Interactive Exploration of Larger Pattern Collections: A Case Study on a Cocktail Dataset

Daniel Paurat
University of Bonn
daniel.paurat@uni-bonn.de

Roman Garnett
University of Bonn
rgarnett@uni-bonn.de

Thomas Gärtner
University of Bonn and Fraunhofer IAIS
thomas.gaertner@uni-bonn.de

ABSTRACT

We present a general method for employing interactive embedding techniques to enable an analyst to explore a larger collection of local patterns. The common idea among pattern-mining methods is to list descriptions of subsets of a dataset according to some interestingness measure. Because the space of all patterns in a dataset is exponentially large in the number of attributes, most pattern-mining algorithms reduce the output for the analyst to a small set of highly interesting and diverse patterns. However, by discarding most of the patterns, these methods have to make a trade-off between ruling out potentially insightful patterns and possibly drowning the analyst in results. We propose an alternative. To counteract information overload, we mine a rather large set of patterns and study this collection using an interactive embedding technique. Using this interactive, visually driven exploration technique, the analyst can develop an understanding of the patterns, their distribution, the concepts underlying them, and how they interrelate.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Interaction styles*

General Terms

Rule and pattern mining, Exploratory analysis

1. INTRODUCTION

We propose an extension to the classical pattern-mining approach. Our idea is to not focus on condensing the resulting output to a small set of high-quality patterns, but rather to visually explore the distribution of a larger collection of patterns as a whole. To do so, we empower the analyst to actively steer the perspective of a two dimensional projection of the mined patterns. Altering the perspective and seeing how related patterns move, zooming and filtering the collection and inspecting structures of interest closer, lets the analyst keep the overview even on larger pattern collections.

Classical pattern-mining algorithms, like closed frequent item set mining, subgroup discovery, and exceptional model mining—to name just a few—search for patterns of high interest to the analyst in a dataset. The goal is to retrieve a small collection of easy-to-understand patterns that expose main concepts occurring frequently within the dataset. In Section 2 we briefly discuss several pattern-mining algorithms and their main objectives. The formal definition of a pattern, how its interestingness is measured, and how the final result is compiled differs from method to method. In general, one can say that a pattern is a description of a subset of the dataset that should be easy to understand. A very commonly used pattern format, which is also used throughout this paper, is the conjunction of different *attribute=value* assignments. For instance, the pattern *"type=fish and color=blue"* describes all blue fishes in a dataset at hand. The result set that is finally delivered to the analyst is usually determined by considering the support of the patterns, a quality measure, and the redundancy among the patterns of the result set. To keep the result set at a convenient size, classical pattern-mining algorithms have to carefully consider whether the information in each pattern bears insight or might contribute to overload.

We propose an interactive, visually driven extension to the classical pattern-mining procedure that does not discard any discovered patterns before presenting the results to an analyst. The idea is not to deliver a condensed result set, but rather to mine a larger collection of patterns first and then project them into a two-dimensional space, with similar patterns being close to each other. This enables further visual analysis. The insights gained from actively exploring the pattern distribution help the analyst to understand and interpret the results of the classical pattern-mining methods. The exploration of the pattern distribution follows Shneiderman's information-seeking mantra *"Overview first, zoom and filter, then details-on-demand"* [30]. Our proposed approach enables the analyst to grasp the pattern collection as a whole and then to further discover and dig deeper into regions of interest. In earlier publications, we investigated different algorithms that enable direct interaction with an embedding to explore a dataset interactively. The direct visual feedback of seeing how the distribution of all data records changes upon interaction can help the analyst understand the underlying structure of the data and formulate hypotheses. One common way to provide the interface for the interaction is to let the analyst select data points as control points and relocate them in a "drag-and-drop" manner within the embedding. Altering the positions of these control points triggers the embedding technique to recalculate the whole projection, subject to the updated control-point locations. The recalculation can usually be done efficiently, such that the updates resulting from the interaction can be rendered live. For an impression on the update-rate of the here used implementation please have a look at Appendix A.1. Note that there are also other methods of interacting with an embedding, e.g., employing *must-link* / *cannot-link* constraints, filtering and inspecting the sub-selection, or simply highlighting and brushing.

The remainder of the paper is organized as follows. In Section 2 we discuss related research and in Section 3 we introduce a general framework for interactive pattern exploration. Section 4 demonstrates our approach in several scenarios on a cocktail ingredient dataset before we finally conclude in Section 5.

2. RELATED WORK

Related to our work are basically two areas of research, pattern mining and interactive embedding methods. For the pattern-mining methods we have to distinguish whether a label is considered. Probably the most-known pattern-discovery technique that does not consider a label is frequent item set mining. Here all conjunctions of *attribute=value* assignments are listed in decreasing order of the number of data records that support the pattern [1, 15]. Because the set of all 1-frequent patterns of a dataset can be exponentially large in the number of attributes of the dataset, usually only the top-k patterns with a thresholded minimum support are considered. However, often the set of frequent patterns contains redundant descriptions; i.e., the same set of data records is described by different patterns. Closed frequent items-set mining methods [4, 32] counteract this by only listing the closure of each of these sets as a unique descriptor. Other ways to discover interesting patterns in an unlabeled dataset are, e.g. to compile an output set of small size that possesses a high entropy [25] or to find large tiles of 1-assignments in a binary dataset [12].

