Automatic generation of statistical graphics

extended abtract submitted to CMC-95 – International Conference on Multimodal Communication

Massimo Fasciano
Guy Lapalme
Département d'informatique et de recherche opérationnelle
Université de Montréal
CP 6128, Succ Centre-Ville
Montréal Québec Canada
H3C 3J7

December 1994

Abstract

Graphics are often seen as an essential part of a report but their appropriate use has remained quite elusive. This work seeks to define the rules for the automatic generation of the appropriate graphics given an author's intention. We describe a system that implements these rules in the context of statistical reports.

1 What is a statistical report?

Reports are an organized synthesis of data that span a whole array of forms going from tables of numbers to a text summarizing the findings. Statistical reports are particularly interesting because the reader can easily be overwhelmed by the raw data. Whithout an appropriate preliminary statistical analysis to make the important points stand out and, without an efficient organization and presentation, the reader might be lost.

Graphics and text are two different media that have to be well integrated in order to achieve their full potential. A picture shows but a text describes. In a statistical report, graphics show the data that is analyzed in the text. This paper describes an important part of a system, called Postgraphe because it combines with a text generator called Prétexte[2], which generates a report integrating graphics and text from a single set of writer's intentions. The system is given the data in tabular form as might be found in a spreadsheet; also input is a declaration of the types of values in the columns of the table. The user then chooses the intentions to be conveyed in the graphics (e.g. compare two variables, show the evolution of a set of variables ...) and the system generates a report in LaTeXwith the appropriate PostScript graphic files. In this paper, we focus on the graphics generation but the system does not need more information to generate the accompanying text that helps the reader to focus on the important points of the graphics.

2 Writer's intention and the characteristics of graphics

Mackinlay [3] describes an algorithm based on the work of Bertin [1] who characterizes the variables as nominal, ordinal or quantitative. This algorithm specifies graphical methods for each type of variables. We adapted it by integrating some more theoretical results from Tufte [5, 6] and Zelazny [7].

These works classify properties of graphics at the global level and at the level of individual components. The former level relate upon the efficiency of a type of graphic for a kind of data e.g. spatial position are more useful than color for continuous variables; the latter level determines the role of the graphics in the transmission of the message, e.g. horizontal bar charts are useful for comparing values but vertical bar charts and curves are more appropriate for temporal data. We extended these findings to a whole set of graphics discussed in the full paper.

While research has most often been on the reader's model [4], we focus instead on the writer's intention and goals which influence the way the graphics are generated and combined with the text. This problem has not been properly dealt with in the litterature.

3 Implementation of Postgraphe

3.1 Input to the system

In addition to the table of numbers to be reported, three annotations have to be added to help the system determine which information is relevant and how to present it. These annotations are the types of variables, the relational keys of the data and a set of directives describing the writer's intentions.

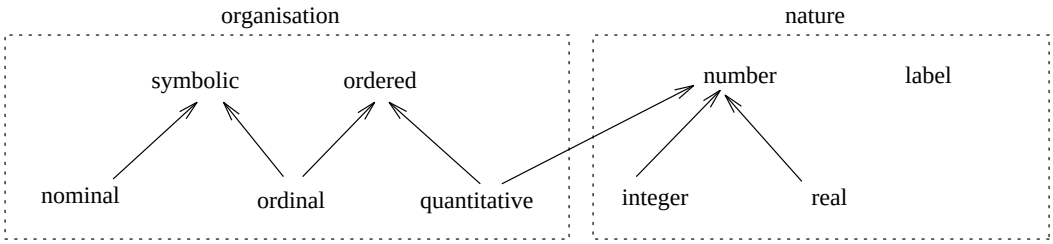


Figure 1: Hierarchies of types

- The types of variables are used to help select the appropriate kind of graphics. They are organized in hierarchies and figure 1 shows the two more important ones (organization and nature). They will be described in more detail in the full paper.
- Relational keys are similar to the notion of the same name in the relational data bases and help determine which variables depend on which other ones. They are also used

for ordering variables in a graphic so that the more important ones are given the more visible positions. Postgraphe computes these keys from the data but the user can also add more constraints.

- Writers' intentions describe what to say and up to a certain point, how to say it. This information is organized in sections that correspond to sections in the report. Each section is a list of intentions that are constraints on the expressivity of the chosen graphics. Postgraphe tries to find the smallest set of graphics that covers the writer's intentions which can be the *presentation* of a variable, the *comparison* of variables or sets of variables, the *evolution* of a variable along another one, the *correlation* of variables and the *distribution* of a variable over another one.

3.2 Planning

From this information, our system determines which graphics can best satisfy the writer's intentions. It builds on the ideas of Mackinlay but extends them in important ways: instead of a list of variables to express, our system starts from a set of intentions to satisfy; in doing so, no a priori ordering of the variables is given; all types of graphics have been assigned a weight for each intention; we can thus build a global quality function to be maximized. But instead of trying all groups of intentions to find the smallest subgroups of variables that best covers the writer's intentions, we use a set of heuristics. This process is divided in four phases that will be described in detail in the full paper:

1. **grouping** to find the intentions that are "compatible" so that each graphic takes into account as many intentions as possible while keeping each one "readable"

