UMR IRISA

Project-Team Pilgrim

Gradedness, Imprecision, and Mediation in Database Management Systems

Lannion

Activity Report

2008

Contents

1 Team

Faculty Member
Olivier Pivert [Team Leader, Faculty member (Professor), Enssat, HdR]
Patrick Bosc [Professor Enssat, HdR]
Allel Hadjali [Associate Professor, Enssat]
Hélène Jaudoin [Associate Professor, Enssat]
Ludovic Liétard [Associate Professor, IUT Lannion]
Daniel Rocacher [Associate Professor, Enssat, HdR]

PhD Student
Mickael Dautrey [since December 1, 2006]
Nadia IbenHssaien [regional grant since October 1, 2004]
Amine Mokhtari [regional grant and Conseil Général 22 grant since October 1, 2007]

Administrative Assistant
Joëlle Thépault [Administrative Assistant, Enssat, 20%]
Nelly Vaucelle [Administrative Assistant, Enssat, 10%]

2 Overall Objectives

2.1 Introduction

In a majority of works undertaken in the area of database management systems (DBMSs), it is assumed that data are perfectly known and that only Boolean queries can be expressed. The aim of the project is to soften this double hypothesis by introducing the notions of imprecision and graduality in information systems. Such concepts impact at least two aspects of database systems:

1. the expression of queries addressed to DBMSs, which become gradual or flexible. The idea is to allow the user to specify preferences instead of Boolean conditions, thus leading to a result whose elements are ordered accordingly, e.g., find the *young* employees who work in a *high* budget department,

2. the description and the manipulation of imprecisely known data. Such data can be expressed in a linguistic way, e.g., a reservoir which is rather *big* or a person who is *young.*

The project reconsiders several aspects of DBMSs by taking imprecision and graduality into account. From a querying point of view, these two characteristics are indeed orthogonal, i.e., one may consider regular queries addressed to databases containing imprecise data, or flexible queries against regular databases. The originality of the project is to rely on a common scientific framework, namely fuzzy sets, to deal with these two sides and then to be able to consider the "joint" issue of flexible queries addressed to databases containing imprecise data.

The issue of flexible querying has received an increasing attention in the last years in the database community and the idea of including preferences inside queries gets more and more acceptance, even if there is a counterpart in terms of performances. The line followed in Pilgrim is to focus on:

1. various types of flexible conditions, including non trivial ones,
2. the semantics of such conditions from a user standpoint,
3. the design of query languages providing flexible capabilities in both relational and object-oriented databases.

Basically, a flexible query involves linguistic terms corresponding to gradual predicates, i.e., predicates which are more or less satisfied by a given (attribute) value. In addition, these various terms may have different degrees of importance, which means that they may be connected by operators beyond conjunction and disjunction. For instance, in the context of a search for used vehicles, one might say that he/she wants a *compact* car *preferably French*, with a *medium* mileage, *around* 6 k$, whose color is *as close as possible* to light grey or blue. The terms appearing in this example must be specified, which requires a certain theoretical framework. For instance, one may think that "*preferably* French car" is meant for a complete satisfaction for French cars, a lower one for Italian and Spanish ones, a still smaller satisfaction for German cars and a total rejection for others. Similarly, "*medium* mileage" can be used to state that cars with less than 40000 km are totally acceptable while the satisfaction decreases as the mileage goes up to 75000 km which is an upper bound. Moreover, it is likely that some of the conditions are more important than others (e.g., the price with respect to the color). In such a context, answers are ordered according to their overall compliance with the query, which makes a major difference with respect to usual queries.

In the previous example, conditions are fairly simple, but it turns out that more complex ones can also intervene. A particular attention is paid to conditions calling on aggregate functions together with gradual predicates. For instance, one may look for departments where *most* employees are *close* to retirement, or where the average salary of *young* employees is *around* $2500. Such statements have their counterpart in regular query language, such as SQL, and the specification of their semantics, when gradual conditions come into play, is studied in the project.

Along this line, the ultimate goal of the project is to introduce gradual predicates inside database query languages, thus providing flexible querying capabilities. Algebraic languages as well as more user-oriented languages are under consideration in both the original and extended relational settings.

Since a few years, another important line of research concerns the querying of databases in the presence of imprecise data. We think that this type of database should receive more and more attention in the future, for instance in the context of automated recognition, data fusion, forecasts or incomplete archives, in which imprecision is intrinsically present. Some works in the 80s have established that null values and disjunctive data have a strong impact on the queries that can be processed in an efficient way. This situation remains true in the context considered in Pilgrim, where the objective is to investigate various types of queries that can be processed efficiently in the relational setting. Although there is no hope for a general family of queries which would include all of the operators of the relational algebra, the identification of its largest subset which fits our requirements is undertaken. Beyond this target, specific queries that make sense and are tractable are sought, such as queries of the type: “to what extent is it possible that tuple t belongs to the result of query Q”, where t is a given tuple and Q is an algebraic query. Such queries are called possibilistic queries (due to the presence of the word “possible” in its statement) and they stem from regular yes/no queries.

Querying imprecise data calls on a given theory of uncertainty. Possibility theory is mainly used in Pilgrim, essentially because it provides a framework which is coherent with the one serving for flexible queries. So doing, it becomes possible to study the issue of flexible queries against imprecise databases. It is worth noticing that the case where probabilistic data are used instead of possibilistic ones, is also a matter of interest.

A new research topic in Pilgrim concerns flexible data integration systems. One considers a distributed database environment where several data sources are available. An extreme case is that of a totally decentralized P2P system. An intermediary situation corresponds to the case where a global schema is available and where the sources can be accessed through views defined on that schema (LAV approach). The problem consists in handling a user query (possibly involving preferences conveyed by fuzzy terms) so as to forward it (or part of it) to the relevant data sources, after rewriting it into the “language” of each selected source. The overall objective is thus to define flexible semantic mapping mechanisms and flexible query rewriting techniques which take into account both the approximate nature of the mappings and the graded nature of the initial query. A large scale environment is aimed, and the performance aspect is therefore crucial in such a context.

2.2 Highlights of the Year

The first highlight of the year concerns the high level of publication of the team in 2008: 10 articles in referred journals and book chapters and the edition of two special issues (Fuzzy Sets and Systems, and Journal of Intelligent Information Systems).

A second fact worthy to mention is the emergence of a new topic in the team, namely flexible route planning, in the domain of Intelligent Transportation Systems. A Ph.D. thesis started at the end of 2007 and some of the results obtained should be implemented in the MOB-ITS platform.

3 Scientific Foundations

The project investigates the issues of flexible queries against regular databases as well as regular queries addressed to databases involving imprecise data. These two aspects make use of two close theoretic settings: fuzzy sets for the support of flexibility and possibility theory for the representation and treatment of imprecise information.

3.1 Fuzzy sets

Fuzzy sets were introduced by L.A. Zadeh in 1965 [Zad65] in order to model sets or classes whose boundaries are not sharp. This is particularly the case for many adjectives of the

[Zad65] L. Zadeh, “Fuzzy sets”, *Information and Control 8*, 1965, p. 338–353.

natural language which can be hardly defined in terms of usual sets (e.g., high, young, small, etc.), but are a matter of degree. A fuzzy (sub)set F of a universe X is defined thanks to a membership function denoted by μ_F which maps every element x of X into a degree $\mu_F(x)$ in the unit interval $[0, 1]$. When the degree equals 0, x does not belong at all to F, if it is 1, x is a full member of F and the closer $\mu_F(x)$ to 1 (resp. 0), the more (resp. less) x belongs to F. Clearly, a regular set is a special case of a fuzzy set where the values taken by the membership function are restricted to the pair $\{0,\ 1\}$. Beyond the intrinsic values of the degrees, the membership function offers a convenient way for ordering the elements of X and it defines a symbolic-numeric interface. The α level-cut of a fuzzy set F is defined as the (regular) set of elements whose degree of membership is equal or over α and this concept bridges fuzzy sets and ordinary sets.

