Quantifying and Comparing Features in High-Dimensional Datasets

Harald Piringer
VRVis Research Center
Vienna, Austria
piringer@vrvis.at

Wolfgang Berger
VRVis Research Center
Vienna, Austria
wberger@vrvis.at

Helwig Hauser
Department of Informatics
University of Bergen, Norway
Helwig.Hauser@uib.no

Abstract

Linking and brushing is a proven approach to analyzing multi-dimensional datasets in the context of multiple coordinated views. Nevertheless, most of the respective visualization techniques only offer qualitative visual results. Many user tasks, however, also require precise quantitative results as, for example, offered by statistical analysis.

In succession of the useful Rank-by-Feature Framework, this paper describes a joint visual and statistical approach for guiding the user through a high-dimensional dataset by ranking dimensions (1D case) and pairs of dimensions (2D case) according to statistical summaries. While the original Rank-by-Feature Framework is limited to global features, the most important novelty here is the concept to consider local features, i.e., data subsets defined by brushing in linked views. The ability to compare subsets to other subsets and subsets to the whole dataset in the context of a large number of dimensions significantly extends the benefits of the approach especially in later stages of an exploratory data analysis. A case study illustrates the workflow by analyzing counts of keywords for classifying e-mails as spam or no-spam.

1. Introduction

For many application domains, the steadily growing amount of collected and generated data bears an enormous potential for gaining knowledge and supporting decision-making. Various technologies address the highly non-trivial issue of extracting useful information from potentially huge datasets in different ways. Statistics have been used for long in order to provide summarized data characteristics. Basic statistical moments like mean, variance or correlation are very common and can be computed extremely fast even for millions of values on today's computers. However, statistics as such – and also most statistics-based techniques in the fields of machine learning – are quite static approaches as they hardly involve the user and typically yield a result without additional context information.

Information visualization follows a user-centric approach particularly suitable for exploratory analysis. Combining visual representations of a dataset with means of interaction has proven powerful for detecting searched and also unexpected patterns, and for conveying an impression of relationships and structures. Brushing refers to the possibility of highlighting such structures directly within the view and is typically a key feature in systems supporting multiple views (like Spotfire [10]).

However, information visualization also faces its special challenges: Visual results are often only qualitative and thus not fully sufficient for many tasks. Additionally, to deterministically parameterize visualizations is a challenge itself when it comes to exploring high-dimensional datasets. Most techniques are inherently limited with respect to the number of dimensions which can simultaneously be displayed (e.g., scatterplots), or do not scale well to a truly large number of attributes (e.g., parallel coordinates become hard to read for more than 10 dimensions). In these cases, the user has to pre-select the displayed dimensions, which may become a difficult task without a-priori knowledge or dedicated support.

Since the pros and cons of statistics and information visualization complement each other very well, recent trends like Visual Analytics [11] aim to unite these technologies in joint approaches. The Rank-by-Feature framework by Seo and Shneiderman [8] is a successful example and addresses the issue of conveying a quick overview about all dimensions at an early stage of the analysis. However, its limitation to global features – statistical measurements are only computed with respect to the entire dataset – clearly restricts its continued application at later stages of the analysis, e.g., after identifying clusters or separating trends from outliers, where the user might be interested in properties of selected data subsets.

The main contribution of this paper is an approach for characterizing and comparing arbitrary subsets by combining the precise information of well-known statistical moments with the expressiveness of visualization. Pursuing the successful concept of Seo and Shneiderman [8, 9], the statistics can be used for ranking small preview visualizations. Coloring allows for quickly assessing differences between dimensions or subsets and numerical values precisely characterize the respective data. We propose a 1D approach for univariate moments of individual dimensions and a 2D approach for bivariate moments of pairs of dimensions.

