Erschienen in: Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies : Graz, Austria, 2012 / Stefanie Lindstaedt and Michael Granitzer (eds.). - New York : ACM, 2012. - Article No. 22. - ISBN 978-1-4503-1242-4

Guided Discovery of Interesting Relationships Between Time Series Clusters and Metadata Properties

Jürgen Bernard
Fraunhofer Institute for Computer Graphics Research
Fraunhoferstr. 5
D-64283 Darmstadt, Germany
juergen.bernard@igd.fraunhofer.de

Tobias Ruppert
Fraunhofer Institute for Computer Graphics Research
Fraunhoferstr. 5
D-64283 Darmstadt, Germany
tobias.ruppert@igd.fraunhofer.de

Maximilian Scherer
TU Darmstadt
Fraunhoferstr. 5
D-64283 Darmstadt, Germany
maximilian.scherer@gris.tu-darmstadt.de

Tobias Schreck
University of Konstanz
Universitätsstr. 10
D-78457 Konstanz, Germany
tobias.schreck@uni-konstanz.de

Jörn Kohlhammer
Fraunhofer Institute for Computer Graphics Research
Fraunhoferstr. 5
D-64283 Darmstadt, Germany
joern.kohlhammer@igd.fraunhofer.de

ABSTRACT

Visual cluster analysis provides valuable tools that help analysts to understand large data sets in terms of representative clusters and relationships thereof. Often, the found clusters are to be understood in context of belonging categorical, numerical or textual metadata which are given for the data elements. While often not part of the clustering process, such metadata play an important role and need to be considered during the interactive cluster exploration process. Traditionally, linked-views allow to relate (or loosely speaking: correlate) clusters with metadata or other properties of the underlying cluster data. Manually inspecting the distribution of metadata for each cluster in a linked-view approach is tedious, especially for large data sets, where a large search problem arises. Fully interactive search for potentially useful or interesting cluster to metadata relationships may constitute a cumbersome and long process.

To remedy this problem, we propose a novel approach for guiding users in discovering interesting relationships between clusters and associated metadata. Its goal is to guide the analyst through the potentially huge search space. We focus in our work on metadata of categorical type, which can be summarized for a cluster in form of a histogram. We start from a given visual cluster representation, and compute certain measures of interestingness defined on the distribution of metadata categories for the clusters. These measures are used to automatically score and rank the clusters for potential interestingness regarding the distribution of categorical metadata. Identified interesting relationships are highlighted in the visual cluster representation for easy inspection by the user. We present a system implementing an encompassing, yet extensible, set of interestingness scores for categorical metadata, which can also be extended to numerical metadata. Appropriate visual representations are provided for showing the visual correlations, as well as the calculated ranking scores. Focusing on clusters of time series data, we test our approach on a large real-world data set of time-oriented scientific research data, demonstrating how specific interesting views are automatically identified, supporting the analyst discovering interesting and visually understandable relationships.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3 [**Information Storage and Retrieval**]: Digital Libraries; I.5 [**Pattern Recognition**]: Clustering; J.2 [**Physical Sciences and Engineering**]: Earth and atmospheric sciences

General Terms

Design

Keywords

Information Visualization, Visual Analytics, Visual Cluster Analysis, Interestingness Measures, Scientific Research Data, Guided Data Exploration

1. INTRODUCTION

Large data measurements are collected in many scientific disciplines like experimental physics, economics, sociology, medical science or the geosciences. Besides analyzing the data itself for certain patterns, researchers are often interested in the relationship between this data content and the conditions under which it was measured, which is described as metadata. For example, consider temperature measurements and the time or location of the measurement; or an EEG measurement and patient data like gender, age, smoker/nonsmoker, etc. Facing large data sets, two data analysis problems arise, namely, (1) understanding the main characteristics of the measurement data, or content data, itself, and (2) the relationships between the content data and the metadata. For (1), cluster

Konstanzer Online-Publikations-System (KOPS)
URL: http://nbn-resolving.de/urn:nbn:de:bsz:352-227057

analysis is popular, as it aggregates many data records to a smaller number of cluster prototypes, which can be more easily compared against each other. Visual cluster analysis supports identification and comparison of clusters, typically by finding (a) appropriate visual representations for the individual clusters, and (b) finding an appropriate 2D layout for the clusters. (2) can be considered a correlation problem, where the degree of variation of one variable (in our case: clusters) with another (in our case: metadata) is considered. For both problems in itself, visualization has shown to be beneficial in exploring data sets, and understanding clusters and correlations. However, considering both problems together has not been addressed extensively to date.

Finding interesting correlations between cluster content and metadata is a challenge, as the data space is often very large. For scientific observation data, usually a multitude of metadata exists that could be relevant for forming hypotheses by the analyst. The metadata in general can take many different forms, e.g. be numerical, nominal, ordinal, textual. Discovering relationships can be supported by supervised machine learning techniques such as regression or classification, or unsupervised clustering, that incorporate metadata properties directly in their learning process. However, these approaches require the definition of a suitable model to combine all these data types. Usually, only domain experts have appropriate insight into the data to devise such models, and yet, few domain experts have a strong technical background in machine learning. Both of which, however would be needed to incorporate these models into the learning algorithms. Therefore, we propose an approach to automatically identify potentially interesting metadata properties of clusters, for exploration by the user. The measure of interestingness is based on the homogeneity of the distribution of the regarded metadata entities for a given clustering, as well as the dispersion of said distribution among the neighborhood of clusters in a 2D arrangement.

