Visualization and Interactive Analysis of Blood Parameters with InfoZoom

Michael Spenke
GMD — German National Research Center for Information Technology
FIT — Institute for Applied Information Technology, http://www.gmd.de/fit
Schloss Birlinghoven, D-53754 Sankt Augustin, Germany
Tel. +49 2241 142642, Fax +49 2241 2065
Michael.Spenke@GMD.de

Abstract

This paper describes the application of the data analysis tool *InfoZoom* to a database containing the results of blood examinations for about 400 patients with a suspect of thrombosis. The main goal was to find correlations between the measurements and the occurrence of a thrombosis. No automatic method for data mining is used. Instead, InfoZoom uses a novel technique to display data sets as highly compressed tables which always fit completely onto the screen. The user can interactively explore animated tabular views of the data. In this way, the user gets a feeling of the data, detects interesting knowledge, and gains a deep understanding of the data set.

Keywords

Information visualization, direct manipulation user interfaces, exploratory data analysis, database queries, data mining.

Introduction

Over the last several years we have been developing an interactive software called *InfoZoom* based on a novel approach for a highly condensed representation of information. Database contents are displayed as extremely compressed tables. These tables are not only used for the visualization of data but also allow the user to perform database queries by direct manipulation of the presented information. For example, the user can zoom into certain areas of the table in order to locate specific data objects. Moreover, powerful operations are available to modify large chunks of a database in a single step by directly editing the compressed representation of the data.

InfoZoom is a generic tool. Like a spreadsheet program it is not restricted to a special application domain, but can be used with virtually any database that can be accessed as a relational table. It covers the whole spectrum from an easy-to-use interface for electronic catalogs over simple and complex database queries up to the mining of large databases.

This paper describes the application of InfoZoom to a database of blood test results. The database was donated for the Discovery Challenge at the European Conference on Principles and Practice of Knowledge Discovery in Databases in 1999 [6]. It contains information about 413 patients which came to the hospital with the suspect of a thrombosis. Besides the normal examination by a physician a special blood test was performed in order to validate the hypothesis that thrombosis is closely related to anti-cardiolipin antibodies. The first data set contains the results of these special blood tests. A second data set contains standard blood parameters measured for the same 413 patients over the years. The task was to find

correlations between the measured blood parameters and thrombosis. As expected, the results of the special examination clearly correlate to thrombosis. However, no really significant correlations where found in the standard examinations. Nevertheless, both data sets are well suited to demonstrate how this kind of data can be analyzed using InfoZoom.

Basic Concepts of InfoZoom

InfoZoom displays database relations as tables with attributes as rows and objects as columns. In Figure 1 the results of the special blood examination are shown. Each column corresponds to a patient who has undergone the examination. The first attributes describe basic data about the patients like age and sex, as well as the diagnosed level of thrombosis. Further down the date of the examination and the measured blood parameters are shown. The attributes are hierarchically ordered like files in a directory.

