Exploring Similarity

Improving Product Search with Parallel Coordinates

Mandy Keck[1], Martin Herrmann[1], Andreas Both[2], Dana Henkens[3], and Rainer Groh[1]

[1] Technische Universität Dresden, 01062 Dresden, Germany
{mandy.keck,martin.herrmann,rainer.groh}@tu-dresden.de
[2] Unister GmbH, Barfußgäßchen 11, 04109 Leipzig, Germany
andreas.both@unister-gmbh.de
[3] queo GmbH, Tharandter Str. 13, 01159 Dresden, Germany
d.henkens@queo-group.com

Abstract. Faceted browsing is an established and well-known paradigm for product search. However, if the user is unfamiliar with the topic and the provided facets, he may not be able to sufficiently reduce the amount of results. In order to increase the understanding of the bidirectional relation between facets and result set, we present an interface concept that allows manifold approaches for product search, analysis and comparison starting with a single product or a summarizing visualization of the entire data set. Moreover, various product features can be analyzed in order to support decision-making. Even without detailed knowledge of a specific topic, the user is able to estimate the range and distribution of characteristics in relation to known or desired features. Conventional list-based search forms do not provide such a quick overview. Our concept is based on two visualization techniques that allow the representation of multi-dimensional data across a set of parallel axes: parallel coordinates and parallel sets.

Keywords: Visual Search Interfaces, Information Visualization, Parallel Coordinates, Motive-based Search, Big Data, E–commerce.

1 Introduction

Although sophisticated algorithms and semantic search approaches exist, product search in large data sets is still a major challenge for web users. Deciding on a product is based on the analysis and comparison of multi-dimensional product data. However, typical web interfaces with simple search masks and result lists do not support the user sufficiently for these tasks. Particularly in the context of financial data, where complex search masks often overstrain non-experts, alternative entry points are required. Fuzzy filters or query-by-example approaches can increase the understanding of the various attributes of product data and can help to improve the precision of the individual search query. To address this challenge, we developed an interface concept based on the visualization technique of parallel coordinates that allows the analysis of various attributes at a glance. Our concept enables the comparison and exploration of

S. Yamamoto (Ed.): HIMI 2014, Part II, LNCS 8522, pp. 160–171, 2014.

similar products to support the evaluation of products and to improve decision-making [1]. The concept combines the search paradigm of faceted browsing and query-by-example to facilitate the search in a bidirectional way. Thus, huge result sets can be quickly narrowed down to smaller subsets or single products and the analysis of products with similar properties can be improved. Furthermore, the distribution of product data is shown to allow the modification of search criteria.

The interactive parallel axis approach that we present in section 3, covers two interface concepts: the first is based on parallel coordinates (see section 2.2) and the second is based on parallel sets (see section 2.3). Both concepts provide insights into patterns and dependencies of multi-dimensional data. To evaluate the suitability for different search tasks, we provide a preliminary user study in section 4.

2 Related Work

This section covers different search and visualization techniques to explore and analyze multi-dimensional data sets.

2.1 Faceted Browsing

A popular interface paradigm to explore a product database is the principle of Faceted Browsing, which many e-commerce sites use (e.g. amazon.com, ebay.com). Faceted Browsing allows multiple access points for the search and the iterative refinement of the result set. Therefore, the products have to be structured using a Faceted Classification [2]. This classification method describes items through a combination of facets, where each facet addresses a different property. In the context of product search, a product can have the facets “product type”, “customer rating”, and “price”. Each facet contains different facet values (e.g. the facet “product type” contains the facet values “books”, “movies”, “music”, etc.) and usually one value per facet describes an item. Additionally, a hierarchy can be used to organize the facets in several subcategories (e.g. hierarchical structuring of the product type in “book”, “textbook”, and “computer science”) [3]. A Faceted Browser allows the navigation in this data structure and the construction of complex search queries by selecting facet values [4]. The user can explore the data collection by restricting or increasing the result set iteratively.