Specifically, we employ unsupervised clustering of the content data, without incorporating the metadata, using the Self-Organizing Map method, which yields an overview of the content data in a two-dimensional layout. Now we guide domain experts to interesting relationships between content data and metadata, by visualizing the distribution of the metadata among the content data clusters. Thus, a user can easily spot clusters which are not only similar concerning the content data but also show a homogeneous distribution of metadata entities. This allows discovery of interesting relationships between the content data (i.e. a temperature measurement over time) and metadata (i.e. geolocation of the measurement).

The remainder of this paper is structured as follows. In Section 2 we discuss related work in several areas. In Section 3 and 4, we motivate our approach and define a couple of measures for automatically screening those metadata properties with a potentially interesting relationship to content data. In Section 5, we apply our implementation to a real-world data set, demonstrating the usefulness of the approach. Finally, the Sections 6 and 7 summarize this paper and discuss future extensions of our approach.

2. RELATED WORK

2.1 Visual Cluster Analysis and High-Dimensional Visual Correlation Analysis

Many clustering methods such as k-Means, DBScan and so forth have been proposed and applied to date [12]. Visual cluster analysis focuses on the visualization of properties between and within clusters. To compare clusters against each other, projection techniques such as PCA or MDS [3] can be applied as a post-processing step after clustering. Alternatively, several methods employ topological restrictions, such as the Self-Organizing Map algorithm [16]. Intra-cluster properties typically are visualized by specific data-dependent representations for the cluster centroids or medoids; in case of clusters in high-dimensional feature space, techniques such as Parallel Coordinates [7] can be used.

Correlation of cluster views against a target variable can be visualized, in the simple case, by means of background color-coding and overlay of glyphs. Linked-view displays allow to select clusters of interest and highlight corresponding data in auxiliary views, see [9] as an example.

Several systems for visual detection of interesting correlations in multivariate data exist. Evaluation and comparison of a set of given regression models in multiple dimensions is described in [20]. The system incorporates sophisticated navigation controls, but is not primarily tailored towards guiding the user through the analysis space.

In [11], a system is proposed that guides the user in identifying well-fitting linear regressions models in multiple dimensions. It is based on a series of heatmap displays, and also supports finding good correlations in a subset of the data. The same authors proposed another framework that supports the analysis of spatio-temporal data [10]. Here, Self-Organizing Maps, parallel coordinate plots and two-dimensional cartographic color design methods are connected in an interactive framework.

In the works of [4] and [1], the authors present cluster results of multidimensional data content arranged with layouts based on categorical metadata. Similar to our approach, the combination of data content and metadata is considered. In contrast, the layout of our data clusters depends on the data content, not on the metadata.

2.2 Interestingness-based Visual Analysis Support

Guiding the user through a potentially large analysis space benefits from a notion of interestingness that is defined in the data or the view space.

Geng and Hamilton give an overview of several ways to determine interestingness in data mining in general [8]. Given such a notion, visual analysis benefits by automatic filtering, highlighting and grouping the data accordingly.

For bivariate data, a scatter plot is an intuitive visualization for humans to judge if the relationship between the two visualized variables is interesting or not. Thus, computational methods to automatically analyze bivariate data for interestingness in a similar way have been proposed [27, 26, 25, 22].The basic idea for unlabeled data is to analyze the distribution of points in a scatter plot, where usually a narrow distribution is considered interesting. The computational techniques to achieve that vary from graph analysis [27], to image-based techniques [26] up to goodness-of-fit parameters from regression [22]. For labeled data, the separability of clusters in a scatter plot is analyzed [25].

In [23], an interestingness-based approach for cluster analysis in pixel-based visualizations is proposed. It is based on measuring the entropy contained in a pixel display of the cluster spaces, and useful for screening a large space of candidate clusterings.

Another example application is presented in [5]. The interestingness definition of data clusters and labelling data is based on a recommendation system. The Wikipedia library is used to enhance the cluster labeling quality of manually labeled clusters.

While several approaches to date allow to measure the interestingness of views, they often are based on heuristics. A rather open challenge is to assess the relation between the used interestingness scores, and the application-dependent interestingness as perceived by the analysis.

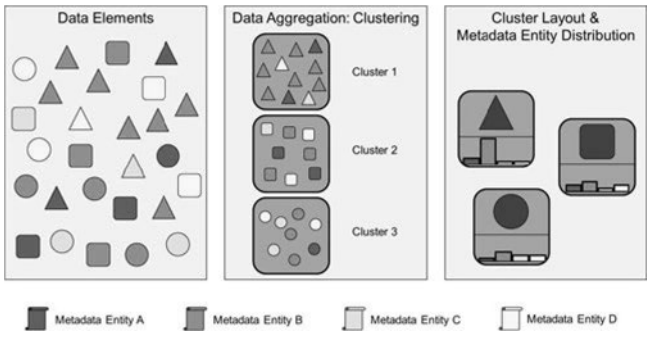

Figure 1: Example data set, aggregated by a clustering algorithm (k=3) and arranged in a particular layout. The metadata entity distribution (A, B, C and D) is considered. Our interestingness functions consider the distribution of metadata categories among clusters, identifying candidate relations.

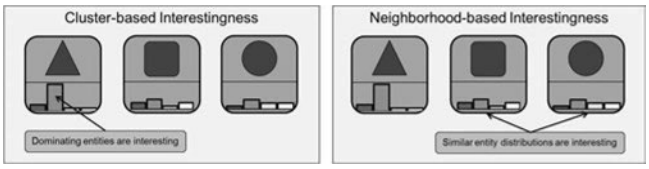

Figure 2: Two interestingness concepts. Left: cluster-based. Single-entity dominated clusters are more interesting compared to clusters with diverse-distributed metadata entities. Right: neighborhood-based. Neighboring clusters with similar metadata entity distributions are interesting.

