Workshop on Mining Complex Data (MCD'05), ICDM, Houston, TX, 2005.

Hierarchical Density-Based Clustering for Multi-Represented Objects

Elke Achtert, Hans-Peter Kriegel, Alexey Pryakhin, Matthias Schubert

Institute for Computer Science, University of Munich
{achtert,kriegel,pryakhin,schubert}@dbs.ifi.lmu.de

Abstract

In recent years, the complexity of data objects in data mining applications has increased as well as their plain numbers. As a result, there exist various feature transformations and thus multiple object representations. For example, an image can be described by a text annotation, a color histogram and some texture features. To cluster these multi-represented objects, dedicated data mining algorithms have been shown to achieve improved results. In this paper, we will therefore introduce a method for hierarchical density-based clustering of multi-represented objects which is insensitive w.r.t. the choice of parameters. Furthermore, we will introduce a theoretical model that allows us to draw conclusions about the interaction of representations. Additionally, we will show how these conclusions can be used for defining a suitable combination method for multiple representations. To back up the usability of our proposed method, we present encouraging results for clustering a real world image data set that is described by 4 different representations.

1. Introduction

In modern data mining applications, the data objects are getting more and more complex. Thus, the extraction of meaningful feature representations yields a variety on different views on the same set of data objects. Each of these views or representations might focus on a different aspect and may offer another notion of similarity. However, in almost any application there is no universal feature representation that can be used to express similarity between all possible objects in a meaningful way. Thus, recent data mining approaches employ multiple representations to achieve more general results that are based on a variety of aspects. An example application for multi-represented objects is data mining in protein data. A protein can be described by multiple feature transformations based upon its amino acid sequence, its secondary or its three dimensional structure. Another example is data mining in image data which might be represented by texture features, color histograms or text annotations.

Mining multi-represented objects yields advantages because more information can be incorporated into the mining process. On the other hand, the additional information has to be used carefully since too much information might distort the derived patterns. Basically, we can distinguish two problems when clustering multi-represented objects, comparability and semantics.

The comparability problem subsumes several effects when comparing features, distances or statements from different representations. For example, a distance value of 1,000 might indicate similarity in some feature space and a distance of 0.5 might indicate dissimilarity in another space. Thus, directly comparing the distances is not advisable.

Other than the comparability problem the semantics problem is caused by differences between the knowledge that can be derived from each representation. For example, two images described by very similar text annotations are very likely to be very similar as well. On the other hand, if the words describing two images are completely disjunctive the implication that both images are dissimilar is rather weak because it is possible to describe the same object using a completely different set of words. Another type of semantics can be found in color histograms. An image of a plane in blue skies might provide the same color distribution as a sailing boat in the water. However, if two color images have completely different colors, it is usually a strong hint that the images are really dissimilar.

To cluster multi-represented objects with respect to both problems, [1] described a multi-represented version of the density-based clustering algorithm DBSCAN. However, this approach is still very sensitive with respect to its parametrization. Therefore, we adapt the density-based, hierarchical clustering algorithm OPTICS to the setting of multi-represented objects. This new version of OPTICS is far less sensitive to parameter selection. Another problem of density-based multi-represented clustering is the handling of the semantic problem for multiple representations. In this paper, we will give a theoretical discussion of possible semantics and demonstrate how to combine the already proposed basic methods for more than two representations. Furthermore, we will explain the need of domain knowledge for judging each representation. To demonstrate the applicability of our approach, we will show the results of several experiments on an image data set that is given by 4 different representations.

The rest of this paper is organized as follows. Section 2 surveys related work in the areas of density-based clustering and multi-represented or multi view clustering. In section 3 the extension of OPTICS to multi-represented data is described. Section 4 starts with a formal discussion of the sematics problems and derives a theory for constructing meaningful distance functions for multi-represented objects. In our experimental evaluation in section 5, it is shown how the cluster quality can be improved using the proposed method. Finally, we will conclude the paper in section 6 with a short summary and some ideas for future work.