For labeled datasets, (closed frequent) subgroup-discovery algorithms [18, 20, 33] find patterns with a significant difference between the label distribution of the whole dataset and the one exposed by the patterns. Exceptional model mining [21], a generalization of subgroup discovery, allows for more-complicated target concepts, like multiple labels. Another generalization applies the theory of relevance [11, 19] to the found subgroups. Relevant subgroup discovery algorithms [11, 13, 24] deliver only patterns that are not covered by any other pattern in the result set. The term 'covering' implies that there is no generalization of a subgroup that extends the subgroup's support set by strictly positively labeled data records. Δ- and ϵ-relevant subgroup discovery methods [14, 23] loosen this tight formulation and allow the considered generalizations of a subgroup to have a controlled amount of additional negatively labeled data records in the support set.

A different approach to the discovery of interesting patterns is to sample from the space of all patterns. Note that pattern sampling does not aim at delivering a condensed result set, but instead samples the patterns with a probability proportional to a given interestingness measure. Possible measures are, e.g., sampling proportional to a pattern's frequency, its squared frequency, its lift, or the area it tiles in the dataset [5]. In addition, pattern sampling can also take labels into account, such that patterns with a high positively and a low negatively labeled share in the support set are more likely to be drawn. The probability of a pattern being drawn can be calculated efficiently [6] by using the sampling technique *coupling from the past*.

Pattern sampling is a good showcase to demonstrate our approach and can be a good option in cases where the space of all patterns is extremely large, such that classical pattern-mining algorithms take too long to terminate. This is especially important if the analyst is on a time budget and the listing strategy of the mining algorithm does not correlate with the relevance of the patterns to the analyst. In this case, sampling from the whole pattern space can yield interesting patterns much earlier. Boley et al. [5] show such an example on the *primary-tumor* dataset, where the patterns that are most discriminating between the labels are among the least-frequent.

Apart from local pattern discovery, there is also related work in the area of embedding data into a lower-dimensional space for visualization and interaction. Many classic techniques are unsupervised and static, like the well known principle component analysis (PCA) [16], multi-dimensional scaling (MDS) [8], isometric mapping [31] and locally linear embedding [29]. These methods consider the distances between the data records in different ways and find lower dimensional embeddings which exhibit similar the distance relations. The projection pursuit method [10] follows a different objective, it searches for interesting projections of the data that display a high degree oy non-gaussianity.

In order to incorporate interaction into the dimensionality-reduction algorithms, the static embedding approaches are typically extended to consider additional user feedback and thus provide an interface with the lower-dimensional embedding of the data to the analyst. There are different approaches for deriving the embedding and incorporating interaction. Some techniques enable the user to relocate selected points within the embedding and incorporate the placement of these control points as constraints or regularization into the optimization problem of a (kernelized) principal component analysis (PCA) [28, 26]. Other techniques embed the data via MDS user-suggested locations of the control points [7, 9, 22]. In contrast to these methods, least squared error projections [27] calculate the embedding solely by considering the control points' original attributes and user-specified embedding locations, ignoring the covariance among the rest of the data records. The interactive embedding technique used in our upcoming study in Section 4 minimizes the uncertainty of the resulting embedding, given a prior belief about it, conditioned on the control points' placements [17]. Throughout this paper we refer to this technique as *most-likely embedding* (MLE). In addition, this method can also be used to actively propose control points to the analyst that minimize the uncertainty about the resulting embedding and thus should be placed next.

Finally, but without a focus on interaction, Berardi et al. proposed to embed collections of patterns in order to discover structures among them by using MDS as the embedding technique [3]. The pairwise similarities between the patterns, required by MDS, were derived by calculating the Jaccard index between two patterns.

3. A GENERAL INTERACTIVE PATTERN EXPLORATION PROCEDURE

Our approach to studying a larger collection of patterns is to embed them into a lower-dimensional space for further interactive visual analysis. Due to the many different ways this can be done, we do not want to propose one particular exploration technique, but rather give a guiding framework on how to gain insights from a larger pattern collection by exploring it interactively. Our proposed procedure comprises the following steps:

1. Mine a large collection of patterns.
2. Represent the patterns in a canonical way as vectors.
3. Embed these vectors with an interactive embedding method and explore the pattern distribution.
4. Inspect the emerging structures of interest deeper.

In our upcoming exemplary study, we utilize a two-dimensional scatterplot for visualization, with each pattern being a point within the plot. Often the initial visualization of the pattern distribution, before any interaction at all, already exhibits interesting structures that invite the analyst to deeper inspection. By further interacting with the embedding by, e.g., selecting single patterns as control

Table 1: Exemplary results of the ten highest quality patterns, delivered by different pattern-mining approaches on the cocktail dataset. Note that here the top-10 frequent item sets are also all closed. The high-lift patterns were sampled according to their *rarity* measure [6]. In case of subgroup discovery, the label indicates whether a cocktail is creamy or not.