2.2 Parallel Coordinates

The visualization technique of parallel coordinates (PC) allows the two-dimensional representation of multi-dimensional data across a set of parallel axes [5]. Each parallel axis represents one attribute of the multi-dimensional data set, whereas all axes are arranged side by side. A polyline represents a single data item and intersects each axis at the appropriate value (see Figure 1 left).

First implementations of parallel coordinates introduced by Inselberg [5] used straight lines to connect each intersection point of a data item. Since multiple items often share the same intersection points, it becomes difficult to trace the path of a

single item. Different approaches have been developed to overcome this "crossing problem". Graham et al. [14] propose adding curvature using continuous gradients, and spreading of close intersection points depending on their positions in the proceeding and following axis to a larger region on the axis.

Another important challenge is to reduce visual clutter. Data mining tasks usually require analysing a large amount of datasets. Due to the overlapping of hundreds of lines, it becomes impossible to identify meaningful patterns. In order to overcome this problem and reveal hidden information, several strategies like colour coding or frequency and density calculations [15] can be applied on the tangle of lines. Bundling of similar multidimensional items is another approach to spatially separate and unravel lines and therefore maintain the user's ability to recognize correlations between the data attributes [13].

Typically, parallel coordinates show their strength in analysing continuous data types, but some fields of application also require analysing categorical data. Rosario et al. [12] propose mapping a class to a single point on an axis and indicate the degree of similarity by the spacing between the points. Teoh et al. [11] and Riehmann et al. [7] extend the point to a vertical area indicating the number of items included in the corresponding category. Riehmann et al. also use this approach to reduce the crossing problem by distributing the lines on this constructed interval [7].

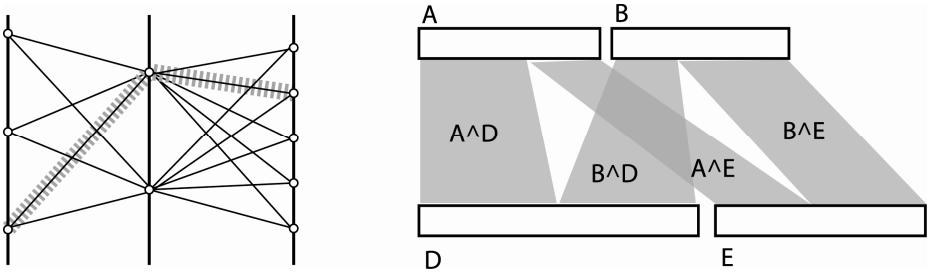

Fig. 1. Parallel coordinates (left), parallel sets (right)

2.3 Parallel Sets

The visualization technique parallel sets (PS) is mostly used for categorical data attributes and is well suited for the visual analysis of large, complex data sets. The basic layout is derived from parallel coordinates, with the axes being replaced by containers representing categories. These containers are scaled according to the frequency of the corresponding category [10]. Instead of single lines, the containers are connected by polygonal streams representing the logical conjunction of the adjacent containers. The size of these streams give an impression of the frequency of items included in the conjunction (see Figure 1 right). Since the complexity of this visualization is independent from the number of regarded items, it is well suited to obtain a fast overview over large-scale data sets.

3 Visual Interface for Product Search

While Faceted Browsing became a common paradigm for product search in the last years, parallel coordinates and parallel sets are mostly restricted to scientific data-mining tasks. With our concept, we try to introduce and evaluate these two concepts in the area of financial data exploration. We propose that users can strongly benefit from the possibility of discovering patterns and interpreting correlations of the manifold characteristics of financial products. Even without detailed knowledge of the topic, the user can get an impression of the range and distribution of characteristics in relation to known or desired features, which cannot be accomplished with conventional list based search forms.

3.1 Data Preparation

To test our visualisation concepts, we use a set of financial products including certificates, leveraged products, and warrants. From the great variety of features, we chose eight of the most meaningful characteristics to describe a single financial product. These features include categorical data (e.g. *underlying value*) as well as continuous data (e.g. *performance*) and ordinal data (e.g. *investment term*). Since parallel sets require categorical data, continuous data needs to be classified. This can be achieved by defining ranges either automatically (with equidistant or logarithmic intervals or by using natural breaks) or by defining meaningful intervals manually.

