Supporting Case-Based Retrieval by Similarity Skylines: Basic Concepts and Extensions*

Eyke Hüllermeier[1], Ilya Vladimirskiy[1], Belén Prados Suárez[2], and Eva Stauch[3]

[1] Philipps-Universität, FB Informatik, D-35032, Hans-Meerwein-Str., Marburg, Germany, {eyke,ilya}@mathematik.uni-marburg.de

[2] Department of Computer Science and Artificial Intelligence, University of Granada, Spain, belenps@decsai.ugr.es

[3] Westfälische Wilhelms-Universität, Historisches Seminar, Robert-Koch-Str. 29, D-48149, Münster, Germany, estauch@uni-muenster.de

Abstract. Conventional approaches to similarity search and case-based retrieval, such as nearest neighbor search, require the specification of a global similarity measure which is typically expressed as an aggregation of local measures pertaining to different aspects of a case. Since the proper aggregation of local measures is often quite difficult, we propose a novel concept called *similarity skyline*. Roughly speaking, the similarity skyline of a case base is defined by the subset of cases that are most similar to a given query in a Pareto sense. Thus, the idea is to proceed from a d-dimensional comparison between cases in terms of d (local) distance measures and to identify those cases that are maximally similar in the sense of the Pareto dominance relation [2]. To refine the retrieval result, we propose a method for computing maximally diverse subsets of a similarity skyline. Moreover, we propose a generalization of similarity skylines which is able to deal with uncertain data described in terms of interval or fuzzy attribute values. The method is applied to similarity search over uncertain archaeological data.

1 Introduction

Similarity search in high-dimensional data spaces is important for numerous application areas. In case-based reasoning (CBR), for example, it provides an essential means for implementing case retrieval, a critical step in case-based problem solving. In case-based retrieval, understood as the application of CBR paradigms to information retrieval tasks [3], similarity search becomes an even more central issue.

A commonly applied approach to case retrieval is nearest neighbor (NN) search. In fact, NN queries as proposed in [4] and their application to similarity search have been studied quite extensively in the past. Despite their usefulness for certain problems, NN methods exhibit several disadvantages. For example, they are usually sensitive toward outliers and cannot easily deal with uncertain

* Revised and significantly extended version of a paper presented at the ICCBR-07 workshop on "Uncertainty and Fuzziness in Case-Based Reasoning" [1].

data. Due to the "curse of dimensionality" [5], the performance of NN methods significantly degrades in the case of high-dimensional data.

Perhaps even more importantly, NN methods assume a *global* similarity or, alternatively, distance function to be specified across the full feature set. The specification of such a measure is often greatly simplified by the "local–global principle", according to which the global similarity between two cases can be obtained as an aggregation of various local measures pertaining to different features of a case [6]. However, even though it is true that local distances can often be defined in a relatively straightforward way, the *combination* of these distances can become quite difficult in practice, especially since different features may pertain to completely different aspects of a case. Moreover, the importance of a feature is often subjective and context-dependent. Thus it might be reasonable to free a user querying a system from the specification of an aggregation function, or at least to defer this step to a later stage.

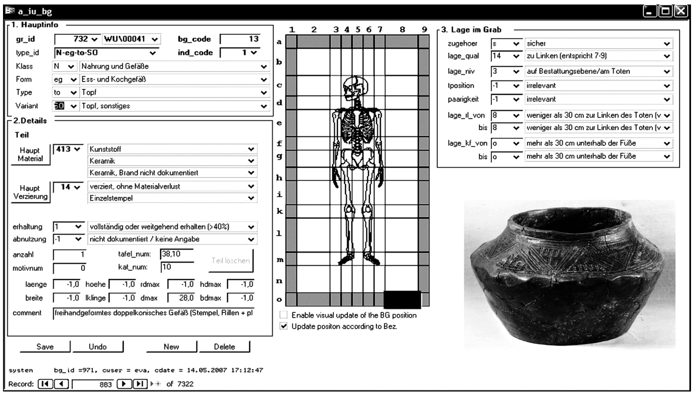
In this paper, we propose a new concept, called *similarity skyline*, for supporting similarity search and case-based retrieval without the need to specify a global similarity measure. Roughly speaking, the similarity skyline of a case base is defined by the subset of cases that are most similar to a given query in a Pareto sense. More precisely, the idea is to proceed from a d-dimensional comparison between cases in terms of d (local) similarity or distance measures and to identify those cases that are maximally similar in the sense of the Pareto dominance relation.

The rest of the paper is organized as follows: Section 2 describes the application that motivates our approach, namely similarity search over uncertain archaeological data. The concept of a similarity skyline is introduced in Section 3. In Section 4, we propose a method for refining the retrieval result, namely by selecting a (small) diverse subset of a similarity skyline. Section 5 is devoted to a generalization of similarity skylines which is able to deal with uncertain data described in terms of interval or fuzzy attribute values. Finally, Section 6 presents some experimental results, and Section 7 concludes the paper.

2 Motivation and Background

Even though the methods introduced in this paper are completely general, they have been especially motivated by a particular application. As we shall report experimental results for this application later on, we devote this section to a brief introduction.

