On Top-*k* Queries over Evidential Data

Fatma Ezzahra Bousnina[1], Mouna Chebbah[1,2], Mohamed Anis Bach Tobji[1,3], Allel Hadjali[4] and Boutheina Ben Yaghlane[1,5]

[1]*Université de Tunis, Institut Supérieur de Gestion, LARODEC, Bardo, Tunisia*

[2]*Université de Jendouba, Faculté des Sciences Juridiques, Economiques et de Gestion de Jendouba, Jendouba, Tunisia*

[3]*Univ. Manouba, ESEN, Manouba, Tunisia*

[4]*Université de Poitiers, Ecole Nationale Supérieure de Mécanique et de l'Aérotechnique, LIAS, Poitiers, France*

[5]*Université de Carthage, Institut des Hautes Etudes Commerciales, Carthage, Tunisia*

Keywords: Evidence Theory, Evidential Databases, Top-*k* Queries, Ranking Intervals, Score Function.

Abstract: Uncertain data are obvious in a lot of domains such as sensor networks, multimedia, social media, etc. Top-*k* queries provide ordered results according to a defined score. This kind of queries represents an important tool for exploring uncertain data. Most of works cope with certain data and with probabilistic top-*k* queries. However, at the best of our knowledge there is no work that exploits the Top-*k* semantics in the Evidence Theory context. In this paper, we introduce a new score function suitable for Evidential Data. Since the result of the score function is an interval, we adopt a comparison method for ranking intervals. Finally we extend the usual semantics/interpretations of top-*k* queries to the evidential scenario.

1 INTRODUCTION

Processing queries over uncertain data received an increasing importance with the emergence of several applications in domains like sensor networks (Considine et al., 2004; Silberstein et al., 2006), moving objects tracking (Cheng et al., 2004a) and data cleaning (Andritsos et al., 2006). Several types of queries deal with uncertain data like uncertain skyline query (Ding and Jin, 2012; Elmi et al., 2014; Elmi et al., 2015), probabilistic top-*k* query (Soliman et al., 2007), uncertain range query (Chung et al., 2009), uncertain threshold query (Cheng et al., 2004b), etc.

In this paper, our interest goes to *ranking queries*, also called *top-k queries*. A top-*k* query reports the *k* objects with the highest scores based on a defined *scoring function*. *Imperfect top-k* queries are different from top-*k* queries over certain data, they focus not only on the value of the scoring function, but also on the degree of objects' uncertainty.

Data uncertainty can be detected in three levels (Tao et al., 2007):

- *The table level uncertainty*, represented with a degree of imperfection about the presence or the absence of the table in the database.
- *The tuple level uncertainty*, represented with a degree of imperfection about the presence or the absence of that tuple.
- *The attribute level uncertainty*, represented by a degree of imperfection about individual attributes.

The tuple and attribute levels are the most studied in the literature. Lots of theories like the probability theory (Laplace, 1812), the fuzzy sets theory (Zadeh, 1965), the possibility theory (Zadeh, 1978) and the belief functions theory, also called the evidence theory (Dempster, 1967; Shafer, 1976) have been introduced in order to handle imperfection as uncertainty, imprecision and inconsistency.

At the best of our knowledge, there is no work that deals with top-*k* queries for *Evidential Data*. Similarly to the probabilistic top-*k* queries (Soliman et al., 2007), *Evidential Top-k queries* should return the *k* answers that respond to an evidential query with the highest scores based on a scoring function that takes into consideration the degrees of imperfection in the database.

Through this paper, we present a new type of uncertain top-*k* queries; the *Evidential Top-k Queries* that we apply over an evidential relation. We recall

Bousnina, F., Chebbah, M., Tobji, M., Hadjali, A. and Yaghlane, B.
On Top-*k* Queries over Evidential Data.
DOI: 10.5220/0006317701060113
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 1*, pages 106-113
ISBN: 978-989-758-247-9

that an evidential database is a database that stores perfect and imperfect data modeled with the theory of evidence (Dempster, 1967; Dempster, 1968). Each object in the evidential database is quantified with an interval of confidence called the *Confidence Level* and denoted *CL* (Bell et al., 1996; Bousnina et al., 2015; Lee, 1992a; Lee, 1992b).

