Project-Team SHAMAN

A Symbolic and Human-Centric View of Data Management

Lannion

Activity Report

2015

1 Team

Head of the team
Olivier Pivert, Professor, Enssat

Administrative assistant
Joëlle Thépault, Enssat, 20%
Vincent Chevrette, Enssat, 10%

Université Rennes 1 personnel
François Goasdoué, Professor, Enssat
Hélène Jaudoin, Associate Professor, Enssat
Ludovic Liétard, Associate Professor, HdR, IUT Lannion
Pierre Nerzic, Associate Professor, IUT Lannion (since February 2015)
Daniel Rocacher, Professor, Enssat
Grégory Smits, Associate Professor, IUT Lannion
Virginie Thion, Associate Professor, Enssat

PhD students
Sara El Hassad, Région Bretagne grant and Lannion Trégor Communauté grant, since October 2014
William Correa Beltran, Région Bretagne grant and Conseil Général 22 grant, since October 2012
Aurélien Moreau, DGA contract, since November 2014
Olfa Slama, DGA contract, since November 2014

Master students
David Mahéo, Enssat, March-July 2014

2 Overall Objectives

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 Shaman 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 be handled. 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 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 Shaman 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. 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 intensional 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 ($\bot$) tied by de Morgan's laws. Then:

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

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.

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

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 considered, 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,
- Łukasiewicz 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]:

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

[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 (for Boolean events) 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 ($\Pi = N = 0$ to $\Pi = 1$ and $N = 0$), then those which are completely possible and somewhat certain (from $\Pi = 1$ and $N = 0$ to $\Pi = N = 1$). 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

[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.

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.

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 Shaman 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:

- bipolar fuzzy queries (including two parts: one viewed as a constraint, the other as a wish),
- the use of a predefined fuzzy vocabulary (which raises the question of its adequacy wrt to the actual content of the database),
- fuzzy extensions of Skyline queries (based on Pareto order),
- implementation and query optimization issues.

[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. 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).

3.4 Query rewriting using views

3.4.1 Data integration

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

[Len02] M. Lenzerini, "Data Integration : A Theoretical Perspective", *in: PODS*, 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.

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 of queries that enable to retrieve data from other peers.

3.4.2 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

[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.

[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] 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

Flexible queries have many potential application domains. Indeed, soft querying turns out to be relevant in a great variety of contexts, such as web search engines, 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.

As for uncertain data management, many potential domains could take advantage of advanced systems capable of storing and querying databases where some pieces of information are imprecise/uncertain: 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.

In the near future, we intend to focus on two application domains:

- Open data management. One of the challenges in web data management today is to define adequate tools allowing users to extract the data that are the most likely to fulfill all or part of their information needs, then to understand and automatically correlate these data in order to elaborate relevant answers or analyses. Open data may be of various levels of quality: they may be imprecise, incomplete, inconsistent and/or their reliability/freshness may be somewhat questionable. An appropriate data model and suitable querying tools must then be defined for dealing with the imperfection that may pervade data in this context. On the other hand, it is of prime importance to provide end-users with simple and flexible means to better understand and analyze open data. The standards of W3C offer popular languages for representing both open and structured data. Another objective is to propose analytical tools suited to these languages

[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.

[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.

through the construction of RDF data warehouses, whereas fuzzy-set-based data summarization approaches should constitute an important step towards making open data more intelligible to non-expert users.

- Environmental information systems. This work will be performed in collaboration with the Biological Station based in Roscoff (Finistère). The general objective is to define an information system architecture (along with an associated "toolbox") suited to the context of marine biodiversity monitoring and environmental protection. We intend to study three main aspects:
  - definition of a data warehouse model suited to this context, capable of dealing with missing values, imprecise information (a situation which often occurs due to the way data is collected and described, through sampling campaigns and human-performed labeling, in particular), uncertain data (uncertainty is unavoidable when data are obtained by means of predictive models, for instance).
  - identification of new needs in terms of query expression: new OLAP operators suitable for the model, making it possible to handle dimensions described by fuzzy concept trees, to manage fuzzy cardinalities, possibility distributions and so on.
  - knowledge discovery: we are notably interested in exploiting a concept that comes from artificial intelligence but has not been applied in the domain of data management yet: that of an analogical proportion, which underlies propositions of the type "A is to B as C is to D". We believe that discovering such "regularities" in a dataset could prove very useful for many purposes connected to environmental monitoring issues, in particular when it comes to predict the evolution of an ecosystem or the population of a species, 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 is coupled with a graphical interface names ReqFlex [SPG13] that makes it easy for an end user to specify his/her fuzzy queries.