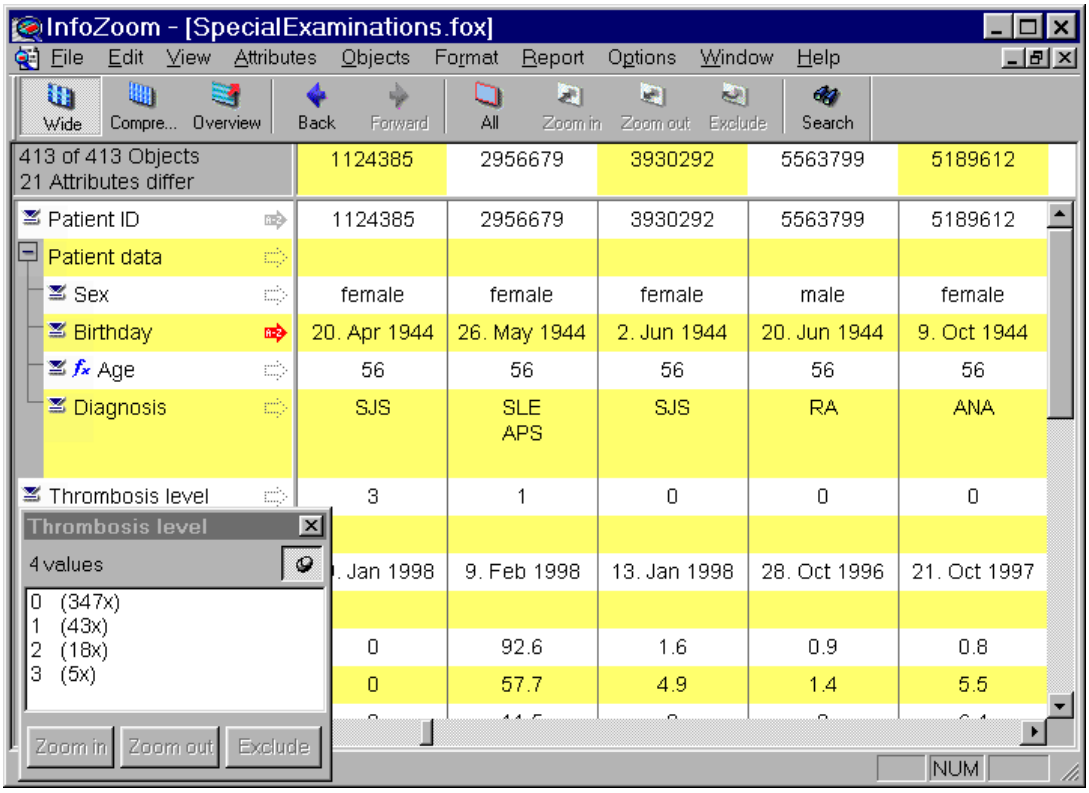

Figure 1: Wide Table Mode

Because the data set contains information about 413 patients, only a small fraction of the table is visible. We can scroll horizontally to see all data but it is impossible to get an overview in this way. Pressing the button left an attribute name shows a list of the possible values and their frequency. This list gives at least an overview of the value range of a single attribute. In Figure 1 it can be seen that there are four different levels of thrombosis. *0* means negative (no thrombosis) and *3* is the most severe level. Selecting and double-clicking one or more values in the list restricts the table to the objects with one of these values. For example, double-clicking level *3* has the effect that only the 5 patients with a severe thrombosis remain in the table. Clicking on the arrow outline right of an attribute sorts the table by this attribute. In Figure 1 the table is sorted by the attribute *Birthday*. This is indicated by a red arrow.

A table can be compressed by clicking the second of the three mode buttons in the upper left corner of the table (above ª413 of 413 objects).

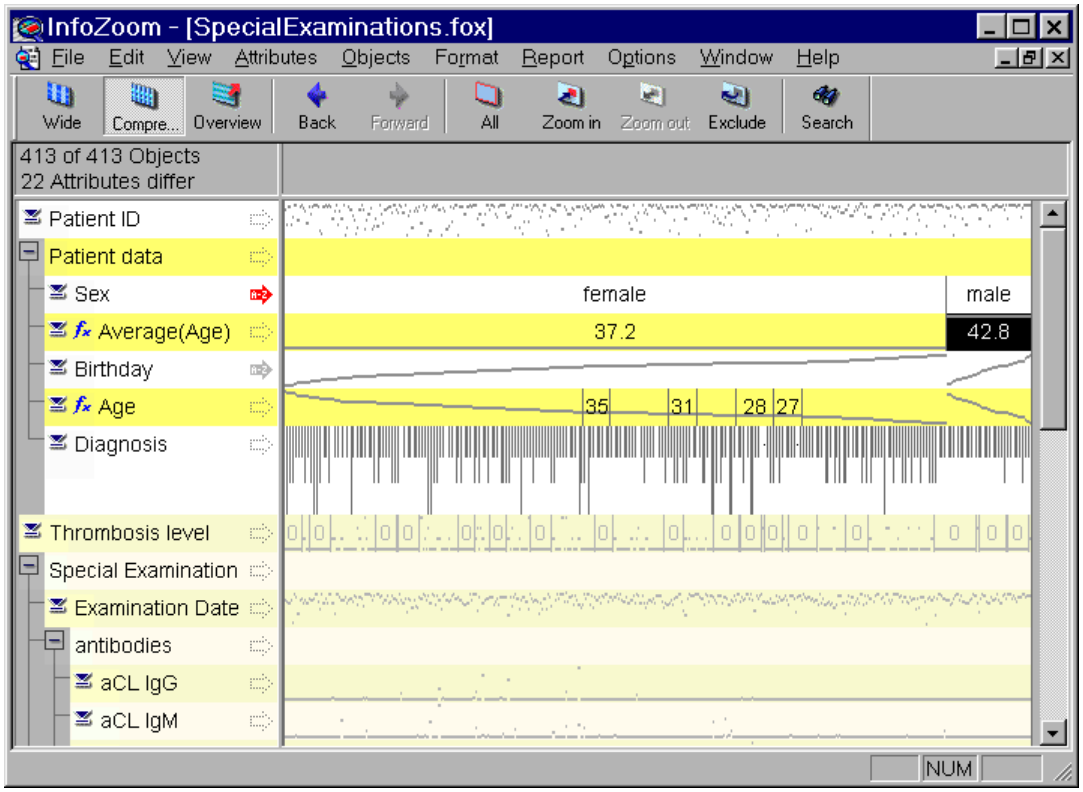

Figure 2: Compressed Table Mode

In *Compressed Mode*, the column width is reduced until all the objects fit into the window so that no horizontal scrolling is necessary. In Figure 2 the column width is about two pixels. In large tables it will be even smaller than one pixel. Some techniques make the table readable in spite of this compression. The most important is that neighboring cells with identical values are combined into one larger cell. Because the table is sorted by the attribute *Sex*, all its values are readable. The width of each cell indicates the number of subsequent objects with this value. So we can see that most of the patients are women. If a cell is to small to display a numeric value, a short horizontal line still indicates its relative height. In this way, it can be seen that *Birthday* and *Age* have complementary values (Figure 2).

Instead of selecting from the list of possible values, an attribute can also be restricted by selecting and double-clicking a value or value-range directly in the table. In a short animation, the clicked cells grow while the others shrink. This looks like *zooming* into the table. For example, double-clicking *male* removes all female patients from the table. The remaining 47 male patients are displayed with an increased column width so that more details can be recognized.

