Towards an Extended SQLf: Bipolar Query Language with Preferences

L. Ludovic, R. Daniel, and S-E Tbahriti

Abstract—Database management systems that integrate user preferences promise better solution for personalization, greater flexibility and higher quality of query responses. This paper presents a tentative work that studies and investigates approaches to express user preferences in queries. We sketch an extend capabilities of SQLf language that uses the fuzzy set theory in order to define the user preferences. For that, two essential points are considered: the first concerns the expression of user preferences in SQLf by so-called fuzzy commensurable predicates set. The second concerns the bipolar way in which these user preferences are expressed on mandatory and/or optional preferences.

Keywords—Flexible query language, relational database, user preference.

I. INTRODUCTION

The amount of information managed by the database management systems (DBMS) becomes increasingly important. As a consequence the interrogation of database should be more and more efficient. This performance can be measured in terms of query time response or a delivered information quality. In particular, taking into account user preferences in query is a key element of relevance. The problem of expressing and managing user preferences has received more and more attention in the last few years [1, 5, 21, 23]. It was shown [7, 13] that the fuzzy set theory provides efficient tools to incorporate user preferences in queries. SQLf language is an example that illustrates this idea. More precisely, “vague conditions” with preferences allow to describe the personalized needs of each user. This paper aims to introduce an extension of SQLf Language to integrate optional user preferences in a bipolar form (anks to a new “THEN” clause) and to present a brief overview of the recent approaches in order to express user preferences in queries.

In relational mode of database, preferences are mainly employed to filter and personalize the information sought by users. Two general approaches are distinguished in the literature to express preferences. The implicit approach in which, each value of attribute is associated with a score. The value is preferred to another if it has a better score. The explicit approach in which the user directly expresses his preferences on the various attribute values. That means the preferences are defined by comparing the attribute values. In addition, these preferences can be seen in a bipolar way, i.e., mandatory preferences (viewed as constraint) and optional preferences –viewed as wishes). In this context, the answers of a query must satisfy absolutely all mandatory preferences and satisfy as possible the optional preferences.

The purpose of this paper is to specify on the one hand, the features of the implicit and explicit approaches like that of bipolarity. The emphasis is put on the consequences of the obtained results, when these techniques are applied, in particular if a total or partial order is obtained. On the other hand, we sketch the extended SQLf language to integrate optional preferences we are currently working on.

In the following Section we study the features of the two approaches to express preferences and a special importance is put on the commensurability assumption and the impact it has on the query responses. In Section 3 we describe bipolar concept of preferences. Section 4 is dedicated to a presentation of the main extension of SQLf language. In Section 5 the main interrogation systems with preferences are investigated and positioned with two axes (preference expression, and the bipolarity). Conclusion summarizes the principal contribution of the paper.

II. PREFERENCES AND ORDER RELATIONS

In the context of relational database, elementary preferences are defined on attribute values then composed to define more sophisticated preferences. Each attribute Aj has associated a domain values Dj. A tuple tj associates to each Aj a value taken from its domain. For a given attribute, two general ways are used to express user preferences (the implicit and explicit way).

In the implicit approach a scoring function is associated to each attributes [1]. An attribute value is preferred to another if it has obtained a better score. As an example, a score can be a distance from an optimal value. The element having the least distance is preferred.

Example 1. Consider the relation Car (cf. Table I). Numerical scores are assigned to the values of “Make” attribute as follows: (BMW = 3, Audi = 2, VW = 1). More the score is high more the make is preferred. Thus, a preference for the “famous makes” car is expressed. ♦
The fuzzy set [24] constitutes a more specific support in which this idea can be instantiated. The following fuzzy set expresses thus the preference for the "famous make": Famous-makes = \{1/BMW, 0.7/Audi, 0.3/VW\}. Each element of this set is associated with a degree. More the degree is high, more the element is preferred. An important effect of the use of score functions if that they induce a total order on the values of each attribute.

In the explicit expression approach, the preferences are explicitly defined on the attribute values (e.g., someone prefers "green" car to "red" car), which is quite natural to the user’s viewpoint. In this case, these attribute values can be totally or partially ordered.

