Project-Team PILGRIM

Graduality, Imprecision, and Mediation in Database Management Systems

Lannion

Activity Report

2012

1 Team

Head of the team

Olivier Pivert, Professor, Enssat

Administrative assistant

Joëlle Thépault, Enssat, 20%
Nelly Vaucelle, Enssat, 10%

Université Rennes 1 personnel

Allel Hadjali, Associate Professor, HdR, Enssat (up to August 31, 2012)
Hélène Jaudoin, Associate Professor, Enssat
Ludovic Liétard, Associate Professor, HdR, IUT Lannion
Daniel Rocacher, Professor, Enssat
Grégory Smits, Associate Professor, IUT Lannion

PhD students

Nouredine Tamani, Région Bretagne grant and Conseil Général 22 grant, since January 2009
Katia Abbaci, Région Bretagne grant and ANR grant, since November 2009

2 Overall Objectives

2.1 Introduction

In database research, the last two decades have witnessed a growing interest in preference queries on the one hand, and uncertain databases on the other hand.

Motivations for introducing preferences inside database queries are manifold. First, it has appeared to be desirable to offer more expressive query languages that can be more faithful to what a user intends to say. Second, the introduction of preferences in queries provides a basis for rank-ordering the retrieved items, which is especially valuable in case of large sets of items satisfying a query. Third, on the contrary, a classical query may also have an empty set of answers, while a relaxed (and thus less restrictive) version of the query might be matched by items in the database.

Approaches to database preference queries may be classified into two categories according to their qualitative or quantitative nature. In the qualitative approach, preferences are defined through binary preference relations. Among the representatives of this family of approaches, let us mention an approach based on CP-nets, and those relying on a dominance relation, e.g. Pareto order, in particular Skyline queries. In the quantitative approach, preferences are expressed quantitatively by a monotone scoring function (the overall score is positively correlated with partial scores). Since the scoring function associates each tuple with a numerical score, tuple t_1 is preferred to tuple t_2 if the score of t_1 is higher than the score of t_2. Well-known representatives of this family of approaches are top-k queries, and *fuzzy-set-based approaches*. The team Pilgrim particularly studies the latter, and the line followed 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 a relational setting.

Basically, a fuzzy 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, a user 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" corresponds to 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.

As to the second topic mentioned at the beginning of this introduction, i.e., uncertain databases, it already has a rather long history. Indeed, since the late 70s, many authors have made diverse proposals to model and handle databases involving uncertain or incomplete data. In particular, the last two decades have witnessed a profusion of research works on this topic. The notion of an uncertain database covers two aspects: i) attribute uncertainty: when some attribute values are ill-known; ii) existential uncertainty: when the existence of some tuples is itself uncertain. Even though most works about uncertain databases consider probability theory as the underlying uncertainty model, some approaches — in particular those proposed by Pilgrim — rather rely on possibility theory. The issue is not to demonstrate that the possibility-theory-based framework is "better" than the probabilistic one at modeling uncertain databases, but that it constitutes an interesting alternative inasmuch as it captures a different kind of uncertainty (of a subjective, nonfrequential, nature). A typical example is that of a person who witnesses a car accident and who does not remember for sure the model of the car involved. In such a case, it seems reasonable to model the uncertain value by means of a possibility distribution, e.g., {1/Mazda, 1/Toyota, 0.7/Honda} rather than with a probability distribution which would be artificially normalized. In contrast with probability theory, one expects the following advantages when using possibility theory:

- the qualitative nature of the model makes easier the elicitation of the degrees attached to the various candidate values;
- in probability theory, the fact that the sum of the degrees from a distribution must equal 1 makes it difficult to deal with incompletely known distributions;
- there does not exist any probabilistic logic which is complete and works locally as possibilistic logic does: this can be problematic in the case where the degrees attached to certain pieces of data must be automatically deduced from those attached to some other pieces of data (e.g., when data coming from different sources are merged into a single database).