- 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

[SPG13] G. Smits, O. Pivert, T. Girault, "ReqFlex: Fuzzy Queries for Everyone", *PVLDB 6*, 12, 2013, p. 1206–1209.

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 [BHPS10].

- 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 which efficiently retrieves the minimal failing subqueries of the initial query (which can then be used to explain the failure and revise the query) [PSHJ12].

- 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 [SP12].

- 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 [SPJP13]. 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.

[BHPS10] P. Bosc, A. Hadjali, O. Pivert, G. Smits, "CORTEX — CORrelaTion-based query EXpansion", *in : Actes des 26e Journées Bases de Données Avancées (BDA'10), session démonstration*, 2010.

[PSHJ12] O. Pivert, G. Smits, A. Hadjali, H. Jaudoin, "LUCIFER : Un système de détection de conflits dans les requêtes flexibles", *in : Actes de la 12e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'12)*, p. 617–620, 2012.

[SP12] 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)*, T. Morzy, T. Haerder, R. Wrembel (editors), *LNCS, 7503*, Springer, p. 357–370, 2012.

[SPJP13] G. Smits, O. Pivert, H. Jaudoin, F. Paulus, "An Autocompletion Mechanism for Enriched Keyword Queries to RDF Data Sources", *in : Proc. of the 10th International Conference on Flexible Query Answering Systems (FQAS'13)*, 2013.

6 New Results

6.1 Possibilistic database modeling and querying

Participants: Olivier Pivert, Ludovic Liétard.

Many works have been undertaken in the area of "fuzzy databases" in the last twenty years. This term is sometimes misused or misleading since it covers both fuzzy querying against regular databases and the handling of databases that are pervaded with imprecision or uncertainty in the data (as opposed to queries).

In [4], we consider relational databases containing uncertain attribute values, when some knowledge is available about the more or less certain value (or disjunction of values) that a given attribute in a tuple may take. We propose a possibility-theory-based model suited to this context and extend the operators of relational algebra so as to handle such relations in a "compact", thus efficient way. It is shown that the model is a representation system for the whole relational algebra. An important result is that the data complexity associated with the extended operators in this context is the same as in the classical database case, which makes the approach highly scalable. A possibilistic logic encoding of the model is also outlined.

In [29, 30], a simplified version of this model is proposed in order to deal with the situation where a database may contain suspect values, i.e. precise values whose validity is not certain, but one is unable to quantify the level of uncertainty attached to such values.

In [11], we consider the modeling and querying of evidential databases, which rely on evidence theory. The semantics of the model proposed is defined in terms of the possible worlds that such an uncertain database is associated with.

In [27], we consider the determination of the satisfaction of a fuzzy predicate by an uncertain data (represented by a possibility distribution). Two indices can be computed to estimate this satisfaction, namely the possibility and the necessity of a fuzzy event. It is shown that these two indices may not always be informative enough and we propose new indices (along with their properties).

6.2 Flexible database querying

6.2.1 Preference queries

Participants: Olivier Pivert, Grégory Smits, Virginie Thion, Ludovic Liétard, Daniel Rocacher.

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

- In [6], we present an implementation strategy for a fuzzy querying system embedded in a regular DBMS. This system relies on the language SQLf that makes it possible to express a great variety of fuzzy queries. Experiments show that this implementation strategy induces performance gains with respect to existing strategies based on a loose (or milder) coupling between a fuzzy querying layer and a DBMS, that necessitate an external postprocessing so as to compute the result in the form of a fuzzy relation. We

also describe a user-friendly interface aimed at helping nonexpert users express their fuzzy queries in an intuitive manner (using graphical tools).

- *Bipolar fuzzy queries.* In [8], we consider a bipolar approach to define database queries expressing user preferences (some corresponding to *constraints*, the others to *wishes*). An algebraic framework for the definition of flexible queries using bipolar conditions of type *and if possible* and *or else* is proposed. We define different ways of performing qualitative thresholding on the results of such queries, through a variety of operators which extend the classical fuzzy notion of α-cut, and study their properties.

- *Graph databases.* Graph databases have aroused a large interest in the last years thanks to their large scope of potential applications (e.g. social networks, biomedical networks, data stemming from the web). In a similar way as what has already been proposed in relational databases, defining a language allowing a flexible querying of graph databases may greatly improve usability of data. In [31], we focus on the notion of fuzzy graph database and describe a fuzzy query language that makes it possible to handle such database, which may be fuzzy or not, in a flexible way. This language, called FUDGE, can be used to express preference queries on fuzzy graph databases. The preferences concern i) the content of the vertices of the graph and ii) the structure of the graph. The FUDGE language is implemented in a system, called SUGAR, that is also described in [31].

