Published in the IASTED International Conference: Computer Graphics and Imaging

August 13-16, 2001 in Honolulu, Hawaii, USA.

VISUALISING THE NON-EXISTENT

CLAIRE KNIGHT AND MALCOLM MUNRO

Visualisation Research Group
Research Institute in Software Evolution
University of Durham
Durham, DH1 3LE, UK.
Tel: +44 191 374 2554, Fax: +44 191 374 2560
{C.R.Knight, Malcolm.Munro}@durham.ac.uk

ABSTRACT

It is hard to see what is not there; but sometimes what is missing is more important than what is present. Visualisation has excelled, as the information visualisation techniques have advanced, as a provider of apparently hidden relationships and patterns in data. It is also important to be able to utilise visualisations for the display of non-existent data. This may sound an odd notion, but the presence (or otherwise) of some data may affect the overall understanding and analysis of a whole data set.

KEY WORDS

Visualisation, Visual Information Systems, Uncertainty

1. INTRODUCTION

Visualisations can often be considered to be of the form overview + detail or focus + context [1], but this doesn't actually address all of the issues of trying to represent large and complex data sets in finite spaces. This is not a new problem, but as the approaches in information visualisation are improving it is necessary to examine the influences that such techniques may have on the understanding and analysis of that data. By obscuring detail in order to be able to use screen real estate for overviews there is a danger that important aspects of the data will be hidden. Since one of the aims of visualisation is to illustrate the less obvious in data sets this is counter productive. That is not to say that overview displays do not have their use. More that the graphics used should be well considered and also integrate effectively with detail and other views so that such problems are minimised.

No visualisation will ever be perfect. Data has to be transformed (at the very least translated in a direct mapping) to the graphics used to create the visualisation but more often than not the graphics then undergo some form of visual transformation; either through the rendering algorithms utilised [2] or through the various other abstractions and views generated. It is also to be remembered that in certain applications (such as visualisations of trends) then blurred distinctions between data items is acceptable. What must, however, be carried out is the acknowledgement of that task dependence, and highlighting why that visualisation technique cannot necessarily be reapplied to any data without thought as to the implications it might have.

Visualisations have in some ways neglected to address all of the issues of trying to represent large and complex data sets in finite spaces. As the approaches in information visualisation are improving it is necessary to examine the influences that such techniques may have on the understanding and analysis of any data. There are several aspects of any data set that may cause problems for visualisations and their associated representations and algorithms. These all relate to the problems of uncertain, missing, or apparently anomalous data and generically can be referred to as problem data. Algorithms and representations, partly through necessity, have to have limits or boundaries of where they are applicable. The real issue with most visualisation is that these boundaries have been set so close to perfect data sets that the appearance of problem data is likely to mean that the visualisations cannot successfully portray it. It may even be the case that the data is truncated so that it fits, or ignored if the value is not what is expected.

The concept of visualising the non-existent is perhaps part of many information visualisations because of the way in which they convert abstract data into visual representations. The difference between non-existent and

abstract is that essentially not only does the visualisation have to create some concrete form for an intangible source but it also has to compensate for missing values.

2. PECULIAR DATA

Data is a problem when it becomes too unwieldy to be converted into information. As any high-school computer studies student knows, data + meaning = information. Unfortunately the process of assigning meaning can become troublesome when the amount and/or complexity of the data exceeds that which can be conceptualised at any one time by the human who needs the information. This would not be a problem if every set of data encountered was *perfect* in that it was all there in a standard format, correctly typed, fitted within given statistical ranges, and so on. In reality the chances of this happening are remote. Whilst some data can be aggregated and analysed through taking the general case (and ignoring the data that falls outside of this) it is generally only applicable when one is interested in overall trends and the likelihood (given the past data) of something happening again. When all of the data is potentially useful, the issue of problem data becomes very important [2, 3].

Data that falls outside the normal ranges is often considered to be in error. This is especially true where data mining techniques are involved, because one of the earliest steps of that process is data cleansing. The problem comes when the analysis is actually concerned with locating not only trends, but also apparent anomalies, or extremes of data. From experience with industrial contacts, showing them what they did not realise was there and also what was not there when it should have been has been one of the major benefits of a visualisation. The problem is that any form of cleansing or smoothing (at a data or graphical level) would not have made this clear.

There is a need to be able to view various classes of *peculiar data*:

- Missing data
    - Reasons for missing data are numerous; population of the data source allows for fields to be left blank; the result of a query or extraction was unable to locate information for certain placeholders. The fact that the data is actually missing is significant, and should not be viewed as a problem with the data or that empty implies valueless.
    - **Example**: a very simple effect of visualising missing data is that it generates different images to that expected. The authors came across just such a situation when carrying out some visualisation work for industry. They expected one picture, especially with their domain knowledge mentally augmenting the process and were not impressed to receive another. When it was pointed out that this was in fact the data that was being worked with, it showed them areas where their dataset was not populated and thus they then created some more input guidelines.

