Ertek, G. (2009). "Visual Data Mining for Developing Competitive Strategies in Higher Education," in Data Mining for Business Applications. Editors: Longbing Cao, Philip S. Yu, Chengqi Zhang and Huaifeng Zhang. Springer.

Note: This is the final draft version of this paper. Please cite this paper (or this final draft) as above. You can download this final draft from http://research.sabanciuniv.edu.

IMPORTANT: Even though this draft erroneously shows the header "Chapter 1", the final version appeared as "Chapter 18 - Visual Data Mining for Developing Competitive Strategies in Higher Education" in the book "Data Mining for Business Applications" edited by Longbing Cao, Philip S. Yu, Chengqi Zhang and Huaifeng Zhang, and published by Springer in 2009.

Developing Competitive Strategies in Higher Education through Visual Data Mining

Gürdal Ertek

Faculty of Engineering and Natural Sciences

Sabanci University

Istanbul, Turkey

Chapter 1
Developing Competitive Strategies in Higher Education through Visual Data Mining

Gürdal Ertek

Abstract Information visualization is the growing field of computer science that aims at visually mining data for knowledge discovery. In this paper, a data mining framework and a novel information visualization scheme is developed and applied to the domain of higher education. The presented framework consists of three main types of visual data analysis: Discovering general insights, carrying out competitive benchmarking, and planning for High School Relationship Management (HSRM). In this paper the framework and the square tiles visualization scheme are described and an application at a private university in Turkey with the goal of attracting brightest students is demonstrated.

1.1 Introduction

Every year, more than 1,5 million university candidates in Turkey, including more than half a million fresh high school graduates, take the University Entrance Exam (Öğrenci Seçme Sınavı- ÖSS) to enter into a university. The exam takes place simultaneously in thousands of different sites and the candidates answer multiple-choice questions in the 3-hour exam that will change their life forever. Entering the most popular departments -such as engineering departments- in the reputed universities with full scholarship requires ranking within the top 5,000 in the exam.

In recent years, the establishment of many private universities, mostly backed-up by strong company groups in Turkey, have opened up new opportunities for university candidates. As the students compete against each other for the best universities, the universities also compete to attract the best students. Strategies applied by universities to attract the brightest candidates are almost standard every year: Publishing past years' placement results, promoting success stories in press -especially

Gürdal Ertek

Sabancı University, Faculty of Engineering and Natural Sciences, Orhanlı, Tuzla, 34956, Istanbul, Turkey, e-mail: ertekg@sabanciuniv.edu

newspapers-, sending high-quality printed and multimedia catalogs to students of selected high schools, arranging site visits to selected high schools around the country with faculty members included in the visiting team, and occasionally spreading bad word-of-mouth for benchmark universities.

Sabancı University was established in 1999 by the Sabancı Group, the second largest company group in Turkey at that time, at the outskirts of Istanbul, the megacity of Turkey with a population of nearly 20 million people. During 2005 and 2006 an innovative framework -based on data mining- was developed at Sabancı University with the collaboration of staff from the Student Resources Unit, who are responsible of promoting the university to high school students, and the author from Faculty of Engineering and Natural Sciences. The ultimate goal was to determine competitive strategies through mining annual ÖSS rankings for attracting the best students to the university. In this paper, this framework and the square tiles visualization scheme devised for data analysis is described.

The developed approach is based on visual data mining through a novel information visualization scheme, namely *square tiles visualization.* The strategies suggested to the managing staff at the university's Student Resources Unit are built on the results of visual data mining. The steps included in visual data mining include performing competitive benchmarking of universities and departments, and establishment of the *High School Relationship Management* (HSRM) decisions, such as deciding on which high schools should be targeted for site visits, and how site visits to these high schools should be planned.

In the study, information visualization was preferred against other data mining methods, since the end-users of the developed Decision Support System (DSS) would be staff at the university and undergraduate students. In information visualization, patterns such as outliers, gaps and trends can be easily identified without requiring any knowledge of the mathematical/statistical algorithms. Development of a novel visualization scheme was motivated by the difficulties in the perception of irregular tile shapes of existing schemes and software.

