to conceptual synopses of semantics is to limit its size such that it contains only the important semantics. This is vital in order to prevent users of the synopses to be lost in or misled by semantics that are actually of subordinate significance, in their effort to understand the nature of the respective social network.

In this paper we deal with the problem of creating a conceptual synopsis for the semantics of a social network employing solely the available schema and mapping information, as well as any optional conceptual clarifications that may be held by the network members. We aim at the minimization of human involvement in this process, as well as to offer tools for conceptual representation that are, on one hand, intuitive and can be manually used in a straightforward manner to express basic human rationale, and, on the other, capable of representing semantics that can be inferred from schemas and mappings. We explore a methodology that allows the deduction of schema and mappings semantics and their unification with additional optional manually expressed clarifications on them. This methodology creates a conceptual synopsis of the respective semantics. We employ the conceptual synopsis in order to construct a global schema that represents adequately the semantics of the respective social network. These can be used as mediating interfaces for sharing of information in the social network.

Furthermore, we consider the practical problem of refining the complete conceptual synopses in order to maintain only the dominant semantics. We solve this problem by proposing a methodology for the compression of the synopses that tracks infrequent semantics, that are also of limited interest to the members, and eliminates them.

Finally, we study thoroughly the quality of the global schemas produced with our methodology experimenting on two use cases.

After briefly discussing related work in section 2, in section 3 we formalize the problem. Section 4 describes the methodology for the deduction of the conceptual synopsis of a social network and section 5 presents the construction of the respective global schema emphasizing on compressed synopses. Section 6 summarizes the experimental study and section 7 concludes this paper.

2 Related Work

The problem of semantic schema merging is generally related to the problems of schema or ontology matching and integration. The recent survey in [13] approaches in a unified way all these problems, since they are basically dealing with schema-based matching. A survey of ontology mapping techniques is presented in [6]. The authors focus on the

current state of the art in ontology matching. They review recent approaches, techniques and tools. Once appropriate mappings between two ontologies have been established, either manually, semi-automatically or automatically, these mappings can be used to merge the two ontologies or to translate elements from one ontology to the other. Examples of tools for ontology merging are OntoMerge [4] and PROMPT [10]. However, creating and maintaining a merged ontology incurs a significant overhead. Moreover, a translation service for OWL ontologies is presented in [8]. The translation relies on a provided mapping between the vocabularies of the two ontologies.

Schema matching is a fundamental issue in the database field, from database integration and warehousing to the newly proposed P2P data management systems. As discussed in [12], most approaches to this problem are semi-automatic, in that they assume human tuning of parameters and final refinement of the results. This is also the case in some recent P2P data management approaches (e.g., [3, 11]). Generally, schema matching [12] and integration [2] are operations that adhere to schema structure in a strict way. Thus, most of the effort is concentrated in detecting and compromising contradictory dependencies and constraints.

Ontology matching/integration is a very similar problem to schema matching/integration. As discussed in [9, 13], both ontologies and schemas provide a vocabulary of terms with a constrained meaning. Yet ontologies and schemas differ in the declaration of semantics: on one hand ontologies specify strict semantics and on the other hand schemas do not specify almost at all explicit semantics. Because of this vital difference and of different aims and usage, ontology matching/integration has to follow a strict semantics structure, whereas schema matching/integration has to obey to strict structural semantic-less constraints. Moreover, different aims lead schema matching/integration to adhere to structural similarity that may or not encompass some similarity of semantics and ontology matching/integration to the opposite. Our work is an effort to complement these approaches by filling their gap. Our focus is the semantics that can be deduced from schemas without being restrained by the schema structure. Instead of making the overly strict assumption that these semantics adhere to an a priori full-fledged ontology, we consider the existence of optional basic clarifications on semantics.

3 Problem Definition

We consider a flat network of nodes (i.e. without super-nodes) that share data stored in a relational DBMS and, thus, that comply to a relational schema. The latter is a set of relations and each relation a set of attributes. The only internal constraints of a schema are foreign key constraints. Pairs of nodes of the social network maintain schema mappings in order to be able to share data. As assumed in other related works [1, 5, 7], these mappings are actually bidirectional inclusion dependencies that match a query on the one schema to a query on the other. Furthermore, each acquaintance may be enhanced with some additional optional clarifications on concept matching using a set of available types of concept correspondences. We would like to deduce the semantics of such a social network employing only the available meta-information on the shared data, i.e. schemas, mappings and correspondences. We want to form this semantics into conceptual synopses that can be used for the better understanding of the participants'

(existing and new-coming) requirements and interests on data and for the fulfillment of them, by constructing a global mediating abstract (i.e. not to be populated) schema. In the following we describe in a formal manner the assumptions of this problem and the characteristics of the pursued solution.

A social network is a pair $(\mathcal{N}, \mathcal{M}, \mathcal{C})$, Here, $\mathcal{N} = (\mathcal{S}, \mathcal{L})$ is an undirected graph, where $\mathcal{S} = \{S_1, \cdots, S_n\}$ is a set of nodes, $\mathcal{L} = \{(S_i, S_j) \mid S_i, S_j \in \mathcal{S}\}$ is a set of acquaintances; each acquaintance (S_i, S_j) is associated with a set $\mathcal{M}_{ij} \in \mathcal{M}$ of mappings and a set of correspondences $\mathcal{C}_{ij} \in \mathcal{C}$.

Each $S \in \mathcal{S}$ is a relational schema, i.e. is a nonempty, finite set $\{R_1[\mathrm{A}_1], \ldots, R_n[\mathrm{A}_n]\}$, where $R_i[\mathrm{A}_i]$, $i = 1, \ldots, n$ denotes a relation R_i over an ordered set A_i of attributes. An *instance* of a relation $I_{R[A]}$ is a (possibly empty) finite set of tuples $\langle t_1, \ldots, t_m \rangle$ where t_i, for $i = 1, \ldots, m$, is an ordered set of constants c with $\mid t_i \mid = \mid A \mid$. A set $K(R_i) \subseteq A$ constitutes a *key* of R.