- *Bipolar fuzzy queries and Qualitative Choice Logic.* 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. 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. On the other hand, Qualitative Choice Logic (QCL) is devoted to a logic expressing preferences for Boolean alternatives. In [3], we lay down the first stones to extend QCL to fuzzy alternatives. In particular, some relationships between QCL and the bipolar expression of preferences queries are emphasized. A new type of bipolar conditions is defined in the Boolean context to express QCL statements. This new type of bipolar conditions is then made gradual and it is shown that this extension can be the basis to the definition of a fuzzy QCL model. In [3], bipolar queries are extended in order to express complex, stratified, wish conditions.

6.2.2 Cooperative answering

Participants: Grégory Smits, Olivier Pivert, Aurélien Moreau, 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.

- *Diagnosing and repairing fuzzy queries that fail.* Telling the user that there is no result for his/her query is not informative and corresponds to the kind of situation cooperative systems try to avoid. Cooperative systems should rather explain the reason(s) of the failure, materialized by Minimal Failing Subqueries (MFS), and build alternative succeeding queries, called maXimal Succeeding Subqueries (XSS), that are as close as possible to the original query. In [5, 36, 34], we consider the issue of failing fuzzy queries and propose an efficient unified approach to the computation of gradual MFSs and XSSs that relies on a fuzzy cardinality-based summary of a part of the database.

- *Explanation of a cluster-based data structure.* On the one hand, clustering methods are of a particular interest to automatically identify the inner structure of a data set. On the other hand, fuzzy partitions are particularly suitable to define a subjective and domain-dependent vocabulary that may then be used to personalize an information system. To help data journalists and communication managers extract knowledge from open data sets, we propose in [35, 32] to generate personalized linguistic and graphical explanations of a cluster-based data structure.

- *Characterization of database query answers.* In [28], we describe an approach providing end-users with more insight to better understand the results of their queries. Using a clustering algorithm, the idea is to form subgroups of answers sharing some properties and to discover explanations for each subgroup. The originality of this work is that the data considered for characterizing each cluster of answers is not limited to the attributes used in the query. The objective is to enable the user to comprehend the structure of the results of his queries, using linguistic labels taken from his own vocabulary.

6.3 Distributed data management

Participants: François Goasdoué, Hélène Jaudoin, Olivier Pivert, Grégory Smits.

- *Enhanced keyword-based search.* In [33], we present a keyword-based constrained query language aimed to improve the expressivity and accuracy of database query interfaces. The idea is to let users explicitly express the intent of their search using keywords linked by meaningful grammatical connectives. Individually, keywords and connectives correspond to textual descriptions attached to components of the database graph schema, and as a whole, a so-called connected keywords query corresponds to a textual description of an SQL query. The translation process of such a query into SQL is mainly composed of two steps: first, the syntactic structure of the keyword query is analyzed to exhibit projection and selection statements using predefined graph patterns, then non explicit joins are deduced to obtain a complete translation of the keyword query as a meaningful connected subgraph. Experimentations show the relevance of the approach in terms of expressivity and efficiency.

- *Social messages classification.* Social messages classification is a research domain that has attracted the attention of many researchers in these last years. In [26], we are mainly interested in the classification of social messages based on their spreading on online social

networks (OSN). We propose a new distance metric based on the Dynamic Time Warping distance and we use it with the probabilistic and the evidential k Nearest Neighbors (k-NN) classifiers to classify propagation networks of messages. The propagation network is a directed acyclic graph that is used to record propagation traces of the message, the traversed links and their types. We tested the proposed metric with the chosen k-NN classifiers on real world propagation traces that were collected from Twitter and we obtained good classification accuracies.

- *Efficient query answering techniques for Semantic Web data. Reformulation-based query answering* is a query processing technique aiming at answering queries under constraints. It consists of reformulating the query based on the constraints, so that evaluating the reformulated query directly against the data (i.e., without considering any more the constraints) produces the correct answer set.