For this purpose, we introduce a new scoring function for evidential data that returns an interval bounded by a belief and a plausibility. To rank the evidential scores, we rely on the method of (Wang et al., 2005). We present also a new imperfect top-k semantics specific to the evidential scenario.

Table 1 presents an example of an evidential table that stores some users' appreciations about books: b_1, b_2, b_3, b_4. This example is a relation with three attributes: The first one is *ID*, it represents the identifier of a specific reader. The second attribute is *BookRate* where the reader expresses its preferred books using the evidence theory[1]. The uncertainty here deals with the attribute level. The third attribute is *CL*, it stores the interval of confidence about the user, and thus about its given appreciations. Here we deal with uncertainty at the tuple level. This example will be used among this paper.

Table 1: Books Appreciations' Table: BAT.

ID	BookRate	CL
1	b_1 0.3 $\{b_2,b_3\}$ 0.7	[0.5;1]
2	b_2 0.5 b_4 0.5	[0.3;0.8]
3	$\{b_1,b_2,b_3\}$ 1	[1;1]
4	b_3 1	[0.5;0.9]

This paper is organized as follows: we recall, in section 2, some definitions and concepts of the top-k Queries, the evidence theory, the evidential databases and some approaches of ranking intervals. In section 3, we present our main contribution about the evidential top-k queries. The conclusion and the future works are held in section 4.

2 BACKGROUND MATERIALS

In this section, some notions about top-k queries and several comparing intervals approaches in the literature are briefly presented. Other fundamental concepts like the evidence theory and the evidential databases are also exposed in this section.

2.1 Top-k Queries

Top-k queries are also known as *Ranking queries*. They represent a powerful tool when we want to order queries' results in order to only give the most interesting answers. Top-k queries were introduced in the multimedia systems (Fagin, 1996; Fagin, 1998). Generally, top-k queries are ranked using a defined *score function* where only the k ($k \geq 1$) most important answers are returned. In other words, only answers with the highest scores are returned.

Ranking queries are needed in real worlds applications. For example movies can be ordered by the most watched ones, music can be ranked by the most listened songs, researchers can be ranked by their H-index, athletes by their race time, etc.

Several top-k processing techniques exist in the literature. Based on data uncertainty, they can be classified into three categories (Ilyas et al., 2008):

- *Exact methods over certain data*, where top-k queries and data are deterministic. The majority of top-k processing techniques are based on exact methods and certain data.
- *Approximate methods over certain data*, where processing top-k queries over certain data produces approximate results (Amato et al., 2003; Theobald et al., 2005).
- *Methods over uncertain data*, where top-k processing techniques deal with imperfect data. The top-k queries are based on different uncertainty models. At the best of our knowledge, only top-k queries' approaches that deal with probabilities exist in the literature (Re et al., 2007; Soliman et al., 2007) but there is no work that deal with other types of imperfect data. Our contribution copes mainly with this category.

2.2 Evidence Theory and Evidential Databases

Evidence theory, also called the Dempster-Shafer theory or the belief functions theory, was introduced by (Dempster, 1967; Dempster, 1968) and was mathematically formalized by (Shafer, 1976). Evidence theory is a powerful tool for the representation of imprecise, inconsistent and uncertain data.

A *frame of discernment* or *universe of discourse* is a set $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$. It is a finite, non empty and

[1]The literature is abundant in term of methods of preference elicitation using the evidence theory. We cite two main works (Ben Yaghlane et al., 2008; Ennaceur et al., 2014)

exhaustive set of n elementary and mutually exclusive hypotheses for a given problem. The *power set* $2^{\theta} = \{\varnothing, \theta_1, \theta_2, ..., \theta_n, \{\theta_1, \theta_2\}, .., \{\theta_1, \theta_2, ..., \theta_n\}\}$ is a set of all subsets of Θ.

A *mass function*, noted m, is a mapping from 2^{Θ} to the interval $[0,1]$. The *basic belief mass* of an hypothesis x is noted $m(x)$, it represents the belief on the truth of that hypothesis x. A mass function is also called *basic belief assignment* (*bba*). It is formalized such that:

$$\sum_{x \subseteq \Theta} m^{\Theta}(x) = 1 \quad (1)$$

If $m^{\Theta}(x) > 0$, x is called *focal element*. The set of all focal elements is denoted F and the couple $\{F, m\}$ is called *body of evidence*.