We assume the existence of a countable finite set of words $\mathcal{D}$ that constitutes the domain of the social network. This means that for each $S \in \mathcal{S}$, for each $R \in S$ the name of R, denoted as $name(R)$ takes value from $\mathcal{D}$, i.e. $name(R) \in \mathcal{D}$. In the same way, for each $A \in R$, $name(A) \in \mathcal{D}$ and for each $c \in t \in I_{R[A]}$, $name(c) \in \mathcal{D}$. Each member of $\mathcal{D}$ constitutes a distinct and possibly non-unique concept. Thus, the function $name(x) = y : \{x \mid x \in S.R, S.R.A, I_{S.R[A]}.t\} \mapsto \mathcal{D}$, gives the concept that corresponds to a schema element or a data value ($S.R$ denotes the R relation of schema S, $S.R.A$ denotes the A attribute of the R relation of schema S).

Considering two schemas S and a S', a mapping between them $M(S, S')$ is the set $\{Equ_M(S, S'), Cond_M(S, S')\}$, where the set of equivalences of concepts $Equ_M(S, S') = \{name(R.A) = name(R'.A') \mid R.A \in S, R'.A' \in S'\}$ holds under the set of conditions $Cond_M(S, S') = \{R_1.A = R_2.B$ or $R_1.A = const | R_1, R_2 \in S$ or $R_1, R_2 \in S'\}$; *const* is a data value.

Figure 1 depicts part of a social network that consists of two universities. The figure shows part of their schemas and some mappings that have been created in order

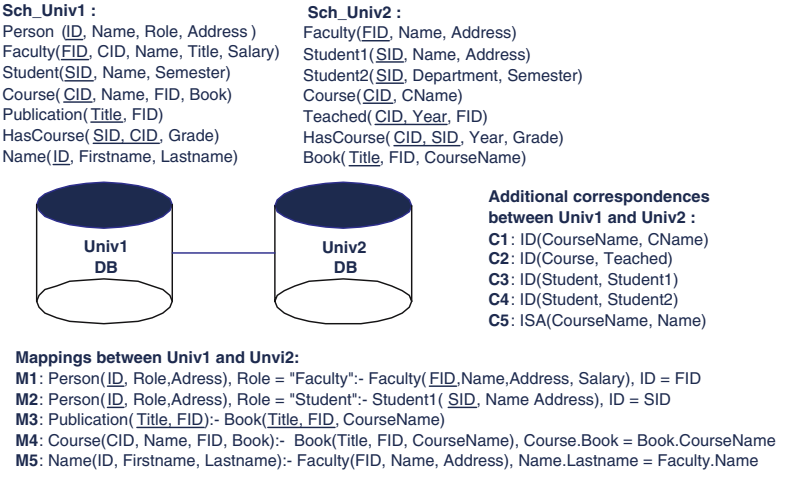

Fig. 1. Parts of the schemas of two universities that collaborate through a social network

to enable schema understanding and data sharing between them. Underlined attribute names refer to attributes that are part of the key of the respective relation. Each mapping corresponds to a conjunctive query on one schema to a conjunctive query on the other.

Beyond mappings, we consider that two acquainted nodes can also declare conceptual correspondences in order to optionally clarify or specify some conceptual relation between two schema elements, i.e. relations, attributes, or attribute values.

A conceptual correspondence CC is a a directed relationship between the concepts that correspond to two schema elements E_1, E_2 (i.e. a relation R, an attribute A, or an attribute value c). The concept of a schema element E is denoted as $name(E)$. Thus a conceptual correspondence CC between E_1, E_2, is declared as $CC(name(E_1), name(E_2))$. Note that the two schema elements do not have to be of the same type. Such a correspondence can be of 4 types: $CC \in \{ISA, ID, HASA, REL\}$. Especially the type ID is bidirectional, i.e. $ID(name(E_1), name(E_2)) \Leftrightarrow ID(name(E_2), name(E_1))$. The interpretations of the correspondence types are pretty straightforward and, thus, very easy to be used by administrators in order to declare some conceptual relations between the concepts of the schema elements E_1 and E_2, i.e. $name(E_1)$ and $name(E_2)$, respectively:

- $ISA(name(E_1), name(E_2))$: $name(E_1)$ is a specialization of $name(E_2)$
- $ID(name(E_1), name(E_2))$: $name(E_1)$ is identical with $name(E_2)$ and vice versa.
- $HASA(name(E_1), name(E_2))$: $name(E_2)$ is part of $name(E_1)$
- $REL(name(E_1), name(E_2))$: $name(E_1)$ is in generally related by an unspecified manner to $name(E_1)$

The ID type of correspondence means that the two members are different textual interpretations of exactly the same concept. The REL type of correspondence is associative. This type can be used by the administrator if she wants to declare that two concepts are related but she does not (a) want to specialize it to one of the other three types, (b) does not know if this relation can be specialized by another type, or (c) believes that this is a kind of relationship that cannot be represented by the other three types.

The correspondence types are not equally strong. The hierarchy of the four types described above is $(ID \succ ISA \succ HASA \succ REL)$, where $cc_j \succ cc_k$ means that cc_j is stronger than cc_k. This means that if there are more than one correspondence links between two schema elements, then the strongest one obliterates the rest.

In Figure 1 some examples of optional correspondences are shown. These can be very easily and intuitively formed in addition to the mappings for the schemas of the two universities by their administrators, as clarifications. For simplicity, in the examples we omit the function $name(.)$ and we denote corresponding concepts and schema elements with the same symbol.

A *conceptual synopsis* of a social network $(\mathcal{N}, \mathcal{M}, \mathcal{C})$ is represented by a directed labeled graph $CG = (V, E)$, where each vertex $v \in V$ is a distinct concept and each edge $e \in E$ is a correspondence. Specifically, each vertex $v \in V$ corresponds to one or more schema elements of the nodes participating in the social network, i.e. $v = \{name(x) \mid x \in S.R, S.R.A, I_{S.R[A]}, S \in \mathcal{S}\}$; also, each edge $e \in E$ corresponds to an element of $\mathcal{C}$ that includes correspondences that have been explicitly expressed and added to $\mathcal{C}$ at the point of acquaintance creations, or correspondences that are deduced in some way from the mappings $\mathcal{M}$. Note that a conceptual synopsis can summarize all or some of the semantics of the social network.