Similarly to a set A which is often seen as a predicate (namely, the one appearing in the intentional definition of A), a fuzzy set F is associated with a gradual (or fuzzy) predicate. For instance, if the membership function of the fuzzy set *young* is given by: $\mu_{young}(x) = 0$ for any x ≥ 30, $\mu_{young}(x) = 1$ for any $x < 21$, $\mu_{young}(21) = 0.9$, $\mu_{young}(22) = 0.8$, ... , $\mu_{young}(29) = 0.1$, it is possible to use the predicate *young* to assess the extent to which Tom, who is 26 years old, is young ($\mu_{young}(26) = 0.4$).

The operations valid on sets (and their logical counterparts) have been extended to fuzzy sets. Their definition assumes the validity of the commensurability principle between the concerned fuzzy sets. It has been shown that it is impossible to maintain all of the properties of the Boolean algebra when fuzzy sets come into play. Fuzzy set theory starts with a strongly coupled definition of union and intersection which rely on triangular norms ($\top$) and co-norms ($\perp$) tied by de Morgan's laws. Then:

$$\mu_{A \cap B}(x) = \top(\mu_A(x), \mu_B(x)) \qquad \mu_{A \cup B}(x) = \perp(\mu_A(x), \mu_B(x))$$

The complement of a fuzzy set F, denoted by $\bar{F}$, is a fuzzy set such that: $\mu_{\bar{F}}(x) = neg(\mu_F(x))$, where *neg* is a strong negation operator and the complement to 1 is almost always used. The conjunction and disjunction operators are the logical counterpart of intersection and union while the negation is the counterpart of the complement.

In practice, minimum and maximum are the most commonly used norm and co-norm because they have numerous properties among which:

- the satisfaction of all the properties of the usual intersection and union (including idempotency and double distributivity), except excluded-middle and non-contradiction laws,
- they still work with an ordinal scale, which is less demanding than numerical values over the unit interval,
- the simplicity of the underlying calculus.

Once these three operators given, other can be extended to fuzzy sets, such as the difference:

$$\mu_{E-F}(x) = \top(\mu_E(x), \mu_{\bar{F}}(x))$$

and the Cartesian product:

$$\mu_{E \times F}(x, y) = \top(\mu_E(x), \mu_F(y)).$$

The inclusion can be applied to fuzzy sets in a straightforward way: $E \subseteq F \Leftrightarrow \forall x, \mu_E(x) \leq \mu_F(x)$, but a gradual view of the inclusion can also be introduced. The idea is to consider that E is more or less included in F. Different approaches can be envisaged, among which one is based on the notion of fuzzy implication (the usual logical counterpart of the inclusion). The starting point is the following definition valid for sets:

$$E \subseteq F \Leftrightarrow \forall x, x \in E \Rightarrow x \in F$$

which becomes :

$$deg(E \subseteq F) = \top_x(\mu_E(x) \Rightarrow_f \mu_F(x))$$

where $\Rightarrow_f$ is a fuzzy implication whose arguments and result take their value in the unit interval. Different families of such implications have been identified (notably R-implications and S-implications) and the most common ones are:

- Kleene-Dienes implication : $a \Rightarrow_{K-D} b = max(1 - a, b)$,
- Rescher-Gaines implication: $a \Rightarrow_{R-G} b = 1$ if $a \leq b$ and 0 otherwise,
- Gödel implication : $a \Rightarrow_{Go} b = 1$ is $a \leq b$ and b otherwise,
- Lukasiewicz implication : $a \Rightarrow_{Lu} b = min(1, 1 - a + b)$.

Of course, fuzzy sets can also be combined in many other ways, such as mean operators, which do not make sense for classical sets.

3.2 Possibility theory

Possibility theory is a theory of uncertainty which aims at assessing the realization of events. The main difference with the probabilistic framework lies in the fact that it is mainly ordinal and it is not related with frequency of experiments. As in the probabilistic case, a measure (of possibility) is associated with an event. It obeys the following axioms [Zad78]:

- $\Pi(X) = 1$,
- $\Pi(\oslash) = 0$,
- $\Pi(A \cup B) = max(\Pi(A), \Pi(B))$,

[Zad78] L. ZADEH, "Fuzzy sets as a basis for a theory of possibility", *Fuzzy Sets and Systems 1*, 1978, p. 3–28.

where X denotes the set of all events and A, B are two subsets of X. If $\Pi(A)$ equals 1, A is completely possible (but not certain), when it is 0, A is completely impossible and the closer to 1 $\Pi(A)$, the more possible A. From the last axiom, it appears that the possibility of $\bar{A}$, the opposite event of A, cannot be calculated from the possibility of A. The relationship between these two values is:

$$max(\Pi(A), \Pi(\bar{A})) = 1$$

which stems from the first and third axioms (where B is replaced by $\bar{A}$).

In other words, if A is completely possible, nothing can be deduced for $\Pi(\bar{A})$. This state of fact has led to introduce a complementary measure (N), called necessity, to assess the certainty of A. $N(A)$ is based on the fact that A is all the more certain as $\bar{A}$ is impossible [DP80]:

$$N(A) = 1 - \Pi(\bar{A})$$

and the closer to 1 N(A), the more certain A. From the third axiom on possibility, one derives:

$$N(A \cap B) = min(N(A), N(B)))$$

and, in general:

- $\Pi(A \cap B) \leq min(\Pi(A), \Pi(B))$,
- $N(A \cup B) \geq max(N(A), N(B))$.

In the possibilistic setting, a complete characterization of an event requires the computation of two measures: its possibility and its certainty. It is interesting to notice that the following property holds:

$$\Pi(A) < 1 \Rightarrow N(A) = 0.$$

It indicates that if an event is not completely possible, it is excluded that it is somewhat certain, which makes it possible to define a total order over events: first, the events which are somewhat possible but not at all certain (from (Π = N = 0 to Π = 1 and N = 0), then those which are completely possible and somewhat certain (from Π = 1 and N = 0 to Π = N = 1). This favorable situation (existence of a total order) is valid for usual events, but if fuzzy ones are taken into account, this is no longer true (because A $\cup$ $\bar{A}$ = X is not true in general when A is a fuzzy set) and the only valid property is: $\forall$ A, $\Pi(A) \geq$ N(A).

The notion of a possibility distribution [Zad78], denoted by π, plays a role similar to that of a probability distribution. It is a function from the referential X into the unit interval and:

$$\forall A \subseteq X, \Pi(A) = sup_{x \in A} \pi(x)$$

[DP80] D. DUBOIS, H. PRADE, *Fuzzy set and systems: theory and applications*, Academic Press, 1980.

In order to comply with the second axiom above, a possibility distribution must be such that there exists (at least) an element x_0 of X for which $\pi(x_0) = 1$. Indeed, a possibility distribution can be seen as a normalized fuzzy set F which represents the knowledge about a given variable. The following formula:

$$\pi(x = a) = \mu_F(a)$$

which is often used, tells that the possibility that the actual value of the considered variable x is a, equals the degree of membership of a in the fuzzy set F. For example, Paul's age may be only imprecisely known as "close to 20", where a certain fuzzy set is associated with this fuzzy linguistic expression.

3.3 Fuzzy sets, possibility theory and databases

The project is situated at the crossroads of databases and fuzzy sets. Its main objective is to broaden the capabilities offered by DBMSs according to two orthogonal lines in order to separate two distinct problems:

- flexible queries against regular databases so as to provide users with a qualitative result made of ordered elements,
- Boolean queries addressed to databases containing imprecise attribute values.

Once these two aspects solved separately, the joint issue of flexible queries against databases containing imprecise attribute values will also be considered. This can be envisaged because of the compatibility between the semantics of grades (preferences) in both fuzzy sets and possibility distributions.

It turns out that fuzzy sets offer a very convenient way for modeling gradual concepts and then flexible queries [1, 2]. It has been proven [BP92] that many ad hoc approaches (e.g., based on distances) were special cases of what is expressible using fuzzy set theory. This framework makes it possible to express sophisticated queries where the semantic choices of the user can take place (e.g., the meaning of the terms or the compensatory interaction desired between the various fuzzy conditions of a query). The works conducted in Pilgrim aim at extending algebraic as well as user-oriented query languages in both the relational and the object-oriented (extended relational in practice) settings. The relational algebra has already been revised in order to introduce flexible queries and a particular focus has been put on the division operation. Current works are oriented towards:

- conditions calling on aggregate functions applying to fuzzy sets, for instance fuzzy quantified statements such as "most employees have a medium salary" which can be expressed in the context of an SQL-like language,
- the handling of fuzzy bags (fuzzy multi sets) and their connection with fuzzy numbers.