The *belief function* noted *bel* is the *minimal* amount of belief committed exactly to x and it is calculated as follows:

$$bel(x) = \sum_{y \subseteq x; y \neq \varnothing} m^{\Theta}(y) \quad (2)$$

The *plausibility function* noted *pl* is the *maximal* amount of belief on x and it is computed as follows:

$$pl(x) = \sum_{y \subseteq \Theta; x \cap y \neq \varnothing} m^{\Theta}(y) \quad (3)$$

An *Evidential database* (*EDB*), also named *Dempster-Shafer database* is a database that stores both perfect and imperfect data. In such databases imperfection is modeled using the theory of belief functions.

Definition 1. *An* EDB *has N objects and A attributes. An* evidential value*, noted V_{la}, is the value of an attribute a ($1 \leq a \leq A$) for an object l ($1 \leq l \leq N$) that represents a basic belief assignment.*

$$V_{la} : 2^{\Theta_a} \rightarrow [0,1] \quad (4)$$

$$\text{with } m_{la}^{\Theta_a}(\varnothing) = 0 \text{ and } \sum_{x \subseteq \Theta_a} m_a^{\Theta_a}(x) = 1 \quad (5)$$

The set of focal elements relative to the bba V_{la} is noted F_{la} such that:

$$F_{la} = \{x \subseteq \Theta_a / m_{la}(x) > 0\} \quad (6)$$

A confidence level*, denoted CL, is a specific attribute used to represent an interval of confidence for each object l in the evidential database. It is a pair of belief and plausibility [bel; pl], that reflects the pessimistic and optimistic believes of the existence of the object in the database (Lee, 1992b; Lee, 1992a; Bell et al., 1996).*

An Evidential Database stores various types of data:

- *Perfect data*: When the focal element is a singleton and its mass is equal to one then the bba is called *certain*. In table 1, the *BookRate* of the fourth object is a certain bba.
- *Probabilistic data*: When focal elements are singletons then the bba is called *bayesian*. In table 1, the *BookRate* of the second object is a bayesian bba.
- *Possibilistic data*: When focal elements are nested, then the bba is called *consonant*.
- *Evidential data*: When none of the previous cases is present the bba is called *evidential*. This is the case of the first and the third objects of the *BookRate* as shown in table 1.

In relational databases, data are stored to be further queried using the relational operators: selection, projection, join, etc. Evidential databases are also interrogated using the extended relational operators like the *extended selection*, the *extended projection*, etc. (Bell et al., 1996; Lee, 1992a; Lee, 1992b).

Applying a query Q over an evidential database *EDB* gives a set of evidential results R. For each answer R_i, a new confidence level is computed denoted CL=[bel_c; pl_c] that quantifies the degree of faith about that answer.

The extended select operator consists on extracting from *EDB* the objects whose values satisfy the condition of a query Q. The result is a relation that obey to a threshold of belief and plausibility. For each object l in the evidential database, the belief b and plausibility p of the condition are multiplied with the confidence level $CL[bel; pl]$ of that object. The computed CL of each resulting object in the relation is defined as follows:

$$CL = [b * bel; p * pl] \quad (7)$$

Example 1. *We take an example of a select query that we process over the evidential table 1, the query is the following:*

*Q_1: SELECT * FROM BAT WHERE (BookRate = $\{b_1\}$)*

Book b_1 appears only in tuples t_1 and t_3 with two confidence levels $t_1.CL_{b_1}$ and $t_3.CL_{b_1}$, summarized in table 2 and computed as follows:

- $t_1.CL_{b_1}$ = *[0.3*0.5 ; 0.3*1] = [0.15 ; 0.3].*

$$with \begin{cases} bel(b_1)=0.3 \text{ and } pl(b_1)=0.3 \text{ for tuple } t_1 \\ bel(t_1)=0.5 \text{ and } pl(t_1)=1 \end{cases}$$

- $t_3.CL_{b_1}$ = *[0*1 ; 1*1] = [0 ; 1].*

$$with \begin{cases} bel(b_1)=0 \text{ and } pl(b_1)=1 \text{ for tuple } t_3 \\ bel(t_3)=1 \text{ and } pl(t_3)=1 \end{cases}$$

Table 2: Result of query Q_1.