The DEADDY project aims at using knowledge discovery techniques to extract valuable information from archaeological databases. The domain under study is the analysis of graves in the Early Middle Ages. The data informs about graves, the persons buried therein, and the grave goods (objects which were put into the grave during the funeral ceremony according to religious rules or traditions typical for the given historical moment). Fig. 1 shows a screen shot of the DEADDY user interface. One can see a data record with information about particular grave goods: type, material, position in the grave, etc.



a_iu_bg

1. Hauptinfo

gr_id: 732 WU\00041 — bg_code: 13
type_id: N-eg-to-SO — ind_code: 1
Klass: N — Nahrung und Gefäße
Form: eg — Ess- und Kochgefäß
Type: to — Topf
Variant: 60 — Topf, sonstiges

2.Details

Teil

Haupt Material: 413 — Kunststoff; Keramik; Keramik, Brand nicht dokumentiert
Haupt Verzierung: 14 — verziert, ohne Materialverlust; Einzelstempel

erhaltung: 1 — vollständig oder weitgehend erhalten (>40%)
abnutzung: -1 — nicht dokumentiert / keine Angabe
anzahl: 1 — tafel_num: 38,10
motivnum: 0 — kat_num: 10
laenge: -1,0 — hoehe: -1,0 — rdmax: -1,0 — hdmax: -1,0
breite: -1,0 — lklinge: -1,0 — dmax: 28,0 — bdmax: -1,0
comment: freihandgeformtes doppelkonisches Gefäß (Stempel, Rillen + pl

3. Lage im Grab

zugehoer: s — sicher
lage_qual: 14 — zu Linken (entspricht 7-9)
lage_niv: 3 — auf Bestattungsebene/am Toten
tposition: -1 — irrelevant
paarigkeit: -1 — irrelevant
lage_rl_von: 8 — weniger als 30 cm zur Linken des Toten (v
bis: 8 — weniger als 30 cm zur Linken des Toten (v
lage_kf_von: o — mehr als 30 cm unterhalb der Füße
bis: o — mehr als 30 cm unterhalb der Füße

□ Enable visual update of the BG position
☑ Update positon according to Bez.

```
system     bg_id =971, cuser = eva, cdate = 14.05.2007 17:12:47
```

Record: 883 of 7322

Fig. 1. Grave Good Form in the DEADDY Database

To demonstrate our approach, we have chosen the graveyard Wenigumstadt, which dates from the Early Middle Ages and is situated in the south of Germany. The inhabitants of a small village were buried in this cemetery from the end of the Roman Empire to the Age of Charlemagne. The data set contains information about 126 graves and 1074 grave goods. Data were extracted from a relational database and put into a joint table containing attributes for graves, individuals and grave goods. In total there are 9 attributes, 3 of which describe a grave, 2 a person, and the remaining 4 the grave goods.

Imagine an archaeologist interested in discovering dependencies between wealth of the grave equipment and the age of the person buried therein. To make a first step in analyzing this question, a system should support similarity searches in a proper way. For example, an archaeologist may choose an *interesting* grave as a starting point and then try to find graves which are *similar* to this one. The techniques developed in this paper are especially motivated by the following experiences that we had with this field of application and corresponding users:

- While local similarity measures pertaining to different attributes or properties of a grave can often be defined without much difficulty, an archaeologist is usually not willing or not able to define a global distance measure properly reflecting his or her (vague) idea of similarity between complete graves.
- Both the data, such as age or spatial coordinates of a grave good, as well as the queries referring to the data are typically vague and imprecise, sometimes even context-dependent.

3 Similarity Search and the Similarity Skyline

We proceed from a description of cases in terms of d-dimensional feature vectors

$$\boldsymbol{x} = (x_1, x_2 \ldots x_d) \in \mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2 \times \ldots \times \mathbb{X}_d, \tag{1}$$

where $\mathbb{X}_i$ is the domain of the i-th feature X_i. A case base CB is a finite subset of the space $\mathbb{X}$ spanned by the domains of the d features. Even though a feature-based representation is of course not always suitable, it is often natural and still predominant in practice [7]. In this regard, we also note that a feature is not assumed to be a simple numerical or categorical attribute. Instead, a single feature can be a complex entity (and hence $\mathbb{X}_i$ a complex space), for example a structured object such as a tree or a graph. We only assume the existence of *local distance measures*

$$\delta_i : \mathbb{X}_i \times \mathbb{X}_i \to \mathbb{R}_+, \tag{2}$$

i.e., each space $\mathbb{X}_i$ is endowed with a measure that assigns a degree of distance $\delta_i(x_i, y_i)$ to each pair of features $(x_i, y_i) \in \mathbb{X}_i \times \mathbb{X}_i$. According to the local–global principle, the distance between two cases can then be obtained as an aggregation of the local distance measures (2):

$$\Delta(\boldsymbol{x}, \boldsymbol{y}) \;=\; A\,(\,\delta_1(x_1, y_1), \delta_2(x_2, y_2) \ldots \delta_d(x_d, y_d)\,)\,, \tag{3}$$

where A is a suitable aggregation operator. As mentioned in the introduction, the specification of such an aggregation operator can become quite difficult in practice, especially for non-experts. Therefore, it might be reasonable to free a user querying a system from this requirement, or at least to defer this step to a later stage.

One may of course imagine intermediary scenarios in which *some* of the local similarity measures can be aggregated into measures at a higher level of a hierarchical scheme. In this scheme, the problem of similarity assessment is decomposed in a recursive way, i.e., a similarity criterion is decomposed into certain sub-criteria, which are then aggregated in a suitable way. In other words, each feature or, perhaps more accurately, *similarity feature* X_i in (1) might already be an aggregation

$$X_i \;=\; A_i(X_{i1}, X_{i2} \ldots X_{ik})$$

of a certain number of sub-features, which in turn can be aggregations of sub-sub-features, etc. Now, our assumption is that a further aggregation of the features $X_1 \ldots X_d$ is not possible, or at least not supported by the user. These (similarity) features, however, do not necessarily correspond to the attributes used to describe a single case. For example, suppose that two cars, each of which might be described by a large number of attributes, can be compared with respect to *comfort* and *investment* in terms of corresponding similarity measures. If a further combination of these two degrees into a single similarity score is difficult, then comfort and investment are the features in (1).