2. Related Work

Mathematically Describing Brushed Subsets – The interaction metaphor of brushing has established itself as proven standard approach to the identification of selected data subsets of interest. Successful systems such as Spotfire [10] and the XmdvTool [12] offer brushing as an integral technology. Recently, Yang et al. [14] introduced a concept for organizing interesting queries (called nuggets) and sharing them with other users. Hao et al. [5] propose an approach to mathematically characterize a focus area for guiding the user to similar parts of the data. Like our work, they apply automated analytical methods to subsets of the data. However, while their intention is to detect other potentially interesting parts of the data based on the description, our approach provides statistical summaries themselves to the user and ranks dimensions accordingly.

High-Dimensional Data Analysis – Most visualization techniques have inherent or practical limitations with respect to the number of dimensions that can simultaneously be displayed. Therefore, exploring truly high-dimensional datasets is a non-trivial yet important research question. Friedman's and Tukey's Projection Pursuit [3] reduces dimensionality by linearly combining attributes which makes the results hard to interpret. Ankerst et al. [1] introduced a measure to place dimensions with alike behavior close to each other. An approach based on a similar idea has more recently been published by Yang et al. [15], who represent dimensions as pixel-oriented glyphs and position them in 2D space according to their properties. However, while these approaches allow for detecting clusters of similar dimensions, they neither consider local features nor do they provide any numerical details, like the work presented in this paper. Friendly [4] describes a technique for reordering correlation matrices and introduces "corrgrams" for visualizing the results. Wilkinson et al. [13] propose graph scagnostics as an alternative to statistical moments to characterize two-dimensional point distributions. This potentially speeds up the identification of interesting plots but requires high abstraction skills from the user. Another approach to improve the selection of displayed dimensions is hierarchical clustering of similar attributes as used by Yang et al. [16].

The Rank-by-Feature Framework [8] is designed to meet the Graphics, Ranking and Interaction for Discovery (GRID) principles: a) "study 1D, study 2D, then find features"; and b) "ranking guides insight, statistics confirm". The user may choose between several statistics displayed in a linked table for ranking preview visualizations of the dimensions. While this approach has proven suitable as an initial guidance to potentially interesting dimensions [9], it is of limited use when it comes to the focused analysis of selected data subsets of interest.

3. Quantifying Brushed Data Features

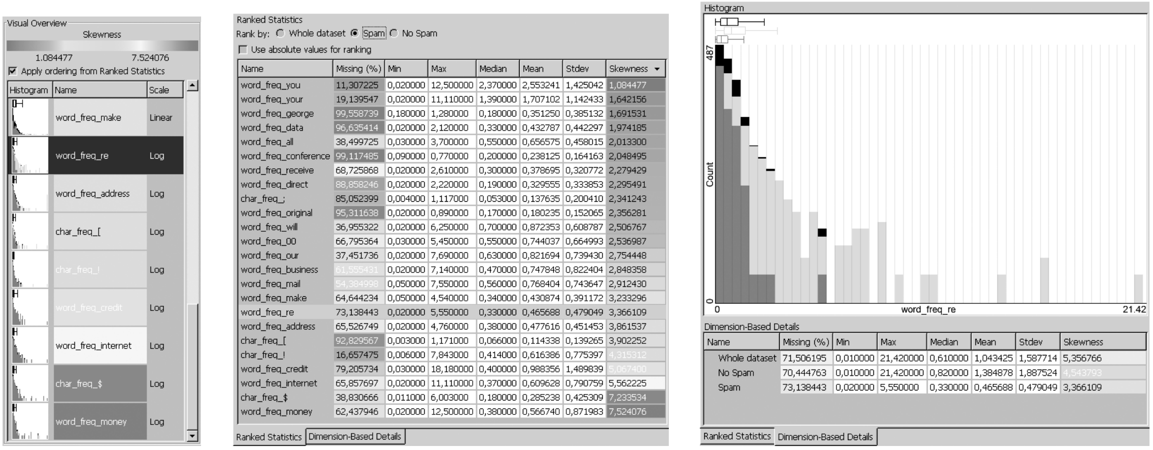
We now present our approach to visualizing, quantifying and comparing data subsets in the context of large numbers of dimensions.