In this paper, a hybrid visualization scheme is proposed and implemented to represent data with categorical and numerical attributes. The visualization that is introduced and discussed, namely square tiles, shows each record in a query's results as a colored icon, and sizes the icons to fill the screen space. The scheme is introduced in Section 1.2. The mathematical model solved for generating the visualizations is presented in Section 1.3 and the software implementation is discussed. In Section 1.4 the analysis of ÖSS data demonstrated with snapshots of the developed *SquareTiles* software that implements square tiles visualization. In Section 1.5 related work is summarized. In Section 1.6 future work is outlined. Finally in Section 1.7 the paper is summarized and conclusions are presented.

1.2 Square Tiles Visualization

Information visualization is the growing field of computer science that studies ways of visually mining high-dimensional data to identify patterns and derive useful insights. Patterns such as trends, clusters, gaps and outliers can be easily identified by information visualization. Keim [10] presents a taxonomy of information visualization schemes based on the data type to be visualized, the visualization technique used, and the interaction and distortion technique used. Recent reviews of information visualization literature have been carried out by Hoffman & Grinstein [6] and de Oliveira & Levkowitz [3]. Many academic and commercial information visualization tools have been developed within the last two decades, some of which are listed by Eick [4]. Internet sources on information visualization include [7] and [11].

The main differences of information visualization from other data mining methods such as association rule mining and cluster analysis are two-folds: Information visualization takes advantage of the rapid and flexible pattern recognition skills of humans [13], and relies on human intuition as opposed to understanding mathematical/statistical algorithms [10].

In the square tiles visualization scheme (Figure 1.1) each value of a selected categorical attribute (such as high schools in this study) is represented as a distinct box, and the box is filled with strictly-square tiles that represent the records in the database based on the value of the categorical attribute. Colors of the tiles correspond to the values of a selected numerical attribute (SS ranking in this study). One can use the names *partitioning attribute* and *coloring attribute* for these attributes, respectively, similar to the naming convention in [9].

Tile visualization has been widely used before, and has even been implemented in commercial software such as Omniscope [12]. However, existing systems either can not use the screen space efficiently, or display the data with the same tile size through irregularly shaped rectangles. The novelty and the advantage that square tiles visualization brings is the most efficient use of the screen space for displaying data when the tiles are strictly square. The problem of "maximizing the utilization of the screen space with each queried record being represented as a square tile" is formulated as a nonlinear optimization problem, and can be solved to optimality in reasonable time through exhaustive enumeration.

Square tiles can be considered as a two-dimensional extension of the well-known Pareto Charts. A Pareto chart is a two-dimensional chart which plots the cumulative impact on the y-axis against the percentage of elements sorted on the x-axis based on their impact. The cumulative impact is typically a non-linear, concave function of the percentage of the elements: A small percentage of the most important elements are typically observed to account for a great percentage of the impacts. In square tiles visualization, the areas of the most important sets and the distribution of the elements in different sets with respect to the coloring attribute can be compared.

The color spectrum used to show the numerical attribute starts from yellow, which shows the largest value, continues to red, and ends at black, which shows the smallest value. This color spectrum allows easy identification of patterns on a grey-scale printout, and has also been selected in [1].
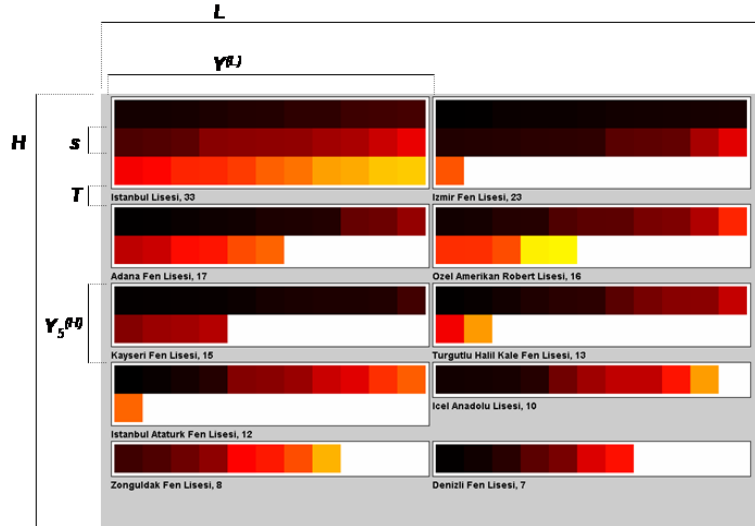
Fig. 1.1 Composition of entrants to a reputed university with respect to high schools (HS_NAME)

The placement of icons within boxes is carried out from left to right and from top to bottom according to the coloring attribute. The layout of boxes within the screen is carried out again from left to right and from top to bottom based on the number of icons within each box. The PeopleGarden system [16] developed at MIT also considers a similar layout scheme.