  In [14], we consider optimizing reformulation-based query answering in the setting of *ontology-based data access*, where SPARQL conjunctive queries are posed against RDF facts on which constraints expressed by an RDF Schema hold. The literature provides query reformulation algorithms for many fragments of RDF. However, reformulated queries may be complex, thus may not be efficiently processed by a query engine; well established query engines even fail processing them in some cases. Our contribution is (i) to generalize prior query reformulation languages, leading to investigating a *space of reformulated queries* we call JUCQs (joins of unions of conjunctive queries), instead of a single reformulation; and (ii) an effective and efficient *cost-based algorithm* for selecting from this space, the reformulated query with the lowest estimated cost. Our experiments show that our technique enables reformulation-based query answering where the state-of-the-art approaches are simply unfeasible, while it may decrease its cost by orders of magnitude in other cases. In [15, 16], we demonstrate this technique.

  In [12, 13], we extend the above results by considering the richer setting of W3C's OWL2 QL, i.e., the lightweight Description Logic constraints of DL-lite$_R$.

- *Parallel RDF data management.* As increasing volumes of RDF data are being produced and analyzed, many massively distributed architectures have been proposed for storing and querying this data. These architectures are characterized first, by their RDF partitioning and storage method, and second, by their approach for distributed query optimization, i.e., determining which operations to execute on each node in order to compute the query answers. In [25], we present CliqueSquare, a novel optimization approach for evaluating conjunctive RDF queries in a massively parallel environment. We focus on reducing query response time, and thus seek to build *flat* plans, where the number of joins encountered on a root-to-leaf path in the plan is minimized. We present a family of optimization algorithms, relying on *n-ary (star) equality joins* to build flat plans, and compare their ability to find the flattest possibles. We have deployed our algorithms in a MapReduce-based RDF platform and demonstrate experimentally the interest of the flat plans built by our best algorithms in [23].

- *RDF analytics.* RDF is the leading data model for the Semantic Web, and dedicated

query languages such as SPARQL 1.1, featuring in particular aggregation, allow extracting information from RDF graphs. We devised a framework for analytical processing of RDF data, published in WWW'14, where analytical schemas and analytical queries (cubes) are fully re-designed for heterogeneous, semantic-rich RDF graphs. In [9], we consider the following optimization problem in the above analytical setting: how to reuse the materialized result of a given analytical query (cube) in order to compute the answer to another analytical query obtained through a typical OLAP operation. We provide view-based rewriting algorithms for these query transformations, and demonstrate experimentally their practical interest.

- *Query-oriented summarization of RDF graphs.* In [17], we study the problem of RDF graph summarization: given an input RDF graph G, find an RDF graph S_G which summarizes G as accurately as possible, while being (most frequently) orders of magnitude smaller than the original graph. Our approach is *query-oriented*, i.e., we use queries to compute the summary, and we aim it as a help for query formulation and optimization. We introduce two summaries: a *baseline* which is compact and simple and satisfies certain accuracy and representativeness properties, but may oversimplify the RDF graph, and a *refined* one which trades some of these properties for more accuracy in representing the structure. In [18, 19], we demonstrate our summarization techniques.

- *Querying inconsistent description logic knowledge bases.* Several inconsistency-tolerant semantics have been introduced for querying inconsistent description logic knowledge bases. In [10], we address the problem of explaining why a tuple is a (non-)answer to a query under such semantics. We define explanations for positive and negative answers under the brave, AR and IAR semantics. We then study the computational properties of explanations in the lightweight description logic DL-lite$_R$. For each type of explanation, we analyze the data complexity of recognizing (preferred) explanations and deciding if a given assertion is relevant or necessary. We establish tight connections between intractable explanation problems and variants of propositional satisfiability (SAT), enabling us to generate explanations by exploiting solvers for Boolean satisfaction and optimization problems. Finally, we empirically study the efficiency of our explanation framework using the well-established LUBM benchmark.

6.4 Analogical proportions in databases

Participants: Hélène Jaudoin, Olivier Pivert, François Goasdoué, William Correa Beltran, Sara El Hassad.

- *Discovery of analogical proportions in relational databases.* In [22, 21], we present an approach aimed at mining a new type of pattern in data, namely analogical proportions. An analogical proportion expresses the equality of the relationships between the attributes of two pairs of structured objects. This notion is investigated in the database context for the discovery of different forms of "parallels" between tuples. First, we give a formal definition of the analogical proportion in the setting of relational databases. Then we

focus on the problem of mining analogical proportions. We show that it is possible to use a clustering approach for building equivalence classes made of pairs of tuples that are bound by the same relationship of analogical proportion.

- *Analogical database queries.* In [20], we introduce a new type of database query based on the concept of analogical proportion. The general idea is to retrieve the tuples that participate in a relation of the form "a is to b as c is to d". We provide a typology of analogical queries in a relational database context, devise different processing strategies and assess them experimentally.