Generally, the aim is to classify the tuples from several elementary preferences. Two directions can be followed; the first point is based on the comparison of the scores, the second based on explicit comparison inter-tuples. We present in the two following sub-sections these directions for preferences expression.

### A. Elementary Preferences Defined by a Scoring Function

In the case where elementary preferences expressed by scoring function [17], each tuple \( t_i \) is associated with a scores vector \( (s_j)_{i=1}^{n} \). Please submit your manuscript electronically for review as e-mail attachments. Each \( s_j \) corresponds to the evaluation of a preference on attribute \( A_j \) of tuple \( t_i \). Two assumptions are considered according to the commensurability of these scores.

If the commensurability assumption holds, all scores of a vector are based on the same scale of satisfaction in order to be compared. Thus, scores can be combined by means of an aggregation function \( f \) (average, weighted average, min, etc.) to give a global evaluation to the vector. Consequently, a total order relation is established between the score vectors. In this case, \( t_i \) is preferable to \( t_j \) if and only if \( f(s_1, \ldots, s_n) \geq f(s'_1, \ldots, s'_n) \). Generally \( f \) is an “ad-hoc” numerical function which can be defined by a user. Obviously \( f \) must take into account the relative senses of the scores which it combines so that the aggregation mechanism is meaningful. It means that \( f \) should make, in somehow, these scores commensurable.

When fuzzy predicates are used, the scores are satisfaction degree in \([0, 1]\) and a logical meaning is allotted to them [4]. The functions to aggregate them use logical extended operators [4, 18] (triangular norm, co-norm, fuzzy implication, etc.).

<table>
<thead>
<tr>
<th>Tuple-id</th>
<th>Make</th>
<th>Price</th>
<th>Cons.</th>
<th>Hp.</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_1 )</td>
<td>BMW</td>
<td>30000</td>
<td>9/100</td>
<td>90</td>
<td>Green</td>
</tr>
<tr>
<td>( t_2 )</td>
<td>Audi</td>
<td>15100</td>
<td>8/100</td>
<td>91</td>
<td>Green</td>
</tr>
<tr>
<td>( t_3 )</td>
<td>VW</td>
<td>15000</td>
<td>7/100</td>
<td>91</td>
<td>Black</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tuple-id</th>
<th>Make</th>
<th>Price</th>
<th>Degree(_{max})</th>
<th>Degree(_{prey})</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_1 )</td>
<td>BMW</td>
<td>30000</td>
<td>0.9</td>
<td>0.2</td>
</tr>
<tr>
<td>( t_2 )</td>
<td>Audi</td>
<td>15100</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>( t_3 )</td>
<td>VW</td>
<td>15000</td>
<td>0.5</td>
<td>0.9</td>
</tr>
</tbody>
</table>

### Example 2

Consider the following query which has to find the “famous make” and “not expensive” cars. Each degree represents a satisfaction with regard to respective conditions, “famous make” and “not expensive” (cf. Table II). These degrees are aggregated according to a triangular norm “min” (expressing a conjunction), which gives the final degrees: 0.2, 0.7 and 0.5. The tuple are classified as follows: \( t_2 \), \( t_1 \) then \( t_3 \), which mean that \( t_2 \) is preferable to \( t_1 \) and the last one is preferable to \( t_1 \).

It is also possible within the fuzzy set field to apply other mechanisms like the Leximin or Discrimin operators [14].

The Leximin operator is based on a permutation of the scores of each vector in order ti be able to compare them. It is defined as follows: if \( s^* \) and \( s^\prime* \) are two permutations of \( s \) respectively so that \( s^*_1 \leq \cdots \leq s^*_n \) and \( s^\prime*_1 \leq \cdots \leq s^\prime*_n \) then \( i < \text{Leximin } s \iff \exists k \leq n, \forall i < k, t_i = s_i \) and \( t_i^\prime > s^\prime*_i \).