A recent 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 several global schemas are available and where the sources can be accessed through views defined on one of these schemas (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 using the views. The overall objective is thus to define 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.

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 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.

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

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 greater than or equal to 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 $x \geq 30$, $\mu_{young}(x) = 1$ 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 generally 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, others 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 may be more or less included in F. Different approaches can be envisaged, among which one is based on the notion of a 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, for instance using 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]:

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

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

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

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:

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

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 to the fuzzy set F. For example, Paul's age may be only imprecisely known as "close to 20", where a given fuzzy set is associated with this fuzzy linguistic expression.

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

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

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. 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 multisets) and their connection with fuzzy numbers.

As to possibility distributions, they are used to represent imprecise (imperfect) data. By doing so, a straightforward connection can be established between a possibilistic database 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 combinatory 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).

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

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 that 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 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 [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 approach suffers from a lack of extensibility as changing or adding new sources affects the global schema. On the 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 [HIM+04] enables sharing heterogeneous data in a distributed and scalable way. 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

[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.

[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.

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 works show that query processing in data integration is related to the general problem of answering queries using views [Hal01, Len02]. In such a setting, the 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 to 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 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] in order to ensure scalability in terms of the number of data sources. A few recent papers also contributed to the development of data integration systems capable of taking into account imprecision or uncertainty. Most of the works along that line use probability theory in order to capture the form of uncertainty that stems from the schema definition process, or that associated with the mere existence of data, or aim at modelling the approximate nature of the semantic links between the data sources and the mediated schema.

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

[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.

particular on the Web (many commercial systems, e.g. Google or Yahoo use a 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, 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 a hard job and the demand is presently not so 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

- PostgreSQLF is a flexible querying prototype that aims at evaluating fuzzy queries addressed to regular databases. It is an extension of PostgreSQL which implements the fuzzy query language SQLf defined in the team. This prototype has been presented at PG Day France 2012 (`http://www.pgday.fr/`), which is the yearly meeting of PostgreSQL's users, and at the demo session of BDA 2012 [28].

- CORTEX (CORrelaTion-based Query EXpansion): Retrieving data from large-scale databases sometimes leads to plethoric answers especially when queries are underspecified. To overcome this problem, we proposed an approach which strengthens the initial query by adding new predicates (cf. Subsection 6.2.4). These predicates are selected among predefined ones principally according to their degree of semantic correlation with the initial query. This way, we avoid an excessive modification of its initial scope. Considering the size of the initial answer set and the number of expected results specified by the user, fuzzy cardinalities are used to assess the reduction capability of these correlated predefined predicates. This approach has been implemented as a research prototype, named CORTEX, to query a database containing 10,000 ads about second hand cars.

- LUCIFER (Leveraging Unveiled Conflicts In FlexiblE Requests): This prototype deals with conjunctive fuzzy queries that yield an empty or poorly satisfactory answer set. It implements a cooperative answering approach [23] which efficiently retrieves the minimal failing subqueries of the initial query (which can then be used to explain the failure and revise the query).

- FALSTAFF (FAceted search engine Leveraging Summaries of daTA with Fuzzy Features): Faced with the difficulty of formulating precise queries to retrieve items from large scale databases, interactive interfaces implementing a faceted search strategy help the users navigate through the data by successively selecting facet-value pairs. This prototype uses a faceted search strategy to construct fuzzy queries. The interactive query construction process relies on precomputed metadata that informs about the data distribution over a predefined vocabulary [29].

- COKE (Connected KEywords): Keyword queries have emerged as the most convenient way to query data sources especially for unexperienced users. Introduced initially for document retrieval on the web, such queries are defined as an enumeration of keywords corresponding to a rough description of what users are looking for. The interpretation process of keyword queries has then been adapted to handle structured data like relational databases or XML documents. Instead of considering queries as an unstructured enumeration of keywords, the approach underlying the COKE system lets users structure their keyword queries using simple but meaningful grammatical connectors. Using the data structure intensively, a COKE query is translated into SQL to retrieve exact answers. An autocompletion strategy is also proposed to help users take advantage of connectors in their keyword queries. An experimentation shows that the COKE system efficiently retrieves more relevant and precise answers than classical queries made of keywords enumerations and offers a good coverage of possible query patterns.