3.1 The Similarity Skyline

Note that a global similarity or distance function, if available, induces a *total* order on the set of all alternatives: Given a query $\boldsymbol{z} = (z_1 \ldots z_d) \in \mathbb{X}$ and two cases $\boldsymbol{x}, \boldsymbol{y} \in$ CB,

$$\boldsymbol{x} \succeq_{\boldsymbol{z}} \boldsymbol{y} \;\overset{\text{df}}{\Longleftrightarrow}\; \Delta(\boldsymbol{z}, \boldsymbol{x}) \leq \Delta(\boldsymbol{z}, \boldsymbol{y}).$$

Instead of requiring a user to define a global distance measure and, thereby, to bring all alternatives into a total order, the idea of this paper is to compare alternatives in terms of a much weaker "closeness" or, say, "preference" relation, namely Pareto dominance: Given a query $\boldsymbol{z}$ and cases $\boldsymbol{x}, \boldsymbol{y}$,

$$\boldsymbol{x} \succeq_{\boldsymbol{z}} \boldsymbol{y} \overset{\text{df}}{\iff} \forall\, i \in \{1, 2 \ldots d\} \,:\, \delta_i(z_i, x_i) \leq \delta_i(z_i, y_i).$$

Thus, $\boldsymbol{x}$ is (weakly) preferred to $\boldsymbol{y}$ if the former is not less similar to $\boldsymbol{z}$ than the latter in every dimension. Moreover, we define strict preference as follows:

$$\boldsymbol{x} \succ_{\boldsymbol{z}} \boldsymbol{y} \overset{\text{df}}{\iff} \boldsymbol{x} \succeq_{\boldsymbol{z}} \boldsymbol{y} \wedge \exists\, i \in \{1, 2 \ldots d\} \,:\, \delta_i(z_i, x_i) < \delta_i(z_i, y_i). \tag{4}$$

When $\boldsymbol{x} \succ_{\boldsymbol{z}} \boldsymbol{y}$, we also say that $\boldsymbol{y}$ is *dominated* or, more specifically, *similarity-dominated* by $\boldsymbol{x}$. Note that the relation $\succeq_{\boldsymbol{z}}$ is only a partial order, i.e., it is antisymmetric and transitive but not complete. That is, two cases $\boldsymbol{x}, \boldsymbol{y} \in \text{CB}$ may (and often will) be incomparable in terms of $\succeq_{\boldsymbol{z}}$, i.e., it may happen that one can neither say that $\boldsymbol{x}$ is "more similar" than $\boldsymbol{y}$ nor vice versa.

However, when $\boldsymbol{x} \succ_{\boldsymbol{z}} \boldsymbol{y}$ holds, $\boldsymbol{x}$ is arguably more interesting than $\boldsymbol{y}$ as a retrieval candidate. More precisely, the following observation obviously holds: $\boldsymbol{x} \succ_{\boldsymbol{z}} \boldsymbol{y}$ implies $\Delta(\boldsymbol{z}, \boldsymbol{x}) < \Delta(\boldsymbol{z}, \boldsymbol{y})$, regardless of the aggregation function A in (3), provided this function is strictly monotone in all arguments. As a result, $\boldsymbol{y}$ cannot be maximally similar to the query, as $\boldsymbol{x}$ is definitely more similar.

Consequently, the interesting candidates for case retrieval are those cases that are non-dominated. Such cases are called *Pareto-optimal*, and the set itself is called the Pareto set. This set corresponds to the set of cases that are potentially most similar to the query: If there exists an aggregation function A such that $\boldsymbol{x}$ is maximally similar to $\boldsymbol{z}$ among all cases in CB, then $\boldsymbol{x}$ must be an element of the Pareto set. For reasons that will become clear in the next subsection, we call the set of Pareto-optimal cases the *similarity skyline*:

$$\text{SSky}(\text{CB}, \boldsymbol{z}) \overset{\text{df}}{=} \{\, \boldsymbol{x} \in \text{CB} \mid \forall \boldsymbol{y} \in \text{CB} \,:\, \boldsymbol{y} \not\succ_{\boldsymbol{z}} \boldsymbol{x} \,\} \tag{5}$$

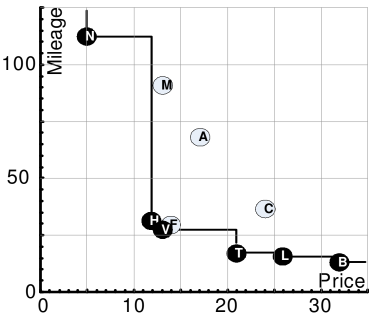
In passing, we note that only the ordinal structure of the local distance measures δ_i is important for this approach, which further simplifies their definition: For the $\mathbb{X} \to \mathbb{R}_+$ mapping $\delta_i(z_i, \cdot)$, it is only important how it orders x_i and y_i, i.e., whether $\delta_i(z_i, x_i) < \delta_i(z_i, y_i)$ or $\delta_i(z_i, x_i) > \delta_i(z_i, y_i)$, while the distance degrees themselves are irrelevant. In other words, the similarity skyline (5) is invariant toward monotone transformations of the δ_i.

3.2 Skyline Computation

The computation of a Pareto optimal subset of a given reference set has received a great deal of attention in the database community in recent years. Here, the Pareto optimal set is also called the *skyline*. A "skyline operator", along with a corresponding SQL notation, was first proposed in [8]. It proceeds from a representation of objects in terms of d *criteria*, i.e., "less-is-better" attributes

C_i, $i = 1 \dots d$, with linearly ordered domains $\mathbb{R}_+$; the corresponding data space is the Cartesian product of these domains, and an object is a vector in this space. In the simplest form, the skyline $\mathrm{Sky}(P)$ of a d-dimensional data set P is defined by the subset of objects $(c_1 \dots c_d) \in P$ that are non-dominated, i.e., for which there is no $(c'_1 \dots c'_d) \in P$ such that $c'_i \leq c_i$ holds for all and $c'_i < c_i$ for at least one $i \in \{1 \dots d\}$.