A global schema *GS* of a social network is a relational schema that is coherent with the respective conceptual synopsis represented by *CG*. This means that each concept and each correspondence in *CG* is represented in a lossless way in *GS*, such that we can use *GS* in order to reconstruct *CG*. *GS* can be employed as a mediating schema for data sharing in the social network.

The conceptual synopsis is a flatter and simpler version of an elaborated ontology; yet it is still very expressive since it allows any kind of four essential kinds of relationship between any two nodes. The conceptual synopsis is an intuitive description of a set of concepts and can be easily constructed by given simple concept correspondences.

In the following sections we will describe algorithms that can construct the conceptual synopsis of a social network by employing existing concept correspondences and deducing correspondences from the schema mappings. Moreover, we will discuss how a conceptual synopsis can be compressed in order to summarize the most frequent concepts. Finally we will present the algorithm for the construction of the global schema that corresponds to a conceptual synopsis.

4 Creation of Conceptual Synopses

In this section we describe the steps for the creation of a conceptual synopsis that represents the complete semantics of a social network. Briefly, a conceptual synopses is created for each individual schema that participates in the social network. These synopses are merged in a serialized way (according to existing acquaintances) employing predefined concept correspondences as well as correspondences that are deduced from the existing schema mappings. Finally the merged conceptual synopsis is refined. The algorithm is summarized in Figure 2 and described in detail in the following.

Creation of the conceptual synopsis
Input: Two relational schemas S_1 and S_2, a set of mappings $\mathcal{M}_{12}$, a set of additional concept correspondences $\mathcal{C}_{12}$ and an existing conceptual synopsis $CG = \{V, E\}$
Output: A global conceptual synopsis $CG' = \{V', E'\}$
Initialization: $CG' = CG$
Step1: Represent S_1 and S_2 as a conceptual synopsis, CG_1 and CG_2, respectively.
Step2: For each mapping $M \in \mathcal{M}$:
- extract the conceptual correspondences C
- add these correspondences to the existing ones: $C \cup \mathcal{C}$
Step3: Merge the conceptual synopses CG_1, CG_2 with CG'
Step4: Refine the CG' by:
- adding the correspondences in $\mathcal{C}$
- removing subsumed correspondences
Step5: Return CG'.

Fig. 2. Algorithm for the creation of the global conceptual synopsis

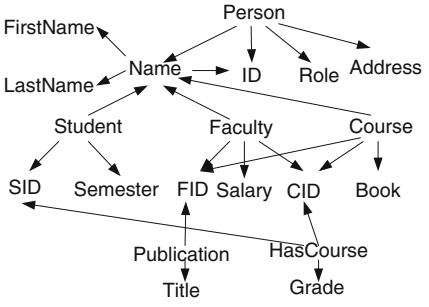

Fig. 3. Conceptual synopsis from Univ1

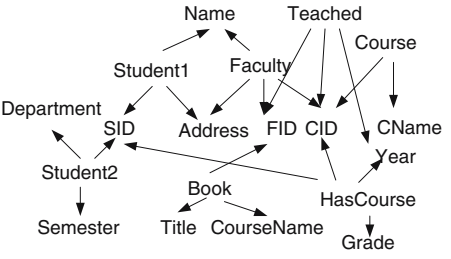

Fig. 4. Conceptual synopsis from Univ2

4.1 Creating the Conceptual Synopsis of a Schema

We use the schema of each node that participates in the social network in order to deduce a relevant conceptual synopsis. This synopsis represents in a concise and intuitive manner the domain for which this node stores and shares data. In order to produce the conceptual synopsis of a schema we create one vertex for each relation in the schema. Formally, for a schema $S = \{R_1,..,R_m\}$ of m relations $CG_S = (V_S, E_S)$ of a schema S, we instantiate the set of vertices as $V_S = \{V_1,\ldots,V_m\}$, such that $V_i = name(R_i)$, for $i = 1,\ldots,m$. For each relation $R_i = \{A_{i1},\ldots,A_{in}\}$ we add to V_S respective vertices for all the attributes A_{ij}, for $j = 1,\ldots,n$. The set of edges E_S is instantiated such that there is one entry E_i in the set for each pair of vertices (V_i, V_{ij}) where V_i is the respective vertex of relation R_i and V_{ij} is the respective vertex of the attribute A_{ij} of R_i. Assuming that foreign key constraints between the i^{th} and k^{th} relations are actually represented as sets of equivalences: $A_{ij} \equiv A_{km}$, for some j and some m in the arity of the respective relations, the conceptual synopsis is connected[1]. Also, note that duplicates in V_S are collapsed, in order for the synopsis to represent a concept uniquely. Hence, after the elimination of the duplicates, each vertex in V_S has a unique name, which is the name of the concept that it represents.

Figures 3 and 4 show the conceptual synopses that are created from the schemas of the two universities of Figure 1. For simplicity, the graphs in these and the following figures do not label the *HASA* correspondences.

4.2 Merging Conceptual Schemas

A set of conceptual schemas are merged sequentially, employing the mappings that they hold in pairs. In order to merge two conceptual synopses there are two coarse steps: (a) first, we add edges that connect semantically the graphs, and (b) second, we collapse vertices with the same name, i.e. vertices that represent the same concept. In order to add inter-graph edges, we employ the knowledge we may have about the semantics interrelations of the schemas that are the origins of these graphs. These interrelations

[1] Note that the conceptual synopsis may be a non-connected graph. This occurs in the rare case that the relations of a schema do not have any foreign key constraints.