Unsupervised pattern-mining methods		Supervised pattern-mining methods	
Frequent (closed) item sets	**Sampled patterns with high lift**	**closed subgroups**	**Δ_1-relevant subgroups**
Vodka	Vodka **&** Cranberry juice	Baileys	Baileys
Orange juice	Vodka **&** Triple sec	Crème de cacao	Crème de cacao
Amaretto	Baileys **&** Kahlúa	Milk	Milk
Pineapple juice	Vodka **&** Gin	Kahlúa	Kahlúa
Grenadine	Vodka **&** Blue curaçao	Baileys **&** Kahlúa	Cream
Gin	Pineapple juice **&** Malibu rum	Cream	Irish cream
Baileys	Vodka **&** Amaretto	Irish cream	Crème de banana
Tequila	Vodka **&** Rum	Vodka **&** Baileys	Butterscotch schnapps
Kahlúa	Orange juice **&** Amaretto	Crème de banana	Whipped cream
Triple sec	Vodka **&** Tequila	Baileys **&** Butterscotch schnapps	Vodka **&** Kahlúa

points and relocating them in a playful manner, the analyst can see how other patterns relate, as they move accordingly. On the other hand, the analyst does not have to 'play' with the embedding, but can also directly express desired similarities among patterns by selecting similar ones and placing them close to each other in the embedding. In this way the analyst can also incorporate domain knowledge into the embedding. The above mentioned structures that occur in the visualization can come in various shapes; clusters of patterns, regions of higher density, outliers, or mirroring shapes can all be fruitful to investigate. Reasoning about the contents of these structures and how they differ from another usually uncovers interesting aspects about the patterns and the original dataset.

4. AN EXEMPLARY STUDY

In this section, we demonstrate the use of our interactive pattern-exploration approach by performing an artificial exemplary knowledge discovery session on a cocktail-ingredient dataset. The data is an excerpt of the drinks presented on the website webtender.com. It can be downloaded, together with our interactive embedding tool from http://kdml-bonn.de/InVis. In the following we give an example of a concrete instantiation of the above introduced framework. This setup is precisely the workflow that we use in our exemplary study in Section 4.1. For the other examples in Sections 4.2 and 4.3, only the first step changes, as the pattern collection is retrieved using different algorithms.

1. Mine the 1000 most-frequent item sets from the cocktail dataset. Here, every cocktail is described as the set of ingredients it contains.
2. Represent each of the 1000 frequent item sets by a binary vector over all occurring items of the pattern collection in lexicographical order.
3. Visualize the pattern vectors, using the *most-likely embedding* technique with an initial PCA embedding as the prior mean and interact with it to shape out interesting structures.
4. Inspect these structures by highlighting patterns that contain certain ingredients and by listing the five most-present single items of the structure in a tag cloud.

A list of the ten highest-quality patterns, found by several classical pattern-mining algorithms, is given as a reference in Table 1. The first three methods, *frequent*, *closed frequent*, and sampled *high-lift* patterns, do not consider label information, but provide us with an overview on the most-striking ingredients and ingredient combinations. The *subgroup-* and *relevant-subgroup-discovery* methods on the other hand do use a label and show us ingredients (and their combinations) that are strongly related to it. For these methods, we manually assigned a label to each cocktail according to whether it is "creamy". In Sections 4.1, 4.3 and 4.2 we will apply our interactive approach on the output of different pattern-mining algorithms with the goal of gaining additional insights into the results of Table 1 and to understand the patterns' relations. In each session we mine 1000 patterns and represent them as binary vectors over all items that occur within the patterns, sorted in lexicographical order. We then visualize the mined patterns using an interactive embedding technique and search for emerging structures in an interactive manner.

In the following studies we employ a variant of Iwata, Houlsby and Ghahramani's *most-likely embedding* technique [17] to interact with the embedding via control points. The general idea behind this method is to customize a matrix that projects the data into the embedding space in a probabilistic way. This projection matrix is assumed to be matrix-normal distributed, a matrix-valued extension to the normal distribution. Ultimately, MLE calculates the embedding with the least uncertainty about the placement of the data records, given a prior belief on the projection matrix and conditioned on the control points' placements as evidence. In contrast to Iwata et al.'s method we do not use the Laplacian of the nearest-neighbour graph, but instead the projection onto the first two principal components as prior belief about the embedding (see Appendix A.1).

Finally, inspecting the structures that emerge when interacting with the embedded patterns can be done in various ways. In our exemplary study we use two simple, yet effective methods. The first is highlighting all the patterns within the embedding that contain an item of interest. Second, we also consider presenting the five most-frequently occurring items in a studied structure in a tag cloud. It is also possible to use more-sophisticated methods to study the pattern distribution. For example, we could perform pattern mining on the previously discovered patterns that form such a structure. Alternatively, we can also find a single well-suited representative pattern of the structure However, as our study shows, it is possible to gain insights and craft hypotheses using only our employed naïve methods.

4.1 Frequent Itemsets

In this section we show our proposed approach in action and demonstrate how the frequent patterns reflect rudimentary properties of the original dataset. Note that investigating the most frequent item sets with our proposed method serves mostly the purpose of a sanity check and demonstrating our approach in action. Figure 1 shows the 1000 most-frequent item sets of the cocktail dataset represented as binary vectors over all items, embedded onto their first two principal components. Immediately, we can see two well separated clusters that resemble roughly in their shape. Investigating these clusters closer reveals that the right one contains only patterns that include the ingredient *Vodka*, the most-frequent single item in the original dataset, whereas the left one doesn't (see Figure 1, left). The second most-frequent ingredient, *Orange juice*, determines whether a pattern is mapped to the top or to the bottom of the embedding (see Figure 1, right).