[BP92] P. Bosc, O. Pivert, "Some approaches for relational databases flexible querying", *Journal of Intelligent Information Systems 1*, 1992, p. 323–354.

As to possibility distributions, they are used to represent imprecise (imperfect) data. So doing, a straightforward connection can be established between a possibilistic database [PT84] and regular ones. Indeed, a possibilistic database is nothing but a weighted set of regular databases (called worlds), obtained by choosing one candidate in every distribution appearing in any tuple of every possibilistic relation. According to this view, a query addressed to a possibilistic database has a natural semantics. However, it is not realistic to process it against all the worlds due to their huge number. Then, the question tied to the querying of a possibilistic database bears mainly on the efficiency, which imposes to obviate the combinatorial explosion of the worlds. The objective of the project is to identify different families of queries which comply with this requirement in the context of the relational setting, even if the initial model must obviously be extended (in particular to support imprecise data).

3.4 Query rewriting using views

Information integration is the problem of combining information residing at disparate sources and providing the user with a unified view of those information. This problem has been a long standing challenge for the database community.

Two main approaches for information integration have been proposed. In the first approach, namely materialization or warehousing, data are periodically extracted from the sources and stored in a centralized repository, called a (data) warehouse. User queries are posed and executed at the warehouse with no need to access to the remote information sources. Such an approach is useful in the context of intra-enterprise integration with few remote sources to integrate. It is, however, not feasible in open environments like the Web where the number of sources may be very large and dynamic.

In the second approach, called mediation or virtual integration, data stay at the sources and are collected dynamically in response to user queries [Len02,Hal03]. Mediation architectures are based on the mediator/wrapper paradigm where native information sources are *wrapped* into logical views through which the underlying sources may be accessed. The views are stored in the mediator component which additionally contains an integrated global schema that provides a single entry point to query the available information sources. The global schema acts as an interface between the user queries and the sources freeing the users from the problem of source location and heterogeneity issues. In such an architecture, user queries posed on the global schema are rewritten in terms of logical views and then sent to the remote sources.

Briefly stated, two main approaches of mediation have been investigated in the literature [Hal01]: the GAV (Global As View) approach where the global schema is expressed as a set of views over the data sources, and the LAV (Local As View) approach where the data sources are defined as views over the global schema. Query processing is expected to be easier in the GAV approach as it can be achieved by a kind of unfolding of original queries. However, this

[PT84] H. PRADE, C. TESTEMALE, "Generalizing database relational algebra for the treatment of incomplete or uncertain information and vague queries", *Information Sciences 34*, 1984, p. 115–143.

[Len02] M. LENZERINI, "Data Integration : A Theoretical Perspective", Madison, Wisconsin, 2002.

[Hal03] A. HALEVY, "Data Integration : A status Report", *in: German Database Conference BTW-03*, Leipzig, Germany, 2003. Invited Talk.

[Hal01] A. Y. HALEVY, "Answering queries using views: A survey", *VLDB Journal 10*, 4, 2001, p. 270–294.

approach suffers from a lack of extensibility as changing or adding new sources affects the global schema. In contrary, the LAV approach is known to be highly extensible in the sense that source changes do not impact the global schema. However, in the context of the LAV approach, query processing is known to be more challenging.

A centralized mediation approach has several drawbacks including scalability, flexibility, and availability of information sources. To cope with such limitations, a new decentralized integration approach, based on a Peer-to-Peer (P2P) architecture, has been proposed. A P2P data management system enables sharing heterogeneous data in a distributed and scalable way [HIM+04]. Such a system is made of a set of peers each of which is an entire data source with its own distinct schema. Peers interested in sharing data can define pairwise mappings between their schemas. Users formulate queries over a given peer schema then a query answering system exploits relevant mappings to reformulate the original query into set of queries that enable to retrieve data from other peers.

Query answering in information integration systems

The problem of answering queries in mediation systems has been intensively investigated during the last decade. In particular, the investigation of this problem in the context of a LAV approach led to a great piece of fundamental theory. Recent work show that query processing in data integration is related to the general problem of answering queries using views [Hal01,Len02]. In such a setting, semantics of queries can be formalized in terms of certain answers [AD98]. Intuitively, a certain answer to a query Q over a global (mediated) schema with respect to a set of source instances is an answer to Q in any database over the global schema that is consistent with the source instances. Therefore, the problem of answering queries in LAV-based mediation systems can be formalized as the problem of computing all the certain answers of the queries. As shown recently this problem has a strong connection with the problem of query answering in database with incomplete information under constraints.

One of the common approaches to effectively computing query answers in mediation systems is to reduce this problem into a query rewriting problem, usually called *query rewriting using views* [Hal01,Len02,TH04]. Given a user query expressed over the global (or a peer) schema, the data sources that are relevant to answer the query are selected by means of a rewriting algorithm that allows to reformulate the user query into an equivalent or maximally subsumed (contained) query whose definition refers only to source descriptions.

The problem of rewriting queries in terms of views has been intensively investigated in the last decade (see [Hal01,Len02] for a survey). Existing research works differ w.r.t. the languages used to express a global schema, views and queries as well as w.r.t. the type of rewriting

[HIM+04] A. Y. Halevy, Z. G. Ives, J. Madhavan, P. Mork, D. Suciu, I. Tatarinov, "The Piazza Peer Data Management System.", *IEEE Trans. Knowl. Data Eng. 16*, 7, 2004, p. 787–798.

[Hal01] A. Y. Halevy, "Answering queries using views: A survey", *VLDB Journal 10*, 4, 2001, p. 270–294.

[Len02] M. Lenzerini, "Data Integration : A Theoretical Perspective", Madison, Wisconsin, 2002.

[AD98] S. Abiteboul, O. Duschka, "Complexity of Answering Queries Using Materialized Views.", *in: PODS*, p. 254–263, 1998.

[TH04] I. Tatarinov, A. Halevy, "Efficient query reformulation in peer data management systems", *in: SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, ACM Press, p. 539–550, New York, NY, USA, 2004.

considered (i.e., maximally contained or equivalent rewriting). In a nutshell, this problem has been studied for different classes of languages ranging from various sub-languages of datalog, hybrid languages combining Horn rules and description logics to semistructured data models. Recently, the problem of rewriting queries in terms of views has been investigated in the context of P2P DBMSs [HIM+04,TH04].

4 Application Domains

As to the aspect dealing with flexible queries, there are several potential application domains. Soft querying turns out to be relevant in many contexts, such as information retrieval, in particular on the Web (many commercial systems, e.g. Google or Yahoo use a hidden technique to rank-order the answers), yellow pages, classified advertisements, image or multimedia retrieval. One may guess that the richer the semantics of stored information (for instance images or video), the more difficult it is for the user to characterize his search criterion in a crisp way, i.e., using Boolean conditions. In this kind of situation flexible queries which involve imprecise descriptions (or goals) and vague terms, may provide a convenient means for expressing information needs.

Even though most of the research works performed in Pilgrim assume relational data, many results can be transposed to other contexts such as information retrieval or multimedia database querying. We are currently working on the specification of a flexible route planning system involving fuzzy preferences (cf. Section 6.4), which should illustrate the utility of fuzzy queries in the context of intelligent transportation systems.

Databases involving imprecise data are not yet common in practice for two reasons: developing DBMSs supporting such data is probably a hard job and the demand is presently not strong. However, many potential domains could take advantage of such advanced systems capable of storing and querying databases where some pieces of information are imprecise: military information systems, automated recognition of objects in images, data warehouses where information coming from more or less reliable sources must be fused and stored, etc.

5 Software

We are currently implementing a flexible querying prototype that aims at evaluating fuzzy queries addressed to regular databases. More precisely, it takes the form of an additional software layer on top of Oracle, whose function is to translate a fuzzy query into a procedural evaluation program including regular SQL queries in order to take advantage of the optimization mechanisms that exist in the DBMS. In its current version, the prototype, called FRIGA (Flexible RetrIeval and GRaded Answers), is able to process "simple" fuzzy queries (i.e., fuzzy queries involving a single block) and we are now extending it so as to make it support nested queries and so-called contextual fuzzy queries (i.e. queries where some fuzzy terms do not have to be explicitly defined by the user but whose interpretation depends on a certain context that can be determined from the query itself).