ID	BookRate	CL
1	b_1 $\{b_2,b_3\}$	[0.15 ; 0.3]
3	$\{b_1,b_2,b_3\}$	[0 ; 1]

The book b_1 appears in the first object with a degree of confidence of [0.15 ; 0.3] and it appears also in the third object with a confidence level of [0 ; 1].

2.3 Methods of Ranking Intervals

Many approaches were introduced to compare or rank intervals: First, (Borda, 1781), managed the ordinal ranking problem and proposed a method to rank candidates in election. Then, (Kendall, 1990) proposed a statistical framework for the ranking problem based on summing ranks assigned to each candidate by the voters. (Arrow, 2012) and (Inada, 1964; Inada, 1969) solved the problem of ranking using the majority rule. (Kemeny and Snell, 1962) used distance measures for the ranking.

(Salo and Hämäläinen, 1992) introduced the decision maker (DM) method that compares intervals in order to get one interval that dominates the other intervals but this method is not always feasible. (Ishibuchi and Tanaka, 1990) used a comparison rule on interval numbers to define an order, however this approach fails when intervals are nested. (Kundu, 1997) defined a fuzzy method that calculates the degree that an interval is superior or inferior to another one. This method requires that all interval numbers are independent and uniformly distributed.

(Wang et al., 2005) developed a ranking method of interval numbers. They proposed a preference aggregation method by combining individual preferences which is a typical group decision making problem like committee decision, voting systems, etc. The final ranking, in preference aggregation, is based on the comparison and the ranking interval numbers. To do so, they used a simple interval ranking approach. That approach will be used and adopted in this paper but other fitted approach can be also used.

3 EVIDENTIAL Top-k QUERIES

Processing queries over evidential databases gives answers, each one quantified with a degree of confidence. That degree reflects the lower and the upper bounds of trust in that response which is calculated from the database. In this case, these given answers are not ranked which don't allow the user to choose the most interesting ones from the set of results according to a defined criteria. In order to give the decision maker the responses that satisfy its request, we need to introduce a new top-k approach specific for evidential databases.

In this section, we present a new formalism to rank evidential results based on a score function. First, we process a query Q over the evidential database EDB. Then, for each generated response an *Evidential Score* is computed. That score is an interval of belief and plausibility, defined as follows:

Definition 2. *Evidential Score: Let R_i be a response generated from processing a query Q over an evidential database EDB, $S(R_i)$ is the score function of that answer R_i and $bel(R_i)$ and $pl(R_i)$ are respectively its belief and plausibility in the table, such that:*

$$S(R_i) = [bel(R_i); pl(R_i)] \tag{8}$$

$$Where\ bel(R_i) = \frac{\sum_{l=1}^{N} bel_l(R_i) * bel_l}{N}$$

$$pl(R_i) = \frac{\sum_{l=1}^{N} pl_l(R_i) * pl_l}{N}$$

The belief of an answer, $bel(R_i)$, is a disjunction of the response's beliefs in each object of the database. The belief of a response in one object l is the product of its belief in the attribute and the belief of that object. Same for the plausibility of an answer, $pl(R_i)$. It is the disjunction of the response's plausibilities in each object of the database where the plausibility of a response in one object l is the product of its plausibility in the attribute and the plausibility of that object (Bell et al., 1996; Lee, 1992a).

Example 2. *Let us process the query Q_2 over the evidential table 1 in order to get the top-2 answers.*

Q_2: The most appreciated books from table BAT.

The score of each item in the relation that may be a response to the query Q_2 is computed as follows:

- *The first possible response is book b_1, it appears in objects l_1 and l_3. Therefore:*

$$\begin{cases} bel(b_1) = \frac{(0.3*0.5)+(0*0.3)+(0*1)+(0*0.5)}{4} \\ pl(b_1) = \frac{(0.3*1)+(0*0.8)+(1*1)+(0*0.9)}{4} \end{cases}$$

Thus:
$S(b_1) = [bel(b_1); pl(b_1)] = [0.0375; 0.325]$

- *The second possible response is book b_2, it appears in objects l_1, l_2 and l_3. Therefore:*

$$\begin{cases} bel(b_2) = \frac{(0*0.5)+(0.5*0.3)+(0*1)+(0*0.5)}{4} \\ pl(b_2) = \frac{(0.7*1)+(0.5*0.8)+(1*1)+(0*0.9)}{4} \end{cases}$$