6 New Results

6.1 Flexible querying of classical databases

6.1.1 Preference queries

Participants: Olivier Pivert, Grégory Smits, Allel Hadjali.

The works presented hereafter deal with different aspects of preference queries (fuzzy and others) in a database context.

- *Fuzzy preference queries to relational databases.* [2] provides a comprehensive study on fuzzy preference queries in the context of relational databases. In particular, this book aims to show that fuzzy set theory constitutes a highly expressive framework for modeling preference queries. It presents a study of the algorithmic aspects related to the evaluation of such queries in order to demonstrate that this framework offers a good trade-off between expressivity and efficiency.

- *Fuzzy connectives.* In most query languages, conjunctive and disjunctive combinations of conditions remain the usual way for aggregation. Fuzzy query languages also offer trade-off operators, such as means in order to compensate between elementary conditions. In [7, 17], we introduce a new type of condition basically founded on the interaction between two predicates, thus enriching the panoply of tools the user is provided with and the power of query languages. In [19], the issue of fuzzy preference aggregation in the context of recommender systems is studied.

- *Outranking.* In [6], we describe an approach to database preference queries based on the notion of outranking, suited to the situation where preferences on different attributes are not commensurable. This model constitutes an alternative to the use of Pareto order whose main drawback is to leave many tuples incomparable in general. Even though outranking does not define an order in the mathematical sense of the term, we describe a technique which yields a complete pre-order, based on a global aggregation of the outranking degrees computed for each pair of tuples, which reflects the global "quality" of a tuple with respect to the others. In [27], an efficient evaluation technique that involves a pre-filtering step is described, whose aim is to limit the number of pairwise tuple comparisons necessary to determine the best answers to a query.

- *Contextual preferences.* Users' preferences have traditionally been exploited in query personalization to better serve their information needs. Most of the time, user preferences depend on context, that is, they may have different values depending on context. In In [9], we propose a fuzzy-rule-based model for the representation of contextual preferences in a database querying framework. We discuss the augmentation of a query with preferences deduced from information regarding the current context of the user. To this end, we make use of an appropriate inference pattern, called *generalized modus ponens*, able to deal with data and knowledge when they are described in a fuzzy way.

- *Reinforced inclusion.* In [18], a fuzzy inclusion indicator is introduced, which is derived from a connective aimed at modulating a fuzzy criterion according to the satisfaction of another one. The idea is to express that one is all the more demanding as to the degree attached to an element x in a set B as this element has a high degree of membership degree to a set A. The use of this reinforced inclusion indicator is illustrated in the context of database querying.

- *Conditions on scores.* [26] deals with database preference queries involving fuzzy conditions which explicitly involve the satisfaction degree obtained by a tuple with respect to a (possibly complex) fuzzy condition. Different forms that such conditions can take are investigated, and some elements about query processing are provided.

- *Query language with symbolic preferences.* In [25], we describe a preference query language based on SQL, which captures preferences modeled by means of symbolic scores expressed in a linguistic manner. The high expressivity of this language is emphasized, and it is shown that some connections between the preference criteria considered and Boolean conditions make it possible to take advantage of the optimization mechanisms offered by classical DBMSs when it comes to query processing.

- *Fuzzy inclusion queries.* In [20], we focus on a specific type of preference queries, based on the concept of fuzzy inclusion, in the context of geographical applications, notably in the biological domain. We deal with imprecise data obtained from samples corresponding to diverse oceanic areas. The approach proposed, based on two fuzzy extensions of the division operator from relational algebra, has been implemented using the spatial database extension of the PostgreSQL system, namely PostGis.