In 2008, we developed FLEXIS, a flexible data mediation prototype using an LAV type of approach. This prototype aims at rewriting user queries in terms of views in a tolerant way, on the basis of the interval constraints involved in both the query and the views. The rewritings obtained are assigned a weight which reflects the probability that any tuple returned by that rewriting satisfies the constraints from the query. FLEXIS can provide the user with either the best k rewritings of the query, or those whose degree is over a certain qualitative threshold. This prototype was presented at the conference BDA'08 [27].

As mentioned in Section 4, we decided to illustrate the interest of flexible querying techniques in the domain of Intelligent Transportation Systems (ITS). Along with two other IRISA teams (namely Cairn and Cordial), Pilgrim is involved in the development of a platform called MOB-ITS in the context of the CPER INVENT'IST 2007-2013. This platform aims at supporting mobile and interactive access to information for ITS applications. In this framework, Pilgrim intends to implement an application that would make it possible for a mobile user to query distributed data sources according to a fuzzy variant of the "pay as you go" [SDH08] type of approach, i.e. without having available a global mediated schema of the sources that can potentially be queried. The flexible route planner outlined in Section 6.4 should also be integrated into that platform.

6 New Results

6.1 Possibilistic database querying

Participants: Patrick Bosc, Nadia IbenHssaien, Olivier Pivert.

Keywords: Ill-known data, possibilistic databases, strong representation system, querying language.

Relational databases including ill-known attribute values represented by weighted sets of more or less possible/probable candidates are considered. The main objective is to study different classes of queries that can be used to retrieve information from such databases, with the constraints that the complexity of their evaluation must be reasonable. In 2008, we paid a special attention to two classes of queries in this context: those involving a division operator, and a special type of so-called "generalized yes/no queries" of the form "to what extent is it possible and certain that the result of Q satisfies the cardinality-based property P". For each of these types of queries, we defined a tractable processing strategy. Another issue dealt with concerned the situation where it is not possible to specify in an exhaustive manner the possible candidates describing an ill-known attribute value.

Some well-known works conducted in the 80's have revealed that the presence of null values (in the sense of existing unknown values) in databases raises several serious problems. In particular, it is not feasible to extend the whole set of relational operators, defined for precise and complete relations, so as to make them work on relations containing imprecise information. Indeed, when null values are involved, the result of some operations (such as the join) cannot

[SDH08] A. D. SARMA, X. DONG, A. HALEVY, "Bootstrapping pay-as-you-go data integration systems", *in: Proceedings of the ACM SIGMOD Conference*, p. 861–874, Vancouver (Canada), 2008.

be represented by means of a "basic" relational table (a more sophisticated model is needed). The same problem obviously arises in the case of more elaborate ill-known values such as possibility or probability distributions. It is thus important to identify classes of queries that can be used in an imprecise database context.

Let us recall that an imprecise database corresponds to a set of interpretations (also called worlds) which are usual databases obtained by choosing a candidate value in each distribution. The combinatorial nature of this mechanism leads to a considerable number of worlds even when the number of ill-known values in the database is relatively small. A crucial objective is thus to define a "compact" processing method for algebraic queries, i.e., a method that does not require to make the interpretations of the database explicit. Consequently, it is mandatory to design a data model that constitutes a strong representation system for the query language considered, i.e., a model that supports a closed set of operations whose results are consistent with a world-based interpretation. In other terms, if $rep(D)$ denotes the set of worlds associated with the imprecise database D, the following property must hold for any operation o of the language: $o(rep(D)) = rep(o(D))$. Such a framework guarantees a sound semantics for the operators of the language, and it makes it possible to envisage a tractable query evaluation process.

In the previous years, we defined a possibilistic database model involving usual possibilistic relations enriched with:

- a degree N associated with every tuple, which expresses the certainty that the tuple has a representative in any world than can be derived from the relation,
- the introduction of nested relations used to represent dependencies between candidate values for different attributes and thus to model possibility distributions over several attributes,

and we showed that this model constitutes a strong representation system for the operations of selection, projection, union and foreign key join (fk-join).

In this model, the result of an algebraic query is a "compact" table representing all the more or less possible worlds resulting from the query. It has been shown that algebraic queries could be processed in a "compact" (then tractable) way at the price of some restrictions as to the operations present in Q (the only "legal" operations are those for which the model constitutes a strong representation system).

In 2008, we paid a particular attention to an additional operator in the framework of this model, namely the division [24]. As for the previous operators of the model, we defined a "compact" processing strategy which avoids computing the entire set of worlds of the imprecise database. The strategy proposed makes it possible to evaluate in a tractable way the division of a relation r of schema $R(A, X)$ by a relation s of schema $S(B)$ in the particular case where X and B are precise attributes and A is an imprecise one. We studied the influence of the model of uncertainty chosen to represent ill-known information and it has been shown in [24] that the possibilistic framework enables to use some optimization techniques that are not possible in the probabilistic case. Indeed, the algorithm in the possibilistic case takes advantage of the non-additive nature of the possibilistic model so as to drastically prune the set of interpretations attached to subrelations of r involved in the evaluation.

In [8], we considered querying possibilistic relational databases by means of so-called generalized yes/no queries whose form is: "to what extent is it possible and certain that the answer to Q satisfies property P". More precisely, we considered cardinality-based generalized yes/no queries (in this case, property P is about cardinality) and proposed a processing technique which avoids computing all the worlds attached to the possibilistic database considered. The evaluation of such queries is based on a two-step mechanism: i) a compact processing of the associated algebraic query, which builds a compact possibilistic relation using the model outlined above, ii) a post-processing which produces the final answer, i.e., the answer to the cardinality-based query, in the form of a pair (possibility degree, certainty degree).

We also extended the possibilistic database model that we defined previously so as to deal with incomplete possibility distributions. The situation considered is that where the data provider does not have a complete knowledge of the attribute domains involved and is only able to specify more or less possible and completely impossible candidates for some attribute values. For instance, let us consider a relation describing images of enemy aircrafts taken by spy satellites. Let us assume that every image represents a unique aircraft whose type may be ill-known due to the imprecision inherent in the automatic recognition process. One wants to take into account the possibility that the reference "catalogue" of airplane types used in the matching phase is incomplete (some aircraft types may be missing). Then, for a given image, some explicit types (from the catalogue) may be recognized as more or less possible, some others (also from the catalogue) may be classified as impossible, and there is a non-zero possibility that the aircraft in the picture matches an unknown type (absent from the catalogue). The model we proposed in [26] is based on the notion of an imprecise twofold value. Such a value is made of a positive part consisting of a weighted set (possibility distribution) which represents the more or less possible candidates, including — if necessary — a special value meaning "unknown", and a negative part consisting of a regular set which represents the totally impossible candidates. We have shown that this framework constitutes a strong representation system for the operations of selection (based on a restricted type of conditions), projection and union.

6.2 Flexible querying of regular databases

6.2.1 Gradual numbers

Participants: Daniel Rocacher, Patrick Bosc, Ludovic Liétard.

Keywords: Gradual integer, gradual rational, quantified statement, flexible querying.

This work takes place in a framework defining a query language which deals with quantification and preferences on data. Our recent work on this topic concerns the evaluation of quantified statements using gradual numbers

This work is at the crossroads of flexible querying of relational databases using fuzzy sets and fuzzy arithmetic based on gradual numbers [Roc03]. Our general objective is to devise new structures and queries able to deal with quantification and preferences over usual databases.

[Roc03] D. Rocacher, "On fuzzy bags and their application to flexible querying", *Fuzzy Sets and Systems 140*, 2003, p. 93–110.