3.1 The General Approach

Our approach distinguishes the subsets as defined by user queries (and also the whole dataset as a "special" subset) and restricts all statistical computations to the according and valid entries of the data. It considers an arbitrary number of dimensions, as selected by the user. All results are automatically updated whenever a query and thus the underlying subset changes (e.g., when the user brushes a linked view), hence providing full linking to all other views. The basic setup consists of three coordinated parts which support different tasks.

The Visual Overview (Fig. 1a and 2a) displays small visualizations with little detail in order to provide an overview of all considered (pairs of) dimensions. Despite their small size, it has proven useful that also these mini-views highlight the subsets as defined by the user. The user may either manually arrange the dimensions or may automatically sort them according to the "active" statistical moment, defined by the Ranked Statistics (see below). If the results are comparable across all dimensions (i.e., if the range of that moment is independent of the scaling of the data), the user may visualize the differences by mapping the results to color. The according transfer function is scaled between the smallest and largest result of the respective statistical moment, unless a "natural" range exists (e.g. -1 to +1 for correlation coefficients).

Ranked Statistics (Fig. 1b and 2b) are structured as table with (pairs of) dimensions as rows and the respective moments as columns. It shows the results, which are optionally computed for the whole dataset or



Visual Overview

Skewness

1.084477 7.524076

☑ Apply ordering from Ranked Statistics

Histogram	Name	Scale
	word_freq_make	Linear
	word_freq_re	Log
	word_freq_address	Log
	char_freq_[	Log
	char_freq_!	Log
	word_freq_credit	Log
	word_freq_internet	Log
	char_freq_$	Log
	word_freq_money	Log

(a)

Ranked Statistics

Rank by: ○ Whole dataset ● Spam ○ No Spam

□ Use absolute values for ranking

Name	Missing (%)	Min	Max	Median	Mean	Stdev	Skewness ▾
word_freq_you	11,307225	0,020000	12,500000	2,370000	2,553241	1,425042	1,084477
word_freq_your	19,139547	0,020000	11,110000	1,390000	1,707102	1,142433	1,642156
word_freq_george	99,558739	0,180000	1,280000	0,180000	0,351250	0,385132	1,691531
word_freq_data	96,635414	0,020000	2,120000	0,330000	0,432787	0,442297	1,974185
word_freq_all	38,499725	0,030000	3,700000	0,550000	0,656575	0,458015	2,013300
word_freq_conference	99,117485	0,090000	0,770000	0,200000	0,238125	0,164163	2,048495
word_freq_receive	68,725868	0,020000	2,610000	0,300000	0,378695	0,320772	2,279429
word_freq_direct	88,858246	0,020000	2,220000	0,190000	0,329555	0,333853	2,295491
char_freq_;	85,052399	0,004000	1,117000	0,053000	0,137635	0,200410	2,341243
word_freq_original	95,311638	0,020000	0,890000	0,170000	0,180235	0,152065	2,356281
word_freq_will	36,955322	0,020000	6,250000	0,700000	0,872353	0,608787	2,506767
word_freq_00	66,795364	0,030000	5,450000	0,550000	0,744037	0,664993	2,536987
word_freq_our	37,451736	0,020000	7,690000	0,630000	0,821694	0,739430	2,754448
word_freq_business	61,555431	0,020000	7,140000	0,470000	0,747848	0,822404	2,848358
word_freq_mail	54,384998	0,050000	7,550000	0,560000	0,768404	0,743647	2,912430
word_freq_make	64,644234	0,050000	4,540000	0,340000	0,430874	0,391172	3,233296
word_freq_re	73,138443	0,020000	5,550000	0,330000	0,465688	0,479049	3,366109
word_freq_address	65,526749	0,020000	4,760000	0,380000	0,477616	0,451453	3,861537
char_freq_[	92,829567	0,003000	1,171000	0,066000	0,114338	0,139265	3,902252
char_freq_!	16,657475	0,006000	7,843000	0,414000	0,616386	0,775397	4,315312
word_freq_credit	79,205734	0,030000	18,180000	0,400000	0,988356	1,489839	5,067400
word_freq_internet	65,857697	0,020000	11,110000	0,370000	0,609628	0,790759	5,562225
char_freq_$	38,830666	0,011000	6,003000	0,180000	0,285238	0,425309	7,233534
word_freq_money	62,437946	0,020000	12,500000	0,380000	0,566740	0,871983	7,524076