- *Discovery of analogical proportions in RDF databases.* In [24], we propose a definition of both analogy and analogical proportions inspired by Gentner's works in cognitive sciences. We then enumerate the related decision and computation problems that arise in a database context. We show that the corresponding algorithms rely on the search of the smallest common generalization of two queries.

6.5 Data quality

Participants: Virginie Thion.

- *Evaluation and Improvement of a Transition Business Process.* A transition in an outsourced IT project is devoted to the transfer of the project from an outgoing project team to an incoming one. It is a complex, risky and challenging building block of importance, identified as being a critical factor in the success of an outsourced project. Ensuring the good quality of a transition is then fundamental. In [7], we present our experience on quality evaluation and improvement of a real transition in a public institution.

7 Other Grants and Activities

7.1 National actions

François Goasdoué is involved in the following projects:

- Datalyse (Investissements d'Avenir, *Big Data / Cloud computing*, 2013–2016). This project deals with Big Data management in a cloud architecture. The consortium is made of industrial partners (Eolas – Business & Decision and Les Mousqueraires), academic partners (Inria, LIFL of Univ. Lille, LIG of Univ. Grenoble, LIRMM of Univ. Montpellier), as well as the city of Grenoble as an open data provider.

- ANR JCJC Pagoda (2013–2017). PAGODA (Practical algorithms for ontology-based data access) is a basic research project whose objective is to improve the efficiency and robustness of ontology-based data access by developing scalable algorithms for query answering in the presence of ontologies as well as pragmatic approaches to handling inconsistent data. Partners are from LIG of Univ. Grenoble, LIRMM of Univ. Montpellier, and LRI of Univ. Paris-Sud.

- ANR ContentCheck, whose other partners are INRIA Saclay, LIMSI (Orsay), LIRIS (Lyon) and the team in charge of the blog "Les Décodeurs" associated with the newspaper Le Monde (http://www.lemonde.fr/les-decodeurs/). This project has been accepted in August 2015 and is to start in the last quarter of 2015.

François Goasdoué, Hélène Jaudoin, Olivier Pivert, Grégory Smits, and Virginie Thion are involved in the DGA project ODIN (Open Data INtelligence) which started in November 2014. The other partners involved are Semsoft and INRIA Saclay. The ODIN project aims to propose a data management and business intelligence solution for big data, i.e., large-scale heterogeneous and imperfect data distributed over several sources. For doing so, we intend to conceive a data processing and multidimensional analysis chain suitable for RDF data, taking into account the data quality aspect.

Grégory Smits and Olivier Pivert are involved in the project 360 Predict (Projet PME du pôle Images et Réseaux), which aims at developing a web tool for predictive scoring. The other partners are two start-ups: PredicSis (Lannion) and Semsoft (Rennes).

7.2 International actions

- Grégory Smits gave a Master's course about Fuzzy Preferences Queries at the Hanoi University of Science and Technology (HUST) in January 2015.

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).

8.2 Scientific activities

8.2.1 Highlights of the year

- François Goasdoué, along with Ioana Manolescu and Alexandra Roatis, gave a tutorial about "RDF Data Management: Reasoning on Web Data" at the 31st IEEE International Conference on Data Engineering (ICDE 2015).

- François Goasdoué, along with Marie-Christine Rousset, gave an invited talk entitled "Web sémantique : beaucoup de données, quelques connaissances et un peu de raisonnement" at the Journées Bases de Données Avancées (BDA 2015).

- Olivier Pivert served as PC Co-Chair of the 22nd International Symposium on Methodologies for Intelligent Systems (ISMIS'15) [2].

- Olivier Pivert served as PC Co-Chair of the 11th Conference on Flexible Query Answering Systems (FQAS 2015) [1].
- Start of the DGA project ODIN.
- Acceptance of the ANR project ContentCheck.

8.2.2 Program committees

François Goasdoué served as a member of the following program committee:

- 29th AAAI Conference on Artificial Intelligence (AAAI), Austin, USA, January 25-30, 2015;
- 12th European Semantic Web Conference (ESWC), Portoroz, Slovenia, May 31-June 4, 2015;
- 27th International Conference on Tools with Artificial Intelligence (ICTAI), Vietri sul Mare, Italia, November 9-11, 2015;
- 24th International Joint Conference on Artificial Intelligence (IJCAI), Buenos Aires, Argentina, July 25-31, 2015;
- 18th International Workshop on the Web and Databases (WebDB'15) at SIGMOD 2015, Melbourne, Australia, May 31, 2015.