2. Related Work

DBSCAN [2] is a density-based clustering algorithm where clusters are considered as dense areas that are separated by sparse areas. Based on two input parameters ($\varepsilon \in \mathbb{R}$ and $k \in \mathbb{N}$), DBSCAN defines dense regions by means of core objects. An object $o \in DB$ is called *core object*, if its ε-neighborhood contains at least k objects. DBSCAN is able to detect arbitrarily shaped clusters by one single pass over the data. To do so, DBSCAN uses the fact, that a cluster can be detected by finding one of its core-objects o and computing all objects which are "density-reachable" from o. OPTICS [3] extends the density-connected clustering notion of DBSCAN by hierarchical concepts. In contrast to DBSCAN, OPTICS does not assign cluster memberships but computes a cluster order in which the objects are processed and additionally generates the information which would be used by an extended DBSCAN algorithm to assign cluster memberships. This information consists of only two values for each object, the core distance and the reachability distance. If the ε-neighborhood of an object o contains at least k objects, the core distance of o is defined as the k-nearest neighbor distance of o. Otherwise, the core distance is undefined. The reachability distance of an object p from o is an asymmetric distance measure that is defined as the maximum value of the core distance of o and the distance between p and o. Using these distances, OPTICS computes a "walk" through the data set and assigns to each object o its core distance and the smallest reachability distance w.r.t. all objects considered before o in the walk. In each step, OPTICS selects the object o having the minimum reachability distance to any already processed object. A special order of the database according to its density-based clustering structure is generated, the so-called cluster order, which can be displayed in a reachability plot. A reachability plot consists of the reachability distances on the y-axis of all objects plotted according to the cluster order on the x-axis. The "valleys" in the plot represent the clusters, since objects within a cluster have lower reachability distances than objects outside a cluster.

In [1] the idea of DBSCAN has been adapted to multi-represented objects. Two different methods have been proposed to decide wether a multi-represented object is a core object: the union and the intersection method. The union method assumes an object to be a core object if at least k objects are found within the union of its local ε-neighborhoods of each representation. The intersection method requires that at least k objects are within the intersection of all local ε-neighborhoods of each representation of a core object. In [4] an algorithm for spectral clustering of multi-represented objects is proposed. The author proposes to calculate the clustering in a way that the disagreement between the cluster models in each representation is minimized. In [5] a version of Expectation Maximization (EM) clustering was introduced. Additionally, the authors proposed a multi view version of agglomerative clustering. However, this second approach did not display any benefit against clustering single representations. [6] introduces the framework of reinforcement clustering, which is applicable to multi-represented objects. All of these approaches do not consider any semantic aspect of the underlying data spaces. The proposed approaches result in a partitioning clusterings of the data spaces, which makes the maximization of the agreement between local models a beneficial goal to optimize. However, in a density-based setting, there is an arbitrary number of clusters and no explicit clustering models that can be optimized to agree with each other.

3. Hierarchical Clustering of Multi-Represented Objects

3.1. Normalization

In order to obtain the comparability of distances derived from different feature spaces, we perform a normalization of the distances for each representation. Let $\mathcal{D}$ be a set of n objects and let $R := \{R_1, \dots, R_m\}$ be a set of m different representation existing for objects in $\mathcal{D}$.

The normalized distance between two objects $o, q \in \mathcal{D}$ w.r.t. R_i is denoted by $\mathrm{d}_i(o, q)$ and can be calculated by applying one of the following normalization methods:

- Mean normalization: Normalize the distance with regard to the mean value μ_i^{orig} of the original distance d_i^{orig} in representation R_i. The mean value can be calculated by sampling a small set of objects from the current representation R_i.

$$\mathrm{d}_i(o,q) = \mathrm{d}_i^{orig}(o,q)/\mu_i^{orig}$$

- Range normalization: Normalize the distance into the range of $[0 \dots 1]$.

$$\mathrm{d}_i(o,q) = \frac{\mathrm{d}_i^{orig}(o,q) - \min\limits_{r,s \in \mathcal{D}}\{\mathrm{d}_i^{orig}(r,s)\}}{\max\limits_{r,s \in \mathcal{D}}\{\mathrm{d}_i^{orig}(r,s)\} - \min\limits_{r,s \in \mathcal{D}}\{\mathrm{d}_i^{orig}(r,s)\}}.$$

- Studentize: Normalize the distance around the mean μ_i^{orig} and standard deviation σ_i^{orig} of the distances in representation R_i.

$$\mathrm{d}_i(o,q) = (\mathrm{d}_i^{orig}(o,q) - \mu_i^{orig})/\sigma_i^{orig}.$$

Since the factors can be calculated on database sample, we employed the first method in our experiments.