(b)

Dimension-Based Details

Name	Missing (%)	Min	Max	Median	Mean	Stdev	Skewness
Whole dataset	71,506195	0,010000	21,420000	0,610000	1,043425	1,587714	5,356766
No Spam	70,444763	0,010000	21,420000	0,820000	1,384878	1,887524	4,543793
Spam	73,138443	0,020000	5,550000	0,330000	0,465688	0,479049	3,366109

(c)

Figure 1. The 1D case showing word and character counts for spam classification of e-mails. (a) The Visual Overview displays mini-histograms, box plots, and color-codes the dimensions according to the current ranking criterion "Skewness". (b) The Ranked Statistics list the results of a user-defined set of statistical moments for each dimension. (c) The Dimension-Based Details provide a larger histogram, whisker plots for each layer, and the results of the statistical moments with respect to the active dimension for the whole dataset and each query.

for the subset defined by a query and is thus suitable for simultaneously quantifying one query with respect to multiple dimensions. The rows can be ordered by any column, denoting the respective moment as "active" which also determines the order and color-coding of the Visual Overview. If applicable, each column is color-coded in order to improve the comparability.

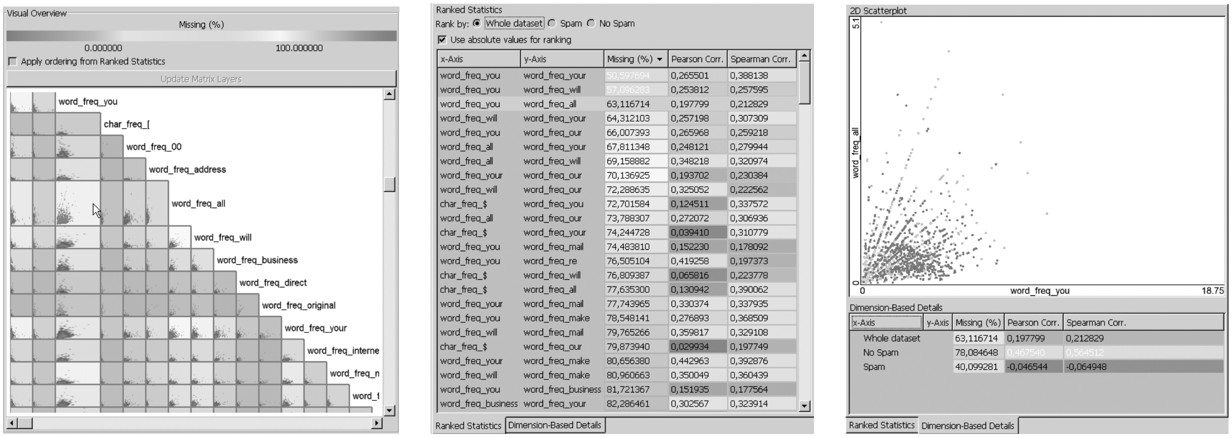
Dimension-Based Details (Fig. 1c and 2c) refer to a single (pair of) dimension(s), which is selected in any of the other parts. The purpose is twofold: First, this part provides a larger visualization with more detail. Second, it displays a table similar to the Ranked Statistics with columns being the selected statistical moments. The difference is that the rows represent the results for subsets defined by the various queries (plus one row referring to the whole dataset) for the active (pair of) dimension(s), allowing for direct comparisons of the characteristics for all queries.