1.3 Mathematical Model

Each square tiles visualization is generated based on the optimal solution of the mathematical model presented below. Let

$\mathscr{I}$: the set of all boxes to be displayed, with $|\mathscr{I}| = n$
N_i: the number of icons in box i.

Let the parameters be defined as follows:

T: text height
B: space between boxes
P: pixel allowance within each box
m: minimum length of each box
$\overline{S}$: maximum icon size for each element
L: length of the screen area
H: height of the screen area

The most important variables are

s: the size (side length) of each icon, and
$x^{(h)}$: the number of horizontal icons placed in each box.

In the solution algorithm the values of these two variables are changed to find the best set of variable values.

Let the other variables be defined as follows:

$x_i^{(v)}$: number of vertical icons in box i
$y^{(L)}$: length of each box
$y_i^{(H)}$: height of box i
$Y^{(L)}$: total length of each box
$Y_i^{(H)}$: total height of box i
$Z^{(h)}$: number of horizontal boxes
$Z^{(v)}$: number of vertical boxes

It should be noted that $s, x^{(h)}, x^{(v)}, y^{(L)}, y_i^{(H)}, Y^{(L)}, Y_i^{(H)}, Z^{(h)}, Z^{(v)} \in \mathbb{Z}^+$, where $\mathbb{Z}^+$ is the set of positive integers.

The mathematical model is given below:

$$max \quad \alpha$$

$$\text{s.t.}$$

$$\alpha = \frac{\sum\limits_{i \in \mathscr{I}} y^{(L)} y_i^{(H)}}{LH} \tag{1.1}$$

$$y^{(L)} = 2P + x^{(h)} s \tag{1.2}$$

$$x_i^{(v)} = \lceil N_i / x^{(h)} \rceil, \ \forall i \in \mathscr{I} \tag{1.3}$$

$$y_i^{(H)} = 2P + x_i^{(v)} s, \ \forall i \in \mathscr{I} \tag{1.4}$$

$$Y_i^{(H)} = y_i^{(H)} + B + T, \ \forall i \in \mathscr{I} \tag{1.5}$$

$$Y^{(L)} = y^{(L)} + B \tag{1.6}$$

$$Z^{(h)} = \lfloor L / Y^{(L)} \rfloor \tag{1.7}$$

$$Z^{(v)} = \left\{ k : \max j \text{ s.t.} \sum_{i=1, Z^{(h)}+1, \dots, jZ^{(h)}+1} Y_i^{(H)} \le H \right\} \tag{1.8}$$

$$\alpha \le 1 \tag{1.9}$$

$$n \le Z^{(h)} Z^{(v)} \tag{1.10}$$

$$m \le y^{(L)} \le L \tag{1.11}$$

$$1 \leq s \leq \overline{S} \qquad\qquad (1.12)$$

The objective in this model is to maximize α, which is defined in (1) as the ratio of the total area occupied by the boxes to the total screen area available. Thus the objective of the model is to maximize screen space utilization. The length of each box $y^{(L)}$ is calculated in (2) as the summation of the pixel allowances $2P$ within that box and the vertical length $x^{(h)}s$ of the icons in that box. (3) calculates the number of vertical icons of box i, namely $x_i^{(v)}$. Calculation of $y_i^{(H)}$, the height of box i in (4), is similar to the length calculation in (2). Calculations in (5) and (6) take the space between boxes B and the text height T into consideration. The number of horizontal boxes $Z^{(h)}$ is calculated in (7). The number of vertical boxes $Z^{(v)}$ is calculated in (8) by finding the maximum j value such that the total height of the boxes does not exceed H, the height of the screen area. (9) states that α can not exceed 1, since it is denoting utilization. (10) guarantees that all the required boxes are displayed on the screen. (11) puts bounds on the minimum and maximum values of $y^{(L)}$, and thus indirectly s. The last constraint (12) bounds the range of s.

To solve the problem to optimality, the variables s and $x^{(h)}$ are changed within bounds that are calculated based on (11), (6), (2) and (12), and the feasible solution that yields the maximum α value is selected as optimum. For determining feasibility of a $(s, x^{(h)})$ combination, the calculations in (2) through (8) are carried out and the feasibility conditions in (9) and (10) are checked. Once the best $(s, x^{(h)})$ combination is determined, the visualization is generated based on the values of the calculated parameters and variables.