An advantage of categorical data representation is that a category can be either generalized or specialized and can be organized into a hierarchy. Therefore, an attribute can be viewed on different levels of abstraction. To allow this semantic zoom ability, we added meta-information about the underlying hierarchy for each axis. For categorical attributes, we identified all possible entities of a feature and added the parent items manually. For continuous attributes, only the top hierarchy levels were defined manually, while lower levels were identified by automatic methods.

3.2 Interface Concept

The interface is divided into a parallel axis view and a list view containing item title and description (see Figure 2). This allows the exploration of the database in a bidirectional way. Starting with the left side the user can get an overview of the underlying data structure and the distribution of product data. The presented result list on the right side can be reduced by various filters. Selecting items in the result list highlights the corresponding elements in the axis view and supports the identification of similar products.

The visual representation of the axis is composed of spatially separated container elements representing the facet values. The height of each container, in reference to the overall height of the axis, shows the distribution of data items within the facets and allows the analysis of predominant values. Additionally, it reduces the overplotting situation at crowded points for the PC view and simplifies the explicit selection and tracing of a single curve or stream [7].

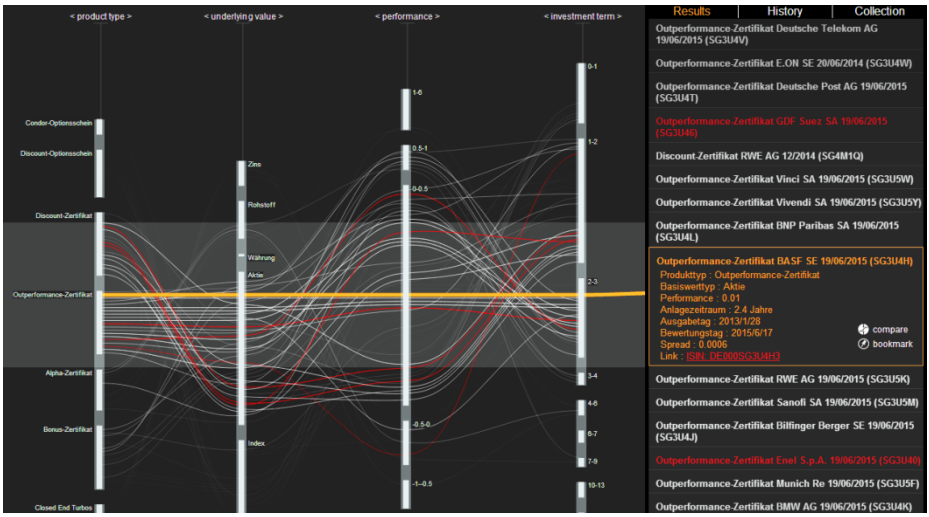

Fig. 2. Interface concept with parallel coordinates (left) and result list (right): the comparing feature centers the selected product; the fuzzy filter (light grey) reduces the amount of polylines

Interaction with Axes. The parallel axes are the main control elements of the visualization and offer a range of interactions for manipulation. Mouse scrolling zooms the selected axis (see Figure 3 left). This semantic zooming allows the user to explore the underlying hierarchy of a facet and thus to generalize or specify his current selection. Connected containers indicate the same parent in the hierarchy. Apart from zooming, each axis can be dragged vertically and horizontally, granting the user full control of the position, ordering, and spacing of axes. The rearrangement of axes allows a better comparison of two axes of interest. To adjust the complexity of the provided facets, axes can be added or deleted.

Facet Filter. Selecting a container activates a filter, leaving only items matching the facet value (see Figure 3 middle & right). Multiple filters can be activated on one or different axes to create complex filters as known from Faceted Browsers (logical disjunction within an axis, logical conjunction on different axes). Showing facet values and products in one representation holds the advantage that the user can immediately see which filters have a strong influence on his current selection and how much he has to alter his query to get better results. It also allows a quick comparison of different products and an in-depth analysis of the database. For instance, the user can determine the best product types depending on the lowest risk and the highest performance.