3.2. Multi-represented OPTICS

The algorithm OPTICS [3] works like an extended DBSCAN algorithm, computing the density connected clusters w.r.t. all parameters ε_i that are smaller than a generic value of ε. Since we handle multi-represented objects we have not only one ε-neighborhood of an object o but several ε-neighborhoods, one for each representation R_i. In the following, we will call the ε-neighborhood of an object o in representation R_i its local ε-neighborhood w.r.t R_i.

Definition 1 (local ε-neighborhood w.r.t R_i)
Let $o \in \mathcal{D}$, $\varepsilon \in \mathbb{R}^+$, $R_i \in R$, d_i the distance function of R_i. The local ε-neighborhood w.r.t. R_i of *o, denoted by $\mathcal{N}_\varepsilon^{R_i}(o)$, is defined as the set of objects around o with distances in representation R_i less or equal than ε, formally*

$$\mathcal{N}_\varepsilon^{R_i}(o) = \{x \in \mathcal{D} \mid d_i(o,x) \le \varepsilon\}.$$

In contrast to DBSCAN, OPTICS does not assign cluster memberships, but stores the order in which the objects have been processed and the information which would be used by an extended DBSCAN algorithm to assign cluster memberships. This information consists of two values for each object, its core distance and its reachability distance. To compute these information during a run of the OPTICS algorithm on multi-represented objects, we must adapt the core distance and reachability distance predicates of OPTICS to our multi-represented approach. In the next two subsections, we will show how we can use the concepts of union and intersection of the local ε-neighborhoods of each representation to handle multi-represented objects.

Union of different representations . The union method characterizes an object o to be density-reachable from another object p, if o is density-reachable from p in at least one of the representations R_i. If an object is placed in a dense region of at least one representation, it is already union density-reachable regardless in how many other representations it is density-reachable. In the following, we adapt some of the definitions of OPTICS in order to apply OPTICS on the union of different representations.

In the union approach, the (global) union ε-neighborhood of an object $o \in \mathcal{D}$ is defined by the union of all of its local ε-neighborhoods $\mathcal{N}_\varepsilon^{R_i}(o)$ in each representation R_i.

Definition 2 (union ε-neighborhood)
Let $\varepsilon \in \mathbb{R}^+$ and $o \in \mathcal{D}$. The union ε-neighborhood *o, denoted by $\mathcal{N}_\varepsilon^{\cup}(o)$, is defined by*

$$\mathcal{N}_\varepsilon^{\cup}(o) = \bigcup_{R_i \in R} \mathcal{N}_\varepsilon^{R_i}(o).$$

Since the core distance predicate of OPTICS is based on the concept of k-nearest neighbor distances, we have to redefine the k-nearest neighbor distance of an object o in the union approach. Assume that all objects $p \in \mathcal{D}$ are ranked according to their minimum distance $\mathrm{d}_{min}(o,q) = \min\limits_{i=1\dots m}(\mathrm{d}_i(o,p))$ in each representation R_i to o. Then, the *union k-nearest neighbor distance of $o \in \mathcal{D}$*, short $\textsc{nn-dist}_k^{\cup}(o)$ is the distance d_{min} to its k-nearest neighbor in this ranking. The union k-nearest neighbor distance is formally defined as follows:

Definition 3 (union k-NN distance)
Let $k \in \mathbb{N}$, $|\mathcal{D}| \ge k$ and $o \in \mathcal{D}$. The union k-nearest neighbors *of o is the smallest set $NN_k^{\cup}(o) \subseteq \mathcal{D}$ that contains (at least) k objects and for which the following condition holds:*

$$\forall p \in NN_k^{\cup}(o), \forall q \in \mathcal{D} - NN_k^{\cup}(o) : \\ \min_{i=1\dots m}\{d_i(p,o)\} < \min_{i=1\dots m}\{d_i(q,o)\}.$$

The union k-nearest neighbor distance of o, *denoted by* $\text{NN-DIST}_k^{\cup}(o)$ *is defined as follows:*

$$\text{NN-DIST}_k^{\cup}(o) = \max\{\min_{i=1\ldots m}\{d_i(o,q)\} \mid q \in NN_k^{\cup}(o)\}.$$

Now, we can adopt the core distance definition from OPTICS to our union approach: If the union ε-neighborhood of an object o contains at least k objects, the union core distance of o is defined as the union k-nearest neighbor distance of o. Otherwise, the union core distance is undefined.