To illustrate, consider a user choosing a car from a used-cars database, and suppose cars to be characterized by only two attributes, namely price and mileage. An example data set and its skyline are presented in Fig. 2. Point A (Acura) is dominated by point H (Honda), because the Honda is cheaper and has lower mileage. The six points (marked black) which are non-dominated by any other point form the skyline.



Car	Price, 1000$	Mileage, 1000km
Acura	17	68
BMW	32	13
Cadillac	24	37
Ford	14	29
Honda	12	33
Land Rover	26	16
Mercedes	13	91
Nissan	5	113
Toyota	21	18
Volkswagen	13	28

Fig. 2. Example of a two-dimensional skyline.

Now, recall the problem of computing a *similarity skyline*, as introduced in the previous subsection: Given a case base CB and a query case $\boldsymbol{z}$, the goal is to retrieve the set of cases $\boldsymbol{x} \in$ CB that are non-dominated in the sense of (4). This problem can be reduced to the standard skyline problem in a relatively straightforward way. To this end, one simply defines the criteria to be minimized by the distances in the different dimensions. Thus, with $\delta_i : \mathbb{X}_i \times \mathbb{X}_i \rightarrow \mathbb{R}_+$ denoting the distance measure for the i-th feature, a case $\boldsymbol{x} = (x_1 \dots x_d)$ is first mapped to a point

$$\boldsymbol{x}' = T_{\boldsymbol{z}}(\boldsymbol{x}) \overset{\mathrm{df}}{=} (\, \delta_1(x_1, z_1),\, \delta_2(x_2, z_2) \dots \delta_d(x_d, z_d)\,) \in \mathbb{R}^d_+ . \tag{6}$$

Geometrically speaking, this transformation is a kind of reflection that, using the reference point $\boldsymbol{z}$ as a center, maps all data points into the positive quadrant (see Fig. 3). The similarity skyline then corresponds to the standard skyline of the image of CB under the mapping $T_{\boldsymbol{z}}$, i.e.,

$$\mathrm{SSky}(\mathrm{CB}, \boldsymbol{z}) \;=\; \mathrm{Sky}(T_{\boldsymbol{z}}(\mathrm{CB})).$$

Computing a skyline in an efficient way is a non-trivial problem, especially in high dimensions (cf. Section 6). In the database field, several main-memory algo-

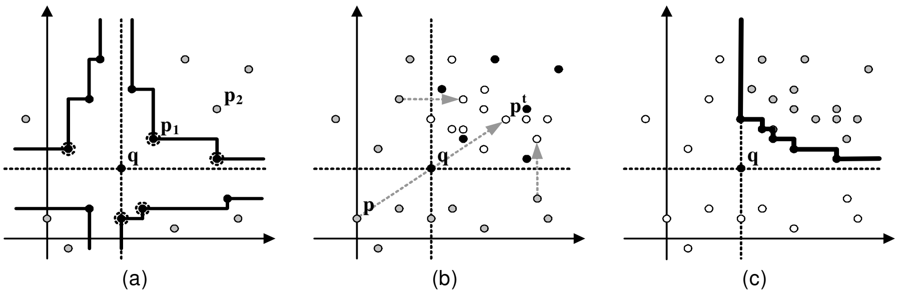


Fig. 3. Using the query point q as a center, the original data points (a) are mapped into the positive quadrant in a distance-preserving way (b). The skyline in the transformed space corresponds to the points that are not similarity-dominated (c).

rithms (for the case where the whole data set fits in memory) as well as efficient methods for computation of skyline points over data stored in the database have been proposed. In our implementation, we used the block nested loop (BNL) algorithm for skyline computation [8]. The most naive way to compute a skyline is to check the non-dominance condition explicitly for each case (by comparing it to all other cases). BNL is a modification of this approach which proceeds as follows: The list of skyline candidate objects (SCL) is kept in the memory, initialized with the first case. Then, the other cases $\boldsymbol{y}$ are examined one by one: (a) If $\boldsymbol{y}$ is dominated by any case in the SCL, it is pruned as it can not belong to the skyline. (b) If $\boldsymbol{y}$ dominates one or more case in the SCL, these cases are replaced by $\boldsymbol{y}$. (c) If $\boldsymbol{y}$ is neither dominated by, nor dominates any case in the SCL, it is simply added to the SCL. We refer to [9] for more details on BNL and a thorough review of alternative skyline computation algorithms. It is also worth mentioning that the concept of *dynamic skyline*, proposed in the same paper, provides a perfect algorithmic framework for implementing similarity skyline computation when the data is stored in an indexed database instead of main memory.

4 Refining Similarity Skylines

The similarity skyline (5) may become undesirably large, especially in high dimensions. A user may thus not always want to inspect the whole set of Pareto optimal cases. A possible solution to this problem is to select an interesting subset from $\mathbb{S} = \mathrm{SSky}(\mathrm{CB}, \boldsymbol{z})$, i.e., to filter $\mathbb{S}$ according to a suitable criterion. Here, we propose the criterion of *diversity*, which has recently attracted special attention in case-based retrieval [10, 11]: To avoid redundancy, and to convey a picture of the whole set $\mathbb{S}$ with only a few cases, the idea is to select a subset of cases which is as diverse as possible.

An implementation of this criterion requires a formalization of the concept of diversity. What does it mean that a set $\mathbb{D} \subseteq \mathbb{S}$ is diverse? Intuitively, it means that the cases in $\mathbb{D}$ should be dissimilar amongst each other. It is important to note that, according to our assumptions, a formalization of this criterion must only refer to the local distance measures δ_i, $i = 1 \ldots d$, and not to a global measure.

