User-Based Visual-Interactive Similarity Definition for Mixed Data Objects - Concept and First Implementation

Jürgen Bernard[1 2]
juergen.bernard
@igd.fraunhofer.de

David Sessler[1 2]
david.sessler
@igd.fraunhofer.de

Tobias Ruppert[1]
tobias.ruppert
@igd.fraunhofer.de

James Davey[1]
james.davey
@igd.fraunhofer.de

Arjan Kuijper[1 2]
arjan.kuijper
@igd.fraunhofer.de

Jörn Kohlhammer[1]
joern.kohlhammer
@igd.fraunhofer.de

[1]Fraunhofer IGD, Germany
[2]TU Darmstadt, Germany

ABSTRACT

The definition of similarity between data objects plays a key role in many analytical systems. The process of similarity definition comprises several challenges as three main problems occur: different stakeholders, mixed data, and changing requirements. Firstly, in many applications the developers of the analytical system (data scientists) model the similarity, while the users (domain experts) have distinct (mental) similarity notions. Secondly, the definition of similarity for mixed data types is challenging. Thirdly, many systems use static similarity models that cannot adapt to changing data or user needs. We present a concept for the development of systems that support the visual-interactive similarity definition for mixed data objects emphasizing 15 crucial steps. For each step different design considerations and implementation variants are presented, revealing a large design space. Moreover, we present a first implementation of our concept, enabling domain experts to express mental similarity notions through a visual-interactive system. The provided implementation tackles the different-stakeholders problem, the mixed data problem, and the changing requirements problem. The implementation is not limited to a specific mixed data set. However, we show the applicability of our implementation in a case study where a functional similarity model is trained for countries as objects.

Keywords: Similarity Measures, User-centered Design, User Feedback, Mixed Data Sets, Feature Selection, Information Visualization, Visual Analytics

1 INTRODUCTION

The definition of similarity between data objects is an important prerequisite to perform data analysis tasks for various data-centered domains. One can assume that the functional definition of similarity for the comparison of data objects is chosen in order to reflect the notion of similarity in the mind of the users. In other words, in most of the existing approaches the users' *mental similarity notion* has to be represented by a functional specification.

However, in many applications the developers of the analytical system are not necessarily the users of the system. In these cases, systems are typically developed in a collaborative effort between *domain experts* and *data scientists*. Due to differing expertise, a knowledge gap exists. Capturing the mental similarity notion of the domain expert and thus defining a meaningful similarity model may be challenging for the data scientist due to a) false assumptions and different vocabulary, b) data complexity, or c) an insufficient number of iterations in the development phase. The problem gets worse if domain experts cannot formalize their similarity notion in the granularity of specific attributes/features. For example, a doctor may not be able to *explicitly* define the functional behaviour of EEG features. But she can *implicitly* identify similar patterns at a glance.

Another challenge for the definition of similarity is based on the data complexity. In many real world data sets numerical, ordinal, categorical and binary attribute types are present; often called mixed data. For mixed data sets a combination of similarity measures for different attribute types is needed in order to cover the attribute space as a whole. However, approaches that deal with similarity on mixed data sets are scarce.

Finally, a problem arises if the mental similarity notion of the domain expert changes over time. In this case, most current systems require an intervention by the data

scientist. The similarity model has to be re-defined, the system has to be re-compiled and re-deployed. In addition, the underlying data set may change over time, which implies similar actions. For both of these cases a 'static' implementation at compile time lacks dynamism at run time.

In this work we present a concept for the visual-interactive definition of similarity for mixed data objects which can directly be accessed by domain experts. Systems based on our concept will be able to tackle the three challenges outlined in the similarity definition process. We describe 15 mandatory steps for the development of such systems. For each step different implementation variants as well as their advantages and disadvantages are presented. This reveals a large design space.

In addition, we present an implementation of the concept in accordance to the 15 mandatory design steps. Domain experts are able to interactively align mixed data objects and thus to express their mental similarity notion visually. Our approach tackles all three described challenges. Finally, we show the benefits of our implementation in a case study. A mixed data set consisting of country attributes is applied to capture a mental similarity notion.

2 RELATED WORK

We identified those fields of research which are related to our work. While many inspiring aspects will be recalled later in the concept section, we first structure the related work according to these fields of research.

2.1 User-Centred Design and Feedback

We present a concept that involves the user in the definition of the similarity operation. User-centred design is a widely recognized practice to improve the effectiveness and efficiency of the user working with software [39]. Visual Analytics guidelines to involve the user in design studies exist (e.g. presented by Tamara Munzner [37]). Furthermore, this work deals with the similarity notion of domain experts. This *mental similarity notion* in the minds of users can also be described as the *mental model* [33]. The objective of our concept is to transform these mental models into functional similarity operations, based on user feedback. There are two possibilities for the collection of user feedback. *Explicit* feedback can be gathered if the user has an in-depth understanding of the attribute space of the provided data set. *Implicit* feedback is provided if the user points out which objects are similar, based on her mental similarity notion without necessarily comprehending the available feature space. This concept is applied by recommender systems [1]. The system then interprets the implicit feedback to create a functional similarity specification based on the features of the objects. While approaches that consider *explicit* user feedback exist [25], [16], [49], incorporating *implicit* feedback in a system is a more difficult task. Our concept is based on implicit feedback, because we do not aim to bias the user's mental model by giving additional information about the underlying attributes. The domain experts should only be aware of handling features on-demand. However, the functional similarity model is based on features. Thus a knowledge gap between the user and developer exists. Bridging this gap [47] is one of the challenges of this work.

