Some problem with current approaches to P2P Information Integration, and a possible research direction

Guido Vetere, IBM Center for Advanced Studies of Rome

September 4, 2007

IBM Confidential

1 Introduction

Peer to Peer (P2P) networks are networks where each system (peer) can act both as data provider and consumer, without hierarchies or (crucial) dependencies on centralized components. P2P information integration takes place when, within some peer with respect to other peers, a dependency between provided and consumed information is established. Peers can either manage their own data schemas or adopt a shared ontology. In any case, they answer queries posed in a given terminology by accessing local data sources and/or by querying other peers basing on a suitable set of mappings. The striking difference between P2P systems and centralized ones is the distribution of the integration logic [5]. While traditional information integration architectures are based on specific components (e.g. *service bus*), P2P integration is, potentially, everywhere. This poses the problem of having *semantic commitments* (i.e. specific interpretations of schemas over data items) distributed all around the network, rather that contained within the boundaries of a single, specific, and controllable system. Hence, handling vagueness, misunderstandings, contradictions, and mistakes must be done in a distributed way, which makes P2P integration much more troublesome than classic, mediator-based information integration. In contrast with its inherent difficulty, P2P integration has a strong industrial value. P2P is the most basic network topology, is modular, flexible, robust, and, above all, reflects the intimate nature of the Web. In many cases, P2P integration is not an option, but a basic functional requirement. In fact, really *loosely-coupled* organizations, such as commerce, scientific, or governmental ones, might not be able (or willing) to set up centralized mediators. For instance, the Italian nationwide Web Service infrastructure currently under development at `CNIPA`[1]

[1] Centro Nazionale per l'Informatica nella Pubblica Amministrazione, http://www.cnipa.gov.it

is basically a P2P network. Research has recently focused on epistemic logics to account for the modular nature of peers in P2P networks [2]. In this light, information integration is generally seen as the *transfer of knowledge* from one peer to another. We will discuss some problem issued with this approach, then we will highlight a possible research direction based on a kind of approach which we will call *doxastic*.

2 Epistemic approach to P2P integration

It has been argued [2] that P2P integration requires *multi-modal epistemic logic* to model peers as providers of *knowledge* with respect to a set of *possible-worlds* (i.e. *interpretations* of a manifest theory) managed by peers in the network. The need of resorting to *multi-modal logics* comes from having different independent systems exchanging information about propositions which can be held for true by a peer and false by another one. In fact, truth-assignments are given by each peer independently, which disallows modeling the network as a single interpretation (*model*) of a global schema. In this setting, the notion of *knowledge* of a peer as *truth in every possible world it can access* replaces the notion of *truth in the unique accessible world* of database systems. Based on this finding, integration has been viewed as the *transfer of knowledge* from one peer to another [2] or as a propagation of queries posed to a peer and basing on its schema-level mappings [4]. Whether peers turn other peers knowledge into their own knowledge, or they limit themselves at reporting what other peers know, we will call *epistemic* this kind of approaches, in that each peer provides *knowledge about facts* with respect to the possible worlds it can access and the peers they are acquainted to. In line with the `HYPER` framework [3], given a set $\mathcal{P}$ of *peer systems* that export a schema (ontology), we consider a P2P mapping assertion as a pair of conjunctive queries $cq_j \rightsquigarrow cq_i$ on the schema of $p_j, p_i \in \mathcal{P}$ respectively, which share a vector of distinguished variables $\overline{\mathbf{x}}$:

$$\{\overline{x} | (\exists \overline{y}. cq_j(\overline{x}, \overline{y}))\} \rightsquigarrow \{\overline{x} | \exists \overline{z}. cq_i(\overline{x}, \overline{z})\}$$