Result List. The right side of the interface provides the result list. The result view is constructed based on the selections made in the visualization on the left side. Brushing visually links a product on the left side to the associated list item. Clicking one item in the list provides detailed product information (see Figure 2).

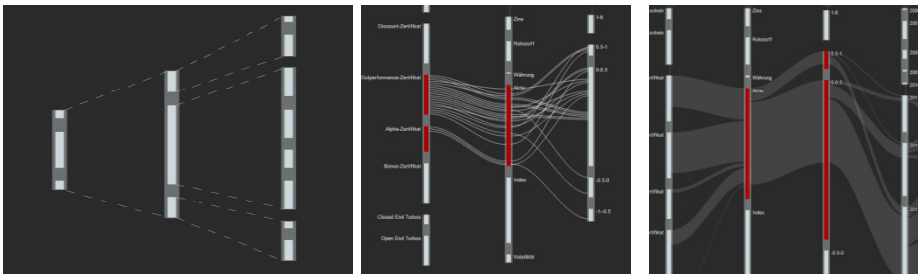

Fig. 3. Zoomable axis to support hierarchical facets (left), facet filter with parallel coordinates (middle) and parallel sets (right)

History. Each explored item is added to the history list for subsequent analysis or comparison. If an item was added to the history, it is highlighted in both views to distinguish them from unexplored or new products. If an item seems particularly interesting, it can be bookmarked and stored in a separate collection. When the user switches to the history or collection, the axes view is synchronized, allowing a direct comparison of the previously selected items.

The visualization of product data is based on two different visualization techniques that are described in section 3.3 and 3.4.

3.3 Parallel Coordinates

As described in section 2.2, the parallel coordinate concept uses lines to display the attributes of each item in the dataset. Intersections points with the axes define the property of each characteristic. Using categorical instead of continuous axes allows the distribution of intersections along the whole container equally and lead to an individual intersection point for each line. Adding curvature to the polylines (see Figure 4) further reduces the crossing problem described in section 2.2 and simplifies the tracking of individual lines.

Apart from the representation of product items, the concept of parallel coordinates supports two other features depending on the particular visualization technique:

Fuzzy Filter. The visual complexity of parallel coordinates increases with the number of displayed items. To address this problem, we introduce a fuzzy filter, represented by a scalable rectangular zone around the median line of the visualization. This filter adjusts the opacity of the lines depending on a calculated relevance value. The value represents the average distance of all intersection points of a line to the median line of the visualization. If the relevance is above a threshold, controlled by the size of the filter zone, a line becomes visible. The opacity further indicates the relevance of the visible lines. In contrast to the facet filter mechanism, the fuzzy filter is implemented to explore subsets with similar properties. Both filters can be combined to address concrete information needs (facet filter) as well as vague ideas (fuzzy filter).

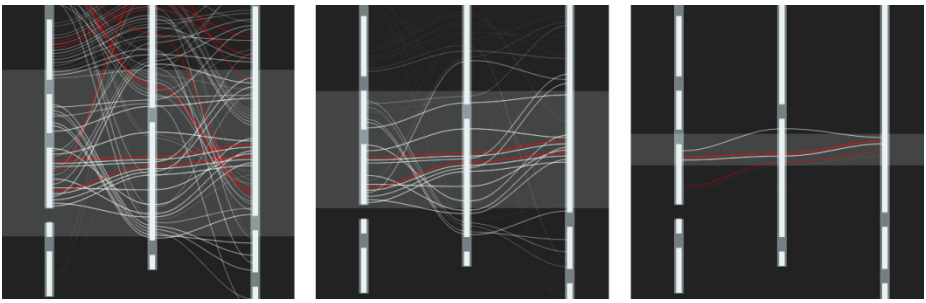

Fig. 4. Fuzzy Filter for data exploration

Filter operations can be executed either by dragging the desired facet values onto the filter zone or by scaling the filter zone to broaden or narrow the scope. The first method allows quick "scanning" of one dimension of the data set with regard to other selected attributes. The second option supports focusing on a range of interest and reduces visual clutter (see Figure 4). The visual attribute "opacity" indicates the relevance of items and is mapped to each polyline in the visualization and to every item in the result list.