Related approaches applying feature selection aim to improve the prediction performance and to reduce the number of features to best candidates for further analysis of the data set. An introduction to feature selection as well as an extensive survey is given by Guyon and Elisseeff [19]. Visual analysis and interactive refinement of automatic techniques for feature subset selection is presented in SmartStripes [35]. Our concept to calculate similarity is primarily based on selecting and weighting appropriate features of the data. By discarding features with weight zero or a weight smaller than a threshold a feature subset selection can be achieved.

Feedback concepts are used in many domains for different tasks. Relevance feedback [41] is a prominent procedure in the information retrieval. It is used to improve query formulations interactively and iteratively. In this way, search queries can be optimized. Direct feedback to improve the quality of search results is applied in *Pixolution* [38] a tool for visual similarity search. *Pixolution* speeds up the process of finding images that are visually similar to a sample image. Different feedback strategies like the weighting of image colours or textual search are applied. In contrast to our concept, the weighting strategy is based on explicit feedback. Indirect feedback, gained by monitoring the behaviour of users, is used by recommender systems. Online platforms for music, film or news, etc., use this technique to suggest similar products. A survey on recommender systems is presented by Adomavicius and Tuzhilin [1]. Active learning approaches use feedback to annotate unlabelled data, which then is used to train machine learning models [43]. In our concept we suggest the use of implicit user feedback generated by object positioning in a 2D area to identify the arranged objects and their pairwise distances.

2.2 Mixed Data and Similarity Metrics

Dealing with mixed data is a non trivial task. Our concept deals with mixed data sets from an algorithmic and visual-interactive perspective. Several approaches that deal with mixed data sets algorithmically exist. Clustering of mixed data sets is presented by Jie et al. [26]. Applications for the quantification of categorical data and the exploration of data sets including both categorical and numerical variables exist [28] [27]. Visual

approaches that incorporate categorical and numerical data are often based on enhanced parallel coordinates metaphor. Examples are *ParallelSets* [29] for categorical and *VisBricks* [32] for mixed data sets. Approaches to explore relations between categorical data not relying on parallel coordinates are presented in the *Contingency Wheel++* [2] and a content-based metadata layout technique [8]. Finally, relations in mixed data can be calculated by statistical dependency tests and be explored in a visual-interactive way [9].

Our concept provides a combined similarity model for mixed data types by unifying the results of prominent measures for individual attribute types. A survey of binary similarity and distance measures is presented by Choi et al. [15]. More binary and numerical similarity metrics are presented by Lesot et al. in their survey [31]. In the work of Boriah et al. categorical distance measures are discussed [12]. For our concept we identify a benefit in the use of weighted distance measures. Approaches to enhance binary distance metrics to provide weighting functionality are described in [14]. A comparative evaluation for weighted categorical distance and similarity metrics is provided in [12].

2.3 Active and Semi-Supervised Learning

In contrast to fully automated approaches to the definition of similarity, like the Topology Matching of 3D Shapes [22], our concept enables the user to take an active role. In the classification domain various supervised and semi-supervised approaches exist. Visual-interactive systems where users can define decision trees are presented by Ware et al. [49] and Ankerst et al. [3]. EnsembleMatrix [46] involves multiple classifiers with user interaction to support machine learning. Semi-supervised clustering encourages the user to define constraints to influence the clustering outcome [5]. Clustering based on supervised dimension reduction is presented in the *iVisClassifier* approach [16]. However, our concept focuses on similarity metric learning, not on data aggregation.

Various supervised approaches for learning distance metrics exist [51]. For numerical data, the Mahalanobis metric can be trained as presented by Weinberger and Saul [50]. Similar to our concept, *Dis-Function* [13] incorporates user feedback in the distance metric learning by weighting features individually. However, no applicability to mixed data sets is illustrated. Moreover, our concept aims at providing feedback based on data subsets, not on the whole data set. A visual-interactive nearest-neighbour definition approach is presented by Mamani et al. [34]. Similar to our concept, the system projects a subset of objects into 2D with respect to user interest. Then, the user aligns this subset according to her similarity notion. While providing some inspiration for the object alignment concept, the approach focuses on image retrieval features.

3 REQUIREMENT ANALYSIS

To the best of our knowledge, a visual-interactive interface for the similarity definition of mixed data objects that solves all described challenges has not been presented before. However, in the past inspiring contributions have addressed some of these challenges. We aim to introduce a concept that covers most aspects of the work we presented. We consolidate this idea with our expertise gained in design studies and hope to contribute a concept that is generalizable, accepting that specific contributions from related works might be neglected. From these considerations, we sketch requirements to break down the high-level challenges presented in previous sections to a more precise and technical level. The objective of this requirement analysis is to concretize the (interface) design space.

- $\mathbf{R}_1$ *No need for data scientist*. The definition of the similarity model should be applicable for domain experts without the presence of a data scientist.
- $\mathbf{R}_2$ *Continuous distance measure between objects*. The distance between any two data objects should be represented by a continuous numerical value, based on object properties, not by class or cluster affiliation.
- $\mathbf{R}_3$ *Implicit in favour of explicit feedback*. Domain experts are not required to quantify similarity based on individual attributes. They can operate at the object level. It should be a matter for the system to identify appropriate features. The visual object representation should focus on unique identifiers. Attribute information should only be provided on demand.
- $\mathbf{R}_4$ *Handling mixed data*. Implementations of the concept should be able to cope with mixed data. Systems should not depend on specialized sub-types.
- $\mathbf{R}_5$ *Adapting to changing requirements*. While using the system, domain experts should be able to adapt the current similarity model in the case of a changing mental similarity notion. There should be no need for re-implementation and re-deployment.
- $\mathbf{R}_6$ *Visual result overview*. While defining the similarity model, domain experts should be supported by a visual overview of the current similarity result.
- $\mathbf{R}_7$ *History*. Implementations of the concept should provide a history which enables the domain experts to observe the work flow and step back to past states.