- *Flexible Skyline.* Given a set of points in a space, a skyline query retrieves the points that are not dominated by any other in the sense of Pareto order. When the number of dimensions on which preferences are expressed gets high, many tuples may become incomparable. In [10], this problem is addressed and an approach based on the use of fuzzy quantifiers is proposed.

6.1.2 Cooperative answering to flexible database queries

Participants: Grégory Smits, Olivier Pivert, Allel Hadjali, Hélène Jaudoin.

The practical need for endowing information systems with the ability to exhibit cooperative behavior (thus making them more "intelligent") has been recognized at least since the early 90s. The main intent of cooperative systems is to provide correct, non-misleading and useful answers, rather than literal answers to user queries. Different aspects of this problem are tackled in the works presented hereafter.

- *Plethoric answers.* Seeking data from large-scale databases often leads to a plethoric answer problem. A possible approach to reduce the set of retrieved items and to make it more manageable is to constrain the initial query with additional predicates. The approach presented in [5], relies on the identification of correlation links between predicates related to attributes of the relation of interest. Thus, the initial query is strengthened by additional predicates that are semantically close to the user-specified ones.

- *Empty answers.* In [24], we deal with conjunctive fuzzy queries that yield an empty or unsatisfactory answer set. We propose a cooperative answering approach which efficiently retrieves the minimal failing subqueries of the initial query (which can then be used to explain the failure). The detection of the minimal failing subqueries relies on a prior step of fuzzy cardinalities computation. The main advantage of this strategy is to rely on a single scan of the database. Moreover, the fuzzy cardinalities which describe the data distributions easily fit in main memory. The research prototype presented in [23] is an implementation of this cooperative approach.

- *Faceted search.* Faced with the difficulty of formulating precise queries to retrieve items from large scale databases, interactive interfaces implementing a faceted search strategy help the users navigate through the data by successively selecting facet-value pairs. In [29], we present a faceted search strategy to construct fuzzy queries. The interactive query construction process relies on precomputed metadata that inform about the data distribution over a predefined vocabulary.

- *Detection of suspect answers.* In practice, dirty data are quite common, especially when a database results from the integration of multiple data sources. In the presence of inconsistent information, two classical approaches consist in (i) cleaning the information by means of an automated process, for instance by performing a minimal set of updates aimed at restoring consistency, (ii) returning only the answers that are certain with respect to a given query, as in consistent query answering. In [22, 21], we propose an alternative approach, somewhat inspired by artificial intelligence works, which is aimed at warning the user about the presence of suspect answers (i.e., answers involved in the violation of a functional dependency) in a query result. Roughly speaking, the idea is that such elements can be identified inasmuch as they can be found in the answers to contradictory queries. This idea may also be refined by introducing some gradedness in terms of cardinality or similarity.

6.2 Fuzzy preferences in intelligent transportation systems

Participants: Ludovic Liétard, Nouredine Tamani, Allel Hadjali, Olivier Pivert, Daniel Rocacher.

Four years ago, we started investigating a new topic, namely the application of fuzzy set theory to the specification of a route planning system involving sophisticated user preferences. In 2012, the results obtained concern the following issues:

- *Route planning.* In [8], we propose a contribution for a new generation of route planners able to deal with complex and sophisticated preferences. Fuzzy set theory is advocated as a framework for modelling preferences. First, a typology of user preferences that make sense in the context of unimodal route planning is investigated. The bipolar nature of such preferences is discussed as well. The foundations of both a formal language and an SQL-like language dedicated to bipolar route planning queries are then presented and illustrated with different examples. The basic components of the architecture of the system proposed are described and deep details about query evaluation are provided. Finally, the approach is evaluated by means of a set of experiments.