The Discrimin ordering between two score vectors. It is defined as follows: if \( D(u, v) = |i, u_i \neq v_i| \) is the set of index for which the corresponding values in the scores vectors \( u \) and \( v \) are different, \( u > \text{Discrimin } v \iff \min_{i \in D(u, v)} u_i > \min_{i \in D(u, v)} v_i \).

In the no commensurability assumption the scores allotted to the various attributes of a tuple are not comparable. Consequently these scores can not be aggregated and only a partial order can be defined on the tuples. In particular, the Pareto order can be used.

### Pareto Order

We want to compare two tuples, \( u \) and \( v \) such as \( v = (v_1, \ldots, v_n) \), \( u = (u_1, \ldots, u_n) \). It is defined as follows [14, 15]: \( v > \text{Pareto } u \iff (\forall i, v_i \geq u_i, \exists j, v_j > u_j) \).

### Example 3

Let us consider the following query which has to find the cars of price around 15,000, with a consumption (Cons.) around 7/100 and a horse power (Hp.) around 90. The preferences on these three criteria are qualified by distances (cf. Table III). Since the commensurability assumption is not considered here, the scores on the values of various attributes can not be compared and combined. Consequently, the score vector of \( v(t_1) = (15000, 0, 0) \) can not be compared with the other vectors of \( v(t_2) = (100, 1, 1) \) and \( v(t_3) = (0, 0, 1) \). By using the Pareto Order, the result of example 3 is given as follows: \( v(t_3) > \text{Pareto } v(t_2) \) but \( v(t_1) \) can not be compared neither with \( v(t_2) \) nor with \( v(t_3) \). Thus, \( t_2 \) is thus preferred to \( t_3 \) and the \( t_1 \) can not be compared neither with \( t_3 \) nor with \( t_2 \).
B. Elementary Preferences Defined Explicitly

It has been observed at a glance that implicit approach to express preferences, has a limited expressive power, since they can not be used to model more complex patterns of preferences. For example, if a user wants to indicate his preferences over paint, it is easier to compare them one by one.

Preference over a relation database is expressed by a collection of pairs of tuples [12]. Each pair specifies the preference of one tuple over another one. In this case, a collection of pairs of tuples \[12\]. Each pair specifies the preferences. For example, if a user wants to indicate his they can not be used to model more complex patterns of express preferences, has a limited expressive power, since they can all be satisfied at all times. Constraints and wishes it has a "green" color. The constraint "not expensive" allows determining acceptable cars set (whose price is \(<15000\), then the condition on the color expresses a wish, which if it is satisfied supports the associated answer.

The fundamental property of the constraints and the wishes is that the set of desired values is a subset of acceptable values \((D \subseteq A)\). Indeed, it is incoherent to wish non-acceptable values (a constraint defined by “an European car” is incoherent with the wish: “a Japanese car”).

It is also possible to consider constraints and wishes are defined by fuzzy set. In this context, the condition of inclusion \(D \subseteq A\) is rewritten: For each tuple \(t_i\), \(\mu_A(t_i) \leq \mu_A(t_i)\), such as \(\mu_B(t)\) (respectively \(\mu_A(t_i)\)) is a satisfaction degree of \(t_i\) on \(D\) (respectively on \(A\)).

Constraints and wishes are different in nature (for example, a non-satisfied wish, does not reject a tuple, unlike a non-satisfied constraint). Consequently, the degrees expressing constraints and wishes are non commensurable and can not be combined in a logical expression. They must be treated independently. The constraint being imperative, it is possible to order the tuples by using a lexicographical order [2] on the constraints and the wishes. Thus, the wishes allow to differentiate between the tuples who are equal with respect to the constraints and a total order can be obtained on \(A\) and \(D\).

For the example 5, a lexicographical method classifies tuples satisfying the constraint “not expensive”, and favoured among of them, those, which satisfy the wish “green color”. So, \(t_i\) is classified before another \(t_j\) if \(\mu_A(t_i) \leq \mu_A(t_j)\) or \((\mu_A(t_i) = \mu_A(t_j) \land \mu_B(t_i) \geq \mu_B(t_j))\).

