Exception-Tolerant Skyline Queries

Hélène Jaudoin, Olivier Pivert, and Daniel Rocacher

Université de Rennes 1, Irisa
Technopole Anticipa 22305 Lannion Cedex, France
{jaudoin,pivert,rocacher}@enssat.fr

Abstract. This paper presents an approach aimed at reducing the impact of exceptional points/outliers when computing skyline queries. The phenomenon that one wants to avoid is that noisy or suspect elements "hide" some more interesting answers just because they dominate them in the sense of Pareto. The approach we propose is based on the fuzzy notion of typicality and makes it possible to distinguish between genuinely interesting points and potential anomalies in the skyline obtained.

Keywords: skyline query, exception, gradual approach.

1 Introduction

In this paper, a qualitative view of preference queries is chosen, namely the *Skyline* approach introduced by Börzsönyi *et al.* [2]. Given a set of points in a space, a skyline query retrieves those points that are not dominated by any other in the sense of Pareto order. When the number of dimensions on which preferences are expressed gets high, many tuples may become incomparable. Several approaches have been proposed to define an order for two incomparable tuples, based on the number of other tuples that each of the two tuples dominates (notion of k-representative dominance proposed by Lin *et al.* [12]), on a preference order of the attributes (see for instance the notions of k-dominance and k-frequency introduced by Chan *et al.* [3,4]), or on a notion of representativity ([14] redefines the approach proposed by [12] and proposes to return only the more representative points of the skyline, i.e., a point among those present in each cluster of the skyline points). Other approaches fuzzify the concept of skyline in different ways, see e.g. [9]. Here, we are concerned with a different problem, namely that of the possible presence of *exceptional* points, aka outliers, in the dataset over which the skyline is computed. Such exceptions may correspond to noise or to the presence of *nontypical* points in the collection considered. The impact of such points on the skyline may obviously be important if they dominate some other, more representative ones.

Two strategies can be considered to handle exceptions. The former consists in removing anomalies by adopting cleaning procedures or defining data entry constraints. However, the task of automatically distinguishing between odd points and simply exceptional points is not always easy. Another solution is to define an approach that is *tolerant to exceptions*, that highlights representative points of the database and that points out the possible outliers. In this paper, we present

A. Laurent et al. (Eds.): IPMU 2014, Part III, CCIS 444, pp. 120–129, 2014.

such an approach based on the fuzzy notion of typicality [15]. We revisit the definition of a skyline and show that it i) makes it possible to retrieve the dominant points without discarding other potentially interesting ones, and ii) constitutes a flexible tool for visualizing the answers.

The remainder of the paper is structured as follows. Section 2 provides a refresher about skyline queries and motivates the approach. Section 3 presents the principle of exception-tolerant skyline queries, based on the fuzzy concept of typicality. Section 4 gives the main implementation elements of the approach whereas Section 5 presents preliminary experimental results obtained on a real-world dataset. Finally, Section 6 recalls the main contributions and outlines perspectives for future work.

2 Refresher about Skyline Queries and Motivations

Let $\mathcal{D} = \{D_1, \ldots, D_d\}$ be a set of d dimensions. Let us denote by $dom(D_i)$ the domain associated with dimension D_i. Let $\mathcal{S} \subseteq dom(D_1) \times \ldots \times dom(D_d)$, p and q two points of $\mathcal{S}$, and $\succ_i$ an order on D_i. One says that p *dominates* q on $\mathcal{D}$ (p is better than q according to Pareto order), denoted by $p \succ_{\mathcal{D}} q$, iff

$$\forall i \in [1, d] : \ p_i \succeq_i q_i \text{ and } \exists j \in [1, d] : \ p_i \succ_i q_i.$$

A skyline query on $\mathcal{D}$ applied to a set of points $\mathcal{S}$, whose result is denoted by $\textsc{Sky}_{\mathcal{D}}(\mathcal{S})$, according to order relations $\succ_i$, produces the set of points that are not dominated by any other point of $\mathcal{S}$:

$$\textsc{Sky}_{\mathcal{D}}(\mathcal{S}) = \{p \in \mathcal{S} \mid \nexists q \in \mathcal{S} : q \succ_{\mathcal{D}} p\}$$

Depending on the context, one may try, for instance, to maximize or minimize the values of $dom(D_i)$, assuming that $dom(D_i)$ is a numerical domain.

In order to illustrate the principle of the approach we propose, let us consider the dataset *Iris* [8], graphically represented in Figure 1.