Definition 4 (union core distance)
The union core distance of object $o \in DB$ w.r.t. $\varepsilon \in \mathbb{R}^+$ and $k \in \mathbb{N}$ *is defined as*

$$\text{CORE}_{\varepsilon,k}^{\cup}(o) = \begin{cases} \text{NN-DIST}_k^{\cup}(o) & \text{if } |\mathcal{N}_\varepsilon^{\cup}(o)| \geq k \\ \infty & \text{else.} \end{cases}$$

The union reachability distance of an object $p \in \mathcal{D}$ from $o \in \mathcal{D}$ is an asymmetric distance measure that is defined as the maximum value of the union core distance of o and the minimum distance in each representation R_i between p and o.

Definition 5 (union reachability distance)
The union reachability distance of an object $o \in DB$ relative from another object $p \in \mathcal{D}$ w.r.t. $\varepsilon \in \mathbb{R}^+$ and $k \in \mathbb{N}$ *is defined as*

$$\text{REACH}_{\varepsilon,k}^{\cup}(p,o) = \max\{\text{CORE}_{\varepsilon,k}^{\cup}(p), \min_{i=1\ldots m}\{d_i(o,p)\}\}$$

Intersection of different representations . The intersection method denotes an object o to be density-reachable from another object p, if o is density-reachable from p in all representations. The intersection ε-neighborhood $\mathcal{N}_\varepsilon^{\cap}(o)$ of an object $o \in \mathcal{D}$ is defined by the intersection of all of its local ε-neighborhoods $\mathcal{N}_\varepsilon^{R_i}(o)$ in each representation R_i.

Definition 6 (intersection ε-neighborhood)
Let $\varepsilon \in \mathbb{R}^+$ *and* $o \in \mathcal{D}$. *The* intersection ε-neighborhood o, *denoted by* $\mathcal{N}_\varepsilon^{\cap}(o)$, *is defined by*

$$\mathcal{N}_\varepsilon^{\cap}(o) = \bigcap_{R_i \in R} \mathcal{N}_\varepsilon^{R_i}(o).$$

Analogously to the union method, we define the intersection k-nearest neighbor distance of an object o. In the intersection approach all objects $p \in \mathcal{D}$ are ranked according to their maximum distance $\mathrm{d}_{max}(o,q) = \max_{i=1\ldots m}(\mathrm{d}_i(o,p))$ in each representation R_i to o. Then, the *intersection k-nearest neighbor distance of* $o \in \mathcal{D}$, short $\text{NN-DIST}_k^{\cap}(o)$ is the distance d_{max} to its k nearest neighbor in this ranking.

Definition 7 (intersection k-NN distance)
Let $k \in \mathbb{N}$, $|\mathcal{D}| \geq k$ *and* $o \in \mathcal{D}$. *The* intersection k-nearest neighbors *of o is the smallest set* $NN_k^{\cap}(o) \subseteq \mathcal{D}$ *that contains (at least) k objects and for which the following condition holds:*

$$\forall p \in NN_k^{\cap}(o), \forall q \in \mathcal{D} - NN_k^{\cap}(o): \max_{i=1\ldots m}\{d_i(p,o)\} < \max_{i=1\ldots m}\{d_i(q,o)\}.$$

The intersection k-nearest neighbor distance of o, *denoted by* $\text{NN-DIST}_k^{\cap}(o)$ *is defined as follows:*

$$\text{NN-DIST}_k^{\cap}(o) = \max\{\max_{i=1\ldots m}\{d_i(o,q)\} \mid q \in NN_k^{\cap}(o)\}.$$

In the following, we define the intersection core distance and the intersection reachability distance analogously to the union method. If the intersection ε-neighborhood of an object o contains at least k objects, the intersection core distance of o is defined as the intersection k-nearest neighbor distance of o. Otherwise, the intersection core distance is undefined.