3. MEASURING RELATIONSHIPS BETWEEN CLUSTERS AND METADATA

3.1 Problem Definition and Basic Idea

Finding interesting relationships between clusters and associated metadata is an important challenge. For example, in the natural sciences, experimental data is often annotated with the conditions under which it was collected. Automatic machine learning techniques to find and model the relationships between cluster data and metadata are only suitable to prove or disprove specific hypotheses. The number of hypotheses for relationships between metadata and content data is potentially very large, and cannot efficiently be manually specified or explored.

Thus, techniques to guide users in efficiently finding interesting relationships between content data and metadata are important.

We denote data which is to be clustered as *content data*. In our application, we consider time-oriented measurement data as content data. The daily time series data elements are additionally described by *metadata* of nominal or numerical type. All content data comprises a metadata record of the same schema, and the metadata is not part of the clustering method (see a generalized example in Figure 1).

Our approach relies on a *visual clustering solution*, and in particular, a projection of the cluster prototypes to a 2D regular grid. We propose the following approach to identify and visualize potentially interesting relations between content data and metadata:

1. First, we aggregate the content data using a cluster algorithm. In our approach, we use the Self-Organizing Map algorithm (SOM) [16] for its convenience with respect to visualization. The method creates a two-dimensional organization of the cluster prototypes on a regular grid which is the basis for further informative overlays. Note that our approach is not limited to SOM. Alternative clustering algorithms (e.g., k-Means, DBScan, etc.) combined with a projection (e.g., PCA, MDS, etc.) of the cluster prototypes can be used.

2. Then, we compute a metadata entity distribution histogram for each data cluster. Two specific interestingness measures defined on the histograms (cf. Section 3.2) are computed and shown to the user, for assessing the interestingness of the selected metadata property for the cluster result. On the one hand, our *cluster-based* interestingness measure identifies single sharp peaks of metadata entities within each cluster (cf. Section 3.2.1). Furthermore, the analyst is guided to discover interesting *neighborhood-based* metadata distributions among neighboring clusters. Neighboring clusters (in terms of similar data content) that contain similar metadata distributions are expected to be interesting (cf. Section 3.2.2).

3. We provide an interactive visualization which allows to visually explore the space of possibly interesting metadata properties, based on the calculated scores. The cluster prototypes are visualized for an overview, together with the interestingness scores (via color mapping) and the distribution of the metadata entities (in form of a bar chart). See Section 4.

4. Both interestingness measures can be used to automatically survey the space of possibly relevant metadata properties, and present to the user a ranking of the most interesting metadata properties (cf. Section 4.3). This ranking is useful when facing large metadata spaces, which cannot be explored manually.

3.2 Definition of Interestingness Measures

We propose two definitions of interestingness to guide the visual exploration - cluster-based and neighborhood-based interestingness measures (see also Figure 2). In the following, we will define these measures, taking into account the data type (nominal or numerical) of the corresponding metadata property. The chosen measures are an initial set of measures proposed heuristically, and expected to be a good starting point (see also Section 6 for a comparison of the measures). They could easily be extended by more specific measures, based on application needs.

3.2.1 Cluster-based Interestingness Measure

The main objective of our cluster-based interestingness measures is to describe the *diversity of the metadata entity distribution* in each individual cluster in order to precisely describe clusters in the context of additional metadata properties. Obviously, a higher interestingness between the cluster and the metadata property is obtained, if only small variations in the metadata entity distribution of the considered cluster exist, hence a small diversity (cf. Figure 2 left).

Reflecting this characteristic for nominal metadata properties, we choose Simpson's index (see equation 1), referred in the literature as one of the most prominent measures of diversity [24], as an interestingness measure. The Simpson's index, taking values between 0 and 1, becomes 1, if all elements of a distribution lie in one bucket. A lower index reflects more diversity of the distribution, which is considered less interesting from an inner cluster-based perspective. To enable the comparison of numerical and nominal metadata properties, we use the same index for both data types. Since the Simpson's index is an appropriate measure for discrete

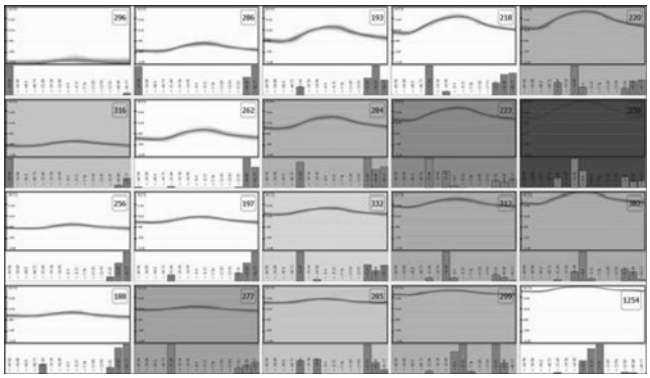

Figure 3: Cluster-based interestingness on a daily temperature clustering: the 'Latitude' emerges to be the best metadata property for the discovery of cluster-based interestingness (cf. Figure 5).