IV. TOWARDS AN EXTEND OF SQLF

This section is mainly concerned with the extension of language SQLF capabilities. First of all, we briefly set out the base structure of SQLF language, then we present some main extensions for this language along two axes:

1. In order to envisage a greater flexibility and delivered response quality, we define user preferences via the set of fuzzy predicates \(P= P_A \cup P_B\) where \(P_A\) expresses mandatory preferences and \(P_B\) expresses optional preferences.

2. Concerning the optional and mandatory preferences, the aim is that SQLF be able to distinguish these two types of preferences. These two axes are orthogonal and can be treated independently.

A. SQLF

The SQLF language extends the SQL language in order to allow the user to formulate queries on atomic conditions defined by fuzzy sets \([6, 7]\). Each attribute of a tuple is associated with a satisfaction degree \(\mu\) in \([0, 1]\). The semantic of degrees is the same, what implies that the criteria are commensurable.

A query in SQLF language has the following syntax:

```
SELECT [distinct][n|t|n,t] <attributes>
FROM <crisp relation>
WHERE <fuzzy condition>
```
Where fuzzy condition contains mandatory preferences expressed by a set of commensurable fuzzy predicates \( P_\alpha \). The parameters \( n \) and \( t \) of the select block limit the number of the answers by using a quantitative condition (the best answers) or a qualitative condition (data which satisfy the query according to a level higher than \( t \)).

### B. Extended SQL\(_f\)

As we have seen in Section 3, constraint and wish can not be combined in a same logical expression. In addition, the processing of constraint must be independent of that of the wish. To guarantee this effect in SQL\(_f\), we consider a bipolar query in SQL\(_f\) as \( A \text{ then } D \) where the satisfaction of the wishes in \( D \) is only used for ordering tuples. For this purpose, we extend SQL\(_f\) by introducing a new clause “THEN” to express optional preferences in query, as follows:

```
SELECT [distinct][n|t|n,t] <attributes>
FROM <crisp relations>
WHERE <fuzzy condition>
THEN <optional preferences>
```

The “THEN” clause may involve both Boolean and fuzzy predicates in PD to express optional preferences (wish) combined by several kinds of connectors. Two cases are considered to process optional preferences:

1. **PD** is a set of commensurable predicates of the same importance level. In this case, we can utilise the Leximin or Discrimin operator to keep, in each class (which corresponds to a certain value of degree relative with \( A \)), only the undominated tuples. These operators provide an order on a response set but they do not provide scores.

2. **PD** is a set of hierarchical commensurable predicates. We can use an operator of semantics “if possible then” level of importance.

**Example 6.** Let us consider again an instance of the relation car (cf. Table IV) and the following query: Find preferably, among the “not expensive” car, those which are “famous make”. The predicate “not expensive” allows to express mandatory preferences \( A \) defined by fuzzy set. While predicate “famous make” permit to describe an optional preference \( D \), which if it is satisfied, supports the associated results of query. In particular, for the tuples of the employee relation we have:

<table>
<thead>
<tr>
<th>not-expensive</th>
<th>{0.2/30000, 0.7/15100, 0.8/15000, 0.9/9000}</th>
</tr>
</thead>
<tbody>
<tr>
<td>famous make</td>
<td>{1/BMW, 0.8/Audi(0), 8/VW, 0.3/Fiat}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tuple-id</th>
<th>Make</th>
<th>Price</th>
<th>Degree_make</th>
<th>Degree_price</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_1 )</td>
<td>BMW</td>
<td>30000</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>( t_2 )</td>
<td>Audi</td>
<td>15100</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>( t_3 )</td>
<td>VW</td>
<td>15000</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>( t_4 )</td>
<td>Fiat</td>
<td>9000</td>
<td>0.3</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Each tuple \( t_i \) is associated with a vector which represents its situation with respect to the atomic conditions (see section 2.1). The result is evaluated on two steps:

1. First, we selected tuples satisfying the mandatory preferences \( A \) (“not expensive”).

2. Then we obtain: \( t_1(0.9) > t_2(0.8) > t_3(0.7) > t_4(0.2) \) which means that the tuple \( t_1 \) is preferable on \( A \) to \( t_2 \), and last one is preferable to \( t_4 \) and so on. We keep tuples having a degree \( \alpha_d \) higher than 0.5 (for example). Secondly, among the tuples satisfying the mandatory preferences \( A \) with respect \( \alpha_d \), we select only those which satisfy the optional preferences with a degree \( \alpha_d = 0.7 \) (for example). Thus, we obtain two classes in which all results are totally ordered, such as class 1 that contains the tuples satisfying \( A \) and \( D \) and a class 2 that contains the tuples satisfying only \( A \). The final result is as follows: Since \( t_1 \) does not satisfy optional condition “famous make”, thus it is not selected, class 1 = \{\( t_1 > t_2 \)\}, that means \( t_i \) is preferred to \( t_1 \). class 2 = \{\( t_i > t_2 > t_4 \)\}.

Beyond the extension of SQL\(_f\) optional preferences itself, an important question puts relate to performances of a SGBD accepting such queries with preferences, and thereafter an evaluation mechanism of these queries. To enable query processing and optimization, we present flexible query with preferences by means of a system of transformation algebraic rules. We will present in the next research papers the performance experiments.

In conclusion, the predicate selection of the clause where permits to select acceptable results that satisfying the mandatory preferences, while “then” clause allows to express optional preferences (wishes) and to order the selected tuples. With this way, extended SQL\(_f\) is be able to process user preferences within a bipolar framework. In this context, such preferences are taken into consideration through the expression of commensurable fuzzy predicates, modelled by fuzzy set of more or less satisfactory values and the selection of the results is totally ordered.

### V. QUERY LANGUAGES WITH PREFERENCES

User preferences can be embedded into database query languages in several different ways. In this section, we briefly present the principal propositions to integrate user preferences in queries. We situate these propositions with respect to implicit and/or explicit preferences and bipolarity.

#### A. Preference SQL

Queries in Preference SQL [19, 20] are mainly made of two parts:

- a WHERE clause aiming at selecting tuples (using Boolean conditions, also called conditions of type must),
- a PREFERRING clause to specify preferences (also called light conditions) in order to make a discrimination between tuples. The typical query block in Preference SQL is then:

```sql
SELECT [distinct][n|t|n,t] <attributes>
FROM <crisp relations>
WHERE <fuzzy condition>
PREFER <optional preferences>
```
Example 7. The query “find hotels from Paris with a price around 100 and a high category” is expressed in Preference SQL by:

```
SELECT *
FROM Hotel
WHERE City = Paris
PREFERRING around_100(Price)
and high(Category);
```

where the relation Hotel is given by Table V.

<table>
<thead>
<tr>
<th>Tuple-id</th>
<th>Price</th>
<th>Category</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>t₁</td>
<td>200</td>
<td>***</td>
<td>Paris</td>
</tr>
<tr>
<td>t₂</td>
<td>100</td>
<td>**</td>
<td>Paris</td>
</tr>
<tr>
<td>t₃</td>
<td>150</td>
<td>**</td>
<td>Paris</td>
</tr>
<tr>
<td>t₄</td>
<td>50</td>
<td>*</td>
<td>Lyon</td>
</tr>
</tbody>
</table>

The preference price around 100 is modeled by a distance (between the price and top value 100), while the preference high category is represented by a level (from 1 to 3, depending on the category). This query discards hotel $t₂$ (since not located in Paris) and evaluates the preferences (in the form of a vector (distance, level)) for hotels $t₁$, $t₂$, and $t₃$ respectively associated to (100, 1), (0, 2) and (50, 2). The Pareto ordering gives: $t₂ \succ t₃ \succ t₁$, while $t₂$ is not comparable with $t₁$ and $t₃$. As a consequence, the non-dominated tuples correspond to $t₂$ and hotel $t₁$ which are presented to the user.