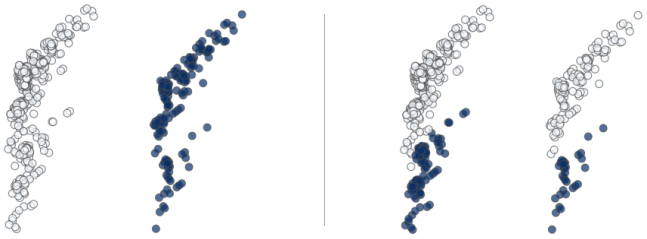

Figure 1: The 1000 most-frequent item sets of the cocktail dataset, embedded onto their first two principal components, labeled by the presence of *Vodka* (left) and *Orange juice* (right).

Interacting with the embedding by relocating two control points, as shown in Figure 2, unravels the blending of the patterns that contain *Orange juice* and the ones that don't. The resulting four clusters clearly separate the patterns by their presence or absence of the ingredients *Vodka* and *Orange juice*.

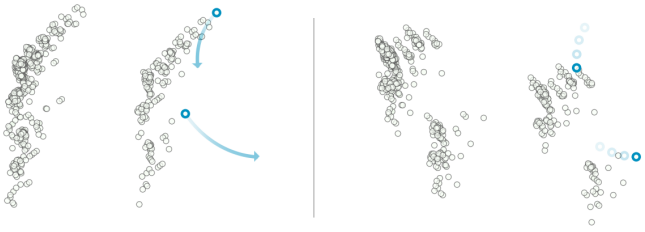

Figure 2: Dragging two control points (emphasized in blue) to new locations, reveals a structure that was previously hidden in the PCA embedding. The four clusters indicate the presence or absence of the two ingredients *Vodka* and *Orange juice*.

Figure 3 inspects one of these emerging structures, the top-right "*Vodka* and no *Orange juice* cluster" from Figure 2, in a closer manner.
With a glance at the top-left picture of Figure 3 we can see that the corresponding patterns containing *Vodka* but no *Orange juice* also frequently contain other strong alcohols, especially *Rum*, *Gin*, and *Triple sec*. We can also observe a sub-cluster structure within this particular embedding, which is determined by the presence or absence of the ingredients *Rum* (top-right, highlighted in green) and *Gin* (bottom-left, highlighted in blue). The ingredient *Triple sec*

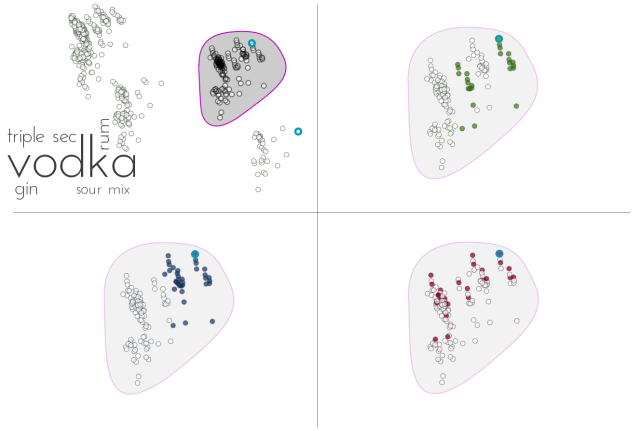

Figure 3: A closer look at the top-right cluster of Figure 2 reveals the ingredients that the patterns from the "*Vodka* and no *Orange juice* cluster" are frequently mixed with (top-left). The other three pictures indicate the presence of *Rum* (highlighted in green), *Gin* (blue), and *Triple sec* (red).

(bottom-right, highlighted in red), although frequent within this cluster, seems not to contribute to the sub-structure, but can be found in all of the sub-clusters. This is an interesting finding, as *Triple sec* is much more frequent than *Rum*. In fact, *Rum* does not even occur among the ten most-frequent ingredients, yet it has a striking influence on the structure of this cluster. Note that this is an insight that could not have been drawn purely from the results of Table 1. In the following sections we will perform similar studies with pattern collections that were drawn according to more-sophisticated interestingness measures than frequency of occurrence.

4.2 Sampled Patterns

A fruitful way to quickly draw patterns from a dataset according to different interestingness measures is to sample. Although sampling itself provides diversity among the drawn patterns, sorting them by the measure and listing only the top-k ones can reintroduce a certain amount of redundancy. On the other hand, diversity is not impaired when exploring the set of all sampled patterns in our proposed way and the analyst is further enabled to discover the different concepts among the patterns. In this study, we sampled 1000 patterns from the cocktail dataset, according to their rarity measure, a variant of the lift measure which promotes patterns containing items that are statistically dependent (see Appendix A.2). The samples were drawn using the *direct local pattern sampling tool* which was provided to us by Boley et al. [6] and can be downloaded from http://kdml-bonn.de/?page=software_details&id=23.

The retained collection of the sampled patterns demonstrates well how our proposed approach benefits from the use of interactive embedding techniques. The plain PCA embedding of the frequent patterns in the previous Section 4.1 already exhibited a clear structure, which directly invited the analyst to further explore it. For this particular set of sampled patterns, however, this is not the case. Figure 4 shows the sampled rare patterns embedded into two dimensions, using different techniques, namely PCA, Isomap, and locally linear embedding.[1]

[1] The latter two techniques estimated the assumed lower-dimensional manifold via the 10-nearest-neighbour graph.

Figure 4: 1000 patterns sampled from the cocktail dataset, according to the *rarity* measure [6] and embedded, using different techniques: principal component analysis (left), locally linear embedding (middle), and isometric mapping (right).