We therefore define the diversity of a subset $\mathbb{D}$ of cases by the vector $\mathrm{div}(\mathbb{D}) = (v_1, v_2 \ldots v_d)$, where

$$v_i \overset{\mathrm{df}}{=} \min\{\, \delta_i(x_i, y_i) \,|\, \boldsymbol{x} = (x_1 \ldots x_d),\ \boldsymbol{y} = (y_1 \ldots y_d) \in \mathbb{D}\,\}$$

is the diversity in the i-th dimension. In principle, it is now again possible to apply the concept of Pareto optimality, i.e., to define a preference relation $\succeq$ on subsets of cases by $\mathbb{D} \succeq \mathbb{D}'$ iff $\mathrm{div}(\mathbb{D}) \geq \mathrm{div}(\mathbb{D}')$, and to look for Pareto optimal subsets of $\mathbb{S}$. However, this Pareto set will also include subsets that are very dissimilar in some dimensions but not at all dissimilar in others. From a diversity point of view, this is not desirable. To find subsets that are as "uniformly" diverse as possible, we therefore propose the following strategy: Suppose that a user wants to get a diverse subset of size K, which means that the set of candidates is given by the set of all subsets $\mathbb{D} \subseteq \mathbb{S}$ with $|\mathbb{D}| = K$. Moreover, for dimension i, consider the ranking of all candidate subsets $\mathbb{D}$ in descending order according to their diversity v_i in that dimension, and let $r_i(\mathbb{D})$ be the rank of $\mathbb{D}$. We then evaluate a candidate subset $\mathbb{D}$ by

$$\mathrm{val}(\mathbb{D}) \overset{\mathrm{df}}{=} \max\{\, r_i(\mathbb{D}) \,|\, i = 1 \ldots d\,\},$$

and the goal is to find a subset minimizing this criterion. Note that the latter is a minimax-solution, that is, a subset which minimizes its worst position in the d rankings; Fig. 4 gives an illustration. Interestingly, the above idea has recently been proposed independently under the name "ranking dominance" in the context of multi-criteria optimization [12].

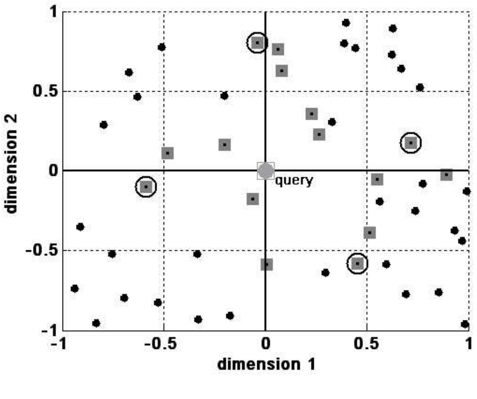


Fig. 4. A set of cases represented as points, the similarity skyline (boxes), and a diverse subset of size 4 (encircled boxes).

Algorithmically, we solve the problem as follows. For every pair of cases $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}$ and for each dimension i, one can precompute the rank $r_i(x_i, y_i)$ of their distance $\delta_i(x_i, y_i)$. For a fixed $v \in \mathbb{N}$, define a graph G_v as follows: the node set is $\mathbb{S}$, and for each $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}$, an edge is inserted in G_v if $r_i(x_i, y_i) \leq v$. Obviously, a subset $\mathbb{D}$ with $\mathrm{val}(\mathbb{D}) \leq v$ corresponds to a K-clique in G_v. The optimization problem can thus be solved by finding the minimal $v \in \mathbb{N}$ such that G_v contains a K-clique.

Unfortunately, the K-clique problem is known to be NP-hard [13]. Nevertheless, there exist good heuristics. In our approach, we use a method similar to the one proposed in [14]. Moreover, to find the minimal value v, we employ the bisection method with lower bound 1 and upper bound v_{max}, where v_{max} is guessed at the beginning (and probably increased if $G_{v_{max}}$ does not contain a K-clique). Essentially, this means that the number of search steps is logarithmic in v_{max}.

We conclude this section by noting that a diverse subset $\mathbb{D}$ can be taken as a point of departure for "navigating" within a similarity skyline. For example, a user may identify one case $\boldsymbol{x} \in \mathbb{D}$ as being most interesting. Then, one could "zoom" into that part of the skyline by retrieving another subset of cases from the skyline that are as similar to $\boldsymbol{x}$ as possible, using a criterion quite similar to the one used for diversity computation. Such extensions are being investigated in ongoing work.

5 Similarity Skyline for Uncertain Data

Motivated by our main application scenario, we have extended the concept of a similarity skyline to the case of uncertain data. In fact, the problem of uncertain and imprecisely known attribute values is quite obvious for archaeological data, though it is of course not restricted to this application field. Besides, note that the query itself is often imprecise. For example, consider a user looking for a case which is maximally similar to an "ideal" case, which is given as a query. This ideal case can be fictitious, and the user may prefer to specify it in terms of imprecise or fuzzy features like "a prize of about 1,200 dollars".

5.1 Uncertainty Modeling

Perhaps the most simple approach to handling imprecise attribute values is to use an interval-based representation: Each attribute value is characterized in terms of an interval that is assumed to cover the true but unknown value. For example, the unknown age at death of a person could be specified in terms of the interval $[25, 45]$.