- *Multimodal route planning.* Today's web applications dedicated to trip planning are mainly limited to only one kind of transport service (urban transport, air transport, transport by rail and sea, etc.). The user is then compelled to manually build his/her trip by combining available transport services. This type of querying can lead to process huge volumes of data, and can deliver plethoric responses. In this context, user preferences are considered in the process of best trip retrieval. In [4], user preferences are modelled by fuzzy bipolar conditions which associate essential preferences and minor preferences, and expressed within a relational framework based on a bipolar relational algebra and a relational query language called *bipolar SQLf*. A reasoning mechanism made of a description logic extended to m-ary relations and fuzzy bipolar conditions makes it possible to target only the most relevant data when evaluating user queries.

6.3 Flexible access to Web services

Participants: Allel Hadjali, Ludovic Liétard, Daniel Rocacher, Katia Abbaci.

With the development of Service oriented Computing (SOC), more and more functionally similar Web services are deployed over the Web. Quality of service (QoS) aspects (e.g., availability, response time, etc.) are thus crucial for selecting among functionally similar Web services. Moreover, skyline queries have been proven to be an important concept for selecting Web service based on QoS. The works presented hereafter deal with preference queries aimed at selecting the most appropriate Web services wrt a given user's need.

- *Ranking composite services.* In [12, 13], a system that supports preference query answering over a set of data Web services is described. This system is able to rank-order the query results in the presence of fuzzy preferences. To do so, different software components organized into two main modules are provided. The first module provides the top-k service compositions. It is mainly based on (i) query rewriting techniques to generate relevant services and compositions, (ii) fuzzy dominance relationship to rank both individual and composite services. The second module adopts a fuzzy database approach to provide a graded service composition execution engine ranking the data results.

- *Collective skyline.* In [14], a novel concept called *collective skyline* is introduced, which aims to deal with the problem of multiple users preferences in the context of web service selection. The approach is based on the notion of collective dominance: a service s_i collectively dominates another service s_j if and only if s_i dominates s_j for more than half of users. A set of experiments that demonstrate the effectiveness of this method are presented in [14]. A variant is presented in [16], where a Web service selection framework based on QoS, called WS-Sky, is presented.

- *Dealing with uncertain QoS attributes.* In a dynamic Web service environment, the QoS delivered by a Web service is inherently uncertain. In [15], we tackle the problem of skyline on uncertain QoS. We represent each QoS attribute of a Web service using a possibility distribution and introduce two skyline extensions on uncertain QoS called *pos-dominant skyline* and *nec-dominant skyline*. Appropriate algorithms have been developed to efficiently compute both the results of these extended skyline queries.

- *Bipolar fuzzy preferences on QoS attributes.* In [11], a method for handling preferences on QoS parameters in a bipolar way, (i.e., distinguishing between negative and positive preferences) is proposed. A set of experiments based on real data has been conducted to demonstrate the relevance and scalability of this approach.

7 Other Grants and Activities

7.1 National actions

- Ludovic Liétard, Allel Hadjali, and Daniel Rocacher participated in the ANR project "AOC" (2009-2012), which dealt with the definition of matching methods for complex objects (graphs in particular). The other teams involved were from IRIT (Toulouse), PRISM (Versailles), LIRIS (Lyon), LIESP (Lyon).

- The team Pilgrim is involved in a PME project of the "pôle Images et Réseaux", named IntelSearch, in collaboration with Semsoft, Swid, and Ensai. This project is about flexible data integration system and the leader for Pilgrim is Hélène Jaudoin.

7.2 International actions

- Salma Ben Harrath, Master's student from the University of Tunis (Tunisia), co-supervised by Allel Hadjali, spent five month in our team from March 8, 2012 to July 1, 2012.
- Khaoula Mabrouki, Ph.D. student from the University of Tunis (Tunisia), co-supervised by Olivier Pivert and Grégory Smits, spent one month in our team, from March 19, 2012 to April 22, 2012.
- Irène Abi-Zeid, Professor at Laval University (Québec, Canada), spent one month in our team from June 4, 2012 to June 29, 2012.