Depending on the analysis task, the targeted data set, and the application domain the requirements for specific implementations of the concept may vary.

4 A CONCEPT FOR USER-BASED SIMILARITY DEFINITION

In this section we present a concept for systems that enable domain experts to express mental similarity notions for mixed data objects in a visual-interactive way. A schematic overview is shown in Figure 1.

The visual-interactive components of the concept are targeted towards visual analytics technology. Thus, implementations allow for an encapsulation of backend (black box) functionality based on data mining and machine learning capability. This means that implementations will not require the presence of data scientists when domain experts execute the systems. We recommend the implementation of the visual components with respect to the design-study methodology known from the visual analytics domain [47] [37].

In the following we describe the four main components of the concept. In the *User Feedback View* domain experts give similarity-based feedback with respect to a set of selected data objects (see Section 4.1). The *Feedback Model* interprets the user feedback and calculates a weighting of the mixed data attributes (see Section 4.2). The *Similarity Model* calculates pairwise object distances based on the weighted attributes (see Section 4.3). Finally, the *Result View* visualizes results of the Similarity Model (Section 4.4). Figure 3 emphasizes the data and control flow of the concept. We suggest the use of distance matrices and attribute weightings as data exchange formats between components. However, the use of other (data) interfaces is possible, if appropriate.

Finally, we contribute 15 mandatory design steps in the work flow. The degrees of freedom and implementation variants reveal a large design space. Our concept is developed based on related work, essential requirements, and the experience of past design studies.

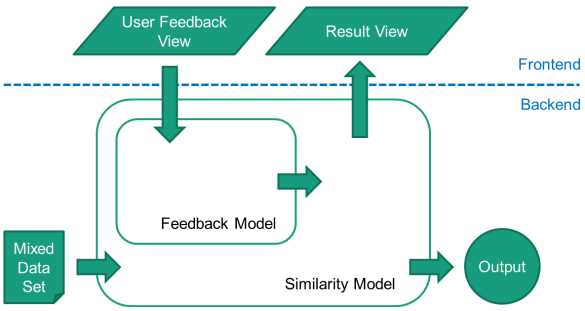

Figure 1: Overview of the concept. User feedback given in a view is interpreted by a Feedback Model. A Similarity Model converts the feedback to similarity values and passes the results to the Result View.

4.1 User Feedback View

We begin with a detailed description of the visual interface in which similarity can be expressed through user interaction. We assume that domain experts can only provide feedback for small data subsets, which influences the available design space. In the following, we discuss three objectives for visual feedback interfaces.

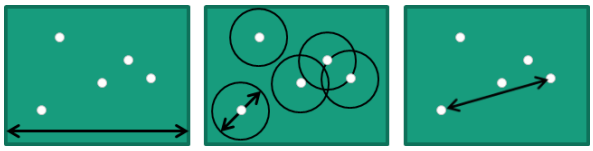

Figure 2: Object Alignment Strategies. Left: *absolute*. Centre: *orbital*. Right: *relative*.

S_1 Object Alignment Strategy

The most important design decision is the alignment strategy of data objects for the expression of mental similarity notions. Following $\mathbf{R}_2$ we focus on techniques that enable the calculation of continuous pairwise object distance values. Thus, we neglect ranking-based techniques and discrete class assignment strategies [49]. While the idea of a 'back-projection' of aligned 2D object geometry is inspiring [34], we see difficulties in the back projection of mixed data sets. We identify three different strategies for the alignment of objects in 2D to provide similarity feedback, which differ in the maximum distance definition (see Figure 2). As a first variant, the borders of the rectangular display serve as a global maximum distance provider (*absolute* mode). Pairwise distances of all aligned objects are calculated with respect to the global maximum distance. For the second feedback concept the maximum distance is defined by a constant radius around each object. In this *orbital* feedback mode objects are only considered for the similarity calculation if they are aligned within a specified distance radius. In a third (*relative*) variant the objects are aligned by means of a user-defined topology where the most distant objects define the maximum distance (similar to [50] [34] for illustrations). Our experience shows a tendency towards the orbital and the relative feedback mode. However, we must draw attention to the possibly high cognitive load of the orbital mode.

S_2 Guidance in the Choice of Objects

Since data sets may be large, a meaningful subset of data objects should be selected for feedback generation in order to adequately represent the mental similarity notion and to reduce over-estimation [21]. Systems could provide guidance concepts to suggest interesting objects (e.g. through relevance feedback [41]), or by assessing the amount of 'untouched' information using entropy measures [24]. Another variant is to provide the best possible overview of all objects to enable domain experts to select the most meaningful objects. If the system is used by multiple users then recommender system strategies [1] might be interesting.