Implementation

The *SquareTiles* software has been developed to create the visualizations, and adopted to the analysis of a particular data set. The software is implemented using Java under Eclipse Integrated Development Environment (IDE) [5]. The data is stored in a Microsoft Access database file, and is queried from within the Java program through ODBC connection. The software developed allows a user without any prior knowledge of a querying language (such as SQL) to create queries that generate visualizations. The software does not have a user-interface at the moment and this is an important requirement for it to be adopted as a decision support system in the long run.

Example Layout

Figure 1.1 displays the parameters L, H, T, and the variables $s, Y^{(L)}, Y_5^{(H)}$ on a sample output. These refer to the number of pixels. From the figure, we can also deduce that the number of horizontal and vertical boxes are $Z^{(h)} = 2$ and $Z^{(v)} = 5$, respectively. The number of icons in boxes 1, 2, ... are $N_1 = 33, N_2 = 23, ...$. The number of horizontal and vertical icons in box 1 can be counted as $x^{(h)} = 11$ and $x_1^{(v)} = 3$. In the

model $x^{(h)}$ and s are the most important variables, whose values are determined such as to maximize screen space utilization. The values of other variables are calculated based on these two.

1.4 Framework and Case Study

In this section, the framework developed for analyzing and understanding the ÖSS data through square tiles will be presented and demonstrated. The selected data set contains information on the top ranking students in ÖSS for a selected year. The selected ÖSS data set includes 5,965 records, covering students within the top 5,000 with respect to two types of scores. The attributes (dimensions) in the data set include HS_NAME (high school name), HS_TYPE_TEXT (high school type in text format), UNIV_NAME (university name), UNIV_DEPT (university department), RANK_SAY (the student's rank according to score type *sayısal* (science and mathematics based)). All of these attributes are categorical, except the rank attribute, which is numerical.

Sabancı University is a newly established private university which accepts students mostly from the top 1% of the students that take ÖSS. Traditionally (until 2007) the Student Resources Unit at Sabancı University assembled the data on top 5,000 students in the exam and analyzed it using spreadsheet software. However, only basic graphs which provide aggregate summaries were generated using the yearly data sets.

The ÖSS data set provided by the Student Resources Unit had to be cleaned to carry out the analysis with square tiles visualization. The main problems were multiple entries for the same value, and missing attribute values for some records. A taxonomy of dirty data and explanation of the techniques for cleaning it is presented by Kim et al. [2]. According to this taxonomy, the issues faced in here all "require examination and repair by humans with domain expertise".

The *SquareTiles* software allowed a range of analysis to be carried out -by freshmen students with no database experience- and interesting and potentially useful insights to be derived. A report was prepared for the use of Student Resources Unit at Sabancı University that contains competitive benchmarking for 7 selected universities and guidelines for developing strategies in managing relationships with 52 selected high-schools. The study suggested establishment of a new approach for High School Relationship Management (HSRM), where the high schools are profiled through information visualization.

Several suggestions were received from staff within the Student Resources Unit during discussions: One suggestion was the printing of the number of icons in each box (thus, the cardinality of each set). This suggestion was implemented within the software.

The proposed framework consists of three main types of analysis described in the below subsections.

1.4.1 General Insights and Observations

The visualizations can be used to gain insights into the general patterns. One example is given in Figure 1.2, which displays the distribution of top 5,000 students with respect to top 10 high school types.
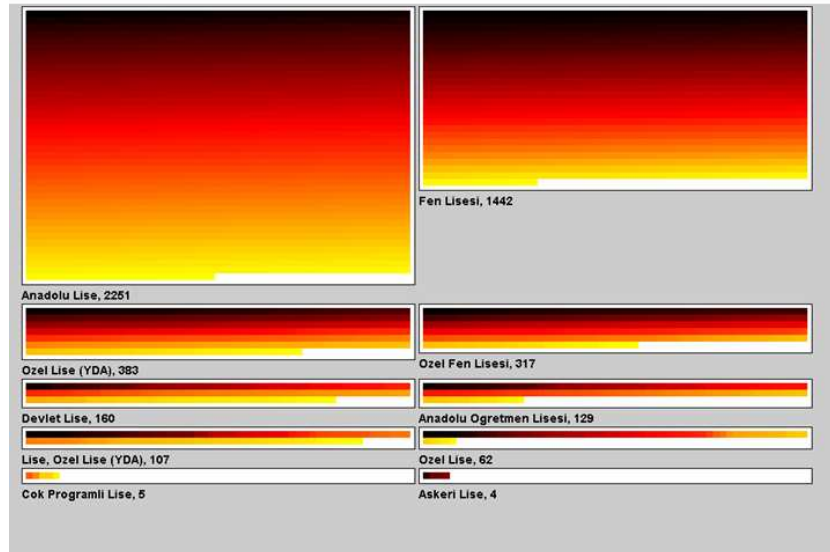