Extracted correspondences from mappings between Univ1 and Univ2 :

C6: ISA (Faculty, Person)	**C9**: ISA(Book, Publication/Book)
C7: ISA(Student, Person)	**C10**: REL(Course, CourseName)
C8: ISA(Publication, Publication/Book)	**C11**: REL(Faculty, LastName)

Fig. 5. Conceptual correspondences extracted from the mappings

are denoted by situations such as: identical relation or attribute names, identical relation keys, value conditions on attributes, etc. Depending on the presence of these situations conceptual correspondences between schema elements can be deduced. Such rules can be derived from basic and intuitive rationale as well as from studies of use cases. We have concluded with the following set of rules that guide the procedure of the deduction of concept correspondences from the mappings between two schemas.
For a mapping M between two schemas S_1, S_2:

- If there is a relation $R_1 \in S_1$ and a relation $R_2 \in S_2$ for which $name(R_1) = name(R_2)$ and the share the same key: $\forall A_1 \in K(R_1), \exists A_2 \in K(R_2)$, s.t. $name(A_1) = name(A_2)$, and vice versa, then the respective vertices are joined with an ID correspondence.
- If there is a relation $R_1 \in S_1$ and a relation $R_2 \in S_2$ that share all their attributes, i.e. $R_1(A_1,\ldots,A_k)$ and $R_2(A_1,\ldots,A_k)$, where $name(R_1.A_i) = name(R_2.A_i)$, for $i = 1,\ldots,k$, then the respective vertices of R_1 and R_2 are joined with an ID correspondence.
- If there is a relation $R_1 \in S_1$ and a relation $R_2 \in S_2$ that share the same key and there is a value condition on one of them, e.g. $R_1.A_j =< constant >$, then a $ISA(V_2,V_1)$ correspondence is added for V_2, V_1 which are the corresponding vertices of R_2, R_1, respectively.
- If there is a correspondence between two attributes of the two involved schemas: a relation attribute $A_1 \in R_1 \in S_1$ corresponded in the mapping with a relation attribute $A_2 \in R_2 \in S_2$ and $name(R_1) = name(A_2)$, then we add $REL(V_1,V_2)$, where V_1, V_2 are the corresponding vertices of R_1 and A_2, respectively.
- If there is a relation $R_1 \in S_1$ and a relation $R_2 \in S_2$ that share the same key then we add a new vertex V and we add the correspondences $ISA(V_1,V)$ and $ISA(V_2,V)$, where V_1, V_2 are the corresponding vertices of R_1 and R_2, respectively.

Figure 5 shows the correspondences that can be deduced from the schema mappings of Figure 1 employing the described set of rules. Using these correspondences, as well as the optional correspondences defined by the administrators (see Figure 1), the conceptual synopses of the two university schemas (see Figures 3 and 4) can be merged. Figure 6 shows the first step of merging, where only ID correspondences have been processed. We remind that $HASA$ correspondences are not labeled.

4.3 Refining the Global Conceptual Synopsis

After the global conceptual synopsis is produced, it is often the case that there are redundant edges between pairs of vertices of the graph. Thus, the latter is refined so that it contains only one edge between each pair of vertices. The following simple steps are taken:

- merging of vertices that are linked with an *ID* correspondence
- eliminating all subsumed correspondences
- eliminating $ISA(V_1, V_k)$ correspondences, if there exist also the correspondences $ISA(V_i, V_{i+1})$, for $i = 1, \ldots, k-1$
- substituting $HASA(V_i, V_j)$ correspondences, if there is a set of correspondences $ISA(V_i, V_{i+1})$, for $i = 1, \ldots, k-1$ with the correspondence $HASA(V_k, V_j)$

It is evident that the role of *ID* correspondences in the synopsis is associative, since they actually denote that there is a duplication of a concept. In order to simplify the picture of the synopsis, we eliminate the *ID* correspondences; these can be entered in an accompanying dictionary, which can be referenced later by joining members of the social network. Moreover, it is often the case that after the merging of individual conceptual synopses, there are multi-linked vertices. Hence, we eliminate all the edges between two vertices except one: the one with is over the others in the hierarchy of correspondences. Also, we eliminate redundant *ISA* correspondences . Finally, we substitute *HASA* correspondences to specialized (through *ISA* ones) with the *HASA* correspondences towards the most generalized respective concepts. Figure 7 shows the refined merged conceptual synopsis for the schemas of the two universities of Figure 1.

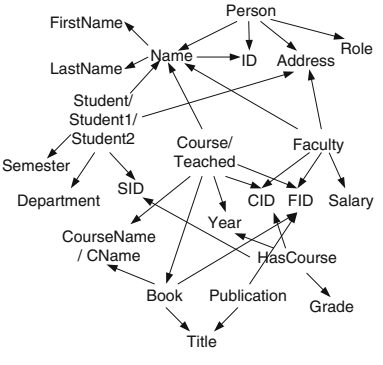

Fig. 6. Merged conceptual synopsis after adding ID correspondences

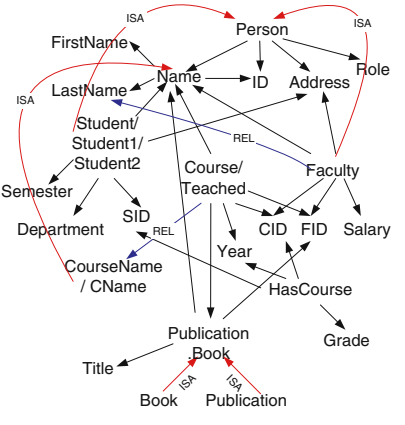

Fig. 7. Final global synopsis

5 Creation of the Global Schema

The conceptual synopsis can be employed in order to produce a global abstract schema that can be used as a mediator for the data sharing of the members of the social network. In the following we discuss how this global schemas can be constructed from the complete conceptual synopsis, but, also, from compressed versions of the latter. The complete conceptual synopsis encapsulates the total of the concepts that are included in the semantics of all the participants in the social network. Yet, it is often the case that the individual semantics of each member, comprise also concepts that are only of local interest; such concepts are not representative of the common network semantics and, therefore, it is suitable to eliminate them from the respective conceptual synopsis.