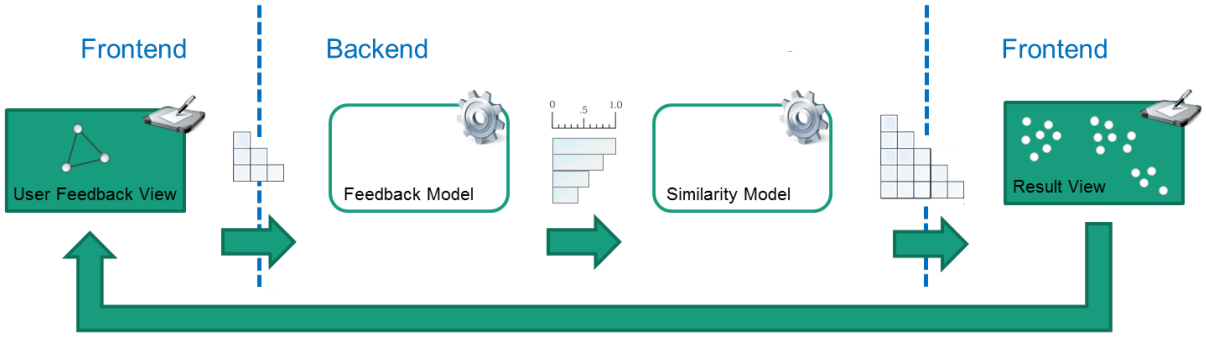

Figure 3: Interface design and functional support.

S_3 Visual Representation of Single Data Objects

In order to ensure usability, the visual representation of data objects should be as intuitive as possible. A unique identifier enables domain experts to recognize well known data objects and to effectively provide feedback with respect to their mental similarity notion ($\mathbf{R}_3$). We suggest a compact visual representation to reduce overplotting in large data sets. Promising alternatives are iconic or glyph-based techniques for the visual representation of single data objects [11].

4.2 Feedback Model

The goal of the Feedback Model is to generate weights for each data attribute, similar to feature selection approaches [19]. Attribute weights are passed to the Similarity Model (see Figure 3).

S_4 Object Geometry Interpretation

Regardless of the chosen *object alignment strategy* in the User Feedback View, the 2D geometry needs to be 'matched' with a possibly high-dimensional attribute space. One solution is to use the object coordinates directly and to apply regression models [36] or other machine learning techniques [50] in order to identify a functional dependency with the provided attributes. Alternatively, pairwise object distances could be used. These distances could then be compared with distances 'caused' by individual attribute candidates, (e.g. using Pearson's correlation coefficient [30] or Spearman's rank correlation [40]). An extension may be to apply tuples instead of comparing attributes with the object geometry independently. For example, the tuple consisting of the *Latitude* and the *Longitude* should be highly appropriate when the mental similarity notion is based on 2D geo-information. However, the complexity of the Feedback Model computation would increase.

S_5 Iterative Weighting Strategy

Domain experts may want to integrate various sets of objects to optimize the functional representation of their mental similarity notions. A variety of update strategies are conceivable for the adaptation of attribute weightings within subsequent feedback steps. The simplest variant is to completely discard 'old' feedback which may be suitable in some cases. However, we suggest the implementation of a ratio function in order to combine old and new feedback.

S_6 Termination of Feedback Process

Another interesting aspect regards the termination of the feedback process. Similar to the entropy-based *guidance in the choice of objects*, the impact of new training iterations might decrease if the degree of remaining 'information' tails off. Alternatively a constant decrease of learning weight can be implemented.

4.3 Similarity Model

The Similarity Model is the component in which continuous similarity values between any two data objects are calculated ($\mathbf{R}_2$). The input is a) the mixed data set (see Figure 1) and b) the attribute weighting based on user feedback (see Figure 3). A schematic overview of the algorithmic work flow is shown in Figure 4.

S_7 Data Input

We have seen a variety of data input variants in the past, spanning from fully automated to supervised approaches. In fact the data input variant may be a critical step when implementations of our concept aim to work without data scientist involvement $\mathbf{R}_1$, or without re-compiling ($\mathbf{R}_5$). Importing data appropriately is a problem in its own right which we do not aim to solve in this work. However, we suggest the use of predefined input file formats, such as WEKA's ARFF-format [20], since the data is then structured in an interpretable format and can be imported automatically.

S_8 Data Preprocessing

We identify a general need for more visual-interactive data preprocessing tools. Many data sets, analysis tasks, etc. expect individual preprocessing guidelines which are difficult to cover at run time at a glance. However, some promising visual analytics approaches for data preprocessing exist [23] [7]. To guarantee the

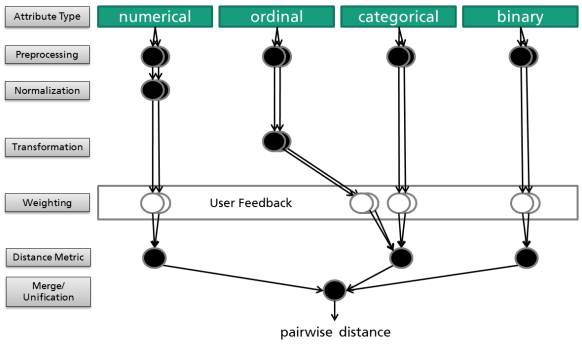

Figure 4: Similarity Model overview. We suggest a data transformation pipeline which is combined with attribute weight information based on user feedback. The output of the Similarity Model is a continuous distance value for every pair of data objects.

data quality expected by later steps of the pipeline we missing values must be handled. Moreover, coping with outliers is critical for subsequent similarity calculation capability. If domain knowledge about attributes is available it might be appropriate to exclude less interesting attributes in order to increase scalability. Other feature selection techniques may be applied [19].