For example, we may be interested in expressing and evaluating requests about quantities, as in: *find the best five companies where most of the young employees are well-paid* or *find the best five companies in which the number of young employees is equal to or greater than the number of well-paid employees.*

A gradual natural integer (which belongs to the set N_f), is a fuzzy set of integers corresponding to the fuzzy cardinality of a fuzzy set A. This approach has been followed up by extending Z_f to Q_f (the set of fuzzy rational numbers). These new frameworks provide an arithmetic basis where difference and ratio between gradual quantities can be exactly evaluated [12]. Moreover, the application of a gradual predicate P on a gradual number x gives a specific fuzzy truth value expressing the satisfaction of each α-cut of x with respect to the predicate P.

Fuzzy arithmetic allows to evaluate quantified statements of type "Q X are A" and "Q B X are A" [14] as well as conditions with aggregates [9, 13]. We have shown that the evaluation of a quantified statement can be either a fuzzy truth value or a scalar value obtained by a defuzzification of the fuzzy truth value [14]. Two types of scalar values can be distinguished: the first one corresponds to a quantitative view of the fuzzy value, the second one to a qualitative view. When dealing with statements of the form "Q X are A, the two scalar values are respectively generalizations of the OWA based interpretation and the Sugeno integral based interpretation. Our work constitutes the first attempt to set the evaluation of fuzzy quantified statements in the framework of an extended arithmetic and algebra. This aspect is important since the algebraic framework guarantees some properties which give a sound semantics to the interpration based on it.

6.2.2 Extended division operators

Participants: Patrick Bosc, Allel Hadjali, Olivier Pivert.

Keywords: Relational division, graded inclusion, quotient, proximity relation, ordinal preferences, anti-division.

The role and properties of the division are very well-known in the context of queries addressed to regular relational databases. This operator can be extended in several directions when preferences come into play. In particular, it may apply to fuzzy or ordinal relations and some tolerance to exceptions may be taken into account. These issues have been investigated along with the relationship between division and inclusion on the one hand and a derived operator called anti-division on the other hand.

In the context of the relational data model, the division operator is somewhat similar to the integer division. The division of r by s whose respective schemas are (A, X) and (B) where A and B are attributes defined over the same domain, delivers a relation t whose schema is X. An element x belongs to the result as soon as it is associated in r (the dividend) with at least all the values appearing in s (the divisor). In other words, for x to be satisfactory, the presence of a value b in s implies that the pair (b, x) is in r. It turns out that t is a quotient, i.e., the largest relation whose Cartesian product with the divisor is included in (or equal to) the dividend. This operator can be extended in several independent directions and a few years ago we focused on the case where operand relations become fuzzy. In this context, the division of fuzzy relations was defined on the basis of the use of fuzzy implications (which generalize the usual material implication). It was shown that the resulting relation is still a quotient (in the sense of a maximal relation whose Cartesian product with the divisor satisfies an inclusion condition), provided that : i) either R-implications or S-implications are used in the extended division, and ii) the Cartesian product is founded on an appropriate conjunction which is closely related to the chosen implication.

More recently, we have addressed a complementary issue, namely the introduction of some tolerance in the division mechanism itself. Indeed, even when fuzzy relations come into play, the complete dissatisfaction of the implication for a single element of the divisor, leads to discard the value under consideration (x). In 2008, the investigation of this question was pursued and four main lines were identified and studied [6, 5]:

i) quantitative exceptions: a certain amount of exceptions is allowed and the objective of the related tolerant division is to look for the elements which are associated with almost all the values of the divisor; in other words, some missing or low satisfactory associations may be more or less ignored depending on the modeling of the relaxed quantifier "almost all",

ii) qualitative exceptions: it may happen that the situation is close to that of full satisfaction for a given implication, though leading to a low satisfaction degree; in such a case, called a low-intensity exception, an upgrade takes place and the tolerant division is based on the fact that the divisor is almost included in the set of A-values connected with x in the dividend,

iii) resemblance over the values of the attribute of the dividend, which comes down to enlarging this operand with tuples which are somewhat similar (according to the resemblance relation); for instance, if the tuple $< x, red, 0.9 >$ is present in the dividend and carmine is assumed to be similar to red at the degree 0.6, the tuple $< x, carmine, 0.6 >$ is added to the dividend; as a consequence, x gets more chance to qualify for the division if carmine (and not red) appears in the divisor,

iv) reduction of the divisor relation according to the importance of the values involved; for instance, values associated with too low levels are discarded or the levels are decreased on the basis of the transformation of the initial fuzzy predicate P into *very* P.

For all of these views, the quality of quotient of the result delivered by the tolerant division was characterized [18].

Due to the fact that a division calls on an inclusion, the first two types of extension of the division (relying on exceptions) were used as a basis for defining tolerant inclusion operators [10, 19, 25]. Moreover, the ability of the extended inclusion to capture the behavior of information retrieval systems is under consideration in collaboration with the research team TexMex.

In 2008, we also started the study of another type of division, namely a stratified division [21]. The basic idea is to deal with ordinal preferences (as it is done in [28] in another framework) concerning solely the divisor, which is explicitly given by the user as a hierarchy of sets of values. The general expression of queries is: retrieve the best k elements which are associated with set-1 and if possible set-2 ... and if possible set-n. This can be seen as a refinement of the regular division where set-1 is mandatory and the following sets allow for breaking ties. One interest of such an approach is to get rid of membership functions and numeric grades, which makes it fairly similar to what the user is asked in non-fuzzy approaches to preference queries (e.g., winnow or PreferenceSQL). This division is proved to return a quotient. Experiments were undertaken in order to assess the extra cost induced by the handling of preferences. It turns out that the overhead varies from 50 to 600% depending on the implementation chosen.

Last, we worked on an operator somewhat close to the division, called the anti-division. While a division looks for elements connected with a given set of values, the anti-division aims at retrieving the elements which are associated with none of the elements of a given set. The properties of this operator (which is not the complement of the division, but more its dual) were studied in particular when the operands are fuzzy relations [23] and a tolerant version of the anti-division was proposed [22].

6.2.3 Cooperative answering to flexible database queries

Participants: Patrick Bosc, Mikael Dautrey, Allel Hadjali, Olivier Pivert.

Keywords: Flexible queries, cooperative responses, query weakening, query intensifying, proximity relation, fuzzy relation..

The aim of this work is to propose intelligent cooperative techniques for dealing with some practical problems that could arise in a flexible database querying context. Two common problems are addressed: the empty answers problem and the overabundant answers problem. A unified approach which is based on a convenient proximity relation is proposed to overcome the shortcomings stemming from both kinds of answers. Such a relation allows for relaxing/intensifying the fuzzy constraints involved in users queries in a controlled iterative way. The relaxation/intensification process results in modified queries which are semantically close to the initial one.

Retrieving desired data over large-scale databases, especially those accessible via the Web, has become a ubiquitous task. In their web data retrieval, (ordinary) users are faced with two common problems: overabundant (or too many) answers and empty answers. In the former, the user is provided with an avalanche of responses that satisfy his/her query, while in the latter, no data is returned to the query asked. It is worthy to note that these kinds of answers are sometimes informative but we assume that users are interested in the values of answers, rather than in the cardinality of the set of answers.

In the context of flexible queries, similar problems could still arise. In this context, the *empty answer problem* is defined in the same way as in the Boolean case. The *fuzzy counterpart* of the *overabundant answer problem* can be stated as follows: there are too many data in the database that *fully satisfy* the user query. This means that *satisfaction degrees* of all retrieved data are *equal to* 1. Facing this problem, users' desires are mainly to reduce this very large set of answers and keep a manageable subset that can be easily examined and exploited.

Our solution basically relies on modulating the fuzzy conditions involved in the user query by applying appropriate transformations based on a tolerance relation [17, 20]. According to the problem at hand, this operation leads to a relaxation or an intensification of the user query. In the case of relaxation, the transformation acts only on the support of the fuzzy set associated with the failing query [4], while in the case of intensification, the transformation shrinks only the core of the query at hand [7]. Both transformations result in modified queries that are close to the initial one semantically speaking [3].

Our solution is database independent and requires no user feedback. It only leverages the attributes specified in the query of interest and the set of data retrieved. Neither human involvement nor any knowledge about the data distribution in the target database is required. A part of the solution proposed is under implementation in order to carry out some experimental studies in a near future.

6.2.4 Bipolar fuzzy queries

Participants: Ludovic Liétard, Daniel Rocacher.