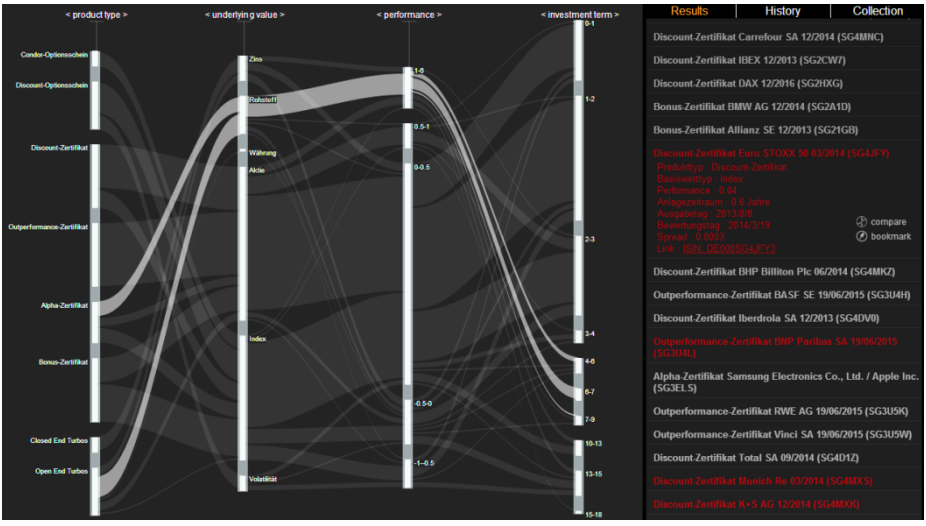

Fig. 5. Interface concept with parallel sets: selecting a product in the result list highlights the corresponding streams between the parallel axes

Comparing. When an interesting item is found in the result list, the comparing feature can be used to identify products with similar properties. This function automatically centers all interpolation points (intersections) to the median line of the visualization (see Figure 2). Furthermore the result list is rearranged to match the order of displayed lines and to put similar objects next to the centered list object. The method

is well suited for a search-by-example approach where the user starts his search by selecting an item of interest matching his expectations in one or more characteristics. Previously selected and examined products are indicated as red polylines which allows the visual separation of uninspected products in the focused area and supports a feeling of finiteness.

3.4 Parallel Sets

Parallel Sets are well suited to visualize large sets of categorical data as described in section 2.3. Therefore, we adapted this technique to our flexible axes framework. The streams between the axes represent subsets of the whole dataset possessing the two properties of the connected facet values. Mouse-over highlights all subsets, which contain the same items of the current subset. Clicking on a particular subset present the contained products in the result list on the left side. The facet filter, presented in section 3.2, can be used to reduce the result set and to hide mismatched subsets as well. Using mouse-over in the result list highlights all subsets containing the focused item and allows the identification of its properties (see Fig. 5).

4 Evaluation

In a preliminary user study, we evaluate the suitability of both interface (see Fig. 2 & Fig. 5) concepts for three different tasks (analysis, comparison, and search tasks). We were interested in measurable values (solution time for different search tasks and error rate) as well as user feedback regarding to the acceptance of both interface concepts and the provided features. Therefore, we developed two prototypes implemented in JavaScript using D3.js[1] to create interactive SVG visualizations. Both prototypes visualize a subset of a real-world data set containing 120 financial products and use four axes to describe the following exemplary characteristics: *type of product* (categorical data divided into 3 high-level and 8 low-level categories), *underlying asset* (categorical data divided into 6 high-level and 50 low-level categories), *investment term* (continuous data classified into long-term, medium-term and short-term investment on first hierarchy level, and divided into sorted annual categories on second hierarchy level), and *performance* (continuous data divided into 5 sorted high-level and 20 low-level categories). Both interfaces offer the presented features in section 3: facet filter, zooming & panning of all axes, rearrangement of the axes, and result list with history and collection. Parallel coordinates additionally provide the introduced comparison feature and fuzzy filter.