S_9 Attribute Normalization

We suggest normalizing numerical attributes in order to provide comparability. Possible attribute normalization strategies are, among others, *zero-max*, *min-max*, or *z-score*. However, our experience shows that if the target value interval of each attribute is $[0...1]$ a 'unification' of attributes in subsequent steps of the work flow becomes more straightforward ($\mathbf{R}_4$).

S_{10} Handling Mixed Data Attributes

A meaningful treatment of mixed data attributes ($\mathbf{R}_4$) is particularly important. One solution is to transform all attributes to categorical attributes, (e.g. by binning) or to 'quantify' all available attributes [28]. However, we suggest the individual treatment of different attribute types to better preserve their inherent properties. A variety of similarity/distance measures exist for numerical [31], categorical [12] and binary [15] attributes. We consider the similarity definition for ordinal data as future work. We recommend the incorporation of measures which enable weighting in order to handle the attributes to the degree of 'suitability' [14] [12]. Finally, a degree of freedom is the decision, whether measures are implemented as static, or chosen at runtime, possibly through automated goodness-of-fit comparison.

S_{11} Attribute Transformation

As mentioned, similarity definition for ordinal data is difficult. We suggest the treatment of ordinal attributes as categorical, or numerical. For the latter variant care should be taken with ordinal attributes that cannot (automatically) be treated as numerical. Another consideration is the combined treatment of categorical and binary attributes, which may be reasonable in some cases.

S_{12} Choice of Distance Measures

Finally, the attribute weighting information and the data provided need to be merged. One design choice could be the use of different distance measures for individual data subsets [50]. The alternative would be to treat individual attribute types differently. While a variety of alternatives exist [51], the weighted Euclidean distance may be an appropriate distance metric for numerical attributes. For categorical attributes a good choice may be the weighted Goodall distance [12]; since it is sensitive to the probability of attribute values, less frequent observations are assigned higher 'scores'. For binary attributes the weighted Jaccard distance [14] might be a good starting point. Similar to other measures this variant is based on a contingency table. The Jaccard distance neglects negative matches (both false) which might be advantageous for many similarity concepts [45]. Based on the attribute weighting the 'impact' of each attribute on the final similarity calculation can be considered in a final unification step where all attribute-based object distance information are condensed to a single distance matrix.

4.4 Result View

We suggest the provision of direct feedback from the Similarity Model to the domain expert. In the Figure 3 we show how the visual result representation concept can be used to close the feedback loop. Thus the Result View may serve as an overview ($\mathbf{R}_6$ $\mathbf{R}_3$), as well as a pool for object selection. Again, a compact *visual data object representation* should be chosen to support the implicit feedback strategy ($\mathbf{R}_3$) and to reduce overplotting.

S_{13} Representation of Pairwise Distances

Especially for large and/or high-dimensional data sets a visual scalability problem may occur regarding the visual representation of pairwise object distances. Regardless of the technique applied, we suggest the inclusion of interactive drill-down functionalities, such as zooming and panning to facilitate the identification of local structures. The distance matrix data structure can be visualized directly, with the drawback that data objects are reduced to a pixel-based display (see [8]). However, large amounts of distance information may be encoded visually. Another idea is to layout the objects in a node-link structure, known as the complementary means to distance matrix visualization [18] [9]. Force-directed algorithms may be applied to layout data objects in 2D, alternatively projection-based techniques

may be used. If such 'map metaphors' are chosen we recommend providing interactive map rotation to better exploit the position information. However, errors may be introduced by layout algorithms. This can be a challenging problem since it distracts the domain expert from the actual distance information calculated by the system. To tackle this we suggest visually supporting the layout with the results of neighbourhood preservation measures (see [48]).

S_{14} Visual Scalability
We suggest the use of data aggregation to cope with large data sets. The visual representation of aggregated data can improve visual scalability a lot [17] [10]. A prominent class of aggregation techniques is (unsupervised) clustering (see e.g., [26]). The quality assessment of clustering/classification results is a non-trivial problem which might be tackled by visual analytics techniques [16] [42], [6]. Another variant to reduce the object space interactively is faceted search. On the one hand, a meaningful choice of facets may support domain experts in drilling down in the data space. On the other hand, the mental similarity notion may be influenced.

S_{15} Providing a History
Or final suggestion affects the integration of a history ($\mathbf{R}_7$). Past (iterative) similarity definition steps should be visually comprehensible [44]. A history functionality may provide both a lookup of past steps and an undo capability. The visual representation of the history may be based on the object feedback geometry, or by the result view showing the overall data set as a 'fingerprint'.

5 A FIRST IMPLEMENTATION

We present an implementation of the described concept provided as a visual-interactive system. Domain experts are able to express object feedback, according to their mental similarity notion. The implementation considers the 15 crucial steps for the development of appropriate systems and thus, tackles the described challenges. While not being limited to a specific mixed data set, in the next Section 6, we apply our implementation to a data set consisting of countries as objects.

5.1 User Feedback View

We chose the relative object alignment strategy. Thus domain experts are able to build object topologies ($\mathbf{S}_1$). We implemented a drag-and-drop mechanism to enable domain experts to explicitly add the most appropriate data objects from the Result View in the User Feedback View ($\mathbf{S}_2$). Based on the chosen data set (see Figure 7), we chose the *Name* attribute as unique identifier, as well as a flag icon ($\mathbf{S}_3$). The visually encoded information items complement each other for a quick lookup of countries.