Like formulas in a spreadsheet program, derived attributes can be defined which are automatically updated by InfoZoom whenever necessary. For example, in Figure 2 the *Average(Age)* of male and female patients will be recomputed after each zoom-operation for the remaining patients, while the *Age* has to be recomputed only if the *Birthday* is modified.

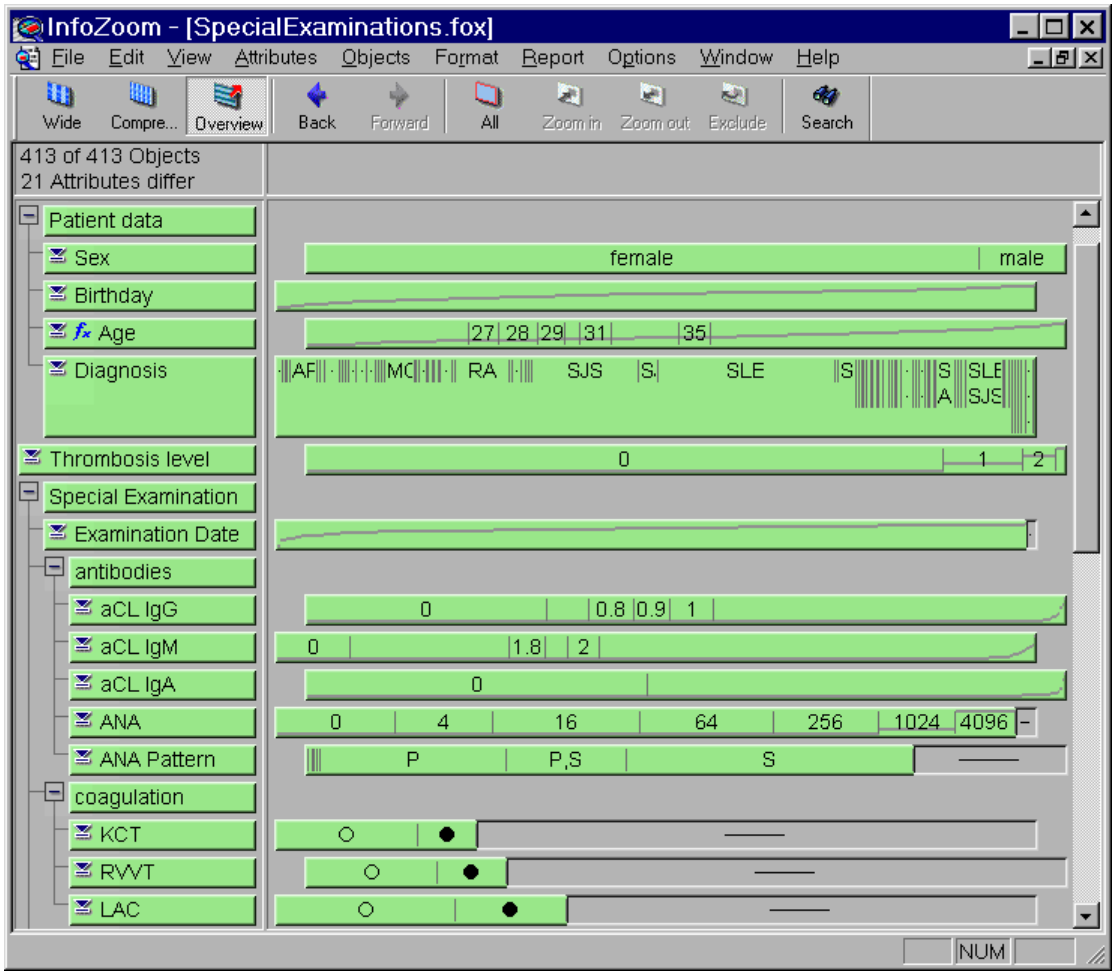

Figure 3: Overview Mode

In *Overview Mode* the values of each attribute are sorted independently, so that the value ranges and distributions of all attributes can be seen at glance (Figure 3). Instead of a table, independent bar charts are displayed.

The analysis of a data set typically starts in the Overview Mode. In Figure 3 one can directly see the percentage of male patients, the uniform distribution of the age of patients, and the frequencies of the three thrombosis levels. The three *aCL* (anti-Cardiolipin antibody) attributes have a rather uneven distribution: Most of the values are quite small, but some are extremely high. The attributes *KCT*, *RVVT*, and *LAC* indicate for the degree of coagulation. They were only measured for about each third patient.

Additional information is displayed when the mouse is moved over cells which are too small to display a value. In Figure 4 the feedback during a select-operation is shown.

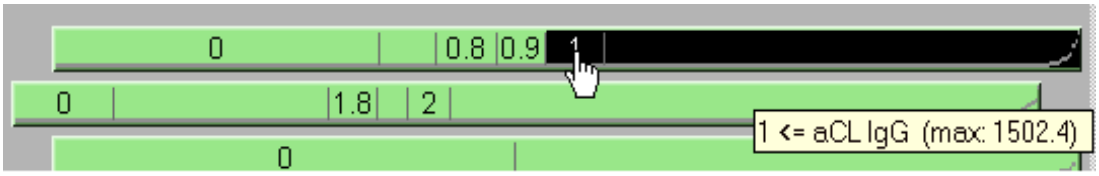

Figure 4: Feedback while dragging a selection

In order to get more details we can select a section of the curve and zoom on it by a double-click. Alternatively, the value list dialog box can be opened.