8 Dissemination

8.1 Teaching

Project members give lectures in different faculties of engineering, in the third cycle University curriculum: ”Bases de données avancées“ in the speciality ”Interaction intelligente avec l'information “ of the Master's degree in computer science at University of Rennes 1, and at Enssat (third year level cursus).

A. Hadjali gave a Master's course entitled "Requêtes à préférences" at the University USTHB of Algiers (Algeria) in September 2012.

8.2 Scientific activities

8.2.1 Highlights of the year

- Ludovic Liétard defended his “Habilitation à Diriger des Recherches” [3] on December 7, 2012.
- Nouredine Tamani defended his Ph.D. thesis [4] on April 23, 2012.
- Publication by Imperial College Press of a handbook authored by O. Pivert and P. Bosc entitled “Fuzzy Preference Queries to Relational Databases” [2].

8.2.2 Program committees

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

- 23$^{\text{rd}}$ International Conference on Database and Expert Systems Applications (DEXA 2012), Vienna, Austria, September 3-7, 2012.
- 20$^{\text{th}}$ International Symposium on Methodologies for Intelligent Systems (ISMIS 2012), Macau, China, December 5-7, 2012.
- Colloque sur l'Optimisation et les Systèmes d'Information (COSI 2012), Tlemcen, Algeria, May 12-15, 2012.

- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2012), Compiègne, France, November 15-16, 2012.

L. Liétard served as a member of the following program committees:

- 27^{th} ACM Symposium on Applied Computing (SAC 2012), Riva del Garda, Italy, March 26-30, 2012.

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

- 27^{th} ACM Symposium on Applied Computing (SAC 2012), Riva del Garda, Italy, March 26-30, 2012.
- 20^{th} International Symposium on Methodologies for Intelligent Systems (ISMIS 2012), Macau, China, December 5-7, 2012.
- 6^{th} International Conference on Scalable Uncertainty Management (SUM 2012), Marburg, Germany, September 17-19, 2012.
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2012), Compiègne, France, November 15-16, 2012.

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

- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2012), Compiègne, France, November 15-16, 2012.

8.2.3 Editorial boards

Allel Hadjali is a member of the following editorial board:

- Journal of Advanced Computing Technologies.

Olivier Pivert is a member of the following editorial boards:

- Journal of Intelligent Information Systems,
- Fuzzy Sets and Systems,
- International Journal of Fuzziness, Uncertainty and Knowledge-Based Systems,

8.2.4 Steering committees

O. Pivert is as a member of the following steering committees:

- International Conference on Flexible Query-Answering Systems (FQAS).
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA).

8.2.5 Edition of special issues

- Allel Hadjali and Olivier Pivert are among the guest editors of a special issue of the international journal Fuzzy Sets and Systems, devoted to "Advances in Soft Computing Applied to Database and Information Systems" [1], following a special session on this topic at the IFSA/EUSFLAT'09 international joint conference.

9 Bibliography

Major publications by the team in recent years

[1] P. Bosc, A. Hadjali, O. Pivert, "Incremental Controlled Relaxation of Failing Flexible Queries", *Journal of Intelligent Information Systems 33*, 3, 2009, p. 261–283.

[2] P. Bosc, L. Liétard, O. Pivert, D. Rocacher, *Gradualité et imprécision dans les bases de données*, Ellipses, 2004.

[3] P. Bosc, O. Pivert, "Modeling and Querying Uncertain Relational Databases: A Survey of Approaches Based on the Possible Worlds Semantics", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 18*, 5, 2010, p. 565–603.

[4] 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.

[5] 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.

[6] P.Bosc, O. Pivert, "About Possibilistic Queries and their Evaluation", *IEEE Transactions on Fuzzy Systems 15*, 1, 2007, p. 439–452.

[7] O. Pivert, P. Bosc, *Fuzzy Preference Queries to Relational Databases*, Imperial College Press, London, UK, 2012.