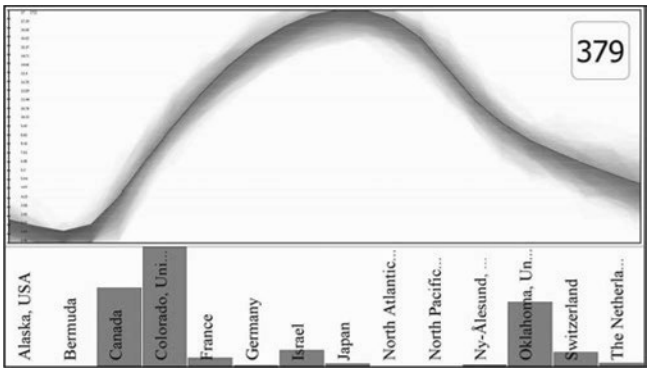

Figure 4: Enlarged view of a single cluster. The cluster centroid (measurement: 24h operating temperature range) is displayed with a black line, data elements are illustrated with transparency bundles, a bar chart represents the metadata distribution. The background color describes the interestingness, the respective colormap is presented in Figure 5.

distributions, the numerical attributes have to be discretized. This is done via a binning of the numerical values with a fixed number of buckets.

$$l = 1 - \sum_{i=1}^{Z} \frac{n_i(n_i-1)}{N(N-1)} \quad (1)$$

Here, n_i denotes the number of entities regarding the i^{th} metadata property and N is the overall number of entities of the cluster. Depending on the data type of the metadata property, for each cell of the visual cluster representation, the Simpson's index as a cluster-based interestingness measure is computed.

3.2.2 Neighborhood-based Interestingness Measure

In addition to cluster-based interestingness, we introduce neighborhood-based interestingness measures. Based on the content-based similarity of clusters, the *homogeneity of metadata distributions among a visual cluster representation* is evaluated. As an assumption, a metadata property is related to a cluster result (and thus a good interestingness candidate), if the regarding metadata entity distributions in the vicinity of a cluster are similar (cf. Figure 2 right).

Hence, we are looking for *similar distributions of the metadata entities for similar clusters* to (a) find metadata properties that have a global relation to the cluster result and (b) to discover neighbored clusters with similar metadata histograms in order to guide the reduction of the number of clusters in further aggregation enterprises.

Regarding numerical metadata properties, for the comparison of distributions we again use the discretized metadata entities of section (3.2.1). For the neighborhood-based scores, the relative frequencies are computed. The resulting distribution vectors are used to compute the similarity between two distributions via the earth mover distance (EMD) [21]. We choose this metric, since the discretization of numerical values might result in an inaccurate bucketing of similar values, and the EMD calculates the minimal costs that are needed to transform one bucketing signature into another, including 'cross-bin' comparisons. In contrast, for nominal metadata properties, we use the Euclidean distance between the distribution vectors, containing the relative frequencies of the categories, as a measure of similarity between two metadata entity distributions. This is a well known 'bin-by-bin' comparison method. Since only relative frequencies are used, both measures lie within the interval [0;1], with 0 resulting from an identical distribution, while higher values indicate decreasing similarity. The similarities are computed between a distinct cluster and its most similar clusters. Then, for each individual cluster an average is calculated, weighting the neighbors with respect to their distance to the regarded cluster.

Like the cluster-based interestingness measures, the neighborhood-based average scores are computed for each visual cluster. We point out that due to the nature of this neighborhood-based measures, not all metadata properties proposed as interesting have a sensitive relation to the clustering in a subtle manner (in the worst case think of an entire uniform histogram distribution that is also homogeneous), but all metadata properties with subtle relation (implying topological histogram orderings) will be recognized as interesting. Practically speaking, metadata properties with heterogeneous histogram distributions are classified uninteresting as expected. We are aware that the effectiveness of visually assessing neighborhood-based measures depends on the layout technique. The choice of the projection method should to be considered with care (cf. Sections 4.2 and 6).

4. VISUALIZING CLUSTERS AND METADATA PROPERTIES

In the following, we introduce the visual design for the interactive exploration of interesting metadata properties.

4.1 Clustering and Layout Algorithm

After ability tests for cluster layouts including PCA, MDS and the SOM, we use the SOM in our approach (see e.g. Figure 3). A beneficial criterion is the ability of the algorithm for both: clustering and projection. Furthermore, the regular grid structure reduces mutual cluster occlusion problems to a minimum. In contrary to PCA, the projection is not restricted to the degree of information of the first two principal components and in contrary to PCA and MDS, the SOM tends to equally distribute the data on the entire display space (we do not want to generalize this finding, but in consideration of our data and our analysis task this criterion is crucial). An existent cause for discussion is the property that often not single but distinct cells together form one data cluster in a more classical cluster notion, usually measured by compactness and separation indices. Despite this, the SOM aggregates the data content in a topology-preserving way, and this is precisely our starting point