Preference SQL follows the bipolar model, since the must predicates represent constraints while the light predicates represent wishes (a tuple which does not satisfy the must predicates is discarded and, if possible, the non-dominated tuples with respect to the light predicates are returned to the user), but it is limited to Boolean constraints and non-commensurable preferences since elementary preferences are distances or levels.

Example 8. Relation Persons (name, age, weight, height) gathers information about people. A top-k query on relation Persons is “find the 5 best persons according to the preference to be over-weighted and to be young.” The over-weight of a person described by a tuple $t_i$ can be calculated by the following function: $f(t_i) = t_i\text{-weight} - (t_i\text{-height} - 100)$, while a fuzzy set Young indicates the extent to which it is young. In this case, the ranking function $f$ must make commensurable the overweight and the score $\mu\text{Young}(t_i\text{-age})$, and it possible to define: $f(t_i) = (t_i\text{-weight} - (t_i\text{-height} - 100)) \cdot (t_i\text{-weight} + \mu\text{Young}(t_i\text{-age}))$, which means that the overweight is all the more important as it is associated to a light weight. Function $f$ is evaluated for each person and the 5 best ones are returned.

The approach advocated by top-k queries delivers a total order since elementary preferences are considered commensurable and aggregated in the ranking function. However, some elementary preferences may not be commensurable and the definition of function $f$ may leads, in this case, to a result which can be difficult to justify.

C. Preference Queries

This approach provides an algebraic framework to formulate query with preferences and an algebraic operator “winnow” [10, 11]. This one picks the set of tuples which are not dominated, according to a given preference relation $\succ$. It is defined by the following formula: if $R$ is a relation schema and $\succ$ preference relation over $R$, then winnow operator is written as $w_\succ(r)$ and for every instance $r$ of $R$, $w_\succ(r) = \{ t_i \in r \mid \exists t_j \in r, t_j \succ t_i \}$. A special case of winnow is called Skyline [3] where preferences are predefined and limited to a set of operations.

The interest of winnow is to be justified when the preference relation delivers a partial order. It allows to select elements undominated and can not be compared between them. However, it was shown [25] that a bipolar query where the constraints $A$ and wishes $D$ are Boolean conditions can be defined by means of the winnow operator. Indeed, $w_\sigma(T)$ such as $T$ is the tuples set and $R'$ is the preference relation defined by: $R’(t_i, t_j)$, $P(t_i) \cap \neg P(t_j)$, where $\sigma$ is a classical selection and $P(.)$ is predicates corresponding to the preferred conditions.
VI. CONCLUSION

In this we have studied and presented two families of approaches to express user preferences: the implicit and the explicit. In the implicit approach, an elementary preference is defined by a score (provided by a function). The aggregation of several scores is possible only when the commensurability assumption holds and leads to a total order of query answers. On the other hand if the preferences are non-commensurable a partial order is obtained since some tuples may be not comparable. In the explicit approach the preferences are specified by binary relation of preferences and in the majority of the cases, a partial order is obtained on the tuples. In addition, the preferences can be considered as constraints (mandatory preferences) and wishes (optional preferences).

In this paper, we have presented an extension of SQLf language in order to integrate optional preference according to bipolar form. The extended SQLf language we are currently working on, uses the fuzzy set theory in order to define the preferences and consider the commensurability assumption. This language provides a founded framework to combine mandatory and optional preferences. We have dealt also with the main interrogation systems which support preferences. The processing of such systems has been discussed and positioned with respect to two aspects (preferences expression and bipolarity). The approaches advocated by the systems Preference SQL and Preference Queries are based on a partial order, consequently, they release to the user only the undominated tuples. Preference SQL incorporates a concept of bipolarity in the Preferring clause. In top-k queries system, relatively little attention has been devoted to the design of appropriate scoring functions, a problem of critical importance since the quality and usefulness of the top-k answers for a query are highly dependent upon the underlying quality of the scoring technique.

Further studies can be made on the results provided by this paper. First, it will be of interest to define more sophisticated operators to determine the best answers in case of non commensurable preferences. We have presented a first step to an extended SQLf language with preferences.

REFERENCES