Books and Monographs

[1] P. Bosc, G. De Tré, J. Dujmovic, A. Hadjali, O. Pivert, R. Ribeiro, S. Zadrożny (editors), *Fuzzy Sets and Systems – Special Issue on Advances in Soft Computing Applied to Databases and Information Systems*, *196*, Elsevier, 2012.

[2] O. Pivert, P. Bosc, *Fuzzy Preference Queries to Relational Databases*, Imperial College Press, London, UK, 2012.

Doctoral dissertations and "Habilitation" theses

[3] L. Liétard, *Apports de la théorie des ensembles flous aux bases de données relationnelles*, PhD Thesis, Habilitation à Diriger les Recherches – University of Rennes 1, 7 décembre 2012.

[4] N. Tamani, *Planification d'itinéraire multimodal : Une approche bipolaire floue*, PhD Thesis, University of Rennes 1 – Ecole doctorale MATISSE, April 23, 2012, supervised by L. Liétard and D. Rocacher.

Articles in referred journals and book chapters

[5] P. Bosc, A. Hadjali, O. Pivert, G. Smits, "An Approach Based on Predicate Correlation to the Reduction of Plethoric Answer Sets", *in: Advances in Knowledge Discovery and Management vol. 2*, F. Guillet, G. Ritschard, and D. Zighed (editors), *Studies in Computational Intelligence*, Springer, 2012, p. 213–234.

[6] P. Bosc, O. Pivert, G. Smits, "An Approach to Database Preference Queries Based on an Outranking Relation", *International Journal of Computational Intelligence Systems 5*, 4, 2012, p. 789–804.

[7] P. Bosc, O. Pivert, "On Four Noncommutative Fuzzy Connectives and their Axiomatization", *Fuzzy Sets and Systems 202*, 2012, p. 42–60.

[8] A. Hadjali, A. Mokhtari, O. Pivert, "Expressing and Processing Complex Preferences in Route Planning Queries: Towards a Fuzzy-Set-Based Approach", *Fuzzy Sets and Systems 196*, 2012, p. 82–104.

[9] A. Hadjali, A. Mokhtari, O. Pivert, "On a Fuzzy-Rule-Based Model for Handling Contextual Preference Queries", *International Journal of Computational Intelligence Systems 5*, 4, 2012, p. 775–788.

Publications in Conferences and Workshops

[10] K. Abbaci, A. Hadjali, L. Liétard, D. Rocacher, "Un opérateur skyline flexible pour l'interrogation de bases de données : Premières explorations", *in: Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'12)*, Cépaduès, 2012.

[11] K. Abbaci, F. Lemos, A. Hadjali, D. Grigori, L. Lietard, D. Rocacher, M. Bouzeghoub, "A Bipolar Approach to the Handling of User Preferences in Business Processes Retrieval", *in: Proc. of IPMU'12*, S. Greco, B. Bouchon-Meunier, G. Coletti, M. Fedrizzi, B. Matarazzo, R. R. Yager (editors), *Communications in Computer and Information Science*, *297*, Springer, p. 400–409, 2012.

[12] S. Amdouni, M. Barhamgi, D. Benslimane, A. Hadjali, K. Benouaret, R. Faiz, "Answering Fuzzy Preference Queries over Data Web Services", *in: Proc. of ICWE'12*, M. Brambilla, T. Tokuda, R. Tolksdorf (editors), *Lecture Notes in Computer Science*, *7387*, Springer, p. 456–460, 2012.

[13] S. Amdouni, D. Benslimane, M. Barhamgi, A. HadjAli, R. Faiz, P. Ghodous, "A Preference-Aware Query Model for Data Web Services", *in: Proc. of ER'12*, P. Atzeni, D. W. Cheung, S. Ram (editors), *Lecture Notes in Computer Science*, *7532*, Springer, p. 409–422, 2012.