Fig. 1.2 Distribution of top 5,000 students with respect to top 10 high school types

From Figure 1.2 it can be seen that the high school types `Anadolu Lise` (Anatolian high schools which teach foreign language, and are preferred by many families for this reason) and `Fen Lisesi` (science high schools) are the most successful. When comparing these two school types, one can observe that the number of darker icons are approximately the same for both. This translates into the fact that science high schools have a greater proportion of their students in the high ranks (with dark colors).

From the figure, once can also observe a pattern that would be observed in a Pareto Chart: The two high school types account for more than half of the screen area; that is, these two high school types impact the Turkish Education System much more significantly than others by accounting for more than half of the top 5,000 students in ÖSS.

`Ozel Lise` (private high schools) and `Ozel Fen Lisesi` (private science high schools) follow the first two high school types. One pattern to notice is the low success rate of `Devlet Lise` (regular public high schools). Even though regular public high schools outnumber other types of high schools by far in Turkey, their success rate is very much below the high school types discussed earlier.

1.4.2 Benchmarking

Benchmarking High Schools

Figure 1.1 gives the composition of entrants from within top 5,000 to a reputed university with respect to top 10 high schools. This figure highlights a list of high schools that Sabancı University should focus on. Top performing high schools, such as `Istanbul Lisesi` and `Izmir Fen Lisesi` should receive special attention and active promotion should be carried out at these schools. One striking observation in the figure is that almost all of the significant high schools are either `Anadolu Lise` (Anatolian high schools) or `Fen Lisesi` (science high schools). The only private high school in the top 10 is `Ozel Amerikan Robert Lisesi`, an American High School that was established in 1863.

Detailed benchmarking analysis of selected universities revealed that there can be significant differences between the universities with respect to the high schools of the entrants. One strategy suggested to the staff of the Student Resources Unit at Sabancı University was to identify high schools that send a great number of students to selected other universities, and carry out a focused publicity campaign geared towards attracting students of these high schools.

Benchmarking Departments

Figure 1.3 gives the distribution of entrants from within 5,000 to top 10 departments of the discussed university. One can visually see and compare the capacities for each department. From the color distributions it can be deducted immediately that the departments `Bilgisayar` (Computer Engineering), `Endustri` (Industrial Engineering), and `Elektrik-Elektronik` (Electrical-Electronics Engineering) are selected by the higher-ranking students in general. Among these three departments, Electrical-Electronical Engineering has the distribution of students with the highest rankings. `Makine` (Mechanical Engineering) and other departments are selected by lower-ranking students from within the top 5,000. It is worthy to observe that there is one student that entered `Iktisat` (Economics) with a significantly higher ranking than others who entered the same department. The same situation can be observed in the least populated four departments in the figure: There exist a number of higher ranking students who selected these departments, who probably had these departments as their top choices.

1.4.3 High School Relationship Management (HSRM)

Figure 1.4 depicts the department preferences of students from a reputed high school within the top 5,000. This distribution is particularly important when planning publicity activities towards this high school. The most popular selections
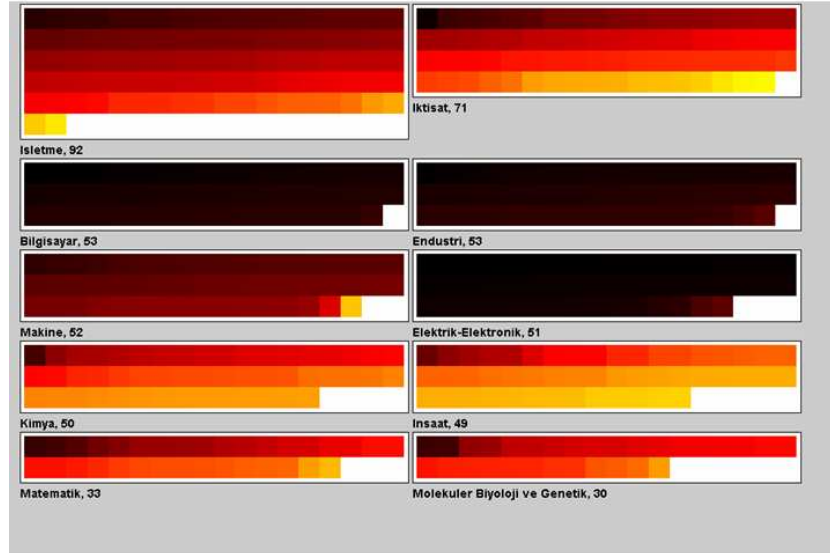
Fig. 1.3 Composition of entrants to a reputed university with respect to top 10 high schools