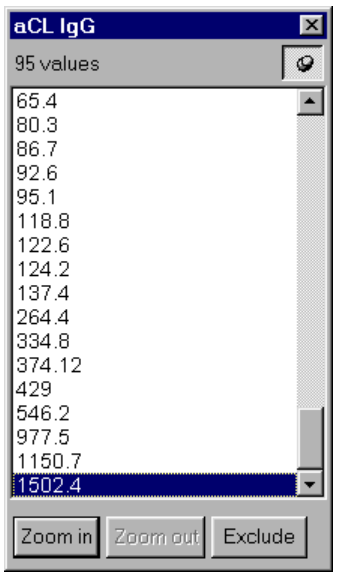

Figure 5: The highest values of attribute *aCL IgG*

Often, typos can be detected by scanning the value lists. For example, the value list dialog for the attribute *Sex* in the original data set revealed a spelling error and two missing entries.

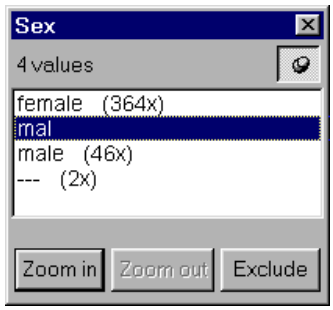

Figure 6: Typos and missing values

Errors can be corrected directly like in a spreadsheet by selecting a cell and typing the correct value. Moreover, a value can be modified for many objects at the same time. For example, in Figure 3 we could select *female* and change it to *woman* for all 364 female patients in a single edit-operation.

Observing Correlations

The main goal of the analysis of the blood test results was to find correlations between the measured parameters and the occurrence of a thrombosis. Therefore, we defined the target attribute *Thrombosis?* which is true if and only if the level of thrombosis is different from 0. It was obtained by duplicating and editing the attribute *Thrombosis level*. Next we defined the derived attributes *Count(Patient ID) per Thrombosis?* and *Percent(Patient ID) per Thrombosis?* which always display how many of the currently displayed patients did have and how many did not have a thrombosis. In Figure 7 only the concatenation of these three attributes is shown. It can be seen that 66 of the 413 patients (16.0%) had a thrombosis.

The easiest technique for the detection of correlations between the target attribute and other attributes is performed in the *Overview Mode* where the value distributions are shown: We zoom-in on the patients with a thrombosis and watch how the value distributions of the measurements change. This is done by double-clicking on the filled circle in the row of the attribute *Thrombosis?* (Figure 7).

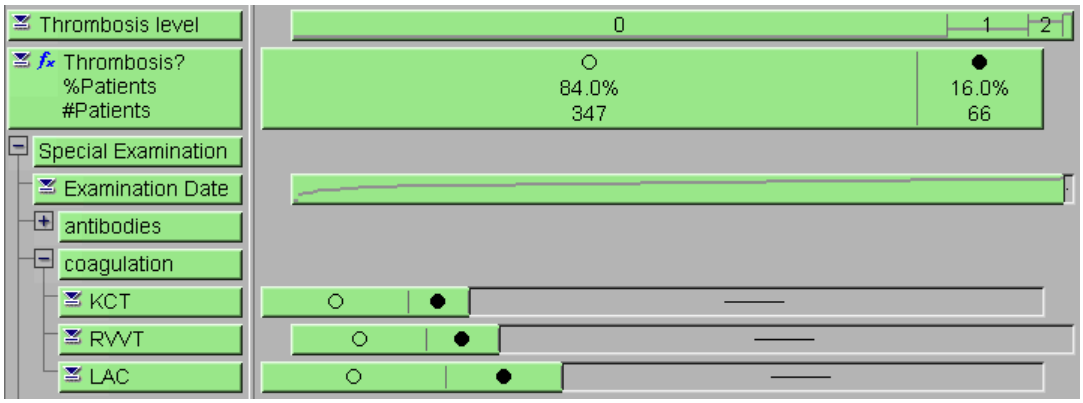

Figure 7: 16% of followed patients have a thrombosis

This starts a short animation where the field with the filled circle is growing and the field with the empty circle becomes smaller. Finally, only the filled circle remains, indicating that only the 66 patients with a thrombosis are displayed now (Figure 8).

Simultaneously during the animation the cell with the filled circle for *LAC* considerably grows. This means that for patients with a thrombosis the LAC more often is positive. So we have found a significant correlation.