An interval of the form $[a, b]$ declares some values to be *possible* or plausible, namely those between a and b, and excludes others as being *impossible*, namely those outside the interval. A well-known and quite obvious disadvantage of the interval-based approach is the abrupt transition between the range of possible

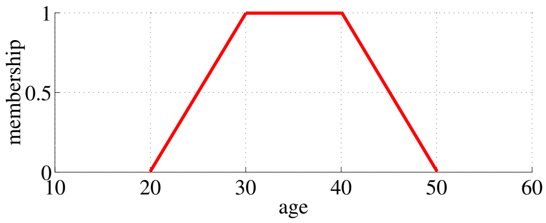


Fig. 5. Example of a fuzzy set modeling the linguistic concept "middle-aged".

and impossible values. In the above example, the age of 45 is considered as fully plausible, while 46 years is definitely excluded.

Another approach to uncertainty modeling, which often appears to be more appropriate, is to characterize the set of possible values of an attribute X_i in terms of a fuzzy subset of the attribute's domain $\mathbb{X}_i$, that is, by a mapping $F : \mathbb{X}_i \rightarrow [0, 1]$. Adopting a semantic interpretation of membership degrees in terms of *degrees of plausibility*, a fuzzy set F can be associated with a possibility distribution π_F: For every $x \in \mathbb{X}_i$, $\pi_F(x) = F(x)$ corresponds to the degree of plausibility that x equals the true but unknown attribute value x_i. A possibility distribution thus allows one to express that a certain value x is neither completely plausible nor completely impossible, but rather possible *to some degree*. For example, given the information that a person was middle-aged, all ages between 30 and 40 may appear fully plausible, which means that $\pi_F(x) = 1$ for $x \in [30, 40]$. Moreover, all ages below 20 or above 50 might be completely excluded, i.e., $\pi_F(x) = 0$ for $x \leq 20$ and $x \geq 50$. All values in-between these regions are possible to some degree. The simplest way to model a gradual transition between possibility and impossibility is to use a linear interpolation, which leads to the commonly employed trapezoidal fuzzy sets (see Fig. 5). According to this model, $\pi_F(25) = 0.5$, i.e., an age of 25 is possible to the degree 0.5.

A possibility distribution π_F induces two important measures, namely a *possibility* and a *necessity* measure:

$$\Pi_F : 2^{\mathbb{X}_i} \rightarrow [0, 1],\ A \mapsto \sup_{x \in A} \pi_F(x)$$

$$N_F : 2^{\mathbb{X}_i} \rightarrow [0, 1],\ A \mapsto 1 - \sup_{x \notin A} \pi_F(x)$$

For each subset $A \subseteq \mathbb{X}_i$, $\Pi_F(A)$ is the degree of plausibility that $x_i \in A$. Moreover, $N(A)$ is the degree to which x_i is necessarily in A. The measures Π_F and N_F are dual in the sense that $\Pi_F(A) \equiv 1 - N_F(\mathbb{X} \backslash A)$. To verbalize, x_i is possibly in A as long as it is not necessarily in the complement $\mathbb{X} \setminus A$.

5.2 Transformation for Fuzzy Attribute Values

As outlined above, a first step of our approach consists of mapping a data point $\boldsymbol{x} = (x_1 \dots x_d) \in \mathrm{CB}$ to the "distance space". According to (6), every attribute value x_i is replaced by its distance $x'_i = \delta_i(x_i, z_i)$ to the corresponding value of the query case $\boldsymbol{z} = (z_1 \ldots z_d)$.

When both x_i and z_i are characterized in terms of fuzzy sets F_i and G_i, respectively, the distance x'_i becomes a fuzzy quantity F'_i as well. It can be derived by applying the well-known extension principle to the distance δ_i [15]:

$$F'_i(d) \;=\; \sup\{\, \min(F_i(x_i), G_i(z_i)) \,|\, \delta_i(x_i, z_i) = d\,\} \tag{7}$$

5.3 The Dominance Relation for Fuzzy Attribute Values

The definition of the skyline of a set of data points involves the concept of dominance. In the case of similarity queries, dominance refers to distance, i.e., a value x_i (weakly) dominates a value y_i if $x_i \leq y_i$. If the data is uncertain, an obvious question is how to extend this concept of dominance to attribute values characterized in terms of intervals or fuzzy sets. This question is non-trivial, since neither the class of intervals nor the class of fuzzy subsets of a totally ordered domain are endowed with a natural order.

Consider two objects (transformed cases) $\boldsymbol{x} = (x_1 \ldots x_d)$ and $\boldsymbol{y} = (y_1 \ldots y_d)$, and suppose that the true distance values x_i and y_i are characterized in terms of fuzzy sets F_i and G_i, respectively (derived according to (7)). The problem is now to extend the dominance relation so as to enable the comparison of two fuzzy vectors $\boldsymbol{F} = (F_1 \ldots F_d)$ and $\boldsymbol{G} = (G_1 \ldots G_d)$.

Let π_{F_i} and π_{G_i} denote, respectively, the possibility distributions associated with the fuzzy sets F_i and G_i. If these distributions can be assumed to be non-interactive, the degree of possibility and the degree of necessity of the event $x_i \leq y_i$ are given, respectively, by

$$\begin{aligned} p_i &= \mathit{\Pi}(x_i \leq y_i) = \sup_{x \leq y} \min(\pi_{F_i}(x), \pi_{G_i}(y)), \\ n_i &= N(x_i \leq y_i) = 1 - \sup_{x > y} \min(\pi_{F_i}(x), \pi_{G_i}(y)) \ . \end{aligned}$$