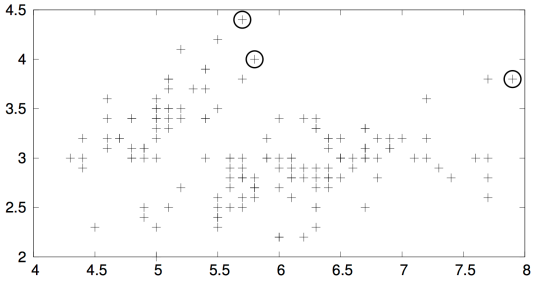

Fig. 1. The dataset *Iris*

The vertical axis corresponds to the attribute *sepal width* whereas the horizontal axis is associated with *sepal length*. The skyline query:

```
select * from iris
skyline of sepallength max, sepalwidth max
```

looks for those points that maximize the dimensions *length* and *width* of the sepals (the circled points in Figure 1).

In this dataset, the points form two groups that respectively correspond to the intervals [4, 5.5] and [5.5, 7] on attribute *length*. By definition, the skyline points are on the border of the region that includes the points of the dataset. However, these points are very distant from the areas corresponding to the two groups and are thus not very representative of the dataset. It could then be interesting for a user to be able to visualize the points that are "almost dominant", closer to the clusters, then more representative of the dataset. The notion of typicality discussed in the next section makes it possible to reach that goal.

2.1 Computing a Fuzzy Set of Typical Values

The *typicality* of an element in a set indicates the extent to which this element is similar to many other points from the set. The notion of fuzzy typicality has been much studied in the contexts of data summaries and approximate reasoning. Zadeh [15] states that x is a typical element of a fuzzy set A iff i) x has a high membership degree to A and ii) *most of the elements of A are similar to x*. In the case where A is a crisp set — as it will be the case in the following —, the definition becomes: x is in A and most of the elements of A are similar to x.

In [7], the authors define a typicality index based on frequency and similarity. We adapt their definition as follows. Let us consider a set $\mathcal{E}$ of points. We say that a point is all the more typical as it is close to many other points. The proximity relation considered is based on Euclidean distance. We consider that two points p_1 and p_2 are close to each other if $d(p_1, p_2) \leq \tau$ where τ is a predefined threshold. In the experiment performed on the dataset *Iris*, we used $\tau = 0.5$. The frequency of a point is defined as:

$$F(p) = \frac{|\{p_i \in \mathcal{E},\, d(p, p_i) \leq \tau\}| - 1}{|\mathcal{E}|}. \tag{1}$$

This degree is then normalized into a typicality degree in $[0, 1]$:

$$typ(p) = \frac{F(p)}{\max_{p_i \in \mathcal{E}}\{F(p_i)\}}.$$

We will also use the following notations:

$$\textsc{Typ}(\mathcal{E}) = \{typ(p)/p \mid p \in \mathcal{E}\}$$

$$\textsc{Typ}_\gamma(\mathcal{E}) = \{p \mid p \in \mathcal{E} \text{ and } typ(p) \geq \gamma\}.$$

$\textsc{Typ}(\mathcal{E})$ represents the fuzzy set of points that are somewhat typical of the set $\mathcal{E}$ while $\textsc{Typ}_\gamma(\mathcal{E})$ gathers the points of $\mathcal{E}$ whose typicality is over the threshold γ. An excerpt of the typicality degrees computed over the *Iris* dataset is presented in Table 1.

Table 1. Excerpt of the *Iris* dataset with associated typicality degrees

length	*width*	*frequency*	*typicality*
7.4	2.8	0.0600	0.187
7.9	3.8	0.0133	0.0417
6.4	2.8	0.253	0.792
6.3	2.8	0.287	0.896
6.1	2.6	0.253	0.792
7.7	3.0	0.0467	0.146
6.3	3.4	0.153	0.479
6.4	3.1	0.293	0.917
6.0	3.0	0.320	1.000

3 Principle of the Exception-Tolerant Skyline

As explained in the introduction, our goal is to revisit the definition of the skyline so as to take into account the typicality of the points in the database, in order to control the impact of exceptions or anomalies.