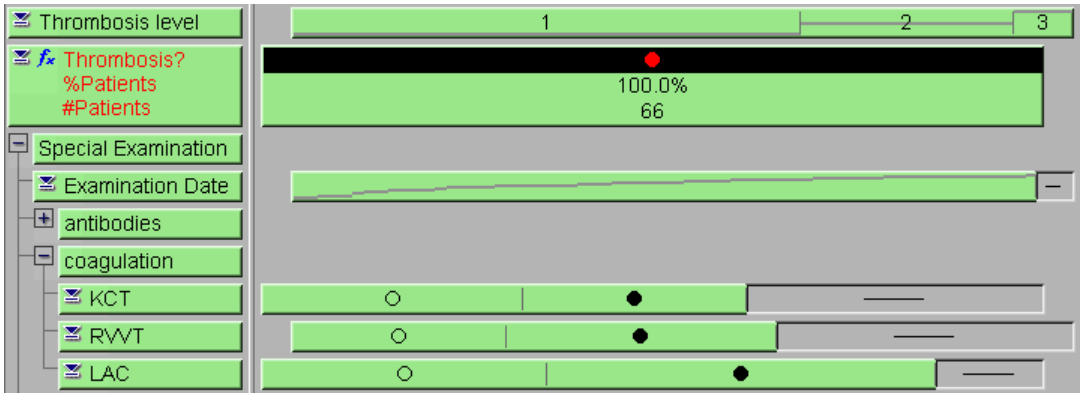

Figure 8: Thrombosis patients more often have a positive LAC

Vice versa, the same correlation can be observed if we zoom-in on the patients with a positive *LAC*: The portion of patients with a thrombosis increases from 16% to 54,1%! This is shown in Figure 9.

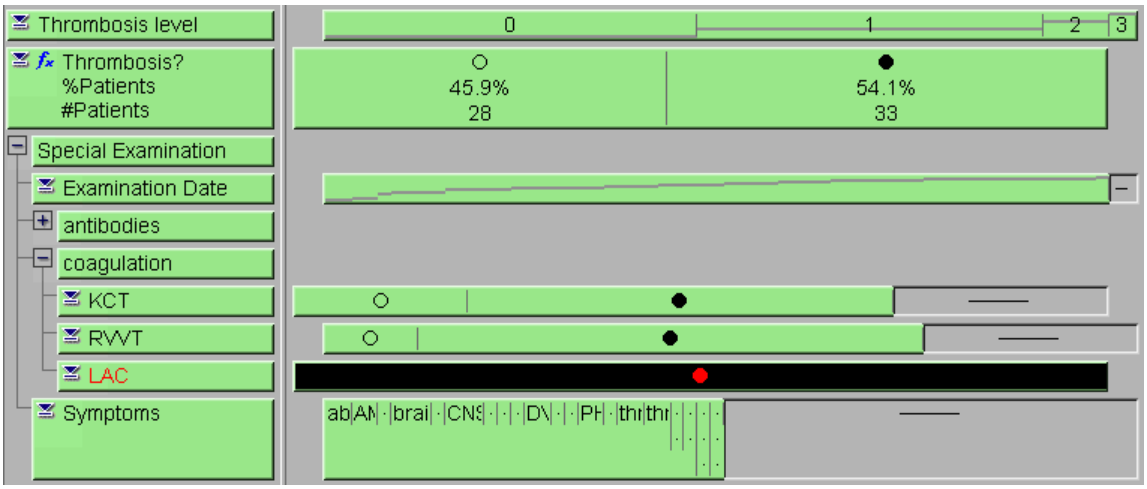

Figure 9: 54.1% of the patients with LAC have a thrombosis

Analogously, one can find out that only 8.5% the male patients had a thrombosis, and that there is no significant difference between old and young patients.

In *Compressed Mode* dependencies can be made obvious by **consecutive sorts**. In Figure 10 we have moved the attribute *LAC* next to the target attribute and sorted the table first by *LAC* and then by the target attribute.

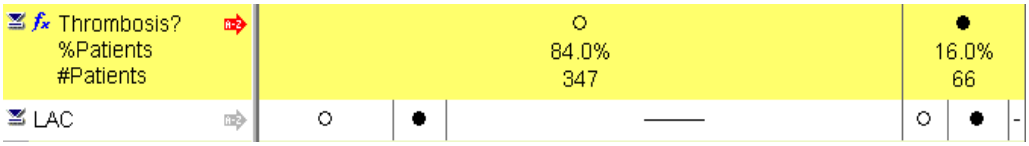

Figure 10: Correlation between LAC and thrombosis

One can observe that for the 66 patients with a thrombosis positive LAC values are measured much more frequently. It also becomes obvious that the LAC was only measured when there was already a strong suspect of thrombosis.

There is also a clear correlation between the three coagulation parameters. This can be seen in Figure 11. The attributes *LAC*, *RVVT*, and *KCT* have been sorted in this order. We can easily see that LAC is positive if either KCT or RVVT are positive.

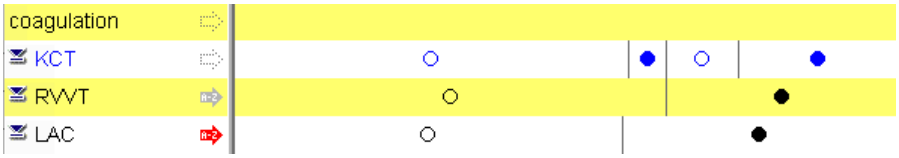

Figure 11: KCT or RVVT positive =>LAC positive

In order to search for correlations with the numeric results of the antibody measurements, the techniques described so far must me slightly extended. While zooming-in on the thrombosis

patients we can observe a general movement of the cells towards the left, mainly because the cells containing *0* shrink. Also, the right end of the curves is distorted and becomes less steep. But it is hard to formulate precise statements.