Although these static embeddings exhibit no structures that immediately raise the analysts attention, relocating just one control point in the interactive embedding reveals clusters that were previously obscured, as Figure 5 (top) shows.

Figure 5: Relocating a control point, using our interactive embedding reveals a clear cluster structure (top). The middle pictures highlight the patterns containing *Vodka* (left) and *Orange juice* (right). The bottom pictures inspect the composition of two of these clusters.

The two middle pictures of the figure highlight the patterns containing *Vodka* (left) and *Orange juice* (right). Clearly we can identify the *Vodka* cluster, but the other clusters come as a surprise. They do not relate to the *Vodka* / *Orange juice* segmentation that was already discovered in Section 4.1, but capture concepts of their own. The two highlighted ones at the bottom of the figure revolve around juicy and *Rum*-heavy cocktails. Because of the initially mentioned redundancy among the highest rated rare patterns, the results from Table 1 mainly exhibit patterns associated with *Vodka*. Our proposed interactive discovery approach, however, was able to overcome this drawback and reveal other, novel concepts among the high-rarity patterns.

4.3 Subgroup Descriptions

Patterns can be discovered according to different measures of interest. In the previous sections we studied pattern sets that were drawn proportional to their measure of frequency or rarity. In some cases, however, the analyst might also want to consider label information. A classic pattern-mining approach that does so is *subgroup discovery*. It ranks the patterns by how much the label distribution of the data records described by the pattern diverges from the label distribution of the whole dataset. In this section we study the top-1000 closed subgroup descriptions from the cocktail dataset, ranked according to the binomial test quality measure [4] (see Appendix A.3). Figure 9 shows the embedding of these 1000 patterns onto their first two principal components.

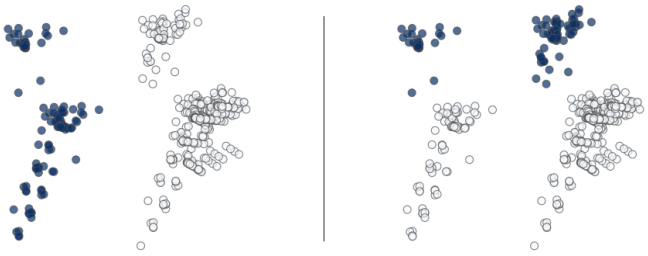

Figure 6: The top-1000 subgroup descriptions associated to the label *creamy*, embedded onto their first two principal components. The four clusters coincide with the presence/absence of the two most striking ingredients among creamy cocktails: *Baileys* (left) and *Kahlúa* (right).

Similar to the embedding of the frequent item sets, but without the help of any interaction, the mined patterns fall directly into four clusters. This time, the clustering goes along with the presence or absence of two other frequently occurring ingredients: *Baileys* (left) and *Kahlúa* (right). From the list of frequent patterns in Table 1 we know that these ingredients are highly frequent, and from the list of subgroups we know that they have a stark impact on the label of a cocktail. In this sense, the observed segmentation doesn't come as a total surprise. However, following the results of Table 1 we might instead have expected *Crème de cacao*, instead of *Kahlúa*. The visualization helps to understand the relations among the listed patterns and invites for further exploration of the exhibited structure. To do so, this time we do not interact with the embedding via the earlier utilized control points, but rather by focusing on a subset of the distribution. We filter the pattern collection to keep only the ones that contain neither *Kahlúa* nor *Baileys* and re-embed them onto their first two principal components. The selection corresponds to the patterns belonging to the bottom right cluster of Figure 6. The re-embedding of these selected patterns can be seen in Figure 7 below.

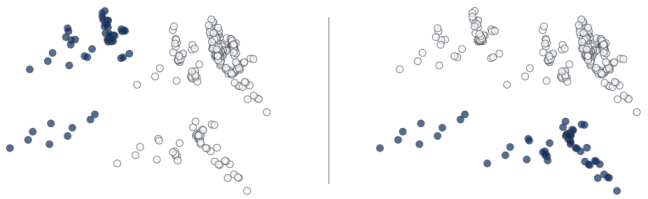

Figure 7: A PCA embedding of the patterns belonging to the bottom right cluster of Figure 6. Again, the embedded patterns can be neatly segmented by the presence of two highly frequent ingredients, this time *Vodka* (left) and *Crème de cacao* (right).

As the re-embedding is not a zoom, but a newly calculated PCA embedding, we are able to discover structures that were previously hidden due to the covariance among the patterns that are now filtered out. Once again we observe that the patterns form four clusters, corresponding to highly frequent ingredients, this time *Vodka* and *Crème de cacao*. Note that this 'four cluster segmentation' is not part of our proposed method, but stems form the sparsity which transactional databases often expose. To achieve a clearer separation of the clusters in the visualization, we use again the placement of a control point, as shown in the following Figure 8.

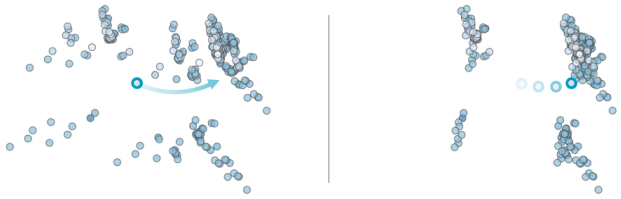

Figure 8: To retrieve a better separation between the clusters, we interact with the embedding by selecting and relocating an appropriate control point.