[14] K. Benouaret, D. Benslimane, A. Hadjali, "Selecting Skyline Web Services for Multiple Users Preferences", *in: Proc. of ICWS'12*, C. A. Goble, P. P. Chen, J. Zhang (editors), IEEE, p. 635–636, 2012.

[15] K. Benouaret, D. Benslimane, A. Hadjali, "Selecting Skyline Web Services from Uncertain QoS", *in: Proc. of the IEEE 9th International Conference on Service Computing (IEEE SCC'12)*, Honolulu, Hawaii, USA, 2012.

[16] K. Benouaret, D. Benslimane, A. Hadjali, "WS-Sky: An Efficient and Flexible Framework for QoS-Aware Web Service Selection", *in: Proc. of the IEEE 9th International Conference on Service Computing (IEEE SCC'12)*, Honolulu, Hawaii, USA, 2012.

[17] P. Bosc, O. Pivert, "À propos de trois connecteurs flous et de leur usage dans des requêtes flexibles", *in: Actes des 28e journées Bases de Données Avancées (BDA'12)*, 2012.

[18] P. Bosc, O. Pivert, "On a Reinforced Fuzzy Inclusion and its Application to Database Querying", *in: Proc. of the 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'12)*, *CCIS*, *297*, Springer, p. 351–360, 2012.

[19] S. Boulkrinat, A. Hadjali, A. Mokhtari, "Vers un système de recommandation basé sur une agrégation des préférences floues", *in: Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'12)*, Cépaduès, 2012.

[20] T. Girault, H. Jaudoin, O. Pivert, "Flexible access to fuzzy data for biodiversity monitoring", *in: Actes des 28e journées Bases de Données Avancées (BDA'12), Session démonstrations*, 2012.

[21] O. Pivert, H. Prade, "Détecter les réponses suspectes en présence de données inconsistantes", *in: Actes des 28e journées Bases de Données Avancées (BDA'12)*, 2012.

[22] O. Pivert, H. Prade, "Detecting Suspect Answers in the Presence of Inconsistent Information", *in: Proc. of the 7th International Symposium on Foundations of Information and Knowledge Systems (FoIKS'12)*, T. Lukasiewicz, A. Sali (editors), *Lecture Notes in Computer Science*, *7153*, Springer, p. 279–298, 2012.

[23] O. Pivert, G. Smits, A. Hadjali, H. Jaudoin, "LUCIFER : Un système de détection de conflits dans les requêtes flexibles", *in: Actes de la 12ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'12), Session démonstrations*, p. 617–620, Bordeaux, France, 2012.

[24] O. Pivert, G. Smits, H. Jaudoin, A. Hadjali, "Détection efficace de conflits dans les requêtes floues", *in: Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'12)*, Cépaduès, 2012.

[25] O. Pivert, G. Smits, "On a Preference Query Language that Handles Symbolic Scores", *in: Proc. of the 16th East-European Conference on Advances in Databases and Information Systems (ADBIS'12)*, *LNCS*, *7503*, Springer, p. 296–309, 2012.

[26] O. Pivert, G. Smits, "On Fuzzy Preference Queries Explicitly Handling Satisfaction Levels", *in: Proc. of the 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'12)*, *CCIS*, *297*, Springer, p. 341–350, 2012.

[27] O. Pivert, G. Smits, "Towards an Efficient Processing of Outranking-Based Preference Queries", *in: Proc. of the 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'12)*, *CCIS*, *297*, Springer, p. 471–480, 2012.

[28] G. Smits, O. Pivert, T. Girault, "PostgreSQLf : un système d'interrogation floue", *in: Actes des 28e journées Bases de Données Avancées (BDA'12), Session démonstrations*, 2012.

[29] G. Smits, O. Pivert, "A Fuzzy-Summary-Based Approach to Faceted Search in Relational Databases", *in: Proc. of the 16th East-European Conference on Advances in Databases and Information Systems (ADBIS'12)*, *LNCS*, *7503*, Springer, p. 357–370, 2012.