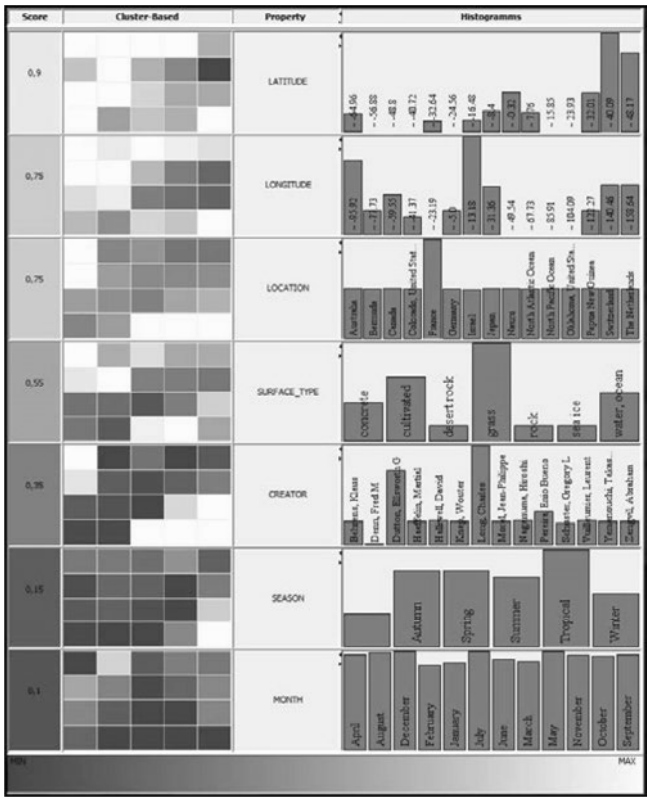

Figure 5: List of metadata properties, ranked by their interestingness (here: cluster-based). Each property is represented by a score, a preview on the colored cluster layout (SOM) and a global metadata entity distribution bar chart. The colormap to depict the interestingness from low (dark green) to high (light yellow) is presented at the bottom.

to apply guided discovery of interesting relationships between data content and metadata.

Despite our choice of the SOM we emphasize, that alternative clustering methods combined with a projection of the cluster prototypes in 2D can be used as a starting point for our approach.

4.2 Visual-Interactive Cluster Design

In the upper part of each cluster visualization (cf. Figure 4) a visual prototype, representing the cluster centroid, is depicted. In the lower part, the distribution of the metadata property in the cluster is visualized via a bar chart. According to the data type of the metadata property, nominal categories or discretized numerical values define the buckets of the metadata bar chart. Hence, the user can get a notion of the homogeneity of the metadata entity distribution in the cluster. Each cluster is colored representing the value of the chosen interestingness measure, cluster-based or neighborhood-based. The color scale ranges from dark green, corresponding to low interestingness, to bright yellow, depicting high interestingness. The color map can be seen in Figure 5. The user can choose between a *local and a global scaling* of interestingness measures. The local variant scales the chosen metadata property according to its respective maximum and minimum and maps it to the color range. In the global variant, scaling is done regarding the minimum and maximum over all metadata properties in the data, and therefore, interestingness can be quantitatively compared across all metadata properties (see Section 4.3).

We provide interactive zooming functionality. Thus, the analyst can get deeper insight into the (overall and cluster-based) histograms and the cluster cells on demand. Rescaling the display and enlarging details is supported until a single cluster is enlarged up to display size (cf. Figure 4). Thereby, we implicitly provide a *single cluster view*. In case of time series data, such as used in our application, an opacity-bands approach [7] can be used, to depict the content data allocated to the visual cluster prototype. The histogram visualization is augmented with labels for the buckets. Individual buckets can be selected and highlighted, those buckets are highlighted with white background color.

4.3 Ranking of Metadata Properties by Correlation to Cluster Data

In addition to the visual-interactive access to interesting candidate views (only one metadata property is explored at a time) we provide an automatic ranking method for all metadata properties based on the interestingness measures. Due to high quantities of metadata it might be infeasible to visually explore every possible metadata view. In that case an automatic ranking of the metadata properties regarding their overall interestingness is useful.

In the following, we propose such a measure. The calculation can be executed on both cluster-based and neighborhood-based interestingness measures. The user can interactively choose, which of the two measures is used for calculating the ranking.

In order to provide a comparable measure, we normalize each interestingness value based on the overall minimum and maximum of all metadata properties. Thereby, we obtain for each cluster cell one global interestingness measure per metadata property. The resulting $n \times m$ measures (n: # of clusters, m: # of metadata properties) are sorted in a list and the median is calculated. Now, we count for each metadata property the number of values that are greater than the median. The ratio of this number and the number of clusters is used as the overall interestingness measure for each property, that defines the ranking.

The result of this ranking is visualized in a list view (see Figure 5). In the first column, the overall interestingness measure for each metadata property is given, ranging from 1 (all clusters have a global interestingness measure above the median, regarding one respective metadata property) to 0. The background color is calculated based on a normalized color map reflecting the ranking score. In the second column a minimized cluster visualization (introduced in Section 4.1) simply reflecting the interestingness values of each cluster is shown. In the third column the title of the respective metadata property is shown. In the last column a histogramm of all metadata entities for the regarded metatdata property is visualized. The applied colormap is displayed at the bottom of Figure 5.

5. APPLICATION

We study the exploratory goal of discovering relations between time series clusters and categorical metadata data in a real-world earth observation data set. We use this real-world data (1) to prove our two interestingness concepts and (2) to give some illustrative examples of how our system could be used to show interesting relations in a challenging and previously undiscovered data set, recognizing that we are not domain experts. We proved the correctness of our data findings by collaborating with domain experts from the Alfred Wegener Institute (AWI) for Polar and Marine Research in Bremerhaven, Germany.