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

- 30th ACM Symposium on Applied Computing (SAC 2015), Salamanca, Spain, April 13-17, 2015;
- IEEE International Conference on Fuzzy Systems (Fuzz-IEEE 2015), Istanbul, Turkey, August 2-5, 2015;
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2015), Poitiers, France, November 5-6, 2015.

O. Pivert served as PC Co-Chair of the following program committees:

- 22nd International Symposium on Methodologies for Intelligent Systems (ISMIS 2015), Lyon, France, October 21-23, 2015;
- 11th Conference on Flexible Query Answering Systems (FQAS 2015), Cracow, Poland, October 26-28, 2015;

and as a member of the following program committees:

- 30th ACM Symposium on Applied Computing (SAC 2015), Salamanca, Spain, April 13-17, 2015;

- IEEE International Conference on Fuzzy Systems (Fuzz-IEEE 2015), Istanbul, Turkey, August 2-5, 2015 – Area Chair of the Fuzzy Database track;
- 16th International Conference on Web Information Systems Engineering (WISE 2015), Miami, Florida, USA, October 18-20, 2015;
- World Congress of the International Fuzzy Systems Association / Conference of the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT 2015), Gijón, Spain, June 30 – July 3, 2015;
- 19th East-European Conference on Advances in Databases and Information Systems (ADBIS 2015), Poitiers, France, September 8-11, 2015;
- 26th International Conference on Database and Expert Systems Applications (DEXA 2015), Valencia, Spain, September 1-4, 2015;
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2015), Poitiers, France, November 5-6, 2015.

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

- Journées Bases de Données Avancées, Porquerolles, France, September 29 – October 2, 2015;
- Conférence en Recherche d'Information et Applications (CORIA 2015), Paris, France, March 18-20, 2015;
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2015), Poitiers, France, November 5-6, 2015.

G. Smits served as a member of the following program committees:

- 22th International Symposium on Methodologies for Intelligent Systems (ISMIS 2015), Lyon, France, October 21-23, 2015;
- IEEE International Conference on Fuzzy Systems (Fuzz-IEEE 2015), Istanbul, Turkey, August 2-5, 2015;
- World Congress of the International Fuzzy Systems Association / Conference of the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT 2015), Gijón, Spain, June 30 – July 3, 2015;
- 17eRencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2015), Caen, France, 22-25 juin 2015.

8.2.3 Editorial boards

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 steering committee of the French-speaking conference "Rencontres Francophones sur la Logique Floue et ses Applications" (LFA).

8.2.5 International advisory boards

O. Pivert is as a member of the international advisory board of the International Conference on Flexible Query-Answering Systems (FQAS).

8.2.6 Invited talks

- François Goasdoué gave an invited talk about "Knowledge base management: Principles and scalability" at the L3I Laboratory (La Rochelle) seminary on March 12, 2015.

9 Bibliography

Major publications by the team in recent years

[1] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Querying Inconsistent Description Logic Knowledge Bases under Preferred Repair Semantics", *in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, p. 996–1002, 2014.

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

[3] D. Colazzo, F. Goasdoué, I. Manolescu, A. Roatis, "RDF analytics: lenses over semantic graphs", *in: 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, p. 467–478, 2014.

[4] F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, I. Manolescu, S. Zampetakis, "Growing triples on trees: an XML-RDF hybrid model for annotated documents", *VLDB J. 22*, 5, 2013, p. 589–613.

[5] 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.

[6] 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.

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

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

Books and Monographs

[1] T. Andreasen, H. Christiansen, J. Kacprzyk, H. Larsen, G. Pasi, O. Pivert, G. De Tré, M. Vila, A. Yazici, S. Zadrozny (editors), *Flexible Query Answering Systems 2015, Proceedings of the 11th International Conference FQAS 2015, Cracow, Poland, October 26-28, 2015*, *Advances in Intelligent Systems and Computing*, *400*, Heidelberg, Germany, Springer, 2015.

[2] F. Esposito, O. Pivert, M.-S. Hacid, Z. Ras, S. Ferilli (editors), *Foundations of Intelligent Systems, 22nd International Symposium, ISMIS 2015, Lyon, France, October 21-23, 2015. Proceedings*, *Lecture Notes on Artificial Intelligence Intelligence*, *9384*, Heidelberg, Germany, Springer, 2015.

Articles in referred journals and book chapters

[3] L. Liétard, A. Hadjali, D. Rocacher, "Une définition pour un modèle QCL graduel", *Revue d'Intelligence Artificielle 29*, 5, 2015, p. 543–567.