Keywords: Flexible query, query evaluation.

The concept of bipolar queries is a particular way to integrate preferences inside queries where mandatory preferences, called *constraints*, are distinguished from optional preferences, called wishes. This year, we particularly studied the adequation of the concept of an intuitionistic fuzzy set for the modeling of bipolar querying.

The issue dealt with is the integration of bipolar conditions into queries addressed to a relational database. Bipolar conditions are made of two different parts and distinguish mandatory preferences, called *constraints*, from optional preferences, called *wishes. Constraints* and *wishes* are respectively defined by a set of acceptable values and a set of desired values. Tuples satisfying the *constraints* and the *wishes* are returned in priority to the user. If such answers do not exist, tuples satisfying only the *constraints* are delivered. *Constraints* are preferred to wishes since wishes are optional in the sense that they may be not fulfilled by the answers provided to the user.

Such a type of conditions can be useful in many contexts, and as an example, we can consider the case of a database containing cars for sale. When querying this database, a bipolar condition can be very helpful: the *constraints* are the buyer's needs (a red car, a price smaller than $1000) and the wishes reflect the seller's needs (a benefit over $300). Cars which do not satisfy the *constraints* are discarded and this behaviour means that a sale is impossible when the user's needs are violated. Otherwise, a sale is possible, even if the seller's needs are not met, but cars which meet both requirements are retained in priority. Such a condition can be rewritten : "find a red car with a price smaller than $1000 and, if possible, with a benefit over $300".

We have considered the case of bipolar conditions where both the *wishes* and the *constraints* are defined by fuzzy sets so as to express preferences (bipolar fuzzy condition). In this case, a bipolar fuzzy condition can be : "find a red car with a *cheap price* and, if possible, with a *large benefit*" where *cheap price* and *large benefit* are vague conditions defined by fuzzy sets.

The theoretical framework of intuitionistic fuzzy sets has been proposed to handle bipolar conditions in flexible queries. We have shown [30, 31] that it is possible to define a bipolar fuzzy condition by an intuitionistic fuzzy set and to combine different bipolar conditions according to extended relational operators. An intuitionistic fuzzy set being a generalization of a fuzzy set, a bipolar fuzzy condition is a generalization of a fuzzy condition and we now consider extending the SQLf language to this context.

6.3 Linguistic summaries of relational databases

Participants: Ludovic Liétard.

Keywords: Linguistic summary, linguistic quantifier, linguistic variable.

Data summarization of a relational database aims at expressing a synthetic description of the content of the database. This description may take different forms and we are interested in linguistic descriptions of data.

The amount of information managed by DBMS's becomes increasingly important and, as a consequence, it is sometimes relevant to summarize this information in order to obtain a compact or synthetic description.

Fuzzy set theory can be used to define linguistic summaries in order to obtain a linguistic description of the data and to offer flexibility in this description.

This year, we have considered R. Yager's proposal which is based on the expression of a quantified statement to express that a quantity of information (defined by a linguistic quantifier) satisfies a gradual property. For example, a relation from the database describing employees can be summarized by the quantified statement "*most of* the employees are *well-paid*" associated with a degree of truth to indicate its validity.

The linguistic summaries proposed by Yager express two aspects of the summarized data: its quantity and its quality. As an example, *the more* the satisfaction of the employees with respect to *well-paid* increases (qualitative aspect) for *most of them* (quantitative aspect), *the more* the linguistic summary "*most of* the employees are *well-paid*" is valid.

We have pointed out that Yager's proposal does not sufficiently stress the relationship between these two aspects and, as a consequence, the interpretation of the degree of validity is not very clear. The only information we get is *the higher the degree, the more valid the summary.* As a consequence, if one considers the following summary "*most of* the employees are *well-paid*" which is true at degree 0.5 on relation R (made of employees from firm1) and true at degree 0.8 on relation R' (made of employees from firm2), it is difficult to get precise explanations about this increase between the two degrees. We only know that, in general, an employee from firm2 has a larger salary than an employee from firm1.

We have proposed [32] a new definition for the degree of validity for linguistic summaries (based on linguistic statements where the universal quantifier is implicit). This degree of truth is computed by extending the value of a regular set-oriented function to a fuzzy set. It has a precise meaning in terms of quantity and quality of the summarized data. More precisely, we have defined linguistic summaries of type :

"tuples from R satisfy C^i and C^{i+1} and ... and C^{i+k}" ,

where C^i, C^{i+1}, ... ,C^{i+k} are fuzzy predicates respectively defined on the domains of attributes A^i, A^{i+1}, ... , A^{i+k} from relation R.

The degree of truth ω of such a summary is defined as the highest percentage δ such that at least δ of tuples from R satisfies “C^i and C^{i+1} and ... and C^{i+k}” *at least* at degree δ. As an example, a value 0.8 for ω allows to state that *at least 80%* of tuples satisfy “C^i and C^{i+1} and ... and C^{i+k}” at least at degree 0.8. If $\omega = 1$ it means the entire relation fully satisfies "C^i and C^{i+1} and ... and C^{i+k}". When ω is small (e.g., 0.1), the summary is not useful because it is impossible to find a better guaranteed minimum for both the quantity and the quality.

This interpretation of the degree of truth allows to justify the increase in the validity of a linguistic summary. A summary is more valid on relation R than on relation R' if relation R contains more tuples which are more satisfactory with respect to the fuzzy constraint. It becomes possible to refine the justification by comparing the α-cuts (with respect to the two degrees of truth) of the fuzzy constraint “C^i and C^{i+1} and ... and C^{i+k}”.

6.4 Personalized route planning involving fuzzy preferences

Participants: Patrick Bosc, Allel Hadjali, Amine Mokhtari, Olivier Pivert.

Keywords: Route planning, fuzzy preferences, personalization, conditional preferences.

This new topic concerns the application of fuzzy set theory to the specification of a route planning system involving sophisticated user preferences. In 2008, we defined a typology of fuzzy preferences that make sense in such a context and proposed a model for conditional competitive fuzzy preferences.

In 2008, we have started to investigate a new topic which, we think, offers an interesting application field for some flexible querying tools that we developed in the past, and which raises some original research issues of its own. The general domain considered is that of Intelligent Transportation Systems, and the problem dealt with, that of personalized route planning involving fuzzy preferences. Basically, we consider the situation where a user wants to go from point A to point B and looks for the route which satisfies the best some criteria that may be expressed in a flexible way.

The aspects involved in the modeling of such a system are: i) fuzzy set theory for the modeling of user preferences, ii) graph theory for the routing aspect, iii) geographic data models. In 2008, we have first established a state of the art of the approaches from the literature about route planning involving user preferences and studied the different geographic data models that could be of interest for our purpose. We retained the GDF (Geographic Data Files) model and defined a typology of fuzzy preferences in the framework of this model. We have also outlined an SQL-like query language enabling to express such preference queries. Among the features of this language is the concept of a bipolar requirement (cf. Section 6.2.4) which makes it possible to express both flexible constraints and wishes.

We have also proposed an approach to the modeling of so-called conditional competitive fuzzy preferences, which could be of interest in such a flexible route planning system. The basic idea is to deal with statements of the form “prefer condition1 (priority 1) or condition2 (priority 0.8) or ...; if condition1 then prefer condition3 (priority 1) or condition4 (priority 0.7) or ..., and so on”. Such a complex statement can be modeled as a tree of conditions where the children of a given node are non-mutually-exclusive fuzzy predicates. In [15, 16], we gave the principle of the evaluation of conditional competitive queries, which is based on an adaptation of the lexicographic order. Among the immediate perspectives of this work, let us mention:

- query optimization
- use of fuzzy rules so as to model contextual queries (e.g., if the user's vehicle is $powerful$ then add to his/her query a default preference for *high-speed* roads)
- specification of the software architecture of a route planning system integrating such flexible requirements.

6.5 Flexibility issues in large-scale data integration systems

Participants: Allel Hadjali, Olivier Pivert, Hélène Jaudoin.

Keywords: Flexible integration system, query rewriting using views, tolerant matching.