are `Endustri` (Industrial Engineering) and `Bilgisayar` (Computer Engineering). The existence of two additional students with high rankings who selected `Endustri (Burslu)` (Industrial Engineering with scholarship) further indicates the vitality of industrial engineering. So when a visit is planned to this high school, the faculty member selected to speak should be from the industrial engineering department who also has a fundamental understanding of computer engineering. It could also be a good strategy to ask this faculty member to emphasize the relationship between industrial engineering and computer science/engineering. Throughout the analysis of 52 selected high schools, significantly differing departmental preferences have been observed, which suggests that publicity activities should be customized based on the profiles of the schools.

1.5 Related Work

The icon-based approach followed in this paper is closest to the approach taken by Sun [13]. The author represents multidimensional production data with colored icons within *smashed tables* (framed boxes). In both [13] and the research here the icons are colored squares which denote elements in a set, with the colors representing values of a numerical attribute. In both papers, *a small multiple design* (Tufte [14], p42, p170, p174) is implemented, where a standard visual design is repeatedly presented side by side for each value of one or more categorical attribute(s).
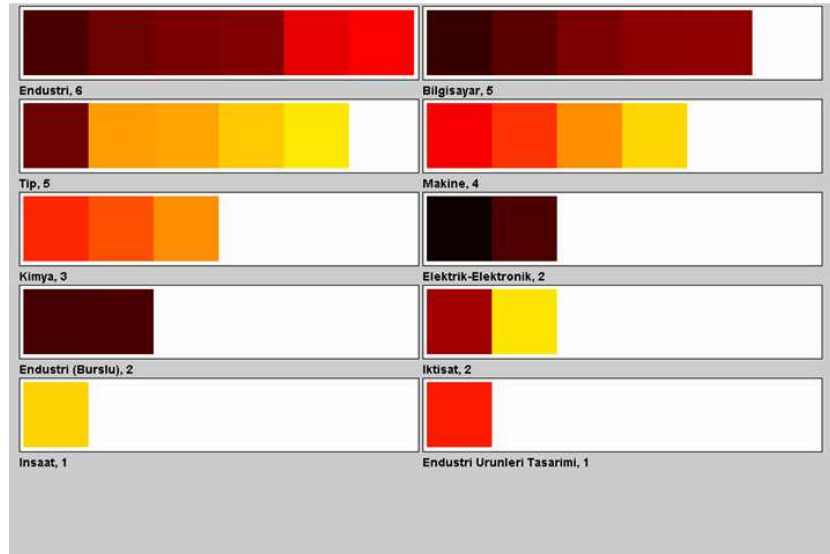
Fig. 1.4 Department (UNIV_DEPT) preferences of students from a reputed high school in Istanbul

Space-filling visualizations seek full utilization of the screen space by displaying attributes of data in a manner to occupy the whole screen space. Square tiles visualization in this paper adjusts the sizes of icons and the layout of the icons to achieve this objective, so it can be considered as a space-filling visualization scheme.

One type of space-filling visualization is pixel-based visualization, where a spacefilling algorithm is used to arrange pixels on the screen space at full space-utilization [8], [9]. Pixel-based visualizations are able to depict up to hundreds of thousands of elements on the screen space, since each pixel denotes an element (such as a person). The research presented in here is very similar to pixel-based visualization research, but also shows one important difference: In [8] and [9] each element of a set is denoted by a single pixel. In here, each element is denoted by a square tile. On the other hand, the research here is also similar to [8] and [9] in the sense that in all these studies a mathematical optimization model, with objective function and constraints, that determines the *best* layout is discussed.