3.2 1D Framework

We now explain, how this general approach can be applied to analyze individual dimensions (1D case) and/or pairs of dimensions (2D case). The intention of the 1D case (see Fig. 1) is to look at the dimensions individually and the offered statistics are therefore univariate. The following set of different statistical moments allows to adapt the analysis to the user task and to the respective properties of the data:

- Minimum and maximum.
- Mean and median.
- Quartiles (1^{st} and 3^{rd}) and standard deviation.
- Trimmed mean and trimmed standard deviation for robust statistics: omits the smallest and largest 10% of the values.
- Skewness, kurtosis and normality: describe and quantify the deviation from normal distribution.
- Entropy: rises with increasing uniformity of the data distribution.
- Number of unique values.
- Value of the biggest gap.
- Percentage of missing entries.

The related visualization approach capitalizes on well-known histograms and box plots [6] to show the distribution of each dimension. The Visual Overview (Fig. 1a) consists of a list of dimensions, where each row contains a small box plot drawn above a histogram, which also highlights all subsets as defined by queries. In order to support multiple queries, which are not necessarily disjunctive, the results of the queries are drawn on top of each other with the "active" subset drawn in front. Furthermore, an attempt is made to determine whether the Y-axis of the histograms should better be scaled linearly or logarithmically in order to guarantee meaningful visualizations, also for distributions where the majority of values lies in a very narrow range – of course, the user may manually override this setting. The Dimension-Based Statistics display a large histogram and a box plot for each query (Fig. 1c).



□ Apply ordering from Ranked Statistics

Rank by: ● Whole dataset ○ Spam ○ No Spam

☑ Use absolute values for ranking

x-Axis	y-Axis	Missing (%) ▾	Pearson Corr.	Spearman Corr.
word_freq_you	word_freq_your	50,597694	0,265501	0,388138
word_freq_you	word_freq_will	57,096283	0,253812	0,257595
word_freq_you	word_freq_all	63,116714	0,197799	0,212829
word_freq_will	word_freq_your	64,312103	0,257198	0,307309
word_freq_you	word_freq_our	66,007393	0,265968	0,259218
word_freq_all	word_freq_your	67,811348	0,248121	0,279944
word_freq_all	word_freq_will	69,158882	0,348218	0,320974
word_freq_your	word_freq_our	70,136925	0,193702	0,230384
word_freq_will	word_freq_our	72,288635	0,325052	0,222562
char_freq_$	word_freq_you	72,701584	0,124511	0,337572
word_freq_all	word_freq_our	73,788307	0,272072	0,306936
char_freq_$	word_freq_your	74,244728	0,039410	0,310779
word_freq_you	word_freq_mail	74,483810	0,152230	0,178092
word_freq_you	word_freq_re	76,505104	0,419258	0,197373
char_freq_$	word_freq_will	76,809387	0,065816	0,223778
char_freq_$	word_freq_all	77,635300	0,130942	0,390062
word_freq_your	word_freq_mail	77,743965	0,330374	0,337935
word_freq_you	word_freq_make	78,548141	0,276893	0,368509
word_freq_will	word_freq_mail	79,765266	0,359817	0,329108
char_freq_$	word_freq_our	79,873940	0,029934	0,197749
word_freq_your	word_freq_make	80,656380	0,442963	0,392876
word_freq_will	word_freq_make	80,960663	0,350049	0,360439
word_freq_you	word_freq_business	81,721367	0,151935	0,177564
word_freq_business	word_freq_your	82,286461	0,302567	0,323914

x-Axis	y-Axis	Missing (%)	Pearson Corr.	Spearman Corr.
Whole dataset		63,116714	0,197799	0,212829
No Spam		78,084648	0,467540	0,564512
Spam		40,099281	-0,046544	-0,064948

(a)

(b)

(c)

Figure 2. The 2D case comparing pairs of word and character counts for the spam classification dataset. (a) The Visual Overview shows a part of the scatterplot matrix, the background color-coded by the percentage of missing entries and the pair of the word counts of "you" and "all" currently highlighted. (b) The Ranked Statistics list the results of the bivariate statistical moments for each pair of dimension, with "Missing (%)" being the current ranking criterion. (c) The Dimension-Based Details show a larger scatterplot and compare the results with respect to the active pair for the whole dataset and each query.