Thus:
$S(b_2) = [bel(b_2); pl(b_2)] = [0.0375; 0.525]$

- *The third possible response is book b_3, it appears in objects l_1, l_3 and l_4. Therefore:*

$$\begin{cases} bel(b_3) = \frac{(0*0.5)+(0*0.3)+(0*1)+(1*0.5)}{4} \\ pl(b_3) = \frac{(0.7*1)+(0*0.8)+(1*1)+(1*0.9)}{4} \end{cases}$$

Thus:
$S(b_3) = [bel(b_3); pl(b_3)] = [0.125; 0.65]$

- *The final response is book b_4, it appears only in object l_2. Therefore:*

$$\begin{cases} bel(b_4) = \frac{(0*0.5)+(0.5*0.3)+(0*1)+(0*0.5)}{4} \\ pl(b_4) = \frac{(0*1)+(0.5*0.8)+(0*1)+(0*0.9)}{4} \end{cases}$$

Thus:
$S(b_4) = [bel(b_4); pl(b_4)] = [0.0375; 0.1]$

The computed evidential scores are shown in table 3.

Table 3: Evidential Score per Item.

item	*EvidentialScore*
b_1	R_1= [0.0375 ; 0.325]
b_2	R_2= [0.0375 ; 0.525]
b_3	R_3= [0.125 ; 0.65]
b_4	R_4= [0.0375 ; 0.1]

Top-*k* queries are based on a defined score function. That function produces precise values, in contrary to the evidential top-*k* queries whose score function produces intervals bounded by belief and plausibility values. (Wang et al., 2005) introduced an approach of ranking intervals based on preference degrees. Their method is the one that we will adopt to rank scores previously generated.

Definition 3. *Preference Degree: Let $S(R_i)$=[bel_i; pl_i] and $S(R_j)$=[bel_j; pl_j] be two evidential scores. Each one is an interval composed of a degree of belief and a degree of plausibility. The degree of one interval to be greater than the other one is called a* degree of preference *and denoted P.*

The degree of preference that $S(R_i) > S(R_j)$ is defined such that:

$$P(S(R_i) > S(R_j)) = \frac{max(0, pl_i - bel_j) - max(0, bel_i - pl_j)}{(pl_i - bel_i) + (pl_j - bel_j)} \quad (9)$$

The degree of preference that $S(R_i) < S(R_j)$ is defined such that:

$$P(S(R_i) < S(R_j)) = \frac{max(0, pl_j - bel_i) - max(0, bel_j - pl_i)}{(pl_i - bel_i) + (pl_j - bel_j)} \quad (10)$$

The different cases of comparing intervals $S(R_i)$ and $S(R_j)$ are as follows:

- *If $P(S(R_i) > S(R_j)) > P(S(R_j) > S(R_i))$, then $S(R_i)$ is said to be superior to $S(R_j)$, denoted by $S(R_i) \succ S(R_j)$.*

- *If $P(S(R_i) > S(R_j)) = P(S(R_j) > S(R_i))$ = 0.5, then $S(R_i)$ is said to be indifferent to $S(R_j)$, denoted by $S(R_i) \sim S(R_j)$.*

- *If $P(S(R_j) > S(R_i)) > P(S(R_i) > S(R_j))$, then $S(R_i)$ is said to be inferior to $S(R_j)$, denoted by $S(R_i) \prec S(R_j)$.*

Theorem 1. *Let $S(R_i)$=[bel_i; pl_i] and $S(R_j)$=[bel_j; pl_j] be two evidential scores such that:*

- *Shortcut 1: if $S(R_i) = S(R_j)$ then $P(S(R_i)) > P(S(R_j)) = P(S(R_i)) < P(S(R_j)) = 0.5$.*

- *Shortcut 2: if $bel_i \geq pl_j$ then $P(S(R_i) > S(R_j)) = 1$.*

- *Shortcut 3: if $bel_i \geq bel_j$ and $pl_i \geq pl_j$ then $P(S(R_i) > S(R_j)) \geq 0.5$ and $P(S(R_j) > S(R_i)) \leq 0.5$.*

In order to detect the dominant interval between the score of relation R_i denoted $S(R_i)$ and the score of relation R_j denoted $S(R_j)$, we need to compute the degree of preference that $S(R_i) > S(R_j)$ and the degree of preference that $S(R_i) < S(R_j)$. The complexity of this computation can be reduced thanks to the *complementarity* of $P(S(R_i) > S(R_j))$ and $P(S(R_i) < S(R_j))$.