4.1 Methodology

Thirteen users (7 females) in the age range of 23 – 60 years (M = 30.23, SD = 9.98) participated in the user study. According to their personal assessment (scale from 1 = extensive experience to 5 = no experience), most of them were not familiar with PC

[1] `http://d3js.org/`, Last accessed: 07.02.2014.

(M = 4.08, SD = 1.15) and PS (M = 4.53, SD = 0.88). Before they started with the experiment, we shortly introduced both visualization techniques, the underlying data set, and the range of features offered by both prototypes. Afterwards, they had some minutes to familiarize themselves with both interfaces. We switched the interface (PC or PS) during the experiment and the order was counterbalanced between participants. The participants had to solve 9 tasks per interface, divided into 3 task types with 3 tasks per type: analysis of the financial data (e.g. "Which product group contains the most products with an investment term between 1 and 2 years?"), comparison of two products (e.g. "Find a similar product with the same type of product and a performance as close as possible") and search for a financial product (e.g. "Find a product with currency as underlying asset and an investment term of approximately 3 years").

A task was solved when the user identified one result that matched his given task. He or she was free to decide when this was the case. This could be a particular facet value (analysis task) or a particular product (comparison and search tasks). During the experiment, we measured completion time (start and end of each task were indicated by the user) and error rates (results were classified in "0 = incorrect", "1 = partial match" and "2 = perfect match"). After each experiment, the participants had to complete a questionnaire to evaluate each interface concept regarding effectiveness, learnability, satisfaction, joy of use, efficiency, and range of functions. Finally, they had to complete a questionnaire which evaluated individual features (facet filter, zooming & panning of all axes, rearrangement of the axes and result list, comparing, and fuzzy filter) offered by PC and PS regarding usefulness and usability. Both questionnaires used a 5-point Likert scale (0 = Strongly Disagree, 4 = Strongly Agree) to point out their personal opinion to the given features.

4.2 Results

Time and errors were subjected to 2 (*system: PC, PS*) x 3 (*task: analysis, comparison, search*) repeated measures ANOVAs. Subjective ratings were compared between both systems with t-tests.

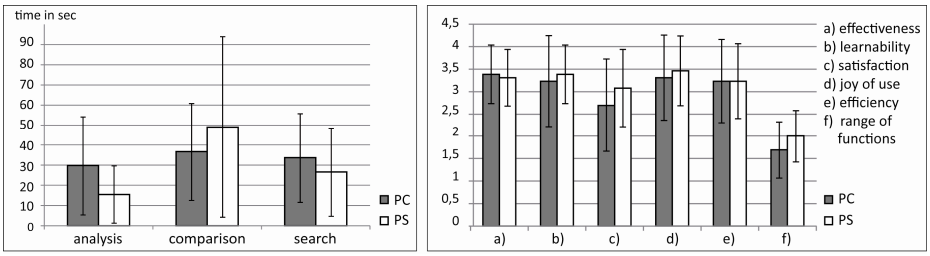

Fig. 6. Solution time for all tasks (left), results of the questionnaire for PC and PS (right) (Error bars represent standard deviations)

In terms of solution time, there was no significant difference between PC and PS, $F(1,12) = 1.71$, $p = .215$ (PC: M = 33.34s, SD = 19.73; PS: M = 30.36s, SD = 23.99). But there was an interaction between *system* and *task*, $F=(2,24) = 10.21$, $p < .001$. The difference between both systems were significant with regard to the tasks: PS was faster than PC for analytical tasks, $p = .017$, and for search tasks, $p = .023$. PC was faster than PS for comparing tasks, $p = .015$ (see Figure 6, left).

We analysed errors to evaluate the precision of each system. With both systems, most tasks have been solved correctly (PC: perfect match = 88.9%, partial match = 7.7%, incorrect = 3.4%; PS: perfect match = 94.9%, partial match = 3.4%, incorrect = 1.7%). But there was no main effect of *system*, $F(1,12) = 2.28$, $p = .157$, and no interaction between *system* and *task*, $F(2,24) = 0.62$, $p = .548$.