Definition 8 (intersection core distance)
The intersection core distance of object $o \in DB$ w.r.t. $\varepsilon \in \mathbb{R}^+$ and $k \in \mathbb{N}$ *is defined as*

$$\text{CORE}_{\varepsilon,k}^{\cap}(o) = \begin{cases} \text{NN-DIST}_k^{\cap}(o) & \text{if } |\mathcal{N}_\varepsilon^{\cap}(o)| \geq k \\ \infty & \text{else.} \end{cases}$$

The intersection reachability distance of an object $p \in \mathcal{D}$ from $o \in \mathcal{D}$ is an asymmetric distance measure that is defined as the maximum value of the intersection core distance of o and the maximum distance in each representation R_i between p and o.

Definition 9 (intersection reachability distance)

The intersection reachability distance of an object $o \in DB$ relative from another object $p \in \mathcal{D}$ w.r.t. $\varepsilon \in \mathbb{R}^+$ and $k \in \mathbb{N}$ *is defined as*

$$\text{REACH}_{\varepsilon,k}^{\cap}(p,o) = \max\{\text{CORE}_{\varepsilon,k}^{\cap}(p), \max_{i=1\ldots m}\{d_i(o,p)\}\}$$

By first normalizing the distances within the representations, we are now able to use OPTICS applying either the union or the intersection method. In the next section, we will discuss the proper choice of one of these basic methods and introduce a heuristic for a meaningful combination of both methods for multiple representations.

4. Handling Semantics

In this section, we will discuss the handling of semantics. Therefore, we will first of all introduce a model that will help us to understand the interaction between different representations.

4.1. A Model for Local Semantics

Since feature spaces are usually not a perfect model of the intuitive notion of similarity, a small distance in the feature space does not always indicate true object similarity. Therefore, we denote two objects that a human user would classify as similar as truly similar. To formalize the semantic problem, we first of all distinguish two characteristics of representation spaces:

Definition 10 (Precision Space) *A precision space is a data space R_i where for each data object o there exists a σ-neighborhood $\mathcal{N}_\sigma^{R_i}(o)$ in which the percentage of data objects in $\mathcal{N}_\sigma^{R_i}(o)$ that are considered to be truly similar exceeds a given value π. Formally, a precision space R_i is defined as:*

$$\exists \sigma \in I\!R^+,\ \forall o \in \mathcal{D} : \frac{|\mathcal{N}_\sigma^{R_i}(o) \cap sim(o)|}{|\mathcal{N}_\sigma^{R_i}(o)} \geq \pi$$

where $sim(o)$ denotes all truly similar objects in $\mathcal{D}$ for object o.

Definition 11 (Recall Space) *A recall space is a data space R_i where for each data object o there exists a σ-neighborhood $\mathcal{N}_\sigma^{R_i}(o)$ in which the percentage of all truly similar data objects among the data objects in $\mathcal{N}_\sigma^{R_i}(o)$ exceeds a given value ρ. Formally, a recall space R_i is defined as:*

$$\exists \sigma \in I\!R^+,\ \forall o \in \mathcal{D} : \frac{|\mathcal{N}_\sigma^{R_i}(o) \cap sim(o)|}{|sim(o)|} \geq \rho$$

where $sim(o)$ denotes all truly similar objects in $\mathcal{D}$ for object o.

A precision space (recall space respectively) is called optimal, iff there exists an σ for which $\pi = 1$ ($\rho = 1$ respectively). Let us note that these definition treats similarity as a boolean function instead of using continuous similarity. However, the density-based clustering employs ε-neighborhoods $\mathcal{N}_\varepsilon^{\mathcal{D}}(o)$ for finding dense regions and within these dense regions the objects should be similar to o. Thus, boolean similarity should be sufficient for discussing density-based clustering. Figure 1 displays a maximal σ_p-neighborhood for object o if R_i would be an optimal precision space. Additionally, the figure displays the minimum σ_r-neighborhood of o if R_i is an optimal recall space as well. Note that the σ_p-neighborhood is a subset of the σ_r in all optimal precision and recall spaces.

Though it is possible that a representation space is as well a good precision as a recall space, most real world feature spaces are usually more suited to fulfill only one of these conditions. An example for a precision space are text vectors. Since we can assume that two very similar text annotations indicate that the described data objects are very similar as well, text annotations usually provide a good precision space. However, descriptions of two very similar objects do not have to use the same words. An object representation that is often well-suited for providing a good recall space are color histograms. If the color histograms of two color images are quite different from each other the images are unlikely to display the same object. On the other hand, two images having similar color histograms, are not necessarily displaying the same motive.