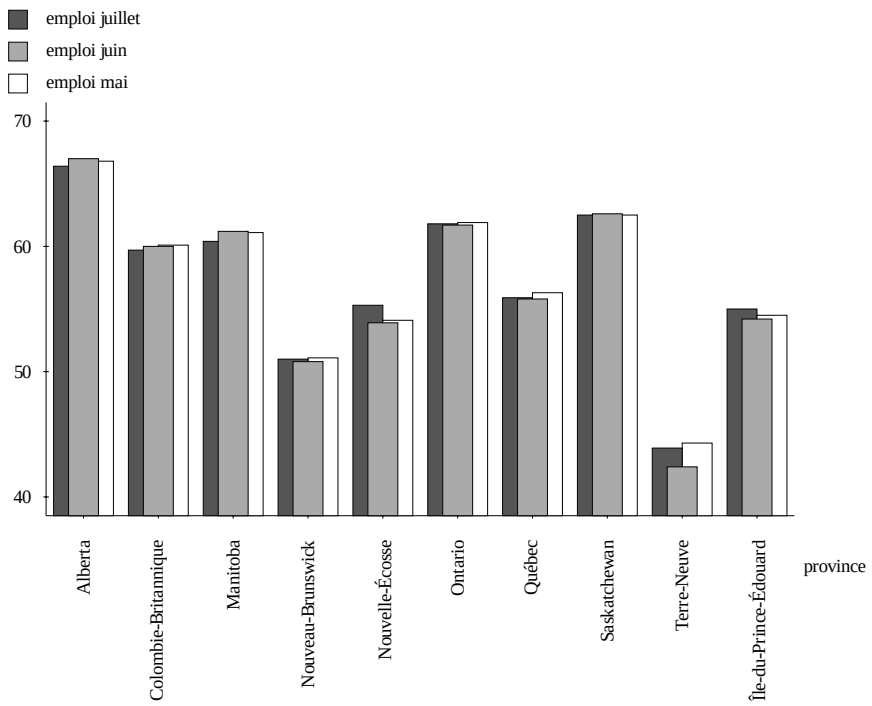
2. **composition** to check if each group is feasible and to determine the best figure to express it

3. **realization** for the low-level generation of graphic primitives; it can be determined at this stage that a figure cannot be generated because of physical reasons: it is too big to fit or not enough grey levels are available. This low level work is quite involved because it has to take into account the 2-D constraints and the limitations of the media. For this we had to develop a Postscript generation system in Prolog in order to determine the exact position of each element (character, line, axis, etc...) of a generated graphic.

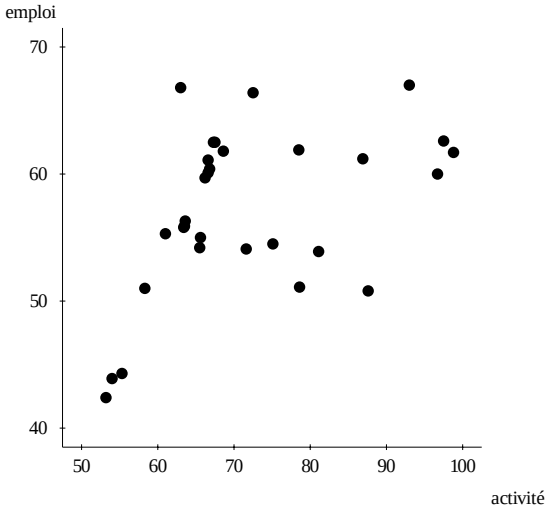
4. **post-optimization** eliminate identical graphics which can occur because the heuristics speed up the system, but in doing so, risk of not finding the optimal solution.

Figure 2 shows a set of graphics that have been generated by Postgraphe from the writer's intentions which are given here as captions[1]. The system is implemented in Prolog and runs on a Unix workstation. It is relatively fast as it takes less than a minute (real time) to deal iwht 13 intentions to generate 7 graphics from a table of 150 values (30 lines by 5 columns).

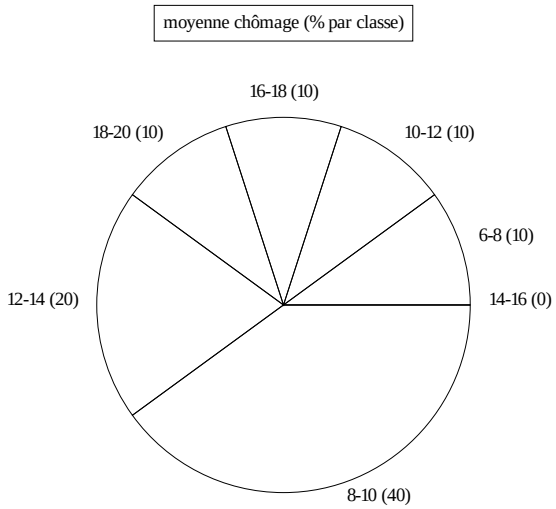
[1]The captions have been manually translated to English from Postgraphe's original French output.



Comparison of employment between provinces and evolution of employment by month.



Correlation of employment and activity ratios.



Fractional distribution of the average unemployment by province.

Figure 2: Example output from Postgraphe

4 Conclusion

Statistical reports make an interesting application for the automatic generation of text and graphics with both media bringing their own contribution. We have described the graphics part of a system that will ultimately generate both from a single set of writer's intentions.

References

[1] Jacques Bertin. *Semiology of Graphics.* The University of Wisconsin Press, 1983. Traduit par William J. Berg.

[2] M. Gagnon. *Expression de la localisation temporelle dans un générateur de texte.* PhD thesis, Département d'informatique et de recherche opérationnelle, Université de Montréal, 1993. Publication 888.

[3] Jock D. Mackinlay. *Automatic Design of Graphical Presentations.* PhD thesis, Computer Science Department, Stanford University, 1986.

[4] C. L. Paris. The role of the user's domain knowledge in generation. *Computational Intelligence*, 7:71–93, 1991.

[5] Edward R. Tufte. *The Visual Display of Quantitative Information.* Graphics Press, 1983.

[6] Edward R. Tufte. *Envisioning Information.* Graphics Press, 1990.

[7] Gene Zelazny. *Dites-le avec des graphiques.* InterÉditions, 1989.