As an example, we pick two of the clusters from Figure 8 and study their compositions. Figure 9 below shows the five most-frequent ingredients within the patterns of these clusters in a tag cloud.

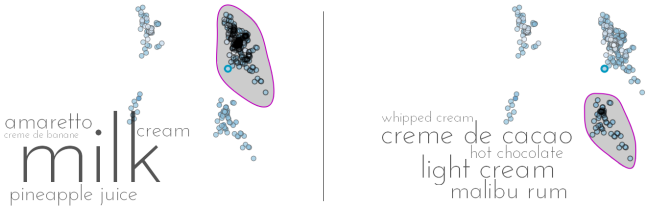

Figure 9: Inspecting the contents of two of the emerging clusters. One interesting finding is the occurring separation between milky and chocolaty patterns. The cluster segmentation stems from the presence of the ingredients *Vodka* and *Crème de cacao*.

We can observe that the inspected regions contain patterns that stem from two different types of creamy cocktails: milky and chocolaty ones. This is an interesting finding, as the strict separation between the clusters does not stem from the milky ingredients within the patterns, but from the ingredients *Vodka* and *Crème de cacao*. However, using our interactive visualization, we were able to craft the hypothesis that milky and chocolaty cocktails form different types of creamy cocktails, offering a good next direction to explore.

4.4 Discussion

Using an interactive embedding of the patterns to visualize and explore it, we were able to remedy the information overload that comes naturally with the consideration of a large pattern collection. Our proposed approach mainly collapses into a two step procedure: (1) mine a large collection of patterns and (2) explore a visualized embedding of the patterns in an interactive way. We demonstrated our approach on pattern collections that resulted from three different mining techniques, namely frequent pattern mining, sampling patterns proportional to their lift, and subgroup discovery. In the second step, we followed the information-seeking mantra and explored the obtained pattern collections in a top-down manner. We started with a visual overview of the whole pattern distribution and dug deeper on striking structures by interacting with the visualization and investigating the emerging structures in different ways, namely by

- reshaping the embedding via relocating control points.
- filtering out and re-embedding the remaining patterns.
- listing the most-frequent items of an inspected structure.
- highlighting all patterns containing an ingredient of interest.

By interactively exploring the pattern collection, we were able to gain some minor insights that we could not draw by purely considering the results of Table 1. To give some examples, from the list of frequent patterns we know that *Vodka* and *Orange juice* are the most-frequent ingredients of the cocktail dataset, but the PCA embedding was able to reveal how much more *Vodka* distinguishes between the cocktails than *Orange juice* does. By inspecting the sub-clusters that emerged from our interaction, we found a surprisingly strong influence of the ingredient *Rum* on the cocktails containing *Vodka* but not those containing *Orange juice*. This discovery is backed up by the high-lift pattern *Vodka & Rum* that we can find in Table 1. However, considering the mirroring of the "no-*Orange juice*-clusters", located at the top in Figure 2, we can also craft a theory about a strong influence of *Rum* among the non-*Vodka* patterns in general. We were also able to discover three strong concepts among the patterns with a high lift: the pattern *Vodka* **&** *Something*, fruity cocktails, and *Rum*-heavy cocktails. This is especially interesting, as *Rum* does not rank among the ten most-frequent ingredients. In addition, we were also able to discover independently from Table 1 that *Kahlúa*, *Baileys*, *Crème de cacao* and *Milk* are mainly responsible for a cocktail being labeled as *creamy*.

However, the strength of our approach lies not in these discoveries, but in the deeper understanding of the relations among the patterns that it provides in combination with the classical pattern-mining methods. By exploring the pattern embedding, interacting with it, exposing interesting structures, and always collating the crafted theories and insights with Table 1, we were able to develop an understanding of the different concepts that the original cocktail data revolves around.

5. CONCLUSION

We proposed an extension to the classical pattern-mining approach that enables the analyst to overcome information overload when browsing and exploring a larger collection of patterns. The goal of our proposed method is to help the analyst understand the underlying distribution of the patterns and additionally to invite them to further exploration. Whereas the classical pattern mining approach focuses on presenting a condensed set of high-quality patterns, our approach uses interactive embedding techniques to visualize and explore the distribution of a larger pattern collection. To do so, we proposed a general four-step approach, where each step can be instantiated in different ways. In our exemplary study, we demonstrated the use of our approach by exploring and interacting with three different pattern collections from a cocktail-ingredient dataset. Collating our findings and the results of different pattern-mining algorithms, we were able to forge and test hypotheses and develop an understanding of the mined patterns and the different concepts that they descend from.

6. REFERENCES

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*. 1996.

[2] A. Asuncion and D. J. Newman. Uci machine learning repository, http://archive.ics.uci.edu/ml, 2007.

[3] M. Berardi, A. Appice, C. Loglisci, and P. Leo. Supporting visual exploration of discovered association rules through multi-dimensional scaling. In *Proceedings of Foundations of Intelligent Systems*. Springer, 2006.

[4] M. Boley and H. Grosskreutz. Non-redundant subgroup discovery using a closure system. In *Proceedings of Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD*, 2009.

[5] M. Boley, C. Lucchese, D. Paurat, and T. Gärtner. Direct local pattern sampling by efficient two-step random procedures. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD*. ACM, 2011.