- Incomplete data
    - Incomplete data can stem from streams of data, especially those being processed in (near) real time.
    - **Example**: any data that is streamed and then transformed/visualised in real time will cause this to be a problem whereby the value(s) may fluctuate, and any aggregation is therefore either an approximation or itself subject to variations in value as the input changes. This problem has been experienced when working with collections of data. Up to now there has been a need to wait until the data has been loaded before laying items out, but ideally this would have operated in real time and updated the display as new items were loaded.

- Uncertain data
    - Because there is some form of confidence value associated with the data, or a range of values that the data item may take means that analysis is complicated through the possible paths this data may force the process to go through. Mean values are not necessarily good indicators as they hide the impact of one uncertain value tending to maximum and other to minimums (for example).
    - Uncertain data can also come from sources that may have trust values associated with them. The weighting of different pieces of evidence can be aggregated with some numeric indication of the confidence in them to provide an overall confidence prediction but this gives a limited view, and a particularly low scoring might then bring the total to a low enough level that it is then removed from any further analysis. Ways to avoid this problem should be considered because the one low confidence value itself might be the value in error; the impact of dataset

errors and assumptions in analysis has far reaching implications.
    - **Example**: since the authors have a background in software visualisation and thus the related disciplines of maintenance and program comprehension this often directs the visualisation work we do. This type of data provides good examples of uncertain data (where a range of possible values exist) even after query. One concrete example from visualising Java based on static analysis facts is that of the calling structure. Because of the object oriented nature of the language and thus the dynamic binding it cannot be known for sure until runtime which object and it's associated method gets called. Analysis at a static level can however provide the list of possible alternatives based on the class hierarchy. This means that a traditional call graph is of no use in a situation like this.

- Apparently anomalous data
    - Allows for challenging of assumptions about the data, and an investigation into why it appears anomalous. This process of discovery can lead to a deeper understanding of that part of the data.
    - **Example**: this was a problem that was illustrated through working with industry. Due to the enormity of the data under analysis, compounded by the number of fields that could be populated with certain information (with varying rules as to when and why this should happen) there were various "errors" in the data. This was only illustrated to the managers when they saw the visualisations and questioned why certain things had been drawn how they had. As the authors did not have the necessary domain knowledge to "correct" the data via the algorithm, an often unconscious act, the data was visualised exactly as given and this meant that the problem of mis-population was highlighted.

- Statistically deviant data
    - If a value is outside of a given boundary then it will affect statistical analysis. Often outliers will contain as much information as the (statistically correct) clusters and distributions when looking at data.
    - **Example**: much of the visualisation that the authors have done has centred on that of abstract data with much of it of a textual nature. Since there has been little focused on clustering with calculated values (rather than actual) then this has not been a direct problem. The one incident that occurred with code metrics was that some visual representations were considered to be wrong until the source code was looked at. This is because the metrics accurately reflected the length and complexity of that piece of code even though software engineering standards suggest code should not be this long!

These are important aspects of any visualisation because of the effects that standard transforms and translations may take. Many visualisations (as with data mining algorithms) will only work with near perfect data. Even if extreme values are encountered they may be ignored in the process of going from data to visualisation. This is especially true of overview where statistical trend information can be ruined through the presence of outliers.

The main issue here is that these perceived outliers might contain the most information for those trying to understand and/or analyse the data. It is important that visualisations do not seek to remove or obscure this information. In doing so the resultant analysis will be flawed, and mis-information is usually worse than no information. As Tufte [4] highlights, graphics and visualisation can be easily used for the process of dis-information.

3. TRANSLATION AND CONVERSION

Obviously the generation of any visualisation requires that the data on which it is based are translated in some way so that the graphical elements that compose and construct the visualisation can be obtained. The concept of abstraction for analysis is not new, and is often used to try and overcome problems of information density. There is also the approach of using levels of abstraction given the level of knowledge required. Either way, the abstractions are a transformation of the original data – a fact lost on some who consider certain representations to be inherently problematic whilst others can be perfectly fine. Whether the data is translated in direct mappings, converted, or transformed in some way is not the most important issue. What is, is that this process needs to be consistent through all of the data and also have some level of integrity for faithfully representing the data.

Visualisations are prone to suffering two levels of transformation; translation of the data to get the visual representations, and then these visual representations may themselves be translated or manipulated. There is also the possibility of multiple concurrent transformations in order to support the generation of various visualisations and views from the same data. Indeed Roberts [5] presents variations on the standard pipeline of data-filter-map-display of visualisations to incorporate multiple filters and also multiple mappings.

Despite the possible errors that these multiple, concatenated, transformations may cause they are inherently necessary if any visualisation is to be generated. Because of the concatenation there is also the problem of then having cascading errors through each level of transform. The aim, therefore, is to be able to minimise any possible error. This requires that not only are the visualisation graphics a major effort of the process of visualising data but also the various filtering and mapping operations that are likely for any given data set.