The evaluation of the questionnaire revealed no significant differences between both systems referring to perceived effectiveness (PC: M = 3.38, PS: M = 3.31), learnability (PC: M = 3.23, PS: M = 3.38), satisfaction (PC: M = 2.69, PS: 3.07), joy of use (PC: M = 3.31, PS: 3.46), efficiency (PC: M = 3.23, PS: M = 3.23) and range of functions (in this case means "0 = too little" and "4 = too much": PC: M = 1.69, PS: M= 2), all $|t| < 2.5$, all $p > .05$ (see Figure 6, right).

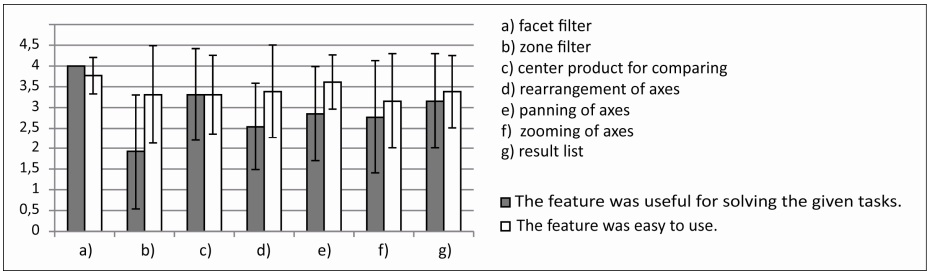

Fig. 7. Evaluation of the features offered by both systems

In the second questionnaire, we evaluated the individual opinion regarding usefulness and usability of the individual features. The facet filter was one feature that was used most frequently to reduce the result set, whereas the fuzzy filter was used less often. All features were easy to use (all mean values between "agree" and "strongly agree") (see Fig. 7). We observed different opinions regarding the fuzzy filter. For some participants it was helpful to reduce the visual clutter and to extend the facet filter (fuzzy filter for vague information need, facet filter for concrete information need). However, during some tasks (especially the analysis task), the fuzzy filter was disturbing because it hides some values in its initial state.

4.3 Discussion

As results have shown, both approaches match the needs of a user while performing a search for financial products. Although the participants of the study were no experts considering financial products, PS as well as the PC enable them to solve the given tasks (perfect match: 88.9% - 94.9%) in a time that is considered as adequate.

Eventually, the positive objective results were underpinned by the subjective feedback by the users considering learnability (M = 3.23 - 3.38) and joy of use (M = 3.31 – 3.46). During the experiments we observed that PC was leading to a faster solution time as soon as a bidirectional interaction was required and a detailed product view and comparison were needed (see Figure 6, left). On the other hand, PS has shown a tendency to provide a higher satisfaction (see Figure 6, right). However, the data was not significant.

Because of a lack of experienced users it is not possible to provide a comparison to the needs of expert users. As there was also no actual real-world search-driven application for financial products available for evaluation it is also not possible to estimate the superiority of PS and PC above current industrial applications. However, experts of the cooperating company have given the feedback that such existing applications are difficult to understand and create no joy of use for unexperienced users. Hence, we can assume that PC and PS are capable of providing a better interface at least for this user group.

5 Conclusion and Future Work

With the presented interface concept, we try to combine useful characteristics of Faceted Browsing and the visualization methods parallel sets and parallel coordinates. Our concept provides many distinctive strategies and approaches for analysis, comparing, and searching in large multi-dimensional datasets. In a preliminary user study, all participants easily solved tasks from all of the three task types. However, the rule “less is more” often applies to interfaces for human-computer interaction. Our user tests showed that users require extensive training to use the full potential of all the features provided by the application in its current state.

Focusing not only on experts but also on casual users, further advancements of the application could include a step-by-step introduction of features or a wizard proposing different approaches depending on the current task.