The problem of answering queries in integration systems has been intensively investigated during the last decade. Recent works show that query processing in the context of LAV (Local As View) mediation systems is related to the general problem of answering queries using views. In such a setting, semantics of queries can be formalized in terms of *certain answers*. One of the common approaches to effectively computing certain answers to queries in mediation systems is to reduce this problem into a query rewriting problem. Given a user query expressed over the global schema, the data sources that are relevant to answer the query are selected by means of a rewriting algorithm that allows to reformulate the user query into an equivalent or maximally *contained* query whose definition refers only to source descriptions, i.e., *views*. The purpose of our work is to define tolerant query rewriting methods and, in a second step, to take into account the fact that data may be pervaded by uncertainty.

In an integration system, in order to enable fine-grained description of views and to reduce the number of non valuable query rewritings, it is interesting to consider value constraints on attributes, i.e., enumeration of possible/authorized values of the attributes, in the description of views and queries. Those constraints allow for specifying queries of the form: "retrieve the individuals whose values on a given attribute cannot be outside of the set of values $\{a_1, ..., a_n\}$". In [11], a sound and complete query rewriting algorithm, i.e., that computes all certain answers to a given query, in the presence of value constraints on attributes has been proposed. The algorithm is based on data mining techniques and hypergraph framework in order to favor scalability of the implementation.

In the context of open environments, like the Web, where data sources are autonomous, integration systems are confronted to the problem of the imperfect mapping between the value domains of the views and the queries. Indeed, descriptions of views may not match exactly the query. It is thus not realistic to assume finding views that totally satisfy domain constraints imposed by the query and that are able to provide certain answers. As an example, let us consider a query Q that asks for *names* of $persons$ whose *age* is in the interval $[28, 38]$ and two views V_1 and V_2 such that:
V_1 supplies *names* of $persons$ whose *age* is in $[25, 35]$ and

V_2 supplies *names* of *persons* whose *age* is in $[36, 46]$.
The views V_1 and V_2 both have an interval constraint on the attribute *age* such that the two intervals are not included in those of the query. Moreover, as V_1 and V_2 supply only names of person, selection on age attribute is impossible, and consequently, V_1 and V_2 cannot provide certain answers to Q.

Contrary to regular query rewriting algorithms, we propose not to eliminate such views although the value domains of views is not contained in those of the queries since the views could return an interesting number of correct answers. Indeed, in our example, if we take an attentive look on the interval constraints of the views, we can observe that V_1 and V_2 can provide correct answers to Q. The problem of answering queries in such a setting brings up at least three problems. The first one is to detect and to assess the approximate mappings between the queries and the views. In the setting of our example, the problem is to determine if V_1 has a better quality than V_2, i.e., if V_1 can provide more relevant answers than V_2. The second one is to define the semantics of the answers thus obtained. The last one is due to the possible huge number of obtained rewritings since we authorize approximate mappings between queries and views. The computing of all the rewritings is then not conceivable. Therefore it becomes necessary to define an algorithm able to generate only the best rewritings.

We focused so far on the problem of computing the query rewritings that have a positive probability to supply correct answers. Such rewritings can be computed on the basis of the interval constraints occurring in the views and in the query. A tuple issued from a query rewriting attached to a degree 0.25 has 25% of chance to be a certain answer. Thanks to these degrees, it is possible to discriminate the rewritings and to evaluate only the best ones. In [27], we propose a LAV-mediation system called FlexIS that computes only the k-best rewritings or those that exceed a given threshold.

As a perspective, we intend to propose a formal framework to the problem presented above and next to study a second way to define approximate matchings between value domains based on the relaxation of some arithmetical constraints involved in the query. The definition of a tolerant graded inclusion [10] constitutes a first step in that direction.

Besides, we have also studied the case where both the views and the queries are specified by means of fuzzy predicates. In such a situation, the tuples from a fuzzy view are attached with a membership degree which expresses the extent to which they fit the fuzzy description of the view. Our objective is to define a mechanism aimed at selecting the views which are guaranteed to deliver only α-certain answers to the fuzzy query, i.e., answers whose satisfaction degree is greater than α. The approach [29] proposed is based on the concept of a graded inclusion involving a fuzzy implication.

7 Other Grants and Activities

7.1 National actions

Patrick Bosc, Allel Hadjali, Hélène Jaudoin and Olivier Pivert participate in the ANR project "FORUM", which is in activity since december 2005 and which deals with the problem of information integration in a large and highly dynamic information space.

7.2 International actions

Carmen Brando, from University Simon Bolivar (Caracas, Venezuela) —with whom we collaborate on a regular basis—, is doing her Master's degree internship in our team since October 2008. Her research topic is about the definition of a query reuse approach for handling failing fuzzy queries.

8 Dissemination

8.1 Teaching

Project members give lectures in different faculties of engineering, in the third cycle University curriculum: ”Bases de données, gradualité et imprécision“ in the speciality ”Intelligence Artificielle et Images“ of the Master's degree in computer science at University of Rennes 1, and at ENSSAT (third year level cursus).

In 2008, Allel Hadjali gave a Master's course entitled "Requêtes à Préférences" at the University of Tlemcen (Algeria).

Patrick Bosc took part in a seminary on fuzzy logics in Grenoble (Univ. Joseph Fourier) in september

8.2 Scientific activities

8.2.1 Program committees

P. Bosc served as a member of the following program committees:

- 23^{rd} ACM Symposium on Applied Computing, Special Track on Information Access and Retrieval, Fortaleza, Brazil, March 16-20, 2008.
- Conférence Logique Floue et Applications (LFA 2008), Lens (France), 16-17 octobre, 2008.
- DEXA 2008 3^{rd} International Workshop on Flexible Database and Information Systems Technology (FlexDBIST'08), Torino, Italy, September 1-5, 2008.
- VLDB 2008 2^{nd} International Workshop on the Management of Uncertain Data (MUD'08), Auckland, New Zealand, August 24-30, 2008.
- 27^{th} International Conference of the North American Fuzzy Information Processing Society (NAFIPS'08), New York, New York, USA, May 19-22, 2008.
- 8^{emes} Journées Francophones "Extraction et Gestion des Connaissances (EGC'08), Sophia Antipolis, France, 29 janvier - 1er février, 2008.
- 4^{th} IEEE International Conference on Intelligent Systems (IEEE IS 2008), Varna, Bulgaria, September 6-8, 2008.

- 12th International Conference on Information processing and the Management of Uncertainty in Knowledge-Based Systems (IPMU 2008), Malaga, Spain, June 22-27, 2008.
- 23emes Journées Bases de Données Avancées (BDA 2008), Guilherand-Granges, France, 21-24 octobre, 2008.
- 1st North American Simulation Technology Conference (Nastec 2008), Montréal, Canada, August 13-15, 2008.
- European Conference on Intelligence and Security Informatics (EuroISI'08), Esbjerg, Denmark, December 3-5, 2008.
- Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on advanced Intelligent Systems (SCIS-ISIS'08), Nagoya, Japan, September 17-21, 2008.
- 26eme Congrès INFORSID, Fontainebleau, France, 27-30 Mai, 2008.
- 17th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'08), Hong Kong, China, June 1-6, 2008.

A. Hadjali served as a member of the following program committees:

- Conférence Logique Floue et Applications (LFA 2008), Lens (France), 16-17 octobre 2008.
- 8eme Conférence Internationale sur les Nouvelles Technologies de la Répartition (NOTERE 2008), Lyon, France, 23-27 juin, 2008.
- 2emes Journées Francophones sur les Ontologies (JFO'08), 1-3 Décembre 2008, Lyon.
- 1st Int. Workshop on Ontologies in Interactive Systems, September 1, 2008, Liverpool, U.K.

O. Pivert served as a member of the following program committees:

- 23rd ACM Symposium on Applied Computing, Special Track on Information Access and Retrieval, Fortaleza, Brazil, March 16-20, 2008.
- Conférence Logique Floue et Applications (LFA 2008), Lens (France), 16-17 octobre 2008.
- DEXA 2008 3rd International Workshop on Flexible Database and Information Systems Technology (FlexDBIST'08), Torino, Italy, September 1-5, 2008.
- VLDB 2008 2nd International Workshop on the Management of Uncertain Data (MUD'07), Auckland, New Zealand, August 24-30, 2008.