Global_Schema_Extraction
Input: A conceptual synopsis $CG = \{V, E\}$
Output: The global schema $GS = \{R_i\}$, for some integer i
Initialization: $GS = \emptyset$
Step1: For each vertex $V_j \in V$ that has outgoing edges of type *HASA* or *ISA* create a new respective relation R_j, i.e. $name(R_j) = V_j$, and add this to GS.
Step2: For each created relation R_j, insert an attribute A_{jk} for each respective concept $V_{jk} \in V$, i.e. $name(A_{jk}) = V_{jk}$, that has an incoming *HASA* correspondence from V_j.
Step3: For each relation R_j that corresponds to a concept V_j that has no outgoing *ISA* edges, determine the subset of the attributes that constitute the key.
Step4: For each relation R_j that corresponds to a concept V_j that has outgoing *ISA* edges, define the key of the relation to be the set of the following attributes:
- add to R_j the attributes that are part of the key of each relation R_k that corresponds to a concept V_k for which there is an edge $ISA(V_k, V_j)$;make these part of the key.
- optionally add more existing attributes of R_j to the key.
Step5: For each relation R_j that corresponds to a concept V_j that has no outgoing *ISA* edges, optionally add more existing attributes of R_j to the key.
Step6: If there are two relations R_i, R_j and there is an attribute of the first, A_{ik}, such that both A_{ik} and R_j correspond to the same concept in V, then:
- if there are no edges $REL(V_j, V_{in})$ such that there is also $ISA(V_{in}, V_i)$, where $V_i = name(A_i)$, substitute A_{jk} with the key of relation R_i,
- else, substitute A_{jk} with the attributes A_{in} that correspond to the concept V_{in}.
Step7: Return GS.

Fig. 8. Construction of the global schema extraction from the conceptual synopsis

5.1 Complete: Keeping All Concepts

The general algorithm that produces a global mediating schema involving all the concepts of a conceptual synopsis is presented in Figure 8. Summarizing, the algorithm creates a relation in the global schema for each concept of the conceptual synopsis that comprises other concepts (outgoing *HASA* correspondences) or are specializations of other concepts (outgoing *ISA* correspondences). Relations that are created from specialized concepts inherit as foreign keys the keys of relations that correspond to the most generalized concepts (respective concepts with no outgoing *ISA* correspondences). We comment here that in order to produce the keys of the relations, we must have some knowledge about at least the basic concepts that are deduced from relation keys in the participating individual schemas [2]. Otherwise keys have to be selected either randomly or based on some heuristic (e.g. number of incoming/outgoing and type of edges); yet, such a method cannot guarantee to produce the most rationally selected keys in the

[2] We omit a full discussion on the determination of keys due to lack of space. We note that knowledge of the eligibility of concepts to produce keys must formally be encoded in the structure of the conceptual synopsis. However, for the sake of simplicity, we have omitted this formality in this paper.

Sch_Global _wo_compression:
Person (ID, Role, Address)
Faculty(FID,ID Lastname, CID, Salary)
Student(SID, ID, Semester, Department)
Course(CID, CourseName, FID, Title, Year)
Publication.Book(Title, FID, ID)
HasCourse(SID, CID, Grade, Year)
Name(ID, Firstname, Lastname)
Book(Title)
Publication(Title)

Sch_Global _w_compression_1:
Person (ID, Role, Address)
Faculty(FID,ID Lastname, CID, Salary)
Student(SID, ID, Semester, Department)
Course(CID, CourseName, FID, Title, Year)
Publication.Book(Title, FID, ID)
HasCourse(SID, CID, Grade, Year)
Name(ID, Firstname, Lastname)

Sch_Global _w_compression_2:
Faculty(FID,ID Role, Address, Lastname , CID, Salary)
Student(SID, ID, Role, Address, Semester, Department)
Course(CID, CourseName, FID, Title, Year)
Publication.Book(Title, FID, ID)
HasCourse(SID, CID, Grade, Year)
Name(ID, Firstname, Lastname)

Fig. 9. Global schema construction from the global conceptual synopsis

general case. Finally, *REL* correspondences are checked in order to specialize the concept representation in the global schema by suitably replacing some relation attributes. Figure 9 presents the global schema constructed from the conceptual synopsis of Figure 7.

5.2 Compressed: Eliminating Infrequent Concepts

Sometimes a global conceptual synopsis is very large since it comprises not only concepts that are frequent among the participants in the social network, but also the seldom ones that interest only very few participants. Hence, there is a need for an algorithm than can produce a global schema that includes only the most frequent, and, therefore, most popular concepts. In order to achieve this, we propose the compression of the conceptual synopsis so that infrequent concepts are eliminated. Then, the summarized global schema is constructed with the algorithm of Figure 8 from the compressed conceptual synopsis. The compression of the latter is performed with the algorithm shown in Figure 10.

The algorithm is guided by the coarse rationale that a global schema is preferred to include fewer relations with more attributes, rather than more relations with fewer attributes. The reason is that this global schema is intended to be used as a mediator with which existing or new participants will have to create schema mappings. The latter are easier to be constructed if there is not much need for joins between relations. Nevertheless, the global schema is not purposed to be populated; thus, there is no fear that the few relations with many attributes will be filled with sparse tuples. Therefore the algorithm chooses to eliminate concepts that do not have outgoing *HASA* but may have incoming *ISA* correspondences. Overall, the algorithm is guided by the logic that elimination of concepts that are specializations or that are multi-linked should be avoided, since this would cause permanent loss of semantics and, as a side-effect, additional loss of more semantics due to probable disconnection of the graph. The algorithm terminates when a pre-specified limit of compression has been reached. This limit refers to the size of the

Conceptual_Synopsis_Compression
Input: A global conceptual synopsis $CG = \{V, E\}$
Output: A compressed conceptual synopsis $CG' = \{V', E'\}$
Initialization: $CG' = CG$
Step1: Concept elimination of is guided by the following set of rules, checked in ascending order. Concept V_i is removed if:
Rule1: There are no outgoing *ISA* edges from V_i and no incoming *HASA*.
Rule2: There are the fewest outgoing *ISA* edges from V_i and no incoming *HASA*.
Rule3: There are no outgoing *ISA* edges from V_i and there are the fewest incoming *HASA*.
Rule3: There are the fewest outgoing *ISA* edges from V_i and the fewest incoming *HASA*.
Rule4: There are the fewest outgoing *ISA* edges from V_i, and the fewest incoming and outgoing *HASA*.
Rule5: There are the fewest outgoing *HASA* correspondences.
Step2: Concepts V_j that had an outgoing *ISA* edge to an eliminated concept V_i, inherit the latter's *HASA* edges.
Step3: Check if the required limit of compression is reached. If no, goto Step1.
Step4: Return CG'.