3.1 Boolean View

A first idea is to restrict the computation of the skyline to a subset of $\mathcal{E}$ that corresponds to sufficiently typical points. The corresponding definition is:

$$\textsc{Sky}_{\mathcal{D}}(\textsc{Typ}_{\gamma}(\mathcal{S})) = \{p \in \textsc{Typ}_{\gamma}(\mathcal{S}) \mid \nexists q \in \textsc{Typ}_{\gamma}(\mathcal{S}) \text{ such that } q \succ_{\mathcal{D}} p\} \quad (2)$$

Such an approach obviously reduces the cost of the processing since only the points that are typical at least to the degree γ are considered in the calculus. However, this definition does not make it possible to discriminate the points of the result according to their degree of typicality since the skyline obtained is a crisp set. Figure 2 illustrates this behavior and shows the maxima (circled points) obtained when considering the points that are typical to a degree ≥ 0.7 (represented by crosses).

Another drawback of this definition is to exclude the nontypical points altogether, even though some of them could be interesting answers. A more cautious definition consists in keeping the nontypical points while computing the skyline and transform Equation (2) into:

$$\textsc{Sky}_{\mathcal{D}}(\textsc{Typ}_{\gamma}(\mathcal{S})) = \{p \in \mathcal{S} \mid \nexists q \in \textsc{Typ}_{\gamma}(\mathcal{S}) \text{ such that } q \succ_{\mathcal{D}} p\} \quad (3)$$

Figure 3 illustrates this alternative solution. It represents (circled points) the objets from the *Iris* dataset that are not dominated by any item typical to the degree $\gamma = 0.7$ at least (represented by crosses).

With Equation (2), the nontypical points are discarded, whereas with Equation (3), the skyline is larger and includes nontypical extrema. This approaches relaxes skyline queries in such a way that the result obtained is not a line anymore but a stripe composed of the regular skyline elements completed with

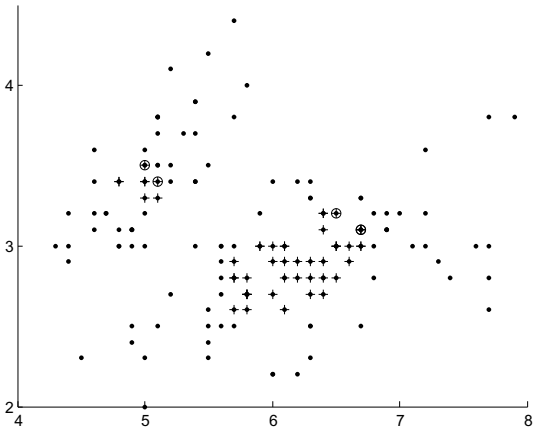

Fig. 2. Skyline of the Iris points whose typicality degree is ≥ 0.7

possible "substitutes". However, the main drawbacks of this definition are: i) the potentially large number of points returned, ii) the impossibility to distinguish, among the skyline points, those that are not at all dominated from those that *are* dominated (by weakly typical points).

3.2 Gradual View

A third version makes it possible to compute a *graded* skyline, seen as a fuzzy set, that preserves the gradual nature of the concept of typicality. By doing so, no threshold (γ) is applied to typicality degrees. The definition is as follows:

$$\begin{aligned} &\textsc{Sky}_{\mathcal{D}}(\textsc{Typ}(\mathcal{S})) = \\ &\quad = \{\mu/p \mid p \in \mathcal{S} \wedge \mu = \min_{q \in \mathcal{S}} (\max(1 - \mu_{\textsc{Typ}}(q),\ deg(\neg(q \succ_{\mathcal{D}} p)))\} \end{aligned} \tag{4}$$

where $deg(\neg(q \succ_{\mathcal{D}} p)) = 1$ if q does not dominate p (i.e., $(q \succ_{\mathcal{D}} p)$ is false), 0 otherwise. A point totally belongs to the skyline (membership degree equal to 1) if it is dominated by no other point. A point does not belong at all to the skyline (membership degree equal to 0) if it is dominated by at least one totally typical point. In the case where p is dominated by somewhat (but not totally) typical points, its degree of membership to the skyline depends on the typicality of these points. Equation (4) may be rewritten as follows:

$$\textsc{Sky}_{\mathcal{D}}(\textsc{Typ}(\mathcal{S})) = \{\mu/p \mid p \in \mathcal{S} \wedge \mu = 1 - \max_{q \in \mathcal{S} \mid q \succ_{\mathcal{D}} p} (\mu_{\textsc{Typ}}(q))\}. \tag{5}$$

With the *Iris* dataset, one gets the result presented in Figure 4, where the degree of membership to the skyline corresponds to the z axis. As expected, the points of the classical skyline totally belong to the graded skyline, along with some additional answers that more or less belong to it. This approach appears