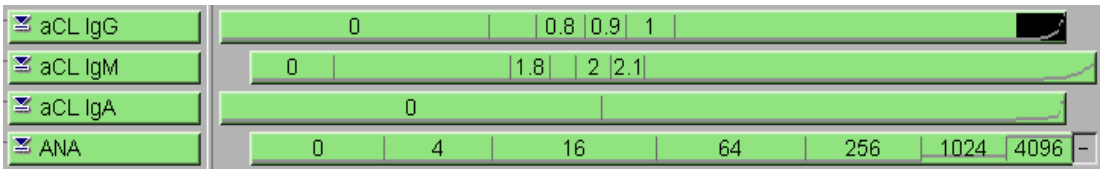

Figure 12: Value distributions for all patients

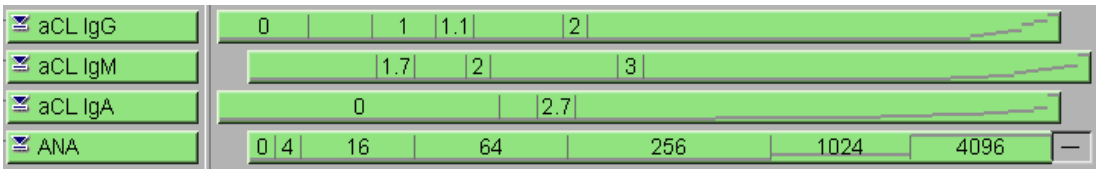

Figure 13: Value distributions for patients with a thrombosis

Vice versa, we can zoom into a range of values, for example the section of the curve selected in Figure 12, and watch how the percentage of thrombosis increases to more than 50%. Different value ranges can quickly be selected and tested.

If we want to do the same analysis more systematically, we perform a *discretization* of the attribute *aCL IgG*. Rounding the values — e.g. to multiples of 10 — does not make sense here because of their unbalanced distribution. Instead we define reasonable ranges manually: First we create a duplicate of *aCL IgG* and rename it to *Range of aCL IgG*. Next we select all values greater than *0* and less than or equal to *1* and type <= 1 . In this way, all selected values are changed to the same string in one edit operation. We perform this step analogously for the other value ranges. The result is shown in Figure 14.

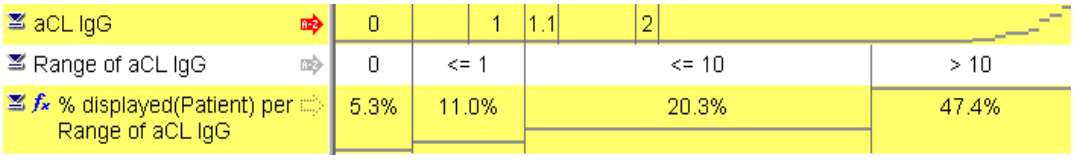

Figure 14: Discretization of a numeric attribute

In Figure 14 we have also defined a derived attribute *% displayed(Patient) per Range of aCL IgG*. This attribute always displays the percentage of patients currently displayed in the table for each value range of *aCL IgG*. As long as the complete table is shown, the result is *100%* for each value range. In Figure 14 we have already zoomed-in on the patients with a thrombosis. Now only 5.3% of the patients with an *IgG* equal to 0 are displayed, but 47.4% of the patients with an *IgG* greater than 10. At a glance, we can see the same percentages we could also obtain by consecutively zooming into each of the four value ranges.

Analyzing Temporal Data

The second data set contains the results of standard laboratory blood tests performed for the same 413 patients who also have undergone the special examination for thrombosis. This data

set is much larger (nearly 20,000 objects) because the patients were tested several times at different days. The number of tests varies between 1 and 367. Figure 15 shows an overview of the data set.