D. Rocacher served as a member of the following program committees:

- Conférence Logique Floue et Applications (LFA 2008), Lens (France), 16-17 octobre 2008.
- XXVI^{eme} Congrès INFORSID, Fontainebleau, France, 27-30 Mai, 2008.

8.2.2 Editorial boards

Patrick Bosc is a member of the following editorial boards:

- International Journal on Fuzziness, Uncertainty and Knowledge-Based Systems,
- Fuzzy Sets and Systems,
- Revue I3.

9 Bibliography

Major publications by the team in recent years

[1] P. BOSC, L. LIÉTARD, O. PIVERT, D. ROCACHER, *Gradualité et imprécision dans les bases de données*, Ellipses, 2004.

[2] P.BOSC, O. PIVERT, D.ROCACHER, "About quotient and division of crisp and fuzzy relations", *Journal of Intelligent Information Systems 29*, 2, 2007, p. 185–210.

[3] P.BOSC, O. PIVERT, "About projection-selection-join queries addressed to possibilistic relational databases", *IEEE Transactions on Fuzzy Systems 13*, 1, 2005, p. 124–139.

[4] P.BOSC, O. PIVERT, "About possibilistic queries and their evaluation", *IEEE Transactions on Fuzzy Systems 15*, 1, 2007, p. 439–452.

Books and Monographs

[1] P. BOSC, A. HADJALI, G. PASI (editors), *Journal of Intelligent Information Systems, Special Issue on Flexible queries in information systems*, 2008. to appear.

[2] O. PIVERT (editor), *From Knowledge representation to Information Processing and Management — Selected Papers from the French Fuzzy Days (LFA 2006)*, *159*, 2008, 1887–2046p.

Articles in referred journals and book chapters

[3] P. BOSC, A. HADJALI, O. PIVERT, "Empty versus overabundant answers to flexible relational queries", *Fuzzy Sets and Systems 159*, 12, 2008, p. 1450–1467.

[4] P. BOSC, A. HADJALI, O. PIVERT, "Incremental controlled relaxation of failing flexible queries", *Journal of Intelligent Information Systems*, 2008, p. 273–287, to appear.

[5] P. BOSC, A. HADJALI, O. PIVERT, "La notion de division tolérante et son intérêt pour remédier aux réponses vides", *Ingénierie des Systèmes d'Information 13*, 5, 2008, p. 131–154.

[6] P. BOSC, A. HADJALI, O. PIVERT, "On the versatility of fuzzy sets for modeling flexible queries", *in: Handbook of Research on Fuzzy Information Processing in Databases*, Information Science Reference, Hershey, PA, USA, 2008, p. 1450–1467.

[7] P. BOSC, A. HADJALI, O. PIVERT, "Overabundant answers to flexible queries — A proximity-based intensification approach", *in: Uncertainty and Intelligent Information Systems*, World Scientific Publishing, 2008, p. 273–287.

[8] P. BOSC, N. I. HSSAIEN, O. PIVERT, "On the evaluation of cardinality-based generalized yes/no queries", *in: Intelligent Techniques and Tools for Novel System Architectures, Series: Studies in Computational Intelligence*, Springer Verlag, 2008, p. 65–79.

[9] P. BOSC, L. LIÉTARD, "Aggregates computed over fuzzy sets and their integration into SQLf", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 16*, 6, 2008, p. 761–792.

[10] P. BOSC, O. PIVERT, "On two qualitative approaches to tolerant inclusion operators", *Fuzzy Sets and Systems 159*, 21, 2008, p. 2786–2805.

[11] H. JAUDOIN, F. FLOUVAT, J.-M. PETIT, F. TOUMANI, "Towards a scalable query rewriting algorithm in presence of value constraints", *Journal on Data Semantics 12*, 2008, to appear.

[12] L. LIÉTARD, D. ROCACHER, P. BOSC, "Compositions de relations d'ordre sur des quantités graduelles et expression de requêtes flexibles", *Technique et Science Informatiques 27*, 2008, p. 51–81.

[13] L. LIÉTARD, D. ROCACHER, "Conditions with aggregates evaluated using gradual numbers", *Control and Cybernetics*, 2008, to appear.

[14] L. LIÉTARD, D. ROCACHER, "Evaluation of quantified statements using gradual numbers", *in: Handbook of Research on Fuzzy Information Processing in Databases*, Information Science Reference, Hershey, PA, 2008, p. 246–269.

Publications in Conferences and Workshops

[15] P. BOSC, A. HADJALI, O. PIVERT, O. SOUFFLET, "Préférences floues compétitives et conditionnelles pour l'interrogation flexible de bases de données", *in: Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA)*, p. 52–59, 2008.

[16] P. BOSC, A. HADJALI, O. PIVERT, "An approach to competitive conditional fuzzy preferences in database flexible querying", *in: Proc. of the 4th IEEE International Conference on Intelligent Systems (IEEE IS)*, 2008.

[17] P. BOSC, A. HADJALI, O. PIVERT, "Cooperative answering to flexible queries via a tolerance relation", *in: Proc. of the 17th International Symposium on International Symposium on Methodologies for Intelligent Systems (ISMIS)*, p. 288–297, 2008.

[18] P. BOSC, A. HADJALI, O. PIVERT, "Exceptions and resemblance: two keys for tolerant division operators", *in: Proc. of the 27th International Conference of the North American Fuzzy Information Processing Society (NAFIPS)*, 2008.

[19] P. BOSC, A. HADJALI, O. PIVERT, "Graded tolerant inclusion and its axiomatization", *in: Proc. of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, 2008.

[20] P. BOSC, A. HADJALI, O. PIVERT, "Une approche coopérative pour le traitement des requêtes flexibles à réponse vide ou pléthorique", *in: Actes des 23èmes Journées Bases de Données Avancées (BDA)*, 2008.

[21] P. BOSC, O. PIVERT, O. SOUFFLET, "Top-k division à préférences ordinales", *in: Actes des 23èmes Journées Bases de Données Avancées (BDA)*, 2008.

[22] P. Bosc, O. Pivert, "A family of tolerant antidivision operators for database fuzzy querying", *in: Proc. of the 2nd International Conference on Scalable Uncertainty Management (SUM)*, p. 92–105, 2008.

[23] P. Bosc, O. Pivert, "On a parameterized antidivision operator for database flexible querying", *in: Proc. of the 19th Int. Conference on Database and Expert Systems Applications (DEXA)*, p. 652–659, 2008.

[24] P. Bosc, O. Pivert, "On the division operator for probabilistic and possibilistic relational databases", *in: Proc. of the 17th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2008.

[25] P. Bosc, O. Pivert, "On the use of tolerant graded inclusions in information retrieval", *in: Actes de la 5ème Conférence en Recherche d'Information et Applications (CORIA)*, p. 321–336, 2008.

[26] P. Bosc, O. Pivert, "Un modèle de base de données possibiliste avec distributions de possibilité incomplètes", *in: Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA)*, p. 18–25, 2008.

[27] P. Colomb, H. Jaudoin, "FlexIS: vers un système d'intégration d'information flexible", *in: Actes des 23èmes Journées Bases de Données Avancées (BDA), Session démonstration*, 2008.

[28] A. Hadjali, S. Kaci, H. Prade, "Database preferences queries — A possibilistic logic approach with symbolic priorities", *in: Proc. of the 5th Inter. Symposium on Foundations of Information and Knowledge Systems (FoIKS)*, p. 291–310, 2008.

[29] A. Hadjali, O. Pivert, "Towards fuzzy query answering using fuzzy views — A graded-subsumption-based approach", *in: Proc. of the 17th International Symposium on International Symposium on Methodologies for Intelligent Systems (ISMIS)*, p. 268–277, 2008.

[30] L. Liétard, D. Rocacher, S.-E. Tbahriti, "Preferences and Bipolarity in Query Languages", *in: Proc. of the 27th International Conference of the North American Fuzzy Information Processing Society (NAFIPS)*, 2008.

[31] L. Liétard, D. Rocacher, "Requêtes bipolaires exprimant des préférences", *in: Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA)*, p. 68–75, 2008.

[32] L. Liétard, "A new definition for linguistic summaries of data", *in: Proc. of the 17th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, p. 506–511, 2008.