Especially when dealing with large amounts of data, the sequence of executed tasks plays a significant role for the success of a decision process. Starting with the full view on all features might not be the best choice. With the help of parallel sets applied to a few chosen attribute features, the user could be encouraged to make a preselection before investigating the subset in detail.

Our concept for fuzzy filtering proved to be convenient in solving comparison task but also confused some users who were trying to solve a search task. While both filters influence the displayed result set, it was often not evident to the user why there are only few displayed results. Additional interface elements can indicate the amount of excluded results for each activated filter.

Further improvements are necessary for the spreading of intersections on each axis. A class-internal reordering of these positions depending on zoom level, filters, attribute value and neighbouring axis could help to reduce visual clutter and enhance the accuracy of the fuzzy filter.

Acknowledgments. This work has been supported by the European Union and the Free State Saxony through the European Regional Development Fund (ERDF). The research presented in this article has been conducted in cooperation of the Chair of Media Design -Technische Universität in Dresden, Unister GmbH from Leipzig and queo GmbH from Dresden, Germany. Thanks are due to Severin Taranko, Viet Nguyen, Marcus Kirsch and Romy Müller for their invaluable feedback and support in this research.

References

1. Keck, M., Herrmann, M., Both, A., Gaertner, R., Groh, R.: Improving Motive-Based Search. In: Streitz, N., Stephanidis, C. (eds.) DAPI 2013. LNCS, vol. 8028, pp. 439–448. Springer, Heidelberg (2013)
2. Ranganathan, S.R.: Elements of Library Classification. Asian Publishing House, Bombay (1962)
3. Hearst, M.: Design recommendations for hierarchical faceted search interfaces. In: ACM SIGIR Workshop on Faceted Search (2006)
4. Polowinski, J.: Widgets for Faceted Browsing. In: Smith, M.J., Salvendy, G. (eds.) HCI International 2009, Part I. LNCS, vol. 5617, pp. 601–610. Springer, Heidelberg (2009)
5. Inselberg, A., Dimsdale, B.: Parallel coordinates: A tool for visualizing multi-dimensional geometry. In: Proc. of IEEE Visualization, pp. 361–378 (1990)
6. Graham, M., Kennedy, J.: Using curves to enhance parallel coordinate visualizations. In: Proc. of the Seventh International Conference on Information Visualization, pp. 10–16 (2003)
7. Riehmann, P., Opolka, J., Froehlich, B.: The Product Explorer: Decision Making with Ease. In: Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI), Capri, Italia, pp. 423–432 (2012)
8. Stefaner, M., Müller, B.: Elastic lists for facet browsers. In: 18th International Conference on Database and Expert Systems Applications (DEXA 2007), pp. 217–221. Regensburg (2007)
9. Ware, C.: Information Visualization. Perception for Design. Elsevier Ltd., Oxford (2004)
10. Bendix, F., Kosara, R., Hauser, H.: Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data. Transactions on Visualisation and Computer Graphics 1 (2006)
11. Teoh, S., Ma, K.: PaintingClass: Interactive construction, visualization and exploration of decision trees. In: Proceedings Knowledge Discovery and Data Mining. ACM Press (2003)
12. Rosario, G.E., Rundensteiner, E.A., Brown, D.C., Ward, M.O., Huang, S.: Mapping nominal values to numbers for effective visualization. In: Proceedings IEEE Information Visualization, pp. 80–95. IEEE CS Press (2003)
13. Heinrich, J., Luo, Y., Kirkpatrick, A.E., Zhang, H., Weiskopf, D.: Evaluation of a Bundling Technique for Parallel Coordinates. In: GRAPP/IVAPP 2012, pp. 594–602 (2011)
14. Graham, M., Kennedy, J.: Using Curves to Enhance Parallel Coordinate Visualizations. In: Proceedings of the Seventh International Conference on Information Visualization, IV (2003)
15. Artero, A.O., Oliveira, M.C.F., Levkowitz, H.: Uncovering Clusters in Crowded Parallel Coordinates Visualizations, Information Visualization. In: IEEE Symposium on INFOVIS 2004 (2004)