Since the dominance relation requires dominance for *all* dimensions, these degrees have to be combined conjunctively. To this end, one can refer to a t-norm as a generalized logical conjunction [16]. Using the minimum operator for this purpose, one eventually obtains two degrees p and n, such that

$$p = \min(p_1 \ldots p_d) \geq \min(n_1 \ldots n_d) = n \ ,$$

which correspond, respectively, to the degree of possibility and the degree of necessity that the first object ($\boldsymbol{x}$) dominates the second one ($\boldsymbol{y}$). Thus, the (fuzzy) dominance relation between $\boldsymbol{x}$ and $\boldsymbol{y}$ is now expressed in terms of a possibility/necessity interval:

$$\mathrm{FDOM}(\boldsymbol{x}, \boldsymbol{y}) \;=\; [n, p] \tag{8}$$

In principle, it would now be possible to use this "fuzzy" conception of dominance to define a kind of fuzzy skyline. More specifically, for each object $\boldsymbol{x}$ one could derive a degree of possibility and a degree of necessity for $\boldsymbol{x}$ to be an element of the skyline. A less complex alternative is to "defuzzify" the dominance relation first, and to compute a standard skyline afterward. Defuzzifying means replacing fuzzy dominance by a standard (non-fuzzy) dominance relation, depending on the two degrees p and n. Of course, this can be done in different ways, for example by thresholding:

$$\boldsymbol{x} \succ \boldsymbol{y} \overset{\text{df}}{\iff} n \geq \alpha \text{ and } p \geq \beta \ , \tag{9}$$

where $0 \leq \alpha \leq \beta \leq 1$. If α is small while $\beta = 1$, this means that $\boldsymbol{x} \succ \boldsymbol{y}$ iff dominance is considered fully plausible and also necessary to some extent. In fact, for $\beta = 1$, (9) has an especially intuitive meaning: A fuzzy interval F_i dominates a fuzzy interval G_i if the $(1-\alpha)$-cut of F_i, which is the interval $[f^l_{1-\alpha}, f^u_{1-\alpha}] = \{x_i \,|\, F_i(x_i) \geq 1-\alpha\}$, dominates the $(1-\alpha)$-cut of G_i, $[g^l_{1-\alpha}, g^u_{1-\alpha}]$, in the sense that the former precedes the latter, i.e., $f^u_{1-\alpha} < g^l_{1-\alpha}$. The dominance relation hence tolerates a certain overlap of the fuzzy intervals, and the degree of this overlap depends on α; see Fig. 6 for an illustration.

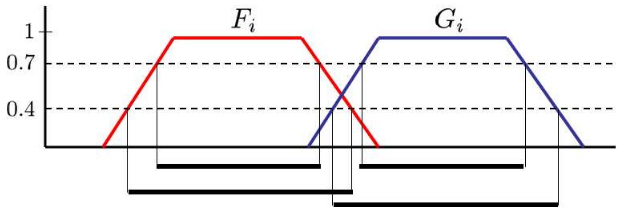


Fig. 6. Example in which the dominance relation (9) holds for $\alpha = 0.3$ (and $\beta = 1$) but not for $\alpha = 0.6$. In the latter case, the $(1-\alpha)$-cuts of F_i and G_i intersect.

As suggested by this example, the thresholds α and β can be used to make the dominance relation more or less restrictive and, thereby, to influence the size of the skyline: If α and β are increased, the dominance relation will hold for fewer objects, which in turn means that the skyline grows. In this regard, also note that α and β must satisfy certain restrictions in order to guarantee that $\boldsymbol{x} \succ \boldsymbol{y}$ and $\boldsymbol{x} \succ \boldsymbol{y}$ cannot hold simultaneously. Since $\text{FDOM}(\boldsymbol{y}, \boldsymbol{x}) = [1-p, 1-n]$, a reasonable restriction excluding this case is $\alpha + \beta > 1$.

6 Experiments

The get a first idea of the efficacy and scalability of our approach, we have conducted a number of experiments. In particular, we investigated how many cases are found to be *similar* to a query depending on the dimensionality of the case base and the strictness of the dominance relation (9), that we used for different values of α (while β was fixed to 1). Moreover, we addressed the issues of run time and scalability. Since the original data in the current version of our archaeological database is interval data, we turned intervals into fuzzy sets with triangular membership functions, using the mid-point of an interval as the core (center point) of the corresponding fuzzy set.

From the original 9-dimensional case base, 22 test sets of different dimension were constructed by projecting to corresponding subsets of the attributes. Each case of a case base CB was used as a query resulting in a total number of $n = |\,\text{CB}\,|$ queries. For the corresponding n answer sets (skylines), we derived the average and the standard deviation of the relative size of answer set (number divided by n); see Fig. 7. Likewise, the average run time and its standard deviation were measured; see Fig. 8. Finally, Fig. 9 shows run time results for the computation of diverse subsets of size 5, depending on the size of the original skyline.

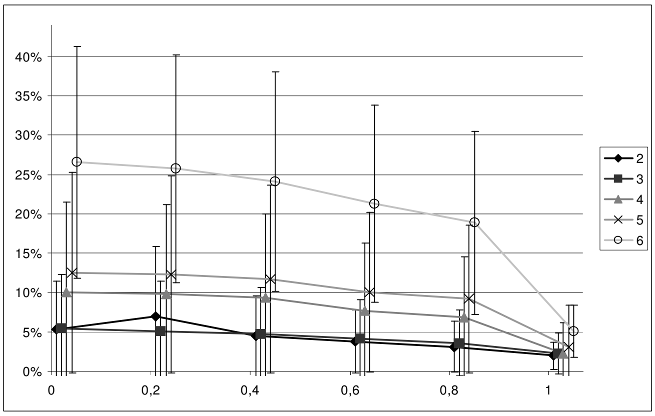


Fig. 7. Mean and standard deviation of the relative size of answer sets (y-axis) depending on the dimension (2–6) and the strictness level α (x-axis).