The complementarity is only feasible when:

$$\begin{cases} S(R_i) \neq S(R_j) \\ bel_i < pl_j \end{cases} \quad (11)$$

Proof. Complementarity:

$$P(S(R_i) < S(R_j)) = \frac{max(0, pl_j - bel_i) - max(0, bel_j - pl_i)}{(pl_i - bel_i) + (pl_j - bel_j)}$$

$$P(S(R_j) < S(R_i)) = \frac{max(0, pl_i - bel_j) - max(0, bel_i - pl_j)}{(pl_i - bel_i) + (pl_j - bel_j)}$$

$$P(S(R_i) < S(R_j)) + P(S(R_j) < S(R_i))$$

$$= \frac{max(0, pl_j - bel_i) - max(0, bel_j - pl_i)}{(pl_i - bel_i) + (pl_j - bel_j)} + \frac{max(0, pl_i - bel_j) - max(0, bel_i - pl_j)}{(pl_i - bel_i) + (pl_j - bel_j)}$$

$$= \frac{max(0, pl_j - bel_i) - 0 + max(0, pl_i - bel_j) - 0}{(pl_i - bel_i) + (pl_j - bel_j)}$$

$$= \frac{pl_j - bel_i + pl_i - bel_j}{pl_i - bel_i + pl_j - bel_j} = 1$$

$$P(S(R_i) < S(R_j)) + P(S(R_j) < S(R_i)) = 1$$

□

Figure 1 summarizes the different cases of evidential scores intervals. It represents also which property from the presented ones to use for each case.

The transitivity property is helpful to achieve a complete ranking order for scores. In (Wang et al., 2005), authors proved that preference relations are *transitive*.

Property 1. *Transitivity*
Let $S(R_i) = [bel_i; pl_i]$, $S(R_j) = [bel_j; pl_j]$ and $S(R_k) = [bel_k; pl_k]$ be three intervals. If $S(R_i) \succ S(R_j)$ and $S(R_j) \succ S(R_k)$ then $S(R_i) \succ S(R_k)$.

Previous definitions provide a total ranking of answers that respond to the proposed top-k query. But how to interpret any evidential answer ?

The top-k queries in deterministic databases are semantically clear. However, the interpretation of top-k queries in imperfect databases are challenging. (Soliman et al., 2007) introduced new semantics relative to probabilistic top-k queries. They defined them as *the most probable query answers*. Their work is based on the possible worlds' model and they proposed interpretations like: (i) *The top-k tuples in the most probable world.* (ii)*The most probable top-k tuples that belongs to valid possible world.*

The interpretations of probabilistic top-k queries can not be considered for evidential top-k queries. Thus, a new specific semantic for *Evidential Top-k Queries* is defined as follows:

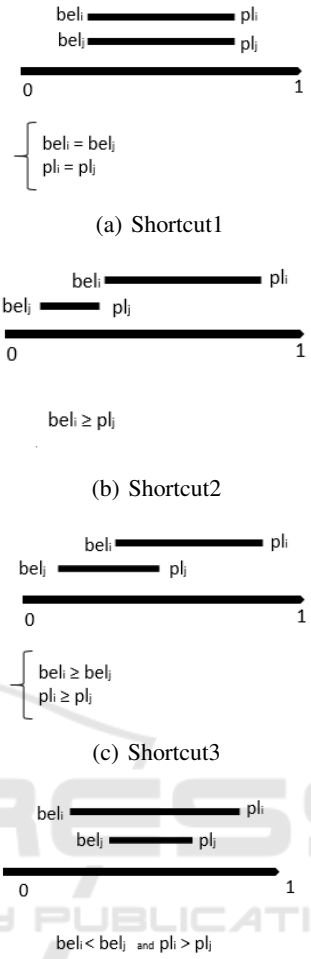

Figure 1: Comparison of Evidential Scores.