5.2 Feedback Model

The relative object feedback is transformed to a Euclidean distance matrix. Each attribute of the mixed data set is interpreted separately, based on the Pearson's correlation coefficient ($\mathbf{S}_4$). A decreasing weight function with the ratio (20% new, 80% old) is chosen as weighting strategy ($\mathbf{S}_5$ and $\mathbf{S}_6$).

5.3 Similarity Model

The incorporation of WEKA's ARFF-file format makes the implementation applicable for a variety of available data sets ($\mathbf{S}_7$). Missing values are removed, an outlier handling strategy is neglected since we do not want to affect the value range of the attributes ($\mathbf{S}_8$). A zero-max normalization preserves the absolute value relations ($\mathbf{S}_9$). We decided to treat mixed data attributes independently ($\mathbf{S}_{10}$). However, ordinal attributes are treated as categorical attributes to omit (quantified) false assumptions ($\mathbf{S}_{11}$). The three individual distance measures for numerical, categorical and binary attributes are chosen as suggested in the concept ($\mathbf{S}_{12}$).

5.4 Result View

We apply an MDS projection to represent objects in 2D with respect to all pairwise distances provided by the Similarity Model ($\mathbf{S}_{13}$). The user can additionally view the projection quality (Trustworthiness measure by Venna and Kaski [48]), visually encoded with object outlines in a continuous colour scale from green (good quality) to red (bad quality) ($\mathbf{S}_{14}$). Due to the comparatively small data set we did not include an additional data aggregation scheme. However, the Result View provides map rotation, zooming and panning to enable users to explore the object space (see Figure 6) and to use the described drag-and-drop functionality. At the bottom of the system we provide a history function ($\mathbf{S}_{15}$). We decided to show the object geometry rather than the Similarity Model result for past iterations.

6 CASE STUDY

In this section we illustrate the applicability of the implementation. We used an enriched version of the *flags data set* [4], consisting of mixed attributes extracted from the flag itself, as well as geographical and demographical attributes of the countries. In the application example the Similarity Model is trained on the basis of an individual mental similarity notion.

Description of the Mental Similarity Notion

We chose the mental similarity notion of topologically correct distances between European countries. Thus, the mental similarity notion is centred on the numerical attributes *Latitude* and *Longitude*, which are in fact

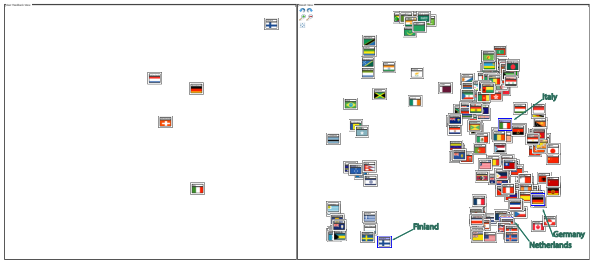

Figure 5: Training Phase. Left: 5 countries generate the topology of Europe (the mental similarity notion). Right: the calculated Similarity Model result in an early stage seems to reflect colour separation.

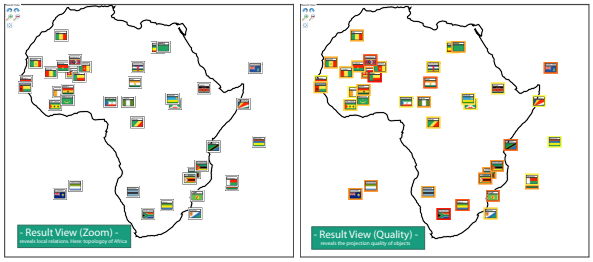

Figure 6: Result View. Rotation, zooming and panning reveals data subspaces (here: Africa). On the right the projection quality can be seen by coloured outlines (green to red).

contained in the data set but not explicitly shown during the feedback process, to facilitate the implicit feedback strategy ($\mathbf{R}_3$). To pass the test, the final Similarity Model of the system had to reflect topologically correct distances of the countries in the data set, and thus, the worlds geographic topology.

Training the Similarity Model

In an initial user feedback step the countries *Germany*, *The Netherlands*, *Switzerland*, *Italy* and *Finland* were arranged in the User Feedback View as shown in Figure 5. We aligned the European countries with respect to the locations of their capitals. The calculated Similarity Model is shown at the right. Objects available in the User Feedback View are highlighted in the Result View with a blue outline and background. It can be seen that *Finland* is located on the left, together with a number of loosely distributed countries. The other four countries are located on the right in a compact cluster that mostly contains flags with red colours. According to our mental similarity notion the five targeted countries should be aligned close to each other. However, it seems to be difficult to represent our mental similarity notion with only five countries. In particular, the colour of the objects (still) seemed to have a strong influence on the calculated Similarity Model. To improve the model quality *Sweden* was picked from the left cluster and *Portugal* from the right. In addition, we dropped *Poland* and *Hungary* into the User Feedback View. A topology based on our mental similarity notion could now be identified. In a final feedback iteration we added the countries *Norway*, *UK*, *Ireland* and *Greece*, mostly to define the outer border of Europe more precisely (on the left of Figure 7).

Analysis of the Similarity Model Result

The final topology on the right of Figure 7 resembles large parts of the geographic topology of the world. We applied the rotation interaction to Similarity Model to align continents as usual. All European objects from the User Feedback View were then aligned in the upper centre of the map. On the left, the North and South American countries were revealed, on the right Asia and Polynesia was located. African countries could be seen in the lower centre. In Figure 6 the cluster of African states can be seen in detail. Again we used zooming, panning and map rotation to better exploit the calculated Similarity Model and thus, to exploit the local structure of the data set. We applied the projection quality indication. Thus we are able to distinguish between the quality of the Similarity Model and the MDS projection error. Coloured outlines of the objects shown represent the result of Venna and Kaskis Trustworthiness measure.