[6] M. Boley, S. Moens, and T. Gärtner. Linear space direct pattern sampling using coupling from the past. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD*. ACM, 2012.

[7] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. In *Proceedings of Visual Analytics Science and Technology, VAST*. IEEE, 2012.

[8] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman and Hall/CRC, 2000.

[9] A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North. Observation-level interaction with statistical models for visual analytics. In *Proceedings of Visual Analytics Science and Technology, VAST*. IEEE, 2011.

[10] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *Transactions on Computers*, 1974.

[11] G. C. Garriga, P. Kralj, and N. Lavrač. Closed sets for labeled data. *Journal of Machine Learning Research*, 9, 2008.

[12] F. Geerts, B. Goethals, and T. Mielikäinen. Tiling databases. In *Proceedings of Discovery science*. Springer, 2004.

[13] H. Grosskreutz and D. Paurat. Fast and memory–efficient discovery of the top–k relevant subgroups in a reduced candidate space. In *Proceedings of Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD*, 2011.

[14] H. Grosskreutz, D. Paurat, and S. Rüping. An enhanced relevance criterion for more concise supervised pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD*, 2012.

[15] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proceedings of Special Interest Group on Management of Data, SIGMOD*, 2000.

[16] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

[17] T. Iwata, N. Houlsby, and Z. Ghahramani. Active learning for interactive visualization. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2013.

[18] W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Proceedings of Advances in Knowledge Discovery and Data Mining*. AAAI, 1996.

[19] N. Lavrač and D. Gamberger. Relevancy in constraint-based subgroup discovery. *Constraint-Based Mining and Inductive Databases*, 2005.

[20] N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5(Feb), 2004.

[21] D. Leman, A. Feelders, and A. Knobbe. Exceptional model mining. In *Proceedings of Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD*. Springer, 2008.

[22] S. C. Leman, L. House, D. Maiti, A. Endert, and C. North. Visual to parametric interaction (v2pi). *PloS one*, 8(3), 2013.

[23] F. Lemmerich and M. Atzmueller. Incorporating exceptions: Efficient mining of epsilon-relevant subgroup patterns. In *Proceedings of the ECML PKDD Workshop LeGo: From Local Patterns to Global Models*, 2009.

[24] F. Lemmerich and M. Atzmueller. Fast discovery of relevant subgroup patterns. In *Proceedings of Florida Artificial Intelligence Research Society, FLAIRS*, 2010.

[25] M. Mampaey, N. Tatti, and J. Vreeken. Tell me what i need to know: succinctly summarizing data with itemsets. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD*, 2011.

[26] D. Oglic, D. Paurat, and T. Gärtner. Interactive knowledge-based kernel pca. In *Proceedings of Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD*, 2014.

[27] D. Paurat and T. Gärtner. Invis: A tool for interactive visual data analysis. In *Proceedings of Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD*, 2013.

[28] D. Paurat, D. Oglic, and T. Gärtner. Supervised PCA for interactive data analysis. In *Proceedings of the NIPS 2nd Workshop on Spectral Learning*, 2013.

[29] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2000.

[30] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 1996.

[31] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2000.

[32] T. Uno, T. Asai, Y. Uchida, and H. Arimura. An efficient algorithm for enumerating closed patterns in transaction databases. In *Proceedings of Discovery Science, DS*, 2004.

[33] S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Proceedings of Principles of Data Mining and Knowledge Discovery, PKDD*. Springer, 1997.

APPENDIX

A.1 The most likely embedding

Our approach for interacting with a lower-dimensional embedding of data makes use of the matrix-normal distribution, an extension of the multivariate normal distribution to matrix-valued arguments. The idea is to find the linear projection from the original data space into the embedding space that is most likely, given a prior belief about the embedding and conditioned on the placement of selected control points. The $(p \times q)$-dimensional matrix normal distribution $\mathcal{MN}_{p,q}(R; M, \Sigma, \Psi)$ has the density function

$$\mathcal{MN}_{p,q}(R; M, \Sigma, \Psi) = (2\pi)^{-\frac{pq}{2}} |\Sigma|^{-\frac{q}{2}} |\Psi|^{-\frac{p}{2}} \exp\Big(-\tfrac{1}{2} \operatorname{tr}\big[\Sigma^{-1}(R-M)\Psi^{-1}(R-M)^\top\big]\Big),$$

where:

- $R \in \mathbb{R}^{p \times q}$ is the matrix-valued argument,
- $M \in \mathbb{R}^{p \times q}$ is the location parameter, and
- $\Sigma \in \mathbb{R}^{p \times p}$, $\Psi \in \mathbb{R}^{q \times q}$ are symmetric positive-definite scale parameters that can be considered the "row" and "column" covariance matrices, respectively.

Why is this useful? Suppose we have a matrix normal distributed belief about a linear embedding matrix R:

$$p(R \mid \theta) = \mathcal{MN}(R; M, \Sigma, \Psi),$$

where θ represents the hyperparameters M, Σ, and Ψ. Now further suppose that we have observed data $X \in \mathbb{R}^{D \times N}$ in a potentially high-dimensional Euclidean space $\mathcal{X} = \mathbb{R}^D$ and that the user has selected a total of m control points $Y \in X$ and has placed them in preferred locations $W \in \mathbb{R}^{2 \times m}$ within the two dimensional embedding. We will write $\mathcal{D}$ to indicate these observed data pairs (Y, W).