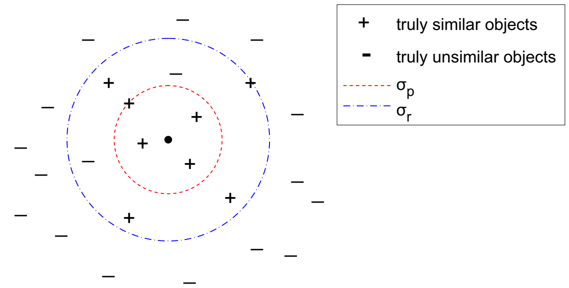


Figure 1. Maximal σ_p-neighborhood and minimum σ_r-neighborhood of an optimal precision and recall space.

When combining optimal precision and recall spaces for density-based clustering our goal is to find maximum density-connected clusters where each object has only truly similar objects in its global neighborhood. In general, we can derive the following useful observations:

1. A data space that is as well an optimal precision as an optimal recall space for the same value of σ is already optimal w.r.t our goal and thus does not need to be combined with any other representation.

2. A set of optimal precision spaces should always be combined by the union method because the union method improves the recall but not the precision. If there is at least one representation for any similar object in which the object is placed in the σ-neighborhood, the resulting combination is optimal w.r.t. recall.

3. A set of recall spaces should always be combined by the intersection method because the intersection method improves the precision but not the recall. If there exists no dissimilar data object that is part of the σ-neighborhoods in all representa-

tions, the resulting intersection is optimal w.r.t. precision.

4. Combining an optimal recall space with an optimal precision space with either union or intersection method does not make any sense. For this combination the objects in the σ-neighborhood of the precision space are always a subset of the objects in the σ-neighborhood of the recall space. As a result, applying the union method is equivalent to only using the recall space and applying the intersection method is equivalent to only using the precision space.

The observations that are made in this model are not directly applicable to the combination of representations using the multi-represented version of OPTICS described in the previous section. To apply the conclusion of our model to OPTICS, we have to fulfill two requirements. The normalization has to be done in a proper way, meaning that the normalization factors should have the same ratio as the σ values in each representation. The second requirement is that $\varepsilon > \sigma_i$ in each representation $R_i \in R$. If both requirements are satisfied, it is guaranteed that there is some level in the OPTICS plot representing the σ-values guaranteeing optimality.

Another aspect is that the derived statements only hold for optimal precision and recall spaces. Since a representation is always as well a precision space as a recall space to some degree, the observations generally do not hold for the non-optimal case. For example, it might make sense to combine a very good recall space with a very good precision space if the recall space has a good quality as a precision space as well at some other σ level. However, the implications to the general case are strong enough to derive useful heuristics.

A final problem for applying our model is the fact that it is not possible to determine π and ρ values for the given representations without additional information about true similarity. Thus, we have to employ domain knowledge when deriving some heuristics for building a well-suited combination of representations for clustering.

4.2. Combining Multiple Representations

Though we might not be able to exactly determine the parametrization for which a representation fulfills the precision and recall space conditions in a best possible way, we can still reason about the suitability of a representation for each of both conditions. Like in our running example of text vectors and color histograms,

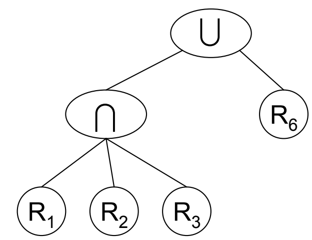


Figure 2. Combination tree of the image data set.

we can derive the suitability of a representation for being a good precision or a good recall space by analyzing the underlying feature transformations. If it is likely that two dissimilar objects have very similar feature representations the data space still might provide a useful recall space. If it is possible that very similar objects are mapped to some far away feature representations the data space might still provide a useful precision space.

The most important implication of our model is that combining a good precision space (recall space respectively) with a rather bad precision space (recall space respectively) will not increase the all over quality of clustering. Considering only two representations, there are only three options: use the union method for two precision spaces, the intersection method for two recall spaces or cluster only the more reliable representation in case of a mixture.

