Finding Successful Queries in a Mediator Context

Alain Bidault, Christine Froidevaux and Brigitte Safar
L.R.I., C.N.R.S & Université Paris-Sud Bâtiment 490, 91405 Orsay Cedex, France

Abstract

In this paper, we study failing queries posed to a mediator in an information integration system and expressed in the logical formalism of the information integration system PICSEL. First, we present the notion of concept generalisation in a concept hierarchy that is used to repair failing queries. Then, we address two problems arising while rewriting a query using views. The first problem concerns queries that cannot be rewritten due to a lack of sources, the second one concerns queries that have only unsatisfiable rewritings.

1 Introduction

Data integration is a problem that has recently received significant attention due to the increasing number of structured and heterogeneous sources. Users and applications today require an integrated access to multiple heterogeneous information systems and many projects (e.g., Information Manifold [15], PICSEL [13], SIMS [1], TSIMMIS [7]) have been proposed. They all provide, in a mediator architecture, a uniform query interface to multiple autonomous data sources, located on the world wide web for example. Users pose queries in terms of a mediated schema, that is, in terms of a set of relations designed to capture the semantics of a given application domain (e.g., tourism). Those relations are virtual in the sense that their instances are not directly available but stored in the sources. As a consequence, answering a query means translating a user's query into a query that refers directly to the relevant sources, which needs a set of source descriptions. In our setting, sources are described by a set of views, for which logical constraints and a logical mapping with domain relations are specified.

The most important advantage of a mediator is that it enables users to focus on specifying their demand, by freeing them from having to find the relevant sources and possibly combine data from multiple sources to obtain answers. Instead, the mediator takes control of the construction of the specialised query plans (expressed in terms of views) to be executed in order to answer the original queries (expressed in terms of the domain model).

1PICSEL is a project supported by the CNET (Centre National d'Études des Télécommunications) under contract number 971B378.
2PICSEL is in collaboration with the travel agency Degriftour, http://www.degriftour.fr/
In the setting of data integration systems, the need for a cooperative query answering process is especially crucial because users do not know the contents of the data sources that are available.

Many works have been done in the field of cooperative answering ([10],[17],[4],[6]), particularly on the problem of failing queries i.e. queries having an evaluation that produces an empty answer set. In that situation, Godfrey [14] says that a query $Q$ may fail because one of its subqueries is evaluated to false in $DB$, $DB \not\models \exists X. Q(X)$: he says that it contains a false presupposition. Furthermore, if one of its subqueries leads to contradiction because of the integrity constraints associated to the database, $DB \models \neg \exists X. Q(X)$, he says that it also contains a misconception. Godfrey proposes to report to the user, in the general case, a minimal failing sub-query (MFS), and in the second case, a minimal conflicting sub-query (MCS). In both cases, he shows that looking for all the problematic sub-queries and finding all their repairs are NP-hard problems, which can become polynomial when exhaustiveness is not required.

Let us remark that MFS cannot be detected without effective access to the data of the sources while MCS can.

Gal [11] confronts the user's query with integrity constraints before consulting the database in order to give the user the integrity constraints that have been violated. As she does not offer any repair, she need not find the MCS.

Motro [18], [19] performs the MFS by testing all the immediate generalisations of the initial query and by iterating this process on all the failing generalisations. The MFS are those for which all generalisations succeed. A query generalisation is done by deleting one of its literals or by varying one of its parameters along a range of predefined values. Motro does not have to detect all the conflicts beforehand, but, instead, he must often consult the effective data of the sources.

Gaasterland [8][9] modifies the query either to get more information, which is relevant for the user, or to repair the query if it fails. She describes how to relax each predicate or constant arguments. As there are many possibilities, relaxations are generated in breadth first, and at each level, they are submitted to the user, who has to decide, interactively, which relaxation he prefers. If the query fails, Gaasterland re-uses Gal's work to propose, first, the relaxations that repair the query.

Our objective is to help a user to reformulate his failing query, without effective access to the contents of the sources. In a previous paper [2], we have considered the problem of repairing queries which do not obtain any answer, due to a violation of constraints. We have shown how to identify all the MCS and to construct minimal repairs, i.e. queries that generalise the user's query and that do not violate any domain constraint.

In this paper, we consider that the user's query, while being meaningful w.r.t the domain model, has no answer because its rewritings lead to specialised query plans that violate the constraints specifying the actual contents of the sources. For instance, the user asks for hotels located in England, and the only sources connected to the mediator provide hotels located in Germany. It is very useful to offer him a new query, called a repair, which is semantically close to the initial one and for