We also assume that the locations chosen for these points, given by the user, represent the correct latent locations for these points, corrupted by iid zero-mean isotropic Gaussian noise. Consider RY, which represents the embedded locations of Y given knowledge of the latent embedding matrix R. Our assumption is that the control points placed by the users are close to their ideal locations:

$$p(W \mid RY, \theta, \sigma^2) = \mathcal{MN}(W; RY, I, \sigma^2 I),$$

which indicates that each of the values in W differs from RY by entrywise iid Gaussian noise with variance σ^2. Henceforth we will include σ^2 in the set of hyperparameters θ.

Now we can reason about the linear projection matrix R that is most likely, given a prior believe about the embedding and conditioned on the observed values W:

$$p(R \mid Y, W, \theta) = \mathcal{MN}(R; M_{R|\mathcal{D}}, \Sigma, \Psi_{R|\mathcal{D}}),$$

where

$$M_{R|\mathcal{D}} = M + (W - MY)(Y^\top \Psi Y + \sigma^2 I)^{-1} Y^\top \Psi;$$
$$\Psi_{R|\mathcal{D}} = \Psi - \Psi Y (Y^\top \Psi Y + \sigma^2 I)^{-1} Y^\top \Psi.$$

In order to retrieve the final most likely embedding of all the data points X, we simply have to calculate the $M_{R|\mathcal{D}} X$.

To utilize this method in a live-update manner, reasonably many updates have to be calculated per second. If the interaction with the embedding is only the movement of control points, then solely $M_{R|\mathcal{D}}$ has to be recalculated and multiplied by X to retrieve the embedding. The following Figure 10 depicts the updates per second for this case, depending on the number of attributes, data records and used control points. However, if the selection of the control points changes, also $\Psi_{R|\mathcal{D}}$ has to be recalculated (which on a regular PC runs in well under a second). As depicted, the update-rate depends the strongest on the number of data records and drops with an increasing amount of them. Using our non-tweaked implementation, a dataset of about 1500 data-records could be interacted with at an update-rate of roughly 10-15 updates per second. The dataset used in this experiment was an excerpt from the *Communities and Crime* dataset, taken from the UCI dataset repository [2].

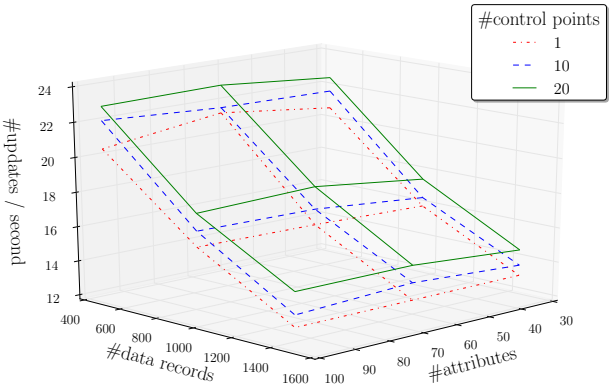

Figure 10: Achieved updates per second for 1,10 and 20 selected control points, depending on the number of data-records and attributes of the dataset.

A.2 The rarity measure

The rarity of a pattern approximates the probability of occurrence of the whole pattern weighted by the probabilities of the single items that build the pattern not occurring. To put it in a formal way, let $\mathcal{D}$ be a transactional database over a fix set of items. Further, let P be a pattern, consisting of k of these items $P = \{p_1, \ldots, p_k\}$. The rarity of P is calculated as

$$\text{rarity}(P, \mathcal{D}) = \text{freq}(P, \mathcal{D}) \prod_{p_i \in P} \big(1 - freq(p_i, \mathcal{D})\big),$$

where $\text{freq}(x, \mathcal{D})$ denotes the observed frequency of occurrence of the pattern x in the database $\mathcal{D}$.

There is a relation to the lift measure of a pattern, which is calculated by

$$\text{lift}(P, \mathcal{D}) = \text{freq}(P, \mathcal{D}) \prod_{p_i \in P} \frac{1}{\text{freq}(p_i, \mathcal{D})}.$$

Whereas *rarity* considers the absence-frequency of the singleton items, *lift* considers the inverse of them.

A.3 Subgroup quality measures

In the context of subgroup discovery, the interestingness of a pattern is measured by a quality function $q(P, \mathcal{D})$ that considers the pattern and the dataset and returns a real-valued number. This function usually combines the size of the support set of the pattern and its unusualness w.r.t. the designated target label in the following

way:

$$q(P, \mathcal{D}) = \text{freq}(P, \mathcal{D})^{\alpha} \cdot \left(\frac{|\mathcal{D}^{+}_{[P]}|}{|\mathcal{D}_{[P]}|} - \frac{|\mathcal{D}^{+}|}{|\mathcal{D}|} \right),$$

where $\mathcal{D}^{+}$ denotes the positively labeled share of the data and $\mathcal{D}_{[P]}$ the portion of the data that supports the pattern P. The coefficient α of the quality function is a constant $0 \leq \alpha \leq 1$, characterizing a family that includes some of the most-popular quality functions. For $\alpha = 1$ it is order-equivalent to the weighted relative accuracy (WRACC) and the Piatetsky–Shapiro quality function. For $\alpha = 0.5$ it corresponds to the binomial test quality function, which is used to mine the subgroup description patterns in Section 4.3.