[4] O. Pivert, H. Prade, "A Certainty-Based Model for Uncertain Databases", *IEEE Transactions on Fuzzy Systems 23*, 4, 2015, p. 1181–1196.

[5] O. Pivert, G. Smits, "How to Efficiently Diagnose and Repair Fuzzy Database Queries that Fail", *in: Fifty Years of Fuzzy Logic and its Applications*, D. Tamir, N. Rishe, and A. Kandel (editors), *Studies in Fuzziness and Soft Computing*, *326*, Springer, Heidelberg, Germany, 2015, p. 499–517.

[6] G. Smits, O. Pivert, T. Girault, "Interrogation floue de bases de données relationnelles : de la théorie à la pratique", *Revue d'Intelligence Artificielle 29*, 5, 2015, p. 569–593.

[7] V. Thion, M. Grim-Yefsah, C. Rosenthal-Sabroux, S. Si-Said Cherfi, "Evaluation and Improvement of a Transition Business Process: A Case Study guided by a Semantic Quality-based Approach", *Information Systems Management*, 2015.

Publications in Conferences and Workshops

[8] J. Akaichi, L. Liétard, D. Rocacher, O. Slama, "On the Qualitative Calibration of Bipolar Queries", *in: Proc. of the 22nd International Symposium on Methodologies for Intelligent Systems (ISMIS'15), LNAI vol. 9384*, p. 88–97, Lyon, France, 2015.

[9] E. Akbari Azirani, F. Goasdoué, I. Manolescu, A. Roatis, "Efficient OLAP Operations for RDF Analytics", *in: Proc. of the ICDE Workshop on Data Engineering meets the Semantic Web (DESWeb'15)*, 2015.

[10] M. Bienvenu, C. Bourgaux, F. Goasdoué, "Explaining Query Answers under Inconsistency-Tolerant Semantics over Description Logic Knowledge Bases", *in: Proc. of the International Workshop on Description Logics (DL'15), extended abstract*, 2015.

[11] F. Bousnina, M. Bach Tobji, M. Chebbah, L. Liétard, B. Ben Yaghlane, "A New Formalism for Evidential Databases", *in: Proc. of the 22nd International Symposium on Methodologies for Intelligent Systems (ISMIS'15), LNAI vol. 9384*, p. 31–40, Lyon, France, 2015.

[12] D. Bursztyn, F. Goasdoué, I. Manolescu, "Efficient Query Answering in DL-Lite through FOL Reformulation", *in: Proc. of the International Workshop on Description Logics (DL'15), extended abstract*, 2015.

[13] D. Bursztyn, F. Goasdoué, I. Manolescu, "Optimizing FOL reducible query answering", *in: Actes des 31es Journées Bases de Données Avancées (BDA'15)*, Île de Porquerolles, France, 2015.

[14] D. Bursztyn, F. Goasdoué, I. Manolescu, "Optimizing Reformulation-based Query Answering in RDF", *in: Proc. of the 18th International Conference on Extending Database Technology (EDBT'15)*, 2015.

[15] D. Bursztyn, F. Goasdoué, I. Manolescu, "Reformulation-Based Query Answering in RDF: Alternatives and Performance", *in: Proc. of the 41st Conference on Very Large Databases (VLDB'15), demo paper*, 2015.

[16] D. Bursztyn, F. Goasdoué, I. Manolescu, "Understanding and Improving Reformulation-Based Query Answering Performance in RDF", *in: Actes des 31es Journées Bases de Données Avancées (BDA'15), session démonstration*, Île de Porquerolles, France, 2015.

[17] S. Cebiric, F. Goasdoué, I. Manolescu, "Query-Oriented Summarization of RDF Graphs", *in: Proc. of the British International Conference on Databases (BICOD'15), work-in-progess paper*, 2015.

[18] S. Cebiric, F. Goasdoué, I. Manolescu, "Query-Oriented Summarization of RDF Graphs", *in: Proc. of the 41st Conference on Very Large Databases (VLDB'15), demo paper*, 2015.

[19] S. Cebiric, F. Goasdoué, I. Manolescu, "Query-Oriented Summarization of RDF Graphs", *in: Actes des 31es Journées Bases de Données Avancées (BDA'15), session démonstration*, Île de Porquerolles, France, 2015.

[20] W. Correa beltran, H. Jaudoin, O. Pivert, "Analogical Database Queries", *in: Proc. of the 11th International Conference on Flexible Query Answering Systems (FQAS'15), Advances in Intelligent Systems and Computing, 400*, Springer, p. 201–213, Cracow, Poland, 2015.