The epistemic approach taken in [2] gives P2P mappings the following semantics:

$$\forall \overline{x}\ (\mathbf{K_j}(\exists \overline{\mathbf{y}}\,(cq_{\mathbf{j}}(\overline{\mathbf{x}}, \overline{\mathbf{y}}))) \rightarrow \mathbf{K_i} \exists \overline{\mathbf{z}}(cq_{\mathbf{i}}(\overline{\mathbf{x}}, \overline{\mathbf{z}})))$$

with $\mathbf{K_j}, \mathbf{K_i}$ denoting the knowledge of $p_j, p_i \in \mathcal{P}$ about the formula cq_j and cq_i respectively, according with *K45n* axiomatization [**?**] [2]. This way, *knowledge* of the peers mentioned in the left side of the assertion becomes knowledge of the peer mentioned in the right side. Then, the answer to a query $q(\overline{x})$ (an open formula with free variables) posed to a peer can be viewed as the set of tuples that satisfy the query according to the epistemic theory made of the P2P mappings in the network ($\mathcal{T}_K(\mathcal{P})$), the actual data stored in local sources ($\mathcal{D}$), and the epistemic state of the inquired peer (say, p_i):

[2] in this axiomatization, $K\phi$ does not imply ϕ, thus the operator $\mathbf{K}$ can also be called *belief*

$$\{\bar{t} | \mathcal{T}_K(\mathcal{P}) \cup DB(\mathcal{D}) \models_{K45n} \mathbf{K}_i q(\bar{t})\}$$

However, since a single peer can be (and typically is) connected with many different ones by way of semantically overlapping mappings, conflicts in the epistemic states of peers in the network can easily lead to inconsistencies which are hard to be dealt with [1]. On the other hand, just providing the union of semantically equivalent queries over the network, as suggested in [4], would not allow a peer to keep control over the information it returns, e.g. filtering unreliable sources in presence of inconsistencies.

A solution proposed in [1] consist in limiting knowledge transfers to those that do not inject inconsistencies in the peer where knowledge is received. It has to be noticed, however, that this strategy might introduce a dependency on the order in which the knowledge is transferred: when knowledge is borrowed from different peers, the first inquired peer could have a greater influence with respect to the following ones. There are many ways to work around this kind of problems (e.g. by setting *peer preferences*), but we want to throw the hypothesis that *knowledge transfer* could be a too strong way of modeling integration in P2P networks, in that it aims at transferring *truth-assignment*s, whereas, if they are really independent the one another, peers should rather borrow mere *opinions*.

3 A doxastic approach

We want to outline now a different approach, which we will call *doxastic* [3], based on a clear separation between *statements* and *facts*. This approach aims at overcoming some of the difficulties of merging different knowledges (e.g. handling inconsistencies) which are typical of knowledge-transfer, and yet avoids the non-committal choice of mere query propagation, which limits the role peers can play. At the basis of a *doxastic* approach is that, for each peer, knowledge of *statements* provided by other peers is kept apart from knowledge of *facts* in their own databases [4]. Thus, information coming from the network is not transferred in what the peer holds for true, nonetheless, this information can be retained and used for further reasoning, rather than being routed to the requester without adding any doxastic connotation.

The approach we want to outline here gives to a P2P mapping $cq_j \rightsquigarrow cq_i$ the following semantics:

$$\forall \overline{x}\ (\mathbf{K_j}(\exists \overline{\mathbf{y}}\ (cq_{\mathbf{j}}(\overline{\mathbf{x}}, \overline{\mathbf{y}}))) \rightarrow \mathbf{K_i}\mathbf{K_j}\exists \overline{\mathbf{z}}(cq_{\mathbf{i}}(\overline{\mathbf{x}}, \overline{\mathbf{z}})))$$

Informally, what is known by a $p_j \in \mathcal{P}$ is *known to be known by* p_j by $p_i \in \mathcal{P}$.

For example, suppose that the peer *Othello* maps its predicate `FAITHFUL` onto the peer *Iago*'s one. Now, *Othello* knows `FAITHFUL(Desdemona)`, while

[3] the term comes from ancient Greek *doxa* (opinion)

[4] this reflects the difference between *relational belief* (*de dicto*) and *notional belief* (*de re*)

Iago knows ¬`FAITHFUL(Desdemona)`. After integration, *Othello* knows that *Iago* knows ¬`FAITHFUL(Desdemona)`.