Interpretation of the Results

We took a closer look at the attribute weighting 'behind' the visual interface. Besides the attributes *Latitude* and *Longitude* the categorical attribute *Colour* had a strong influence on the intermediate Similarity Model, among others. In the course of the training this influence declined due to the increased number of objects in the User Feedback View. Even if our feedback of geolocations might have been imperfect, the weights of the attributes *Latitude* and *Longitude* were comparatively high from the training start, although not perfect. In the course of the training, the two 'target' attributes remained with weights near the maximum, while all other attribute weights of the data set declined.

7 DISCUSSION AND FUTURE WORK

In our implementation we had good experiences with a small data set. In future, we aim to test the concept on large data sets. It will be interesting to see where limitations of scalability exist, depending on the complexity of chosen implementations, applied data sets and chosen mental similarity notions. As pointed out in $\mathbf{S}_{13}$ and $\mathbf{S}_{14}$ the visual scalability of the Result View depends on a meaningful choice of data aggregation, visual encoding and interaction design. We refer to the cited related work to cope with particular visual

Figure 7: The final training result based on 13 objects. Left: objects are aligned to define a topology (Europe). Right: the Result View represents the calculated Similarity Model of the data set (topology of the world).

scalability challenges in this special application context. One way to tackle possible functional scalability aspects is to consistently apply multi-threaded implementation variants. Since many of the 15 mandatory steps can be exploited independently, a parallelization of tasks might be highly appropriate. One might even consider moving specific steps from CPU to GPU execution. However, this design decision must be made with care to avoid bottlenecks caused by data-transfer. Nevertheless we think that visual analytics capabilities to support users with additional guidance concepts may benefit from GPU execution in general. Another future work aspect is our goal of conducting more evaluations on different implementations. We consider the described design space as large and at the moment we are still not aware of the 'best' implementation configuration. Of course analysis goals, data sets and further task-driven aspects have an influence on appropriate design decisions. We aim to apply the presented concept in several research approaches currently envisaged. It will be interesting to see how different domain experts act with respective implementations.

8 CONCLUSION

We presented a concept for the development of visual-interactive systems that enable domain experts to express mental similarity notions for mixed data objects. The direct definition of similarity by domain experts contributes to user-centred design principles and helps to increase both the effectiveness and the efficiency in the object similarity definition process, especially if the attribute space is large and/or unknown. We also contribute to exploratory search tasks and support gaining insights in multivariate, mixed data sets. The concept is sub-divided into 15 mandatory steps. Moreover, we presented an implementation of the presented concept. The provided system contains visual encodings to support domain experts in the similarity definition process. We showed the applicability of the system in a case study with countries serving as mixed data objects. We were able to create a similarity model representing the geo-locations of the world based on similarity feedback of 13 European countries.

9 REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *TKDE*, pages 734–749, 2005.

[2] Bilal Alsallakh, Wolfgang Aigner, Silvia Miksch, and M. Eduard Groller. Reinventing the contingency wheel: scalable visual analytics of large categorical data. *TVCG*, pages 2849–2858, 2012.

[3] Mihael Ankerst, Martin Ester, and Hans-Peter Kriegel. Towards an effective cooperation of the user and the computer for classification. In *SIGKDD*, pages 179–188. ACM, 2000.

[4] K. Bache and M. Lichman. Flags Data Set - UCI machine learning repository, 2013.

[5] Sugato Basu, Mikhail Bilenko, Arindam Banerjee, and Raymond J Mooney. Probabilistic semi-supervised clustering with constraints. *Semi-supervised Learning*, pages 71–98, 2006.

[6] Jürgen Bernard, Tatiana von Landesberger, Sebastian Bremm, and Tobias Schreck. Multi-scale visual quality assessment for cluster analysis with self-organizing maps. In *VDA*, Proceedings of SPIE; 7868, pages 78680N–1–78680N–12. SPIE, 2011.

[7] Jürgen Bernard, Tobias Ruppert, Oliver Goroll, Thorsten May, and Jörn Kohlhammer. Visual-interactive preprocessing of time series data. In *SIGRAD*, pages 39–48, 2012.

[8] Jürgen Bernard, Tobias Ruppert, Maximilian Scherer, Jörn Kohlhammer, and Tobias Schreck. Content-based layouts for exploratory metadata search in scientific research data. JCDL, pages 139–148, New York, NY, USA, 2012. ACM.

[9] Jürgen Bernard, Martin Steiger, Sven Widmer, Hendrik Lücke-Tieke, Thorsten May, and Jörn Kohlhammer. Visual-interactive exploration of interesting multivariate relations in mixed research data sets. In *Computer Graphics Forum, Proceedings EuroVis, 2014, Swansea, Wales, UK*, volume 33, 2014.

[10] Jürgen Bernard, Nils Wilhelm, Björn Krüger, Thorsten May, Tobias Schreck, and Jörn Kohlhammer. Motionexplorer: Exploratory search in human motion capture data based on hierarchical aggregation. *TVCG*, 19(12):2257–2266, 2013.

[11] Rita Borgo, Johannes Kehrer, David H.S. Chung, Eamonn Maguire, Robert S. Laramee, Helwig Hauser, Matthew Ward, and Min Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *EG State of the Art Reports*, pages 39–63. Eurographics Association, 2013.