5.1 The Research Data Set

We consider research data from the scientific data repository PANGAEA [19] operated by the AWI. PANGAEA archives, pub-

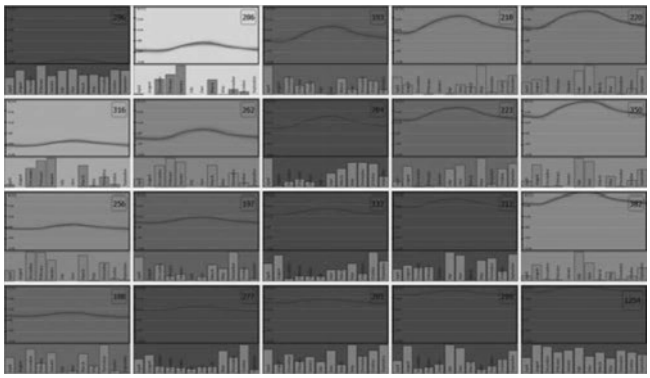

Figure 6: Cluster-based interestingness on a daily temperature clustering: the metadata property 'Month' is not suitable to describe the cluster result in a manner of cluster-based interestingness, since only diverse entity distributions are discovered.

lishes, and distributes geo-referenced scientific earth observation data. Our data set focuses on atmospheric weather measurements, gathered in the scope of the Baseline Surface Radiation Network (BSRN) [18], a PANGAEA compartment. The research data is annotated with metadata information on citations, originating project name, spatial and temporal conditions, parameter description, etc. Such rich available metadata is well suited for our application to discover relationships between these metadata properties and aggregations of the observation (content-based) data.

Our data pool consists of time series measurements from 2006 measured at various stations of BSRN all over the world [2]. We focus on temperature curves, as temperature oscillation is a rather intuitive physical process. After discussion with domain experts from BSRN we created the following application scenario.

Clustering of time series data is a non-trivial task, as during precalculation steps, many degrees of freedom exist and have to be considered. Preprocessing steps like missing value replacement, outlier handling and quantization techniques can be applied. Keogh [15] and Liao [17] give a broad overview to the field, also regarding the calculation of time series descriptors and time series similarity measures. As the number of observations per time series comes to 50000 per file (amounting to about 10 million time stamps in total), we segment the long measurement time series into daily series. Then we apply Piecewise Aggregate Approximation descriptor [14] with 24 bins (one per hour) on the time series, to obtain low-dimensional data representations (feature vectors). The resulting 6430 feature vectors are clustered and layouted with the Self-Organizing Map algorithm, using rules of thumb parameters [6, 16].

We next show results of applying our method to temperature measurement time series from the above mentioned data set. After the clustering step is completed, a visual overview of the data content is presented via the cluster representations. Now the user can interactively select one of the metadata properties, and the distribution of the metadata entities within each cluster cell is visualized. According to the interestingness measures introduced in Section 3.2 a color mapping is computed, either visualizing the cluster-based, or the neighborhood-based interestingness. The attributes the user can select from are presented in a ranked list in descending order of potential interestingness as described before.

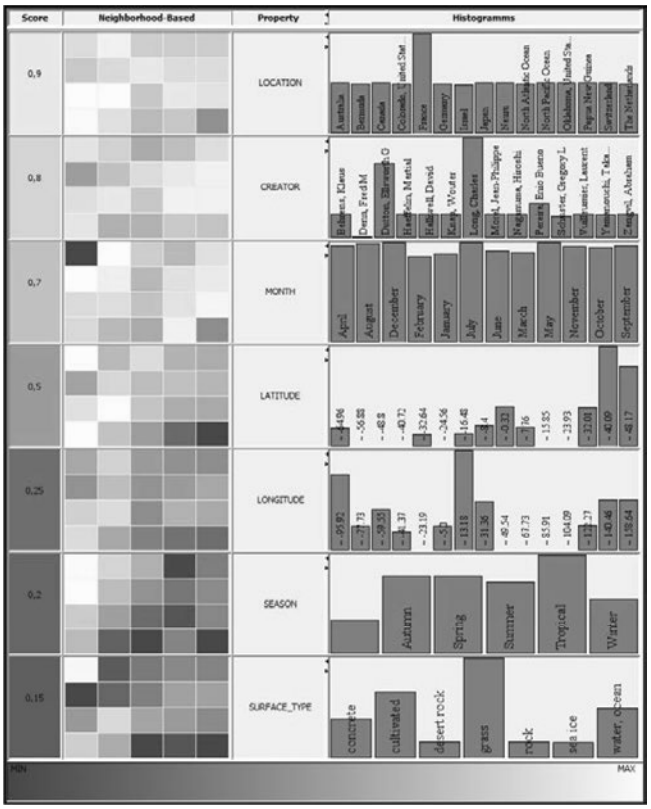

Figure 7: List of metadata properties, ranked by their neighborhood-based interestingness. Properties 'Location', 'Creator' and 'Month' are ranked best; properties 'Season' and 'Surface_Type' remain weak.

5.2 Cluster-based Interestingness

We first focus on our cluster-based interestingness measure, discovering peak distribution histograms of the metadata entities at distinct clusters. Coming back to Figure 5, we use the metadata property ranking list as a global overview and starting point for the available metadata space to discover.

'Latitude' is outlined to be the most interesting metadata property, as its ranking score is 0.9 (90% of its clusters hold interestingness values above the global median). We select the 'Latitude' property and discover a variety of clusters that can be well described by very few metadata entities (cf. Figure 3). For example, measurements of the upper left cluster (coldest of all measurements) are exclusively located at the lowest (-65°) and the highest (50°) latitudes in the data set. Based on this local finding and without effort, the researchers from AWI state a meaningful global relation between the metadata property 'Location' and the cluster result: cold temperature measurements originate from polar regions on earth, warm measurements (at the lower right of the cluster layout) can be found in equator-near regions. This relation is observable in a great variety of clusters.

Besides the discovery of interesting relations between a clustering metadata properties of the data set, we also want to prove our concept. Thus, we explore the 'Month' property in Figure 6 on purpose of evaluating worst interestingness values calculated on the given metadata set. It can be seen that the entity distribution is not suitable to describe the distinct clusters due to high histogram diversities.