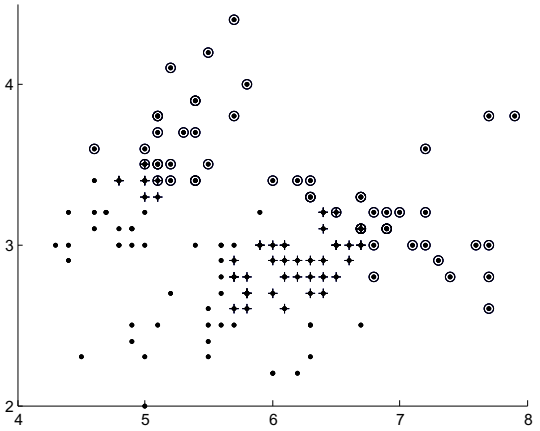

Fig. 3. Points that are not dominated by any other whose typicality degree is ≥ 0.7

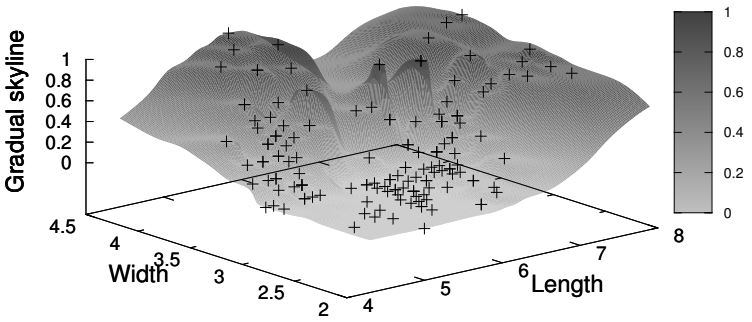

Fig. 4. Graded skyline obtained with the *Iris* dataset

interesting in terms of visualization. Indeed, the score associated with each point makes it possible to focus on different α-cuts of the skyline. In Figure 4, one may notice a slope from the optimal points towards the less typical or completely dominated ones. The user may select points that are not necessarily optimal but that represent good alternatives to the regular skyline answers (in the case, for instance, where the latter look "too good to be true"). Finally, an element of the graded skyline is associated with two scores: a degree of membership to the skyline, and a typicality degree (that expresses the extent to which it is *not* exceptional). One may imagine different ways of navigating inside areas in order to explore the set of answers: a simple scan for displaying the characteristics of the points, the use of different filters aimed, for instance, at optimizing diversity on certain attributes, etc.

4 Implementation Aspects

Two steps are necessary for obtaining the graded result: i) computation of the typicality degrees, and ii) computation of the skyline. Many algorithms have been proposed for processing skyline queries: Block-Nested-Loops(BNL) [2]; Divide and Conquer [2]; a technique exploiting a B-tree or an R-tree [2]; an algorithm based on bitmap indices [13]; an improvement of the BNL method, named Sort-Filter-Skyline [5,6], and a strategy proposed in [1] that relies on a preordering of the tuples aimed to limit the number of elements to be accessed and compared. We have based our implementation on the approach proposed in [13] with which Formula (5) appears the easiest to evaluate.

The data structure underlying the algorithm described in [13] is an array of Boolean values or *bitmap*. A bitmap index is defined for each dimension of the skyline: every column corresponds to a possible value in the dimension considered, and every row references a tuple from the database. Value 1 at the intersection of row l and column c means that the tuple referenced in row l has the value corresponding to column c. Then, every point p of the database $\mathcal{S}$ is tested in order to determine if it belongs to the skyline or not. For doing so, two other data structures are created. The first one, denoted by A, gathers *the tuples that are as good as* p on every dimension, the second one, denoted by B, contains *the tuples that are better than* p on at least one dimension. A and B are defined as two-dimensional tables of Booleans whose columns are associated with the tuples of $\mathcal{S}$. They are initialized using the bitmap indices.

Algorithm 1 constitutes the heart of our prototype and follows the principle described above. We have also used three tables ($\mathcal{T}$, Sky_{grad}, A') where each column corresponds to a tuple of $\mathcal{S}$, that contain real numbers in [0, 1]. In $\mathcal{T}$, these numbers correspond to typicality degrees whereas in Sky_{grad} they represent degrees of membership to the graded skyline. *AND* corresponds to the logical conjunction between the pairs of values ($A[i]$, $B[i]$) so as to check whether the point i is both as good as p on all dimensions and better than p on one dimension. If it is the case, then this point dominates p. *MULT* is used to compute the product of the values $A[i]$ and $\mathcal{T}[i]$, which makes it possible to associate a point i with its typicality degree if it dominates the point p considered. Finally, *MAX* returns the maximal value of the array A. The membership degree of p to the graded skyline is then obtained by means of Formula (5).