[21] W. Correa beltran, H. Jaudoin, O. Pivert, "A Clustering-Based Approach to the Mining of Analogical Proportions", *in: Proc. of the 27th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'15)*, Vietri sul Mare, Italy, 2015.

[22] W. Correa Beltran, H. Jaudoin, O. Pivert, "Découverte de proportions analogiques dans les bases de données : Une première approche", *in: Actes de la 15e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'15)*, Luxembourg, 2015.

[23] B. Djahandideh, F. Goasdoué, Z. Kaoudi, I. Manolescu, J. Quiané-Ruiz, S. Zampetakis, "CliqueSquare in action: Optimizing RDF Queries for Parallel Execution", *in: Proc. of the 31st IEEE International Conference on Data Engineering (ICDE'15), demo session*, 2015.

[24] S. El Hassad, "Interrogation par analogie dans les bases de données", *in: Actes des 31es Journées Bases de Données Avancées (BDA'15), session doctorants*, Île de Porquerolles, France, 2015.

[25] F. Goasdoué, Z. Kaoudi, I. Manolescu, J. Quiané-Ruiz, S. Zampetakis, "CliqueSquare: Flat Plans for Massively Parallel RDF Queries", *in : Proc. of the 31st IEEE International Conference on Data Engineering (ICDE'15)*, 2015.

[26] S. Jendoubi, A. Martin, L. Lietard, B. Ben Yaghlane, H. Ben Hadji, "Dynamic Time Warping Distance for Message Propagation Classification in Twitter", *in : Proc. of the 13th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'15)*, p. 419–428, Compiègne, France, 2015.

[27] L. Liétard, "Indices de confiance pour l'interrogation flexible de données imprécises", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'15)*, Cépaduès, p. 35–42, Poitiers, France, 2015.

[28] A. Moreau, O. Pivert, G. Smits, "A Clustering-Based Approach to the Explanation of Database Query Answers", *in : Proc. of the 11th International Conference on Flexible Query Answering Systems (FQAS'15)*, *Advances in Intelligent Systems and Computing*, *400*, Springer, p. 307–319, Cracow, Poland, 2015.

[29] O. Pivert, H. Prade, "A Certainty-based Approach to the Cautious Handling of Suspect Values", *in : Proc. of the 11th International Conference on Flexible Query Answering Systems (FQAS'15)*, *Advances in Intelligent Systems and Computing*, *400*, Springer, p. 73–85, Cracow, Poland, 2015.

[30] O. Pivert, H. Prade, "Database Querying in the Presence of Suspect Values", *in : New Trends in Databases and Information Systems – ADBIS 2015 Short Papers and Workshops, BigDap, DCSA, GID, MEBIS, OAIS, SW4CH, WISARD, Poitiers, France, September 8-11, 2015. Proceedings*, T. Morzy, P. Valduriez, L. Bellatreche (editors), *Communications in Computer and Information Science*, *539*, Springer, p. 44–51, 2015.

[31] O. Pivert, G. Smits, V. Thion, "Expression and Efficient Processing of Fuzzy Queries in a Graph Database Context", *in : Proc. of the 24th IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'15)*, Istanbul, Turkey, 2015.

[32] O. Pivert, G. Smits, "Réparation efficace de requêtes floues", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'15)*, Cépaduès, p. 27–34, Poitiers, France, 2015.

[33] G. Smits, O. Pivert, V. Thion, "Connected Keywords", *in : Proc. of the 9th IEEE International Conference on Research Challenges in Information Science (RCIS'15)*, p. 112–120, Athens, Greece, 2015.

[34] G. Smits, O. Pivert, "Explications linguistiques et graphiques de groupes de données", *in : Actes des Rencontres Francophones sur la Logique Floue et ses Applications (LFA'15)*, Cépaduès, p. 19–26, Poitiers, France, 2015.

[35] G. Smits, O. Pivert, "Linguistic and Graphical Explanation of a Cluster-Based Data Structure", *in : Scalable Uncertainty Management – 9th International Conference, SUM 2015, Québec City, QC, Canada, September 16-18, 2015. Proceedings*, C. Beierle, A. Dekhtyar (editors), *Lecture Notes in Computer Science*, *9310*, Springer, p. 186–200, 2015.

[36] G. Smits, O. Pivert, "Une approche coopérative d'aide à la réparation de requêtes floues", *in : Actes des 31es Journées Bases de Données Avancées (BDA'15)*, Île de Porquerolles, France, 2015.