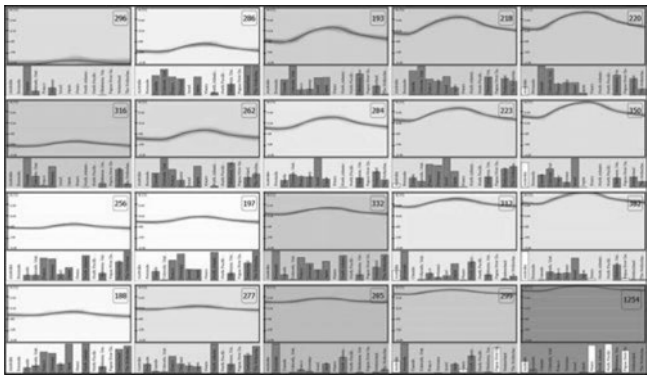

Figure 8: Neighborhood-based interestingness on a daily temperature clustering: The 'Location' property proves to hold the most highest interestingness of all metadata properties. Four selected locations (white bars) only appear in the lower right of the cluster layout. The researchers assign the highest temperature measurements of the data set to this metadata entities.

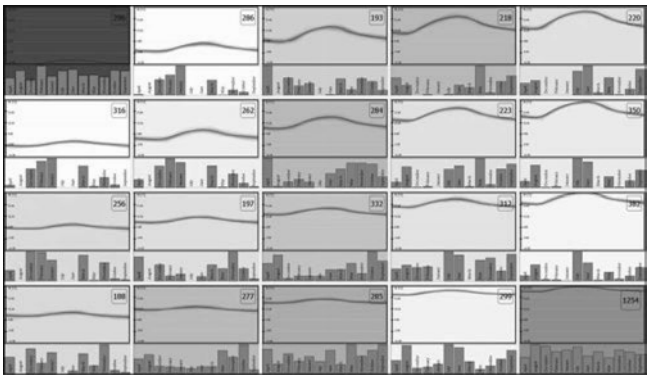

Figure 9: Neighborhood-based interestingness on a daily temperature clustering: The property 'Month' also shows some interesting regions on the cluster layout. Apart from that, a rather uninteresting cluster in the upper left sticks out: the histogram of this cluster does not relate to those of its neighbors.

5.3 Neighborhood-based Interestingness Measure

Facing the neighborhood-based interestingness measure, we again investigate the ranked metadata property list in Figure 7.

We discover relations of 'Location', the best ranked metadata property in Figure 8. We ascertain constant good interestingness values, since most of the histograms correlate quite well to those of their neighbors. In the lower left, four clusters have almost identical cluster values AND metadata histograms, this might be an indicator for recalculating the clustering with a lower number of clusters or merging those similar clusters to a super cluster. Similar observations can also be made at other parts of the cluster layout. Since merging of clusters is not the analysis goal of this work and just a starting point for further approaches, we refer to the discussion in Section 6.

Another interesting property is 'Month' in Figure 9. Besides a few outlier clusters (particularly the cluster at the upper left with the lowest temperature values of the entire data set), all clusters exhibit similar relationships to their neighbors above-average.

We want to take a closer look to the (unordered) distribution of the month entities along the global cluster layout. The researchers state a strong gradient from dominating winter months at the left to dominating summer months at the right of our layout. This shows the relation of the chosen metadata property to the cluster result.

6. DISCUSSION, LIMITATIONS AND POSSIBLE EXTENSIONS

The proposed approach helps users search through a larger space of metadata parameters (a bin, or a category), identifying those which show a strong relationship to a cluster, or neighborhood of clusters. We next discuss in turn the used interestingness measures, baseline clustering, and visual representation.

We proposed two measures of interestingness to guide the user in the task of finding relationships between data clusters and metadata attributes: The *cluster-based measure* (see Section 3.2.1) identifies single clusters who are flagged interesting when they show a peaked metadata distribution. Such a distribution indicates that there is a single clear correspondence between that given cluster and a single metadata property. The size of the cluster for which this relation holds is not considered in the measure. It means, that the number of data entities which support the finding, may also be small, if the underlying cluster is relatively small. This however, could be accommodated by an appropriate setting of the clustering method. The *neighborhood-based measure* (see Section 3.2.2) considers data portions interesting, if a larger neighborhood (a larger support of clusters) shows all similar metadata distributions. The measure is irrespective of the distributions having a peak, but accepts any distribution, as long as it remains similar for a larger neighborhood. Both measures constitute heuristics and imply that the user should inspect the found portions, for validation and explanation. Our measures help to provide a starting point for the analysis, but they also require the human analysis to make sense. We note that our approach is not limited to the described distribution measures, but could accommodate many more rules based on cluster sizes and metadata distributions to search for. An interesting future work would be to define a user editor for efficient specification and adaption of such interestingness rules.

In Section 4.1 we argued that alternative clustering and projection methods can be used as a starting point to our approach. Despite that fact, the interpretability of the outcomes heavily rely on the existence of a meaningful clustering and projection layout (cf. the garbage in-garbage out principle). For the analysis to be meaningful, cluster validation and projection quality measures, e.g., based on stress analysis, could be employed, to support the findings. The applied SOM method is convenient in that it provided clustering and projection in one step. On the other hand, the SOM method, for cases, can fail to meaningfully represent groups in data [13] and alternative clustering methods should be considered. A visual comparison of results based on multiple clustering and layout techniques can be considered in future works.