The computation of the typicality degrees uses a threshold on the distance that depends on the attributes involved in the skyline query. The complexity of the computation is obviously in $\theta(n^2)$ since one has to compute the Euclidean distance between each pair of points of the dataset.

One may think of two ways for computing typicality: either on demand, on the attributes specified in the *skyline* clause of the query, or beforehand, on different relevant sets of attributes. The first method implies an additional cost to the evaluation of skyline queries, whereas the second one makes it necessary to update the typicality degrees when the dataset evolves. In any case, indices such as those used for retrieving k nearest neighbors (for instance kdtrees) may be exploited. Furthermore, since in general the computation of the graded

Algorithm 1. Main algorithm for computing the graded skyline

Require: d distance, n cardinality of the dataset $\mathcal{S}$, the points of the dataset $p \in \mathcal{S}$, the set of dimensions $\{d_i\}$
Ensure: graded skyline: $\forall p \in \mathcal{S},\ Sky_{grad}(p)$
  Preprocessing: creation of the bitmap indices on the d_i's
  Preprocessing: computation of the typicality of the points $\mathcal{T}$: $\forall p \in \mathcal{S},\ Typ(p)$
  for all $p \in \mathcal{S}$ **do**
    // Search for those points that dominate p
    Creation of A
    Creation of B
    $A := A$ AND B
    $A' := A$ MULT $\mathcal{T}$
    $Sky_{grad}(p) := 1 - Max(A')$
  end for

skyline concerns only a small fragment of the database, the extra cost related to typicality should not be too high.

It is worth emphasizing that the algorithm could be parallelized by partitioning the arrays A, B, A' and $\mathcal{T}$. Similarly, the creation of the structures A and B may be parallelized, provided that the bitmap indices and the typicality degrees are distributed or shared.

Table 2. Excerpt of the database and associated degrees (skyline and typicality)

id	*price*	*km*	*skyline*	*typicality*
1156771	6000	700	1	0.247
1596085	5800	162643	1	0.005
1211574	7000	500	1	0.352
1054357	1800	118000	1	0
1333992	500	220000	1	0
1380340	800	190000	1	0
891125	1000	170000	1	0
1229833	5990	10000	1	0.126
1276388	1300	135000	1	0
916264	5990	2514000	0.874	0
1674045	6000	3500	0.753	0.315

5 Experimental Results

The approach has been tested using a subset of the database of 845810 ads about second hand cars from the website *Le bon coin* [1] from 2012. The skyline query used hereafter as an example aims at minimizing both the price and the mileage. In the query considered, we focus on small urban cars with a regular (non-diesel) engine, which corresponds to 441 ads. Figure 5 shows the result obtained. In dark

[1] `www.leboncoin.fr`

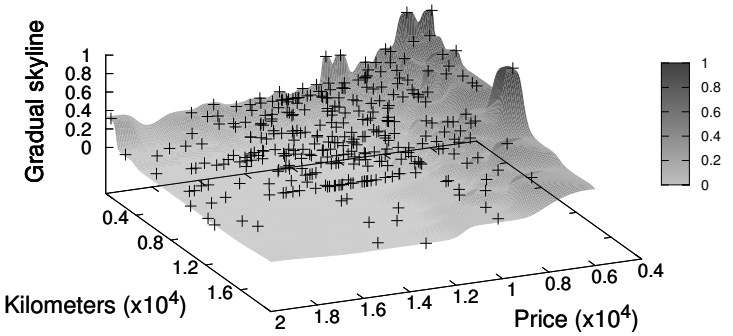

Fig. 5. 3D representation of the graded skyline

Table 3. Excerpt of the area [0.6, 0.8]

id	*price*	*km*	*skyline*	*typicality*
870279	6900	1000	0.716	0.358
981939	6500	4000	0.637	0.363
1022586	6500	7200	0.637	0.258
1166077	7750	2214	0.642	0.532
1208620	6500	3300	0.716	0.363
1267726	6500	100000	0.637	0
1334605	10500	500	0.647	0.642
1366336	7490	4250	0.637	0.516
1529678	7980	650	0.647	0.458
1635437	9900	590	0.647	0.621
1685854	7890	1000	0.642	0.458