The answer to a query q to a peer p_i in a doxastic P2P system can be modeled as:

$$\{\bar{t} | \mathcal{T}_K(\mathcal{P}) \cup DB(\mathcal{D}) \models_{K45n} \mathbf{K}_i q(\bar{t})\} \bigcup_{j \neq i} \{K(\bar{t}, j) | \mathcal{T}_K(\mathcal{P}) \cup DB(\mathcal{D}) \models_{K45n} \mathbf{K}_i \mathbf{K}_j q(\bar{t})\}$$

where K is a suitable reification function that represents peers' knowledge in a set of special symbols which is part of the domain of discourse. Informally, the answer of a peer p_i to a query $q(\overline{x})$ is made of two distinct parts:

- the peer's own knowledge of facts, like in [2], with the provision that this knowledge is restricted to local information

- the peer's knowledge of facts known by other peers, of which the inquired peer is aware according to mappings

Running the example, if inquired on the matter, *Othello* would answer: *I know* `FAITHFUL(Desdemona)`. *Also, I know that Iago knows* ¬`FAITHFUL(Desdemona)`.

An immediate benefit of the doxastic approach is a property that can be called *consistency keeping*. It can be shown that local inconsistence of a peer cannot cause any other peer to be inconsistent, thus cannot be propagated to the network. If *Iago* knew `FAITHFUL(Desdemona)` ∧ ¬`FAITHFUL(Desdemona)` this would not be a problem for *Othello*, since $K_i K_j \phi$ does not contradict $K_i K_j \neg\phi$. Also, for any peer, conflicting knowledge coming from different peers would result in knowledge of the existence of different opinions, rather that inconsistency.

Still, at any time, a peer can decide to revise its own beliefs basing on some *justification*. For example, the peer *Othello* can draw the conclusion that ¬`FAITHFUL(Desdemona)`, based on *Iago*'s knowledge, just because it trusts *Iago* more than itself [5]. Different *theories of justification* are known in philosophy[6], but is not our purpose to discuss them here. We just want to stress that, by adopting a doxastic approach, any kind of belief revision is possible at any time, since any peer can suitably materialize the (reified) knowledge collected by the network, having it available for computation and, at the same time, setting it apart from its own knowledge.

4 Conclusion

P2P integration is of crucial relevance for the development of modern IT distributed systems based on the Web. Multi-modal systems provide the notion of knowledge (belief) which is promising to achieve a suitable model (i.e. decidable and tractable) for this kind of integration. However, the strategy of *knowledge transfer* (which we have called *epistemic*) raises the problem of handling consistency in a strong sense. The *doxastic* approach to P2P information

[5] note that the sad conclusion of Shakespeare's drama could not be modeled with in [1]

[6] to get a general idea, see http://en.wikipedia.org/wiki/Theory_of_justification

integration outlined here enjoys the same good properties of epistemic P2P multi-modal systems. In addiction, by demoting *knowledge transfer* (*de re*) to *opinion awareness* (*de dicto*), it guarantees consistency. Moreover, it allows a wide range of belief revision strategies to be adopted and implemented.

References

[1] D. Calvanese, G. D. Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Inconsistency tolerance in p2p data integration: An epistemic logic approach. In *DBPL*, pages 90–105, 2005.

[2] D. Calvanese, G. D. Giacomo, M. Lenzerini, and R. Rosati. Logical foundations of peer-to-peer data integration. In *PODS*, pages 241–251, 2004.

[3] D. Calvanese, G. D. Giacomo, M. Lenzerini, R. Rosati, and G. Vetere. Hyper: A framework for peer-to-peer data integration on grids. In *ICSNW*, pages 144–157, 2004.

[4] Z. Majkic. Intensional semantics for p2p data integration. pages 47–66, 2006.

[5] G. Vetere and M. Lenzerini. Models for semantic interoperability in service-oriented architectures. *IBM Systems Journal*, 44(4):887–904, 2005.