3.3 2D Framework

Apart from analyzing dimensions separately, users are typically interested in relationships between multiple dimensions. This is true on a global scale (e.g., identifying groups of similar dimensions) and also applies to local features like individual clusters. In order to support such tasks, the 2D framework (see Fig. 2) allows for exploring all combinations of assigned dimensions. Therefore, the investigated items are pairs of dimensions. Concerning the handling of missing data, this implies that only entries are considered valid if they are present in both respective dimensions. Due to the symmetry of all employed techniques, it turned out to be sufficient and fosters the overview to maintain only one pair per combination (i.e., for two dimensions X and Y, either (X, Y) **or** (Y, X)) and to omit all pairs of any dimension with itself.

Any symmetric bivariate statistical moments are suitable for the 2D approach. The currently available moments comprise Pearson's correlation coefficient and Spearman's rank correlation coefficient [7] (plus the percentage of entries considered as missing). While the first one is appropriate for describing linear relationships, the latter provides more robustness and the ability to detect also non-linear dependencies at slightly higher computational costs. Both coefficients range from -1 to +1 independent of the scaling of the data and are thus suitable for coloring.

Relationships between two dimensions are usually visualized with 2D scatterplots, which are also used in our case. While the Dimension-Based Details display a scatterplot with higher resolution showing more details of a selected pair (Fig. 2c), the Visual Overview arranges the potentially large number of plots as a scatterplot matrix (Fig. 2a). Due to exploiting symmetry as explained above, a single plot is drawn for each pair, which reduces the matrix to a triangle and leaves space for printing the names of the dimensions. If the extents of the matrix exceed the available space, it is scaled down to a certain minimal size, before scrollbars are shown. However, the plot beneath the mouse cursor is always zoomed smoothly to its original size. The exact layout is based on a linear order of the dimensions (like in the 1D case). The user can specify this order manually or adopt the order of the Ranked Statistics.

Unlike the 1D case, automatically obtaining an order of single dimensions from ranking dimension pairs is not straightforward and ambiguous. After evaluating several strategies, we employ the following algorithm for this task: First, those two dimensions are selected of which the pair achieves the highest ranking, which specifies the topmost plot. Afterwards, the algorithm selects the dimension, which produces a row of the matrix including the pair with any already assigned dimension, which has the highest ranking. This latter step is repeated until all dimensions have been assigned, thus constructing the matrix line by line. The benefit of automatically ordering the matrix is that similar dimensions tend to be placed close to each other, indicating groups of dimensions more directly.

3.4 Further Aspects of Our Approach

The approach described in this paper has been realized in the context of an application framework for visually supported knowledge discovery in large and high-dimensional datasets. Apart from providing various kinds of visualizations (like 2D and 3D scatterplots, histograms, parallel coordinates, etc.), which can be combined in any constellation, a key aspect is to discriminate multiple queries, which are defined by composite brushing and are highlighted in all views in a linked way. All parts (views or managers for system-wide objects like dimensions, selections, etc.) offer convenience functionality regarding the usability like undo/redo or a consistent way to arrange controls. Another fundamental requirement is to support datasets with millions of entries and thousands of dimensions, which has a major impact on the design of views and necessitates advanced software techniques like multi-threading. The framework explicitly allows for denoting single values as missing, which are expected to be omitted for all visualizations and computations.