Fig. 10. Compression of the conceptual synopsis

C12: ID(CourseName, Name)
M6: Book(Title, FID, CourseName):-Name(ID, Firstname, Lastname) , Book.CourseName = Name.ID

Fig. 11. Solving the problem of misleading declarations about social network semantics

conceptual synopsis and can be expressed either in terms of storage requirements or in terms of the size of the global schema to be constructed from the synopsis. We prefer the second of the two and, specifically, we determine the schema size in terms of the number of included relations, since these are the principal schema features. Figure 9 shows the global schemas after compressing the conceptual synopsis twice and three times.

5.3 Problem Limitations

After presenting our approach for the creation of conceptual synopses and respective mediating schemas, we briefly comment on the natural limitations imposed by the assumptions of the problem. First, the quality of the conceptual synopsis, and therefore, of the mediating schema depends in a straightforward manner on the quality of the mappings and the correctness of the additional conceptual correspondences. Ideally, complete mappings and consistent correspondences between acquainted members can lead to the creation of a representative conceptual synopsis. However, the lack of an a priori default agreement on concept matching, makes it impossible to guarantee the creation of infallible conceptual synopses. For example, observe the additional correspondence $C5$ in Figure 1, which denotes that "CourseName" is a special kind of "Name". If the latter is indeed used in a very broad manner, then this estimation is correct;

however, if "Name" turns out to refer only to people's names, then this estimation is wrong. Wrong or controversial concept matching estimations can be compromised with concept correspondences that are deduced from mappings and subsume the first. Moreover, the *REL* correspondence type can indicate special usage of concepts and can lead to a more correct schema construction. For example, assume that, additionally to the correspondences and mappings of Figure 1, there is correspondence $C12$ and mapping $M6$ in Figure 11. Thus, $C12$ subsumes $C5$, and, from $M6$, the correspondence $C13 : REL(Book, CourseName)$ is decuded. The latter leads to a refinement of the produced respective relation in the global schema: $Publication.Book(\underline{Title}, FID, CourseName)$.

Mistaken or inconsistent estimations on concept matching are possible and even unavoidable as the semantics of the social network refer to a broader domain of life. Naturally, broad concepts that are used with several meanings (such as the concept "Name") are certain to provoke confusion of semantics. Yet, social networks that target a more specific domain of knowledge, e.g. domain of a specific science, profession, sport, etc, are more eligible to use our proposed approach to the problem of conceptual synopsis in the lack of a global default ontology.

6 Experimental Study

In this section we present a summary of the experimental study that we have conducted in order to measure the efficiency of our technique in creating conceptual synopses and global schemas for social networks.

Experimental setup. In all the series of experiments that we have performed we have measured the similarity of the individual schemas of the participant databases in the social group with the global schema that is constructed from the deduced respective conceptual synopsis. The schema similarity comprises three partial similarity metrics; average similarity of (a) relations, (attributes), and (c) relation keys. Due to lack of space we present results only for the metrics (a) and (b). We note that we do not employ an overall schema similarity metric, since we believe that similarity of individual schema features is more informative.

We have conducted three groups of experiments. The first group studies the similarity of the global schema that is constructed from the complete conceptual synopsis, with the individual participating schemas. The second group studies the similarity of the global schema after compression of the conceptual synopsis, with the individual participating schemas. Finally, the third group of experiments studies the role of compression on the major schema features.

Experimental data. The experimental study has been performed for individual schemas of databases that participate in two social networks: (a) a network of hospitals and, (b) a network of universities. Specifically, for each one of these two domains, we have created a big pool of related concepts; we have given the latter to people with good knowledge of the database field and we have asked them to produce a relevant original schema with names of schema features or even data values that come from the respective pool of concepts. After collecting these original schemas, we have artificially produced additional new schema groups in order achieve schema similarities with values approximate

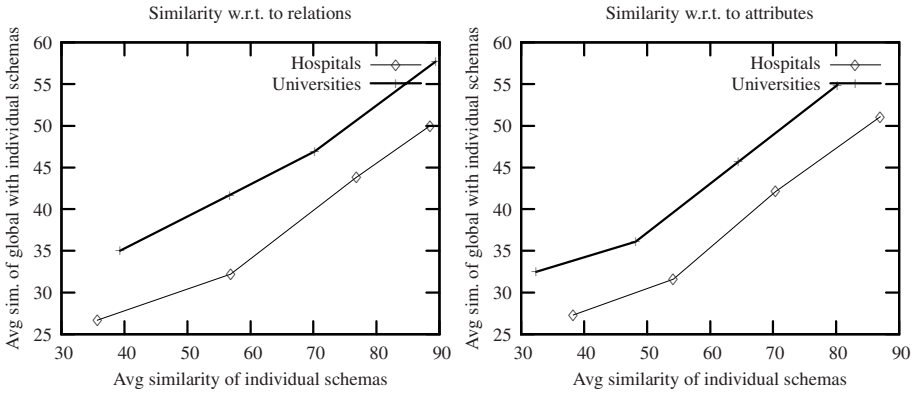

Fig. 12. Results for global schema constructed from the full conceptual synopsis

to the ones required by our experiments. The total of 100 (50 for the domain of hospitals and 50 for that of the universities) individual schemas is the input of our technique; the intermediate output of the latter are the conceptual synopses and the final output are the global schemas. It is interesting to note that the similarity of the original schemas w.r.t. to the relations is compatible with the respective similarity of keys. Overall, we followed this compatibility for the altered schemas.