[12] Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. *SIAM*, 30(2):3, 2008.

[13] Eli T Brown, Jingjing Liu, Carla E Brodley, and Remco Chang. Dis-function: Learning distance functions interactively. In *VAST*, pages 83–92. IEEE, 2012.

[14] Sung-Hyuk Cha, Sungsoon Yoon, and Charles C Tappert. Enhancing binary feature vector similarity measures. 2005.

[15] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics & Informatics*, 8(1), 2010.

[16] Jaegul Choo, Hanseung Lee, Jaeyeon Kihm, and Haesun Park. ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *VAST*, pages 27–34. IEEE, 2010.

[17] Niklas Elmqvist and Jean-Daniel Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *TVCG*, 16(3):439–454, 2010.

[18] Mohammad Ghoniem, Jean-Daniel Fekete, and Philippe Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *Information Visualization*, pages 17–24. IEEE, 2004.

[19] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[20] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD*, pages 10–18, 2009.

[21] Douglas M Hawkins. The problem of overfitting. *Chemical information and computer sciences*, pages 1–12, 2004.

[22] Masaki Hilaga, Yoshihisa Shinagawa, Taku Kohmura, and Tosiyasu L Kunii. Topology matching for fully automatic similarity estimation of 3d shapes. In *SIGGRAPH*, pages 203–212. ACM, 2001.

[23] Stephen Ingram, Tamara Munzner, Veronika Irvine, Melanie Tory, Steven Bergner, and T Moller. Dimstiller: Workflows for dimensional analysis and reduction. In *VAST*, pages 3–10. IEEE, 2010.

[24] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[25] Dong Hyun Jeong, Caroline Ziemkiewicz, Brian Fisher, William Ribarsky, and Remco Chang. ipca: An interactive system for pca-based visual analytics. In *Computer Graphics Forum*, volume 28, pages 767–774, 2009.

[26] Ll Jie, Gao Xinbo, and Jiao Li-Cheng. A csa-based clustering algorithm for large data sets with mixed numeric and categorical values. In *WCICA*, pages 2303–2307. IEEE, 2004.

[27] Sara Johansson. Visual exploration of categorical and mixed data sets. In *SIGKDD*, pages 21–29. ACM, 2009.

[28] Sara Johansson, Mikael Jern, and Jimmy Johansson. Interactive quantification of categorical variables in mixed data sets. In *Information Visualization*, pages 3–10. IEEE, 2008.

[29] Robert Kosara, Fabian Bendix, and Helwig Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *TVCG*, 12(4):558–568, 2006.

[30] Joseph Lee Rodgers and W Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[31] Marie-Jeanne Lesot, Maria Rifqi, and H Benhadda. Similarity measures for binary and numerical data: a survey. *IJKESDP Journal*, pages 63–84, 2009.

[32] Alexander Lex, H Schulz, Marc Streit, Christian Partl, and Dieter Schmalstieg. Visbricks: multiform visualization of large, inhomogeneous data. *TVCG*, pages 2291–2300, 2011.

[33] Zhicheng Liu and John T Stasko. Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *TVCG*, 16(6):999–1008, 2010.

[34] Gladys MH Mamani, Francisco M Fatore, Luis G Nonato, and Fernando V Paulovich. User-driven feature space transformation. In *CGF*, pages 291–299, 2013.

[35] T May, J Davey, and T Ruppert. Smartstripes - looking under the hood of feature subset selection methods. In *EuroVA*, pages 13–16, 2011.

[36] Thomas Mühlbacher and Harald Piringer. A partition-based framework for building and validating regression models. *TVCG*, 19(12):1962–1971, December 2013.

[37] Tamara Munzner. A nested model for visualization design and validation. *TVCG*, 15(6):921–928, November 2009.

[38] pixolution - Visual Search. http://www.pixolution.de/index.php?id=5.

[39] A Johannes Pretorius and Jarke J Van Wijk. What does the user want to see? what do the data want to be? *Information Visualization*, 8(3):153–166, 2009.

[40] Philip H Ramsey. Critical values for spearman's rank order correlation. *Journal of Educational and Behavioral Statistics*, 14(3):245–253, 1989.

[41] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Readings in Information Retrieval*, 24:5, 1997.

[42] Tobias Schreck, Jürgen Bernard, Tatiana Tekušová, and Jörn Kohlhammer. Visual cluster analysis of trajectory data with interactive Kohonen maps. *Palgrave Macmillan Information Visualization*, 8:14–29, 2009.

[43] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52:55–66, 2010.

[44] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *VL*, pages 336–, Washington, DC, USA, 1996. IEEE Computer Society.

[45] Robert R Sokal and Peter HA Sneath. Numerical taxonomy. *Freemont, San Francisco, CA*, 1973.

[46] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S Tan. Ensemblematrix: interactive visualization to support machine learning with multiple classifiers. In *SIGCHI*, pages 1283–1292. ACM, 2009.

[47] Jarke J van Wijk. Bridging the gaps. *Computer Graphics and Applications, IEEE*, 26(6):6–9, 2006.

[48] Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *ICANN*, pages 485–491. Springer-Verlag, 2001.

[49] Malcolm Ware, Eibe Frank, Geoffrey Holmes, Mark Hall, and Ian H Witten. Interactive machine learning: letting users build classifiers. *Human-Computer Studies*, pages 281–292, 2001.

[50] Kilian Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. *NIPS*, page 1473, 2006.

[51] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State University*, 2, 2006.