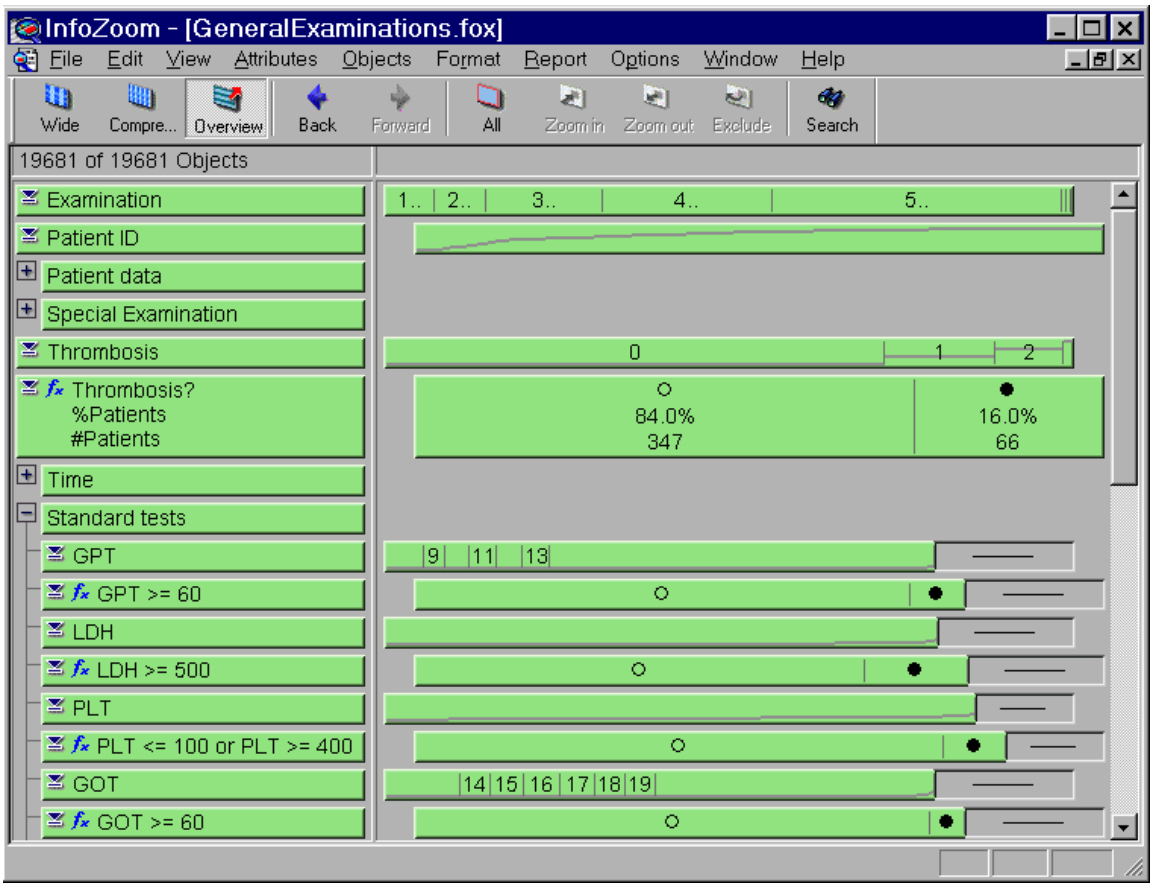

Figure 15: Overview of the general blood examinations

We see the same target attribute as in the first data set. Again, the percentages of patients with and without a thrombosis are displayed. However, the widths of the two cells are not proportional to the percentages because a column in the table does corresponds to a patient but to a single examination.

For each of the numeric blood parameters we have defined a derived Boolean attribute, which indicates whether a value lies within the normal range. There are many more parameters which are not shown in Figure 15.

Again, the goal of the analysis was to find blood parameters (or combinations of them) which correlate with thrombosis. The first idea might be simply to apply the same techniques as for the first data set. For example, we can zoom into the positive *GPT* (glutamic pylvic transaminase) values by double-clicking the filled circle at the attribute *GPT >= 60* and watch the percentage of thrombosis for the remaining patients. These patients had a positive GPT in at least one test, which however might have been performed many years before (or even after) the thrombosis was diagnosed.

In order to be able to formulate more reasonable queries we first defined the derived attribute *Months from Thrombosis test*. It is computed from the difference between the date of the standard blood test and the date of the examination for thrombosis. In Figure 16 we have

already zoomed-in on the tests in the last 12 months before the examination. Moreover, we have defined an attribute *Average(GPT) per Patient ID* and zoomed-in on the values greater than 60. Therefore, only patients whose average GPT exceeded the normal range during the last year before the thrombosis test are shown. For these patients the danger of thrombosis is somewhat higher. However, this finding is not very significant and it is only based on 30 of the 413 patients.

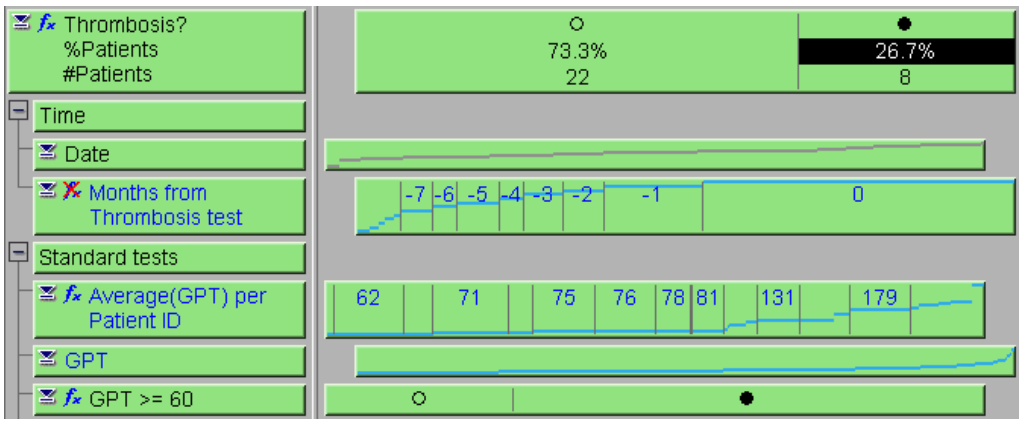

Figure 16: Patients with an increased average GPT in the 12 month before the test

Only medical experts can decide which are the relevant queries that should be performed. The time span to be chosen may vary from parameter to parameter. Instead of the average value, the minimum or maximum value for a patient could be considered, i.e. we could look for patients where all values or none at all exceed a certain threshold.

It is also possible to observe the development of a blood parameter over time. In Figure 17 we have defined the time distance from the thrombosis test in quarters instead of months. The next derived attribute shows the average LDH (lactate dehydrogenase) of all patients at a certain time distance. (3 patients where LDH was never measured are missing here.)