Definition 4. *E-Top-k:*
Let EDB be an evidential database with N objects and A attributes; CL is an attribute where the intervals associated to objects reflects the degrees of confidence about these objects. Let $S(R_i)$ be a score function that maximizes both CL and the interval of belief on each result. Responses R are ordered according to scores.

An E-top$k$ *returns* the k most credible answers from the set of answers *such that:*

$$S(R_i) = Argmax_{R_i \in R}([bel(R_i); pl(R_i)]) \qquad (12)$$

Example 3. *We carry on with the same example of table 1 and we give a total ranking of the resulting evidential scores. We also deduce the top-2 answers and their interpretation.*

(i) Since $bel_{b_1} = bel_{b_2}$ and $pl_{b_2} > pl_{b_1} > pl_{b_4}$

then $b_2 \succ b_1 \succ b_4$

(ii) Since $bel_{b_3} > bel_{b_2}$ *and* $pl_{b_3} > pl_{b_2}$

then $b_3 \succ b_2$

The final ranking deduced from (i) and (ii) is: $b_3 \succ b_2 \succ b_1 \succ b_4$.

The Top-2 appreciated books are:

- b_3 *with a confidence level [0.125 ; 0.65]*
- b_2 *with a confidence level [0.0375 ; 0.525]*

Books b_3 *and* b_2 *are the most appreciated credible answers from the set of results.*

4 CONCLUSION AND FUTURE WORKS

In this paper, we presented a new imperfect top-*k* query called the *evidential top-k query*. It consists in processing top-*k* query over evidential data (data modeled using the theory of belief functions). First, we introduced a new score function that computes an interval of belief and plausibility relative to each answer responding a given top-*k* query. Then, we adopted a preference approach of comparing intervals (Wang et al., 2005). We also presented the proof of complementarity relative to that approach, in order to reduce the complexity of computations while calculating the evidential score. Finally, we introduced a new semantics relative to evidential top-*k*.

As future works, top-*k* queries may be implemented and other types of such a query (like the aggregation, the project and the join uncertain queries for the evidential databases) may be also detailed.

REFERENCES

Amato, G., Rabitti, F., Savino, P., and Zezula, P. (2003). Region proximity in metric spaces and its use for approximate similarity search. *ACM Transactions on Information Systems (TOIS)*, 21(2):192–227.

Andritsos, P., Fuxman, A., and Miller, R. J. (2006). Clean answers over dirty databases: A probabilistic approach. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 30–30. IEEE.

Arrow, K. J. (2012). *Social choice and individual values*, volume 12. Yale university press.

Bell, D. A., Guan, J. W., and Lee, S. K. (1996). Generalized union and project operations for pooling uncertain and imprecise information. *Data & Knowledge Engineering (DKE)*, 18:89–117.

Ben Yaghlane, A., Denœux, T., and Mellouli, K. (2008). Elicitation of expert opinions for constructing belief functions. *Uncertainty and Intelligent Information Systems*, pages 75–88.

Borda, J. C. (1781). Mémoire sur les élections au scrutin. *Translated in the political theory of condorcet. Sommerlad F, Mclean I. Social studies, Oxford, 1989*.

Bousnina, F. E., Bach Tobji, M. A., Chebbah, M., Liétard, L., and Ben Yaghlane, B. (2015). A new formalism for evidential databases. In *22nd International Symposium on Methodologies for Intelligent Systems (ISMIS), Foundations of Intelligent Systems*, pages 31–40. Springer.

Cheng, R., Kalashnikov, D. V., and Prabhakar, S. (2004a). Querying imprecise data in moving object environments. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 16(9):1112–1127.

Cheng, R., Xia, Y., Prabhakar, S., Shah, R., and Vitter, J. S. (2004b). Efficient indexing methods for probabilistic threshold queries over uncertain data. In *13th International Conference on Very Large Data Bases (VLDB)*, pages 876–887. VLDB Endowment.

Chung, B. S., Lee, W.-C., and Chen, A. L. (2009). Processing probabilistic spatio-temporal range queries over moving objects with uncertainty. In *12th International Conference on Extending Database Technology, Advances in Database Technology*, pages 60–71. ACM.

Considine, J., Li, F., Kollios, G., and Byers, J. (2004). Approximate aggregation techniques for sensor databases. In *20th International Conference on Data Engineering (ICDE)*, pages 449–460. IEEE.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multiple valued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339.