4 Demonstration

This section briefly illustrates a potential workflow with our approach by analyzing a dataset that has been used for classifying e-mails as spam or no spam. The dataset is based on 4601 e-mails. It contains the relative frequencies of certain words and characters in the respective message and whether it is regarded as spam, summing up to 57 dimensions. It has been obtained from the UCI Machine Learning Repository [2] and originates from the Hewlett-Packard Labs, where employees collected and classified e-mails in order to build a personalized spam-filter. The goal of this case study is to show that our framework supports the task of assessing words and joint occurrences of words with respect to the relevance regarding spam classification. Note that counts of zero are treated as missing values.

As first step of the analysis after importing the dataset, two queries are created in order to select all e-mails, which are classified as spam (red) and no spam (green), respectively. This is accomplished by interactively brushing a linked view for visualizing such categorical dimensions (see Fig. 3).

As the next step, all dimensions related to counts of words and characters are assigned to the 1D case (see Fig. 1). The histograms show that most dimensions are distinctly exponentially distributed, i.e., most counts have many small and very few large numbers of occurrences. Therefore, the Y-axes of most histograms are logarithmically scaled in order to make also small occurrences visible. Moreover, many dimensions have a lot of missing values, which means that these words do not occur at all in many messages. When looking at the histograms, the distinctive coloring of spam and no spam messages reveals that the distribution per subset is quite different for individual words and characters: Some clearly occur more often in one or the other class and for others, the distribution is more or less equal. Ranking the Visual Overview by the average number of occurrences of spam mails provides a very approximate ordering with respect to the likelihood to indicate spam. Picking one word ("re") as an example of a dimensions, where the overview suggests good indication properties, the Dimension-Based Details (see Fig. 1c) confirm this assumption (e.g., by different box plots and mean values). However, due to the high degree of missing values, it is obvious that single words and characters will not be sufficient for a good classification.

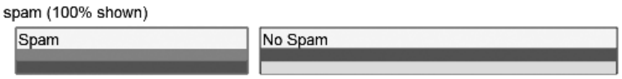


Figure 3. Brushing mails classified as spam (red) and no spam (green) in a linked view for visualizing categorical data.

Therefore, joint occurrences are analyzed in the 2D case, where words with promising indication properties are assigned to (see Fig. 2). Because some words occur together only rarely, the percentage of missing data is mapped to color in order to indicate the significance for each combination. As an example of a comparatively frequent pair (see Fig. 2a), inspecting the combination of the words "you" and "all" in more detail (see Fig. 2c) shows that this pair is missing in 78% of the messages not considered as spam, but only in 40% of the spam messages. In other words, this dataset suggests that encountering both words in one e-mail significantly increases the likelihood for being spam, though it is of course no proof on its own - for this, more pairs would need to be considered together. Furthermore, the Dimension-Based Details also show that the number of occurrences for "you" and "all" are much more correlated for e-mails being no spam (with $\sim$ 47% according to Pearson and $\sim$ 56% according to Spearman). However, probably most interestingly, the scatterplot shows some very distinct "needles", where several messages obviously have perfectly linearly correlated counts of these two words. Such structures are impossible to explain without further knowledge about the e-mails and raise questions regarding the quality and authenticity of the data. If this can be justified, these features suggest the existence of multiple classes of mails, which could be the starting point of a more in-depth analysis.

5. Conclusions and Future Work

In this paper, we have introduced an approach for quantifying and comparing multiple subsets of a dataset by computing and ranking univariate as well as bivariate statistical moments with respect to an arbitrary number of dimensions. The subsets are defined by interactive brushing in linked views. The 1D case supports analyzing multiple dimensions separately, while the 2D case reveals relationships between dimensions. Like the Rank-by-Feature Framework by Seo and Shneiderman [8], our approach is well suited for conveying a quick overview about global properties of the dimensions at an early stage of analysis. However, the aspect of linking the approach to other views significantly extends its applicability also to later stages of analysis – e.g., computing statistics after deselecting identified outliers, characterizing detected clusters, or comparing various categories to each other. It turned out that using well-known statistics leads to faster understanding and acceptance of the approach for domain experts, though extending the set of offered statistics is easily possible.