grey are the points that belong the most to the skyline (membership degree between 0.8 and 1). These points are detailed in Table 2. According to the definition used, points dominated by others that are not totally typical belong to the result. It is the case for instance of ad number 916264 that is dominated by ads numbered 1054357 and 1229833. The identifiers in bold correspond to the points that belong to the regular skyline. One may observe that the points from Table 2 (area $[0.8, 1]$) are not very (or even not at all) typical. Moreover, certain features may not satisfy the user (the mileage can be very high, the price can be very low) and may look suspicious. On the other hand, Table 3, which shows an excerpt of the 0.6-cut of the graded skyline, contains more typical – thus more credible – points whose overall satisfaction remains high. Let us also mention that the time taken by the precomputation of the typicality degrees associated with the selected elements is twice as large (around 0.54 second) as the time devoted to the computation of the graded skyline (about 0.22 second). However, this result must taken carefully as the computation of typicality has not been optimized in the prototype yet.

6 Conclusion

In this paper, we have proposed a graded version of skyline queries aimed at controlling the impact of exceptions on the result (so as to prevent interesting points to be hidden because they are dominated by an exceptional one). An improvement could consist in using more sophisticated techniques for characterizing the points according to their level of representativity as a typicality-based clustering approach [11] or statistical methods for detecting outliers [10]. As a short-term perspective, we intend to carry out a parallel implementation of the algorithm and to use indexes for reducing the processing time devoted to the computation of typicality degrees.

References

1. Bartolini, I., Ciaccia, P., Patella, M.: Efficient sort-based skyline evaluation. ACM Trans. Database Syst. 33(4), 1–49 (2008)
2. Börzsönyi, S., Kossmann, D., Stocker, K.: The skyline operator. In: Proc. of ICDE 2001, pp. 421–430 (2001)
3. Chan, C., Jagadish, H., Tan, K., Tung, A., Zhang, Z.: Finding k-dominant skylines in high dimensional space. In: Proc. of SIGMOD 2006, pp. 503–514 (2006)
4. Chan, C.-Y., Jagadish, H.V., Tan, K.-L., Tung, A.K.H., Zhang, Z.: On high dimensional skylines. In: Ioannidis, Y., Scholl, M.H., Schmidt, J.W., Matthes, F., Hatzopoulos, M., Böhm, K., Kemper, A., Grust, T., Böhm, C. (eds.) EDBT 2006. LNCS, vol. 3896, pp. 478–495. Springer, Heidelberg (2006)
5. Chomicki, J., Godfrey, P., Gryz, J., Liang, D.: Skyline with presorting. In: Proc. of ICDE 2003, pp. 717–719 (2003)
6. Chomicki, J., Godfrey, P., Gryz, J., Liang, D.: Skyline with presorting: Theory and optimizations. In: Proc. of IIS 2005, pp. 595–604 (2005)
7. Dubois, D., Prade, H.: On data summarization with fuzzy sets. In: Proc. of IFSA 1993, pp. 465–468 (1984)
8. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics 7(2), 179–188 (1936)
9. Hadjali, A., Pivert, O., Prade, H.: On different types of fuzzy skylines. In: Kryszkiewicz, M., Rybinski, H., Skowron, A., Raś, Z.W. (eds.) ISMIS 2011. LNCS (LNAI), vol. 6804, pp. 581–591. Springer, Heidelberg (2011)
10. Hodge, V., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review 22(2), 85–126 (2004)
11. Lesot, M.: Typicality-based clustering. Int. J. of Information Technology and Intelligent Computing 12, 279–292 (2006)
12. Lin, X., Yuan, Y., Zhang, Q., Zhang, Y.: Selecting stars: the k most representative skyline operator. In: Proc. of the ICDE 2007, pp. 86–95 (2007)
13. Tan, K.L., Eng, P.K., Ooi, B.C.: Efficient progressive skyline computation. In: Proc. of VLDB 2001, pp. 301–310 (2001)
14. Tao, Y., Ding, L., Lin, X., Pei, J.: Distance-based representative skyline. In: Ioannidis, Y.E., Lee, D.L., Ng, R.T. (eds.) ICDE, pp. 892–903. IEEE (2009)
15. Zadeh, L.A.: A computational theory of dispositions. In: Wilks, Y. (ed.) COLING, pp. 312–318. ACL (1984)