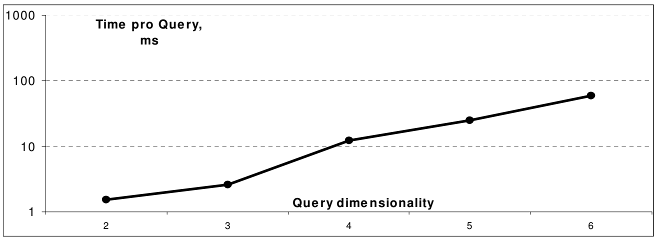


Fig. 8. Run time for skyline computation depending on the dimensionality of the case base.

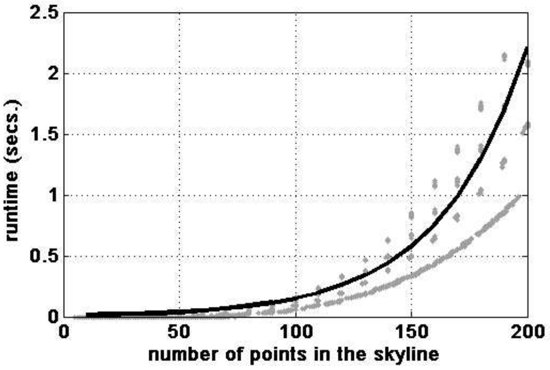


Fig. 9. Run time for the computation of diverse subsets of size 5 and dimensions 2–15 depending on the size of the original skyline.

As it was to be expected, the cardinality of the answer set critically depends on the dimensionality of the case base and the strictness of the dominance relation. Run time increases correspondingly but remains satisfactory even for high-dimensional queries (171 ms on average for a 9-dimensional query). Similar remarks apply to the computation of diverse subsets.

In summary, our results confirm theoretical findings showing that the complexity of skyline computation, like most other retrieval techniques, critically depends on the dimension of a data set, in the worst case exponentially. Still, the results also show that problems of reasonable size (the number of features deemed relevant by a user in a similarity query is typically not very large) can be handled with an acceptable cost in terms of run time.

7 Conclusions

Motivated by an application in the field of archeology, we have proposed a new approach to similarity search. Our method is based on the concept of Pareto dominance and, taking an example case as a reference point, seeks to find objects that are maximally similar in a Pareto sense. It is especially user-friendly, as it does not expect the specification of a global similarity or distance function. Our first experiences are promising, and so far we received quite positive feedback from users.

Again motivated by our application, we have extended the computation of a similarity skyline to the case of uncertain (fuzzy) data. Apart from advantages with respect to modeling and knowledge representation, the fuzzy extension also allows for controlling the size of answer sets: Since one object can dominate another one "to some degree", the (non-fuzzy) dominance relation can be specified in a more or less stringent way. This effect is clear from our experimental results.

We believe that similarity search based on Pareto dominance is of general interest for CBR, and we see this paper as a first step to popularize this research direction. Needless to say, a lot of open problems remain to be solved. For example, as Pareto dominance is a rather weak preference relation, the number of cases "maximally similar" to the query can become quite large. Implementing additional filter strategies, such as diverse subset computation, is one way to tackle this problem. Another direction is to refine Pareto dominance, so that it discriminates more strongly between cases. This is a topic of ongoing work.

References

1. Vladimirskiy, I., Hüllermeier, E., Stauch, E.: Similarity search over uncertain archaeological data using a modified skyline operator. In Wilson, D., Khemani, D., eds.: Workshop Proceedings of ICCBR–07, Belfast, Northern Ireland (2007) 31–40
2. Aizerman, M., Aleskerov, F.: Theory of Choice. North-Holland, Amsterdam (1995)
3. Daniels, J., Rissland, E.: A case-based approach to intelligent information retrieval. In: Proc. 18th International ACM SIGIR Conference, Seattle, Washington, US (1995) 238–245
4. Roussopoulos, N., Kelley, S., Vincent, F.: Nearest neighbor queries. In: Proc. SIGMOD–95, New York, NY, USA (1995) 71–79
5. Weber, R., Schek, H., Blott, S.: A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In: Proc. VLDB–98, San Francisco, CA, USA (1998) 194–205
6. Richter, M.: Foundations of similarity and utility. In: Proc. FLAIRS-20, The 20th International FLAIRS Conference, Key West, Florida (2007)
7. Cunningham, P.: A taxonomy of similarity mechanisms for case-based reasoning. Technical Report UCD-CSI-2008-01, University College Dublin (2008)
8. Borzsony, S., Kossmann, D., Stocker, K.: The skyline operator. In: Proc. 17th International Conference on Data Engineering, San Jose, California, USA (2001) 421–430
9. Papadias, D., Tao, Y., Fu, G., Seeger, B.: Progressive skyline computation in database systems. ACM Transactions on Database Systems **30**(1) (2005) 41–82
10. McSherry, D.: Diversity-conscious retrieval. In Craw, S., Preece, A., eds.: Advances in Case-Based Reasoning. Number 2416 in LNAI. Springer-Verlag, Berlin, Heidelberg (2002) 219–233
11. McSherry, D.: Increasing recommendation diversity without loss of similarity. Expert Update **5** (2002) 17–26
12. Kukkonen, S., Lampinen, J.: Ranking-dominance and many-objective optimization. In: IEEE Congress on Evolutionary Computation, Singapore (2007) 3983–3990
13. Pardalos, P., Xue, J.: The maximum clique problem. Journal of Global Optimization **4**(3) (1994) 301–328
14. Tomita, E., Kameda, T.: An efficient branch-and-bound algorithm for finding a maximum clique with computational experiments. Journal of Global Optimization **37**(1) (2007) 95–111
15. Zadeh, L.: The concept of a linguistic variable and its applications in approximate reasoning. Information Science **8** (1975) 199–251
16. Klement, E., Mesiar, R., Pap, E.: Triangular Norms. Kluwer Academic Publishers (2002)