While the approach scales well with respect to the number of entries in the dataset, there is a certain practical limit concerning the number of simultaneously shown dimensions. Due to the quadratic increase of dimension pairs, this limit is significantly lower in the 2D case. Our experience shows that approximately 35 to 40 dimensions can reasonably be handled in the 2D case, while the 1D case also works well for a few hundred dimensions.

We see at least two directions for potential future work. First, more work would be helpful on how to automatically extract and quantify the characteristics and even semantics of brushes. Second, exploring really high-dimensional datasets with several hundreds or even thousands of dimensions is still a big challenge. As our approach also generates tables with (pairs of) dimensions as rows, applying well-known visualization techniques for multivariate data like parallel coordinates to such tables could be an interesting start.

Acknowledgements

This work was done at the VRVis Research Center in Vienna, Austria, and was partly funded by the Austrian Funding Agency (FFG) in the scope of the project PROVIA (Nr. 814300). Thanks go to Jürgen Platzer for his work on the statistics library used by the implementation of our approach. Furthermore, we thank Helmut Doleisch and Philipp Muigg for help in preparing this paper.

References

[1] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data. In *Proc. of the 1998 IEEE Symposium on Information Visualization*, pages 52–60, 1998.

[2] A. Asuncion and D. J. Newman. UCI Machine Learning Repository (`http://www.ics.uci.edu/~mlearn/MLRepository.html`), January 2008.

[3] J. H. Friedman and J. W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers*, C-23(9):881–890, 1974.

[4] M. Friendly. Corrgrams: Exploratory Displays for Correlation Matrices. *The American Statistician*, 19:316–325, 2002.

[5] M. C. Hao, U. Dayal, D. A. Keim, D. Morent, and J. Schneidewind. Intelligent visual analytics queries. In *Proc. of the 2007 IEEE Symposium On Visual Analytics Science And Technology*, pages 91–98, 2007.

[6] D. Massart, J. Smeyers-Verbeke, X. Capron, and K. Schlesier. Visual presentation of data by means of box plots. *Practical Data Handling, LCGC Europe*, 18(4):215–218, 2005.

[7] D. C. Montgomery and G. C. Runger. *Applied Statistics and Probability for Engineers.* Wiley, 2003.

[8] J. Seo and B. Shneiderman. A Rank-by-Feature Framework for Unsupervised Multidimensional Data Exploration Using Low Dimensional Projections. In *Proc. of the 2004 IEEE Symposium on Information Visualization*, pages 65–72, 2004.

[9] J. Seo and B. Shneiderman. Knowledge Discovery in High-Dimensional Data: Case Studies and a User Survey for the Rank-by-Feature Framework. *IEEE TVCG*, 12(3):311–322, 2006.

[10] Spotfire Inc. Spotfire. `http://spotfire.com/`.

[11] J. J. Thomas and K. A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics.* IEEE Computer Society, 2005.

[12] M. O. Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. In *VIS '94: Proceedings of the 1994 IEEE Conference on Visualization*, pages 326–333, 17-21 Oct. 1994.

[13] L. Wilkinson, A. Anand, and R. Grossman. Graph-Theoretic Scagnostics. In *Proc. of the 2005 IEEE Symposium on Information Visualization*, pages 21–28, 2005.

[14] D. Yang, E. A. Rundensteiner, and M. O. Ward. Analysis guided visual exploration of multivariate data. In *Proc. of the 2007 IEEE Symposium On Visual Analytics Science And Technology*, pages 83–90, 2007.

[15] J. Yang, A. Patro, S. Huang, N. Mehta, M. O. Ward, and E. A. Rundensteiner. Value and Relation Display for Interactive Exploration of High Dimensional Datasets. In *Proc. of the 2004 IEEE Symposium on Information Visualization*, pages 73–80, 2004.

[16] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration Of High Dimensional Datasets. In *Proc. of the 2003 IEEE Symposium on Information Visualization*, pages 105–112, 2003.