For visual inspection of clusters and metadata, we rely on a grid-based view of overlaid glyphs (time series and bar charts). We recognize that appropriate visual design for cluster prototypes is important to support the user in interpreting the differences between clusters and the correlation to the metadata properties. For general high-dimensional content data, this is a challenge in its own. For alternative data types, different glyphs may be useful. It remains to be explored how effective this overlay is for different numbers of clusters and complexities of the respective glyphs. We assume that for growing numbers of clusters and bar chart dimensions, the display may become cluttered and more scalable visual representations may become necessary.

Our presented use cases are only a first step toward a more encompassing application of the method. While we have applied it to earth observation data, discussion of results with domain experts remains to be done. From such discussion, we expect possible improvements and new measures to specify what are interesting relationships between clusters and metadata fields. Our approach currently considers the cluster data in its entirety. An interesting challenge is to support efficient screening of subsets of clusters for which interesting correlations may exist. Also, interestingness measures based on combinations of metadata fields could be possible, yet they will increase the search space. Finally, iterative visual-interactive cluster refinement strategies based on the additional metadata can be considered.

7. CONCLUSIONS

We have considered an approach for supporting the user in navigating possibly large data spaces, searching for interesting correlations between data clusters and associated metadata. As manually inspecting all possible correlations between clusters and metadata properties is infeasible for large data sets, automatic guidance measures are needed. We proposed an approach to rate the interestingness of nominal and numerical parameters with respect to their homogeneity within a given cluster, as well as the homogeneity across a 2D map of clusters. Our approach is applicable to any clustering algorithm that provides a mapping to 2D, and the set of interestingness measures is extensible. We showed the applicability of the approach for discovering interesting relationships between content data and metadata in a large real-world data set.

Acknowledgments

We thank the Alfred Wegener Institute (AWI) in Bremerhaven, Germany, for kindly supporting this research effort. Rainer Sieger, Gert König-Langlo and Hannes Grobe helped us in developing an initial understanding of the data domain, and in selecting an appropriate PANGAEA data subset [2]. We are especially grateful to the many scientists that contributed the data available through BSRN.

8. REFERENCES

[1] J. Bernard, T. Ruppert, M. Scherer, J. Kohlhammer, and T. Schreck. Content-based layouts for exploratory metadata search in scientific research data. In *Proceedings of the int. ACM/IEEE JCDL*, pages 139–148, NY, USA, 2012. ACM.

[2] J. Bernard, T. Ruppert, M. Scherer, T. Schreck, and J. Kohlhammer. Reference list of 265 sources used for the discovery of relationships between data clusters and metadata properties. doi:10.1594/pangaea.785666, 2012.

[3] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling*. Springer, 2nd edition, 2005.

[4] N. Cao, D. Gotz, J. Sun, and H. Qu. Dicon: Interactive visual analysis of multidimensional clusters. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2581–2590, 2011.

[5] D. Carmel, H. Roitman, and N. Zwerdling. Enhancing cluster labeling using wikipedia. In *ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2009. ACM.

[6] G. Deboeck and T. Kohonen. *Visual explorations in finance: with self-organizing maps*, volume 2. Springer, 1998.

[7] Y.-H. Fua, M. Ward, and E. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *Proc. IEEE Conference on Visualization (VIS)*, pages 43–50, 1999.

[8] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *Comput. Surv.*, 38, 2006.

[9] D. Guo, J. Chen, A. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1461–1474, 2006.

[10] D. Guo, J. Chen, A. M. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics*, 12:1461–1474, 2006.

[11] Z. Guo, M. O. Ward, and E. A. Rundensteiner. Model space visualization for multivariate linear trend discovery. *IEEE Symposium on VAST*, pages 75–82, 2009.

[12] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kauffman, 2nd edition, 2006.

[13] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, corrected edition, July 2003.

[14] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, 3(3):263–286, 2001.

[15] E. Keogh and S. Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003.

[16] T. Kohonen. *Self-Organizing Maps*. Springer, 3rd edition, 2001.

[17] T. W. Liao. Clustering of time series data-a survey. *Pattern Recognition*, 38:1857–1874, 2005.

[18] A. Ohmura, E. G. Dutton, B. Forgan, C. Fröhlich, H. Gilgen, H. Hegner, A. Heimo, G. König-Langlo, B. mcarthur, G. Müller, R. Philipona, R. Pinker, C. H. Whitlock, K. Dehne, and M. Wild. Baseline surface radiation network (BSRN/WCRP): New precision radiometry for climate research. *Bull. Amer. Met. Soc.*, 79:2115–2136, 1998.

[19] PANGAEA - Data Publisher for Earth and Environmental Science. http://www.pangaea.de/.

[20] H. Piringer, W. Berger, and J. Krasser. Hypermoval: Interactive visual validation of regression models for real-time simulation. *Comput. Graph. Forum*, 2010.

[21] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[22] M. Scherer, J. Bernard, and T. Schreck. Retrieval and exploratory search in multivariate research data repositories using regressional features. In *Proceedings of the int. ACM/IEEE JCDL*, pages 363–372, NY, USA, 2011. ACM.

[23] J. Schneidewind, M. Sips, and D. A. Keim. Pixnostics: Towards measuring the value of visualization. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, 2006.

[24] E. Simpson. Measurement of diversity. *Nature*, 163(4148):688–688, 1949.

[25] M. Sips, B. Neubert, J. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009.

[26] A. Tatu, G. Albuquerque, M. Eisemann, P. Bak, H. Theisel, M. Magnor, and D. Keim. Automated visual analysis methods for an effective exploration of high-dimensional data. *IEEE Transactions on TVCG, to appear*, 2010.

[27] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proceedings of IEEE Symposium on Information Visualization*, pages 157 – 164, 2005.