1.6 Future Work

The most obvious extension to the research presented here is the innovation or adoption of effective and useful visualization schemes that allow carrying out currently unsupported styles of analysis. One such analysis is the analysis of changes in the queried sets over time. To give an example with the ÖSS data set, one would most

probably be interested in visually comparing the distribution of students from a high school to universities in two successive years.

Ward [15] provides a taxonomy of icon (glyph) placement strategies. One area of future research is placing icons and boxes in such a way to derive the most insights.

One weakness of the current implementation is that it does not allow user's interaction with the visualization. The software presented here can be modified and its scope can be broadened to enable visual querying and interaction.

1.7 Conclusions

In this paper, the application of data mining within higher education is illustrated. Meanwhile, the novel visualization scheme used in the study, namely square tiles was introduced and its applicability was illustrated throughout the case study. The selected data contains essential information on top ranking students in the National University Entrance Examination in Turkey (ÖSS) for a selected year. The software implementation of the visualization scheme allows users to gain key insights, carry out benchmarking, and develop strategies for relationship management. As the number of attributes increases, the potential of finding interesting insights has been observed to increase.

The developed *SquareTiles* software that implements the visualization scheme requires no mathematical or database background at all, and was used by two freshmen students to carry out a wide range of analysis and derive actionable insights. A report was prepared for the use of Student Resources Unit at Sabancı University that contained competitive benchmarking analysis for seven of the top universities and guidelines for HSRM for 52 selected high schools.

Detailed benchmarking analysis of selected universities revealed that there exist significant differences between the universities with respect to the high schools of the entrants. One strategy suggested to the staff of the Student Resources Unit at Sabancı University was to identify high schools that send a large number of students to selected other universities, and carry out a focused publicity campaign geared towards attracting students of these high schools. The analysis for HSRM provided the details of managing relations with the 52 selected high schools.

The described framework was implemented at Sabancı University for one year, but was discontinued due to the high costs of retrieving the data from ÖSYM, the state institution that organizes ÖSS, and due to the difficulties in arranging faculty members to participate in HSRM activities. Still, we believe that the study serves as a unique example in the data mining literature, as it reports discussion of practical issues in higher education and derivation of actionable insights through visual data mining, besides development of a generic information visualization scheme motivated by a domain-specific problem.

References

1. Abello, J., Korn, J.: MGV: A system for visualizing massive multidigraphs. IEEE Transactions on Visualization and Computer Graphics **8**, no.1, 21–38 (2002)
2. Kim, W., Choi, B., Hong E., Kim, S., Lee, D.: A taxonomy of dirty data. Data Mining and Knowledge Discovery. **7**, 81–99 (2003)
3. de Oliveira, M. C. F., Levkowitz, H.: From visual data exploration to visual data mining: a survey. IEEE Transactions on Visualization and Computer Graphics **9**, no.3, 378–394 (2003)
4. Eick, S. G.: Visual discovery and analysis. IEEE Transactions on Visualization and Computer Graphics **6**, no.1, 44–58 (2000)
5. http://www.eclipse.org
6. Hoffman, P. E., Grinstein, G. G.: A survey of visualizations for high-dimensional data mining. In: Fayyad, U., Grinstein, G. G., Wierse, A. (eds.) Information visualization in data mining and knowledge discovery, pp. 47-82 (2002)
7. http://iv.homeunix.org/
8. Keim, D. A., Kriegel, H.: VisDB: database exploration using multidimensional visualization. IEEE Computer Graphics and Applications. September 1994, 40–49 (1994)
9. Keim, D. A., Hao, M. C., Dayal U., Hsu, M.: Pixel bar charts: a visualization technique for very large multi-attribute data sets. Information Visualization. **1** 20–34 (2002)
10. Keim, D. A.: Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics. **8**, no.1, 1–8 (2002)
11. http://otal.umd.edu/Olive/
12. http://www.visokio.com
13. Sun, T.: An icon-based data image construction method for production data visualization. Production Planning & Control. **14**, no.3, 290–303 (2003)
14. Tufte, E. R.: The Visual Display of Quantitative Information. Graphics Press, Cheshire, CT. (1983)
15. Ward, M. O.: A taxonomy of glyph placement strategies for multidimensional data visualization. Information Visualization. **1**, 194–210 (2002)
16. Xiong, B., Donath, J.: PeopleGarden: Creating data portraits for users. Proceedings UIST '99 Conference, ACM 37–44 (1999)