6.1 Results for Similarity of Global and Individual Schemas

Figures 12(a),(b) show the average similarity of the global schema (inferred from the complete global conceptual synopsis) with the individual schemas, versus the average similarity of the individual schemas. Figure 12(a) shows results on the similarity of relations and 12(b) on the similarity of attributes. Both show smooth increasing average similarity of global with individual schemas, as the similarity of the latter increases. This is good since it indicates that the global schema encapsulates the overall semantics of the social network and the more coherent this semantics is, the more of it can be found in the global schema. However, the figures show that the similarity of the global with individual schemas is slightly more influenced by the similarity of the relations than the attributes of the individual schemas. Rationally, relations are considered to be more dominant schema elements and more determinant for the semantics of the schema (for example the lack of a relation influences more the schema semantics than the lack of an attribute, even if the lack of the relation does not entail the lack of attributes, too). Finally, the gradient of the synopses, (which is more abrupt as similarity increases), shows that as the individual schemas are more similar, the global schema naturally turns out to be overall more similar to all of them.

6.2 Results for Similarity of Compressed Global Schema and Individual Schemas

Figures 13(a), (b) show the average similarity of the compressed global schema (i.e. the global schema that is constructed after compression of the global conceptual synopsis)

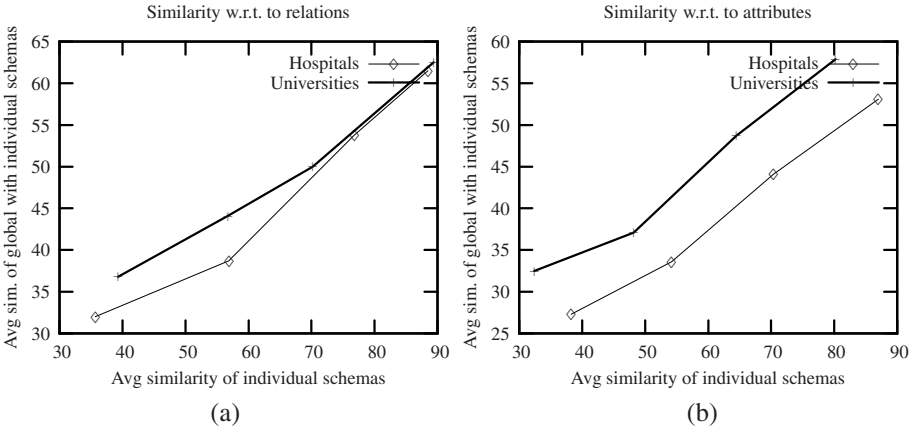

Fig. 13. Results for global schema constructed from the compressed conceptual synopses

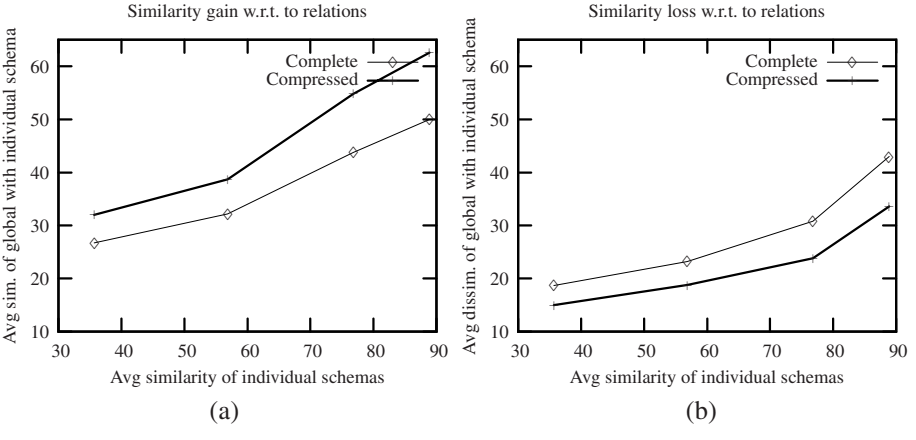

Fig. 14. Results for gain and loss in similarity with the individual schemas w.r.t. relations for complete and compressed global schema

versus the overall similarity of the individual schemas. Note that the degree of compression is constant in this set of experiments. The results are analogous to those of the respective Figures 12(a), (b). Again, it is verified that the role of relations is more critical to schema similarity than the role of attributes and, that, as the overall similarity of the individual schemas increases, their similarity with the compressed global schema increases with a faster rate. Moreover, a comparison between Figures 12, 13 reveals that the compressed schema is slightly more similar in general with the individual schemas than the respective complete global schema. We explore this interesting result more:

As indicated by Figure 13(a), the elimination of relations in the compressed global schema increases the overall similarity of it with the individual schemas. This means that the removed relations are indeed rare ones. As the similarity of individual schemas increases, this effect is even more obvious, since the rate of similarity increment

becomes bigger (this is very evident for the social network of hospitals). Figure 13(b) shows that the similarity of attributes is not so much affected by the compression, since the latter does not influence a lot the attributes.

Figure 14 shows the results for experiments on the gain and loss in similarity of the compressed and the complete schema with the individual schemas focusing on relations. The compressed schema differs from the complete schema in that some relations of the latter are not there in the first.

The gain in similarity is actually derived from the elimination of respective dissimilarity of the global compressed schema with the individual ones, due to the elimination of infrequent schema elements. The relations that are eliminated from the compressed schema, are still there in some individual schemas. This causes increment of the dissimilarity of the first with the latter, and constitutes actually the loss in similarity between them.

Figure 14(a) shows that the compression w.r.t. relations not only increases the similarity of the global schema with the individual ones, but also increases the rate of similarity increment. This means that there is substantial gain in similarity w.r.t. relations for the compressed vs. the complete global schema. Moreover, Figure 14(b) shows that the compression w.r.t. relations decreases also the dissimilarity of the global schema with the individual ones, which verifies that the eliminated relations are rare in the individual schemas. Naturally, the dissimilarity increases as the overall similarity of individual schemas increases, for both the complete and the compressed schema; yet, the rate of dissimilarity increment, and, therefore, the loss in similarity w.r.t. relations, is smaller for the compressed than the complete global schema.