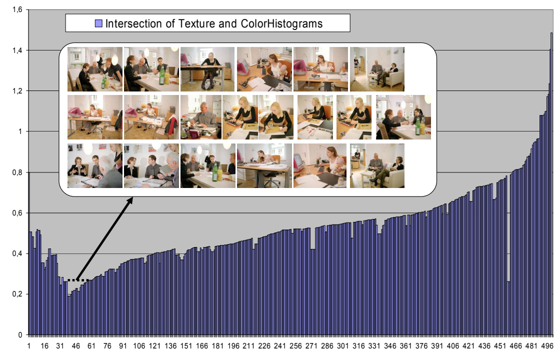
For more than two representations, the combination of precision and recall spaces still can make sense. The idea is to combine these representations on different levels. Since the intersection method increases the precision and the union method increases the recall, we are able to construct recall spaces or precision spaces from a subset of the representations. To formalize this method, we will now define the so-called combination tree:

Definition 12 (Combination Tree) *A combination tree is a tree of arbitrary degree fulfilling the following conditions:*

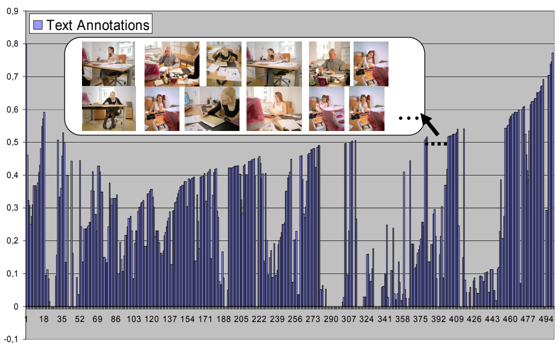
- *The leafs are labeled with representations.*
- *The inner nodes are labeled with either the union or the intersection operator.*

A good combination according to our heuristics can be described by a combination tree where the sons of each intersection node are all reasonable recall spaces and the sons of each union node are all reasonable precision spaces. After we derived the combination tree, we can now modify the core distance and reachabil-

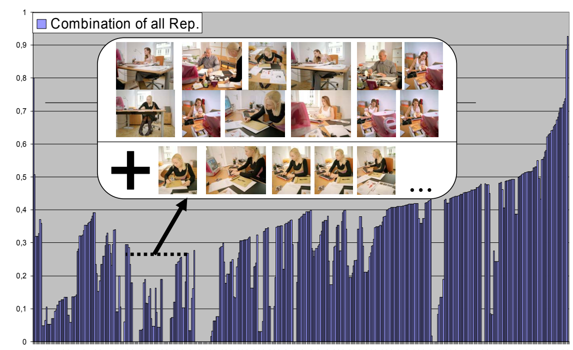
(a) OPTICS plot using only color histograms. Additionally, a representative sample set for one of the clusters is shown.



(b) OPTICS plot when employing the intersection of color histograms and both texture representations. The displayed cluster shows promising precision.



(c) OPTICS plot using only Text annotations. The displayed cluster has a high precision but is incomplete.



(d) OPTICS plot of the combination of all representations. The precise cluster observed in the text representation is completes with similar images.

ity distance of OPTICS in an top-down order to implement the semantics described by the combination tree.

Figure 2 displays the combination tree of the image data set, we used in our experiments. R_1, R_2 and R_3 represent the content based feature repressntations expressing texture features and colors distributions. In each of these representations a small distance between the feature vectors does not necessarily indicate that the underlying image is truly similar. Therefore, we use all of these 3 representations as recall spaces. R_4 consists of text annotations. As mentioned before, text annotations usually provide good precision spaces but may provide good recall spaces. Thus, we use the text annotation as a precision space. The combination of the representation is now done in the following way. We first of all combine our recall spaces R_1, R_2 and R_3 using the intersection method. Due to the increased precision resulting from applying the intersection method, the suitability of the result for being a precision space should be sufficient for applying the union method with the remaining text annotations R_4.