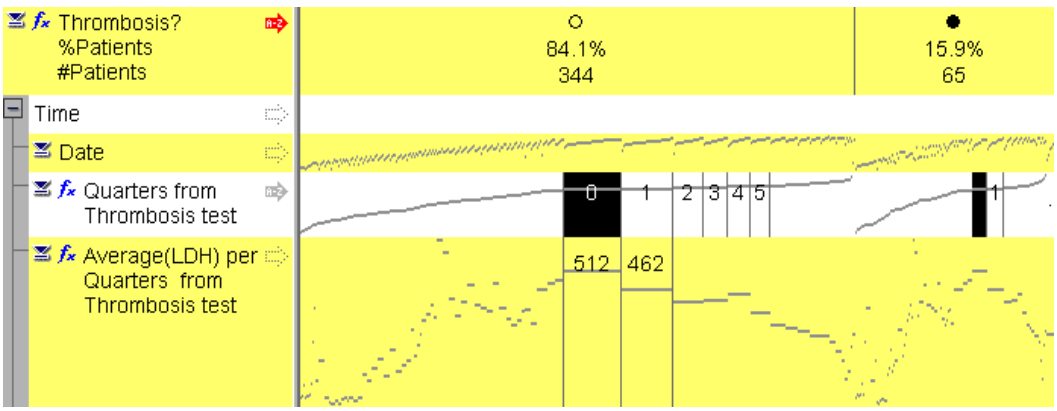

Figure 17: The average LDH increases before

Looking at the right part of table we can observe the general trend that the average LDH increased before the thrombosis and then decreased again. However, a very similar curve is displayed for the patients who did *not* have a thrombosis! Obviously, a correlation between the LDH and the *suspect* of thrombosis exists. It does not seem to be important whether it really comes to a thrombosis.

This last example shows how important it is that the medical experts themselves can explore their data sets interactively, get a feeling of the data, ask the right questions, and judge the relevance of the answers. An automatic data mining algorithm can only be a supplement to this intellectual process of *understanding* the data.

Related Work

The *TableLens* [4] is the only approach we know which also uses the basic idea of compressing database tables until they completely fit on the screen. While InfoZoom displays each record in a column, in *TableLens* each row contains a record. Therefore, the *TableLens* cannot use the technique of uniting neighbored cells with identical values, which is vital to make textual values readable. Also, a counterpart to InfoZoom s *Overview Mode* is missing. The graphical representation of numeric values, however, is quite similar.

InfoZoom follows the principle of *Dynamic Queries* introduced by Ahlberg and Shneiderman [1]: Whenever a zoom-operation restricts the values of one attribute, the remaining possible values of all other attributes are immediately reflected in the table and the value lists. Therefore it is hardly impossible to perform a query returning no solution at all.

Conclusion

We have shown the interactive techniques for visual data analysis supported by InfoZoom. The goal of our approach is not a completely automatic algorithm that searches for interesting results. Instead, InfoZoom enables the user to interactively explore the data set and to get a feeling of the contained information. It introduces a novel visualization technique which displays the whole data set on a single screen. Queries are simply performed by selecting parts of the displayed data. Derived attributes can be defined like in a spreadsheet program and are automatically updated when necessary. We are convinced that using InfoZoom is simple enough to be used by medical experts in order to understand their data and to detect the hidden knowledge.

InfoZoom allows the user to perform queries very easily and quickly, but a systematic search of all possible queries is usually too time consuming. Therefore, we plan to integrate a data mining algorithm into InfoZoom which leads the user to interesting views of the data.

Even if an automatic data mining algorithm is used, a tool like InfoZoom is important for getting a first impression of the data set, checking the value ranges, and finding typos, inconsistencies, and missing values. Moreover, defining the relevant derived attributes is a very important step before running a mining tool.

InfoZoom was initially developed at GMD — the German National Research Center for Information Technology. It is now extended and marketed by the GMD spin-off company humanIT [3]. InfoZoom runs on any standard Windows PC and can read data from flat text files and a broad range of database systems using the standardized ODBC (Open DataBase Connectivity) interface. Databases containing a few hundred thousand or even a million records can be analyzed provided that sufficient main memory is available. A fully functional evaluation version can be downloaded from [3].

References

[1] Ahlberg, C. and Shneiderman, B., Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. In *Proceedings of the ACM SIGCHI*

Conference on Human Factors in Computing Systems (Boston, MA, Apr 24—28, 1994), pp. 313—317.

[2] Beilken, Chr.; Spenke, M., Visual, Interactive Data Mining with InfoZoom — the Medical Data Set. In *Workshop Notes on Discovery Challenge*, 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases, Prague, September 15-18,1999, pp. 49-54.

[3] http://www.humanIT.de — The InfoZoom home page. A free test version of InfoZoom can be downloaded. [2] and [5] are available online.

[4] Rao, R. and Card, S. K., The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+Context Visualization for Tabular Information. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, Apr 24—28, 1994), pp. 318—322.

[5] Spenke, M.; Beilken, Chr.; Berlage, Th., FOCUS: The Interactive Table for Product Comparison and Selection, *Proceedings of the UIST 96 Ninth Annual Symposium on User Interface Software and Technology*, Seattle, November 6 - 8, 1996. ACM 1996, pp. 41-50.

[6] Tsumoto, S.: Guide to the Medical Data Set. In *Workshop Notes on Discovery Challenge*, 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases, Prague, September 15-18,1999, pp. 45-47.