This result is coherent with the intuitive rationale that the importance of schema features, (relations, in the case of these experiments), to the semantics of the social network is proportional to their frequency among the individual participating schemas. Going a step further and taking into consideration the experimental results, this means that the quality in terms of representative semantics of the global schema depends on the similarity of the individual schemas: the more similar they are, the better the complete and the worse the compressed global schema is. Hence, as final outcome, we form the following proposition:

For social networks with very similar participants the complete conceptual synopsis is necessary for the construction of the global schema, as each concept is valuable to the semantics of the network; whereas, for those with dissimilar participants, compression of the conceptual synopsis is a better option, as it tends to maintain the most important/frequent semantics and disregard the rest.

6.3 Results for Similarity for Different Degrees of Global Schema Compression

Figures 15(a), (b) show the results for experiments about various degrees of compression on the conceptual synopsis. These experiments are performed on schemas with 5 relations; Figure 15(a) shows results for the complete global schema as well as for 3 degrees of compression, where each degree corresponds to the elimination of one relation. Thus, the compression reaches up to 60% of the average size of the individual schemas in terms of relations. The average similarity of the global schema with

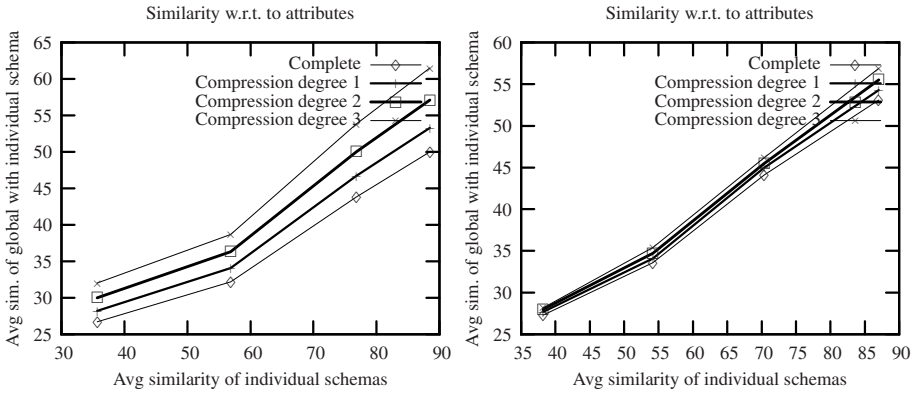

Fig. 15. Results for similarity of the compressed global schema with the individual ones for several compression degrees

the individual schemas increases with compression even up to 60%, even for schemas that are quite similar. With reference to previous experiments on gain and loss of similarity, this means that gain is bigger than loss, even for high degrees of compression w.r.t. relations. Figure 15(b) shows similar the results for compression w.r.t. attributes. In these experiments compression degrees refer to elimination of attributes [3]: degree i means that one attribute is eliminated. Naturally, the elimination of an attribute does not have a big good or bad impact to the average similarity of the global schema with the individual ones, although as this similarity increases, this impact becomes greater.

7 Conclusions

We tackle the problem of creating a conceptual synopsis for the semantics of a social network that shares relational data. We focus on networks that base their communication in schema mappings and that may hold some clarifications about conceptual matching. We propose a methodology for the deduction of conceptual correspondences from the schema mappings and integration of them in a refined way so that a synopsis that represents concept interrelations is produced. Using this synopsis we create a global mediating schema. We elaborate on the problem of producing aschema that maintains the most popular concepts of the social network and eliminates the concepts of limited interest. Finally, we perform an experimental study on the quality of the global schema.

References

1. Arenas, M., Kantere, V., Kementsietsidis, A., Kiringa, I., Miller, R.J., Mylopoulos, J.: The hyperion project: from data integration to data coordination. SIGMOD Record 32(3), 53–58 (2003)

[3] Of course, compression of attributes is achieved after elimination of relations has reached the maximum according to the compression rules: i.e. all the relations that are derived from ISA correspondences in the global conceptual synopsis are eliminated.

2. Batini, C., Lenzerini, M., Navathe, S.B.: A comparative analysis of methodologies for database schema integration. ACM Comput. Surv. 18(4), 323–364 (1986)
3. Bernstein, P.A., Melnik, S., Churchill, J.E.: Incremental schema matching. In: VLDB, pp. 1167–1170 (2006)
4. Dou, D., McDermott, D.V., Qi, P.: Ontology translation on the semantic web. J. Data Semantics 2, 35–57 (2005)
5. Halevy, A., Ives, Z., Suciu, D., Tatarinov, I.: Schema Mediation in Peer Data Management Systems. In: ICDE (2003)
6. Kalfoglou, Y., Schorlemmer, M.: Ontology Mapping: The State of the Art. Knowl. Eng. Rev. 18(1), 1–31 (2003)
7. Kantere, V., Tsoumakos, D., Sellis, T., Roussopoulos, N.: GrouPeer: Dynamic Clustering of P2P Databases. Technical Report TR-2006-4, National Technical University of Athens (2006) (to appear in Information Systems Journal), http://www.dbnet.ece.ntua.gr/pubs/uploads/TR-2006-4
8. Mota, L., Botelho, L.: OWL Ontology Translation for the Semantic Web. In: Proceedings of the Semantic Computing Workshop of the 14th International World Wide Web Conference (2005)
9. Fridman Noy, N.: Semantic integration: A survey of ontology-based approaches. SIGMOD Record 33(4), 65–70 (2004)
10. Fridman Noy, N., Musen, M.A.: Prompt: Algorithm and tool for automated ontology merging and alignment. In: AAAI/IAAI, pp. 450–455 (2000)
11. Ooi, B., Shu, Y., Tan, K.L., Zhou, A.Y.: PeerDB: A P2P-based System for Distributed Data Sharing. In: ICDE (2003)
12. Rahm, E., Bernstein, P.: A Survey of Approaches to Automatic Schema Matching. In: VLDB Journal (2001)
13. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. J. Data Semantics IV, 146–171 (2005)
14. Tatarinov, I., Halevy, A.: Efficient Query Reformulation in Peer-Data Management Systems. In: SIGMOD (2004)