5. Performance Evaluation

In order to show the capability of our method,we implemented the proposed clustering algorithm in Java 1.4. All experiments were processed on a work station with a 2.4 GHz Pentium IV processor and 2 GB main memory. We used a data set containing 500 images manually annotated by a short text. From each image, we extracted 3 representations, namely a color histogram and two textural feature vectors. We used the HSV color space and calculated 32 dimensional color histograms based on 8 ranges of hue and 4 ranges of saturation. The textural features were generated from 16 gray-scale conversions of the images. We computed contrast and inverse difference moment using the cooccurrence matrix [7]. For comparing text annotations, we applied the cosine coefficient and used the Euclidian distance in the rest of the representations. Since OPTICS does not generate a partitioning clustering but only a cluster order, we did not apply a quantitative quality measure. To verify the results of the found clustering, we visually verified the similarity of images in each cluster instead. To demonstrate the results of multi-represented OPTICS with the combination method described above, we ran OPTICS on each single representation. Additionally, we examined the clustering for the combination of color histograms and texture features using the intersection method like proposed in the combination tree. Finally, we ran OPTICS using the complete combination of image and text features. For all clusterings, we used $k = 3$ and $\varepsilon = 10$. Normalization was achieved using the average distances between two objects in the data set.

The result for the text annotations provided a very precise clustering. However, due to the fact that some annotations used different languages for describing the image, some of the clusters were incomplete. Figure 3(c) displays the result of clustering the text annotations. The observed cluster displays only similar objects. The cluster order derived for color histograms, found some clusters. However, though the images within the clusters had similar colors the objects were not necessarily similar. Figure 3(a) displays the cluster order using color histograms and an image cluster containing two similar groups of images and some noise. Let use note that the clustering of the two texture representation performed similarily. However, due to space limitations, we do not display the corresponding plots. In Figure 3(b) the clustering of all 3 image feature spaces using the intersection method is displayed. Though the number of clusters was decreased, the quality of the remaining clusters increased considerably, as expected. The cluster shown in figure 3(b) showed exclusively very similar images. Finally, figure 3(d) displays the result on all representations. The cluster observed for text annotations displayed in figure 3(c) was extended with additional similar images that are described in German instead of English language. To conclude, examining the complete clustering, the all over quality of clustering was improved by using all 4 representations.

6. Conclusions

In this paper, we discussed the problem of hierarchical density-based clustering of multi-represented objects. A multi-represented object is described by a tuple of feature representations belonging to different feature spaces. Therefore, we adapted the hierarchical clustering algorithm OPTICS to multi-represented objects by introducing the union (intersection) core distance and the union (intersection) reachability distance for multi-represented objects. Since union and intersection method might not be suitable to compare an arbitrary large number of representations, we proposed a theoretical model distinguishing so-called precision and recall spaces. Based on these concepts, we observed that the combination of good precision spaces using the union method increases the completeness of clusters and applying the intersection method on good recall spaces increases the pureness of clusters. Finally, we concluded that combining a good precision (recall) space with a bad one results in no benefit. To use these conclusion for combining problems with multiple representation, we introduced combination trees that display valid combination of precision and recall spaces. In our experimental evaluation, we described the improvement of clustering results for an image data set that is described by 4 representations.

For future work, we plan to find a reasonable way to quantify the usability of representations as precision or recall spaces. Additionally, we are currently working an theory for describing optimal combination trees.

References

[1] Kailing, K., Kriegel, H.P., Pryakhin, A., Schubert, M.: Clustering multi-represented objects with noise. In: PAKDD. (2004) 394–403

[2] Ester, M., Kriegel, H.P., Sander, J., Xu, X.: "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: Proc. KDD'96, Portland, OR, AAAI Press (1996) 291–316

[3] Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: "OPTICS: Ordering Points to Identify the Clustering Structure". In: Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99), Philadelphia, PA. (1999) 49–60

[4] de Sa, V.R.: "spectral clustering with two views". In: Proceedings of the ICML 2005 Workshop on Learning With Multiple Views,Bonn, Germany. (2005) 20–27

[5] Bickel, S., Scheffer, T.: Multi-view clustering. In: ICDM. (2004) 19–26

[6] Wang, J., Zeng, H., Chen, Z., Lu, H., Tao, L., Ma, W.: "ReCoM: reinforcement clustering of multi-type interrelated data objects. 274-281". In: Proc. SIGIR 2003, July 28 - August 1, Toronto, CA, ACM (2003) 274–281

[7] Haralick, R.M., K., S., I., D.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics **3** (1973) 610–621