Dempster, A. P. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society, Series B*, 30:205–247.

Ding, X. and Jin, H. (2012). Efficient and progressive algorithms for distributed skyline queries over uncertain data. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 24(8):1448–1462.

Elmi, S., Benouaret, K., HadjAli, A., Bach Tobji, M. A., and Ben Yaghlane, B. (2014). Computing skyline from evidential data. In *8th International Conference on Scalable Uncertainty Management (SUM)*, pages 148–161, Oxford, UK.

Elmi, S., Benouaret, K., HadjAli, A., Bach Tobji, M. A., and Ben Yaghlane, B. (2015). Requêtes skyline en présence des données évidentielles. In *Extraction et Gestion des Connaissances (EGC)*, pages 215–220.

Ennaceur, A., Elouedi, Z., and Lefevre, E. (2014). Multicriteria decision making method with belief preference relations. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 22(04):573–590.

Fagin, R. (1996). Combining fuzzy information from multiple systems. In *15th ACM SIGACT-SIGMOD-SIGART*

symposium on Principles of database systems, pages 216–226. ACM.

Fagin, R. (1998). Fuzzy queries in multimedia database systems. In *17th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 1–10. ACM.

Ilyas, I. F., Beskales, G., and Soliman, M. A. (2008). A survey of top-k query processing techniques in relational database systems. *ACM Computing Surveys (CSUR)*, 40(4):11.

Inada, K.-i. (1964). A note on the simple majority decision rule. *Econometrica, Journal of the Econometric Society*, pages 525–531.

Inada, K.-i. (1969). The simple majority decision rule. *Econometrica, Journal of the Econometric Society*, pages 490–506.

Ishibuchi, H. and Tanaka, H. (1990). Multiobjective programming in optimization of the interval objective function. *European Journal of Operational Research (EJOR)*, 48(2):219–225.

Kemeny, J. G. and Snell, L. (1962). Preference ranking: an axiomatic approach. *Mathematical models in the social sciences*, pages 9–23.

Kendall, M. (1990). Rank correlation methods. *Oxford University Press, 5th edition*.

Kundu, S. (1997). Min-transitivity of fuzzy leftness relationship and its application to decision making. *Fuzzy sets and systems*, 86(3):357–367.

Laplace, P. S. d. (1812). Théorie analytique des probabilités. *Courcier, Paris*.

Lee, S. K. (1992a). An extended relational database model for uncertain and imprecise information. In *18th Conference on Very Large Data Bases (VLDB)*, pages 211–220, Canada.

Lee, S. K. (1992b). Imprecise and uncertain information in databases : an evidential approach. In *8th International Conference on Data Engineering (ICDE)*, pages 614–621.

Re, C., Dalvi, N., and Suciu, D. (2007). Efficient top-k query evaluation on probabilistic data. In *IEEE 23rd International Conference on Data Engineering (ICDE)*, pages 886–895. IEEE.

Salo, A. and Hämäläinen, R. (1992). Processing interval judgments in the analytic hierarchy process.

Shafer, G. (1976). A mathematical theory of evidence. *Princeton University Press*.

Silberstein, A., Braynard, R., Ellis, C., Munagala, K., and Yang, J. (2006). A sampling-based approach to optimizing top-k queries in sensor networks. In *22nd International Conference on Data Engineering (ICDE)*, pages 68–68. IEEE.

Soliman, M. A., Ilyas, I. F., and Chang, K. C.-C. (2007). Top-k query processing in uncertain databases. In *23rd International Conference on Data Engineering (ICDE)*, pages 896–905. IEEE.

Tao, Y., Xiao, X., and Cheng, R. (2007). Range search on multidimensional uncertain data. *ACM Transactions on Database Systems (TODS)*, 32(3):15.

Theobald, M., Schenkel, R., and Weikum, G. (2005). An efficient and versatile query engine for topx search. In *31st International Conference on Very large Data Bases (VLDB)*, pages 625–636. VLDB Endowment.

Wang, Y.-M., Yang, J.-B., and Xu, D.-L. (2005). A preference aggregation method through the estimation of utility intervals. *Computers & Operations Research*, 32(8):2027–2049.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28.