4. NON-EXISTENCE; EXAMINATION AND CHALLENGES

Visualisations are effective ways of dealing with the flood of information. In many situations they are useful for aggregating large and complex data sets into one-screen overviews. Visualisations can of course always resort to the original data set, but visualisations enable the user who is analysing and/or in the processes of understanding the data to discover the middle steps for themselves. There is less of a black-box type view whereby data is presented to some "magic thing" that then produces a very simple and concise answer. The abstractions used at these middle levels can also provide information that would otherwise be missed about possible groupings. In this way the visualisation aids the knowledge gathering aspect of the analysis.

Most visualisations these days have some degree of interactive control. This enables the analyser to change bounds (for example) and see the impact that this has on the rest of the dataset. In doing this type of activity, and in being able to ask "what if…" the knowledge process is enhanced because instead of just fact, the reason why is also evident. This provides stronger mental links for the analyser.

Visualisation is also good for combining multiple data sources to create several highly related views [1]. This augments the information that is available from only one of those sources and makes the analysis and cross referencing tasks easier than if done by hand after putting each data source through individual analysis algorithms. This also supports the insight process in the same way "what if…" interaction does. This extra visual capability may also reveal hidden trends that would not have been made evident from separate analysis; even after recombination at the end. Visualisation is good for handling multi-dimensional data and multi-dimensional analysis. Many tasks, would otherwise be impossible given the high dimensionality involved.

Tufte [6] stressed the need to maximise the data displayed in any given graphic. He also emphasised the need to minimise the *lie factor* of any given graphic. In the presentations provided in his book, with the examination centering mainly on the presentation of the results of statistical analysis, this was well illustrated. Tufte's original directives are:

> *"Maximise the data-ink ration, within reason. Erase non-data-ink, within reason. [6]"*

Certainly these guidelines have great benefit and applicability in many information visualisations. Unfortunately by following these guidelines strongly the concept of non-existence of data is likely to be overlooked or ignored.

In order to create visualisations that do not have high dis-information [4] or lie factor [6] values then it is necessary to be faithful to the original data. This means that to do so there is a need to either represent *nothing* in some way, or to refine the guidelines presented by Tufte to account for the variance likely to be encountered in the type of abstract data set often encountered by information visualisations. Wittenbrink et al. [2] have investigated the concept of using uncertainty glyphs as a way of illustrating that the data values are not single values, or to attach different levels of soundness to the data source providing those glyphs. This tends toward providing representations for nothing, or in actual fact, the provision of representations for possibly anomalous data, given the categories provided in Section 2 of this paper.

It is also important that visual distinctions are made depending on the type of non-existence of standard data. The difference between the categories may not seem much when they are listed, but the actual category and also the reason for that non-existence could provide vital clues when performing analysis and understanding the data. There is a need to be able to be fully informed, or as fully informed as is possible given the huge sizes of data set utilised today, when carrying out analysis. Mis-information leading to flawed analysis, depending on the application, can have possibly huge consequences. This negates any effect of having visualisations as an aid to this process, and also negates the promised benefits of any such tool. In such situations examining the raw data or performing standard queries may produce more effective results; in fact humans are drawn to items that stand out

therefore it is entirely feasible that the analyser would be drawn directly to the *non-existence* in the data set!

Visualisation aims to help this process and to amplify human cognition. Indeed, it can even be used to aggregate and abstract. Therefore it follows that it must not only make clear the hidden trends and relationships, it must also make clear the real data. This includes any data that isn't actually present, data that appears not to conform to identified ranges for that type, and also data that has a variance of range and/or validity associated with it.

5. CONCLUSIONS

This paper has presented the problems of non-existence in data sets to be visualised. Various types of non-existence or variability have been highlighted, and the issues of actually visualising just this type of data presented. The techniques smoothing of values to get statistical trends or the cleaning of data to enable data mining algorithms to produce the most likely result for a maximum percentage of the data is not enough when visualising data. The beauty and power of visualisations is that they provide a representation of the real data. There may be use of aggregation, abstraction, selection, and so on, to provide more meaningful views, but ultimately all of the data is important in forming the visualisations. The use of multiple views and/or multiple representations also provides another dimension for uncovering information and knowledge from the data. This process would become flawed if the visualisation techniques were to discount peculiar data. The scaled transformation of data whereby perceived anomalies are reduced to bring them into line with the rest of the data is even worse than disregarding such data.

It is not claimed that creating visualisations to cope with peculiar or non-existence is an easy problem to solve. Visualisations and representations are hard to design in the first instance, requiring the use and integration of graphic design, task and data understanding, visualisation techniques, and computational knowledge. It is not an easy task to create augmentation and amplification aids that have to deal with a theoretical range of zero to infinity. To then have to deal with a data item that has properties that are oblique to the rest of the data, or has nothing as a value, is an added level of complexity. Despite this, a well-designed and engineered visualisation can provide so many benefits that this is an area that is worth addressing.

ACKNOWLEDGEMENTS

This work has been supported by the EPSRC project VVSRE; **V**isualising **S**oftware in **V**irtual **R**eality **E**nvironments.

REFERENCES

[1] S. Card, J. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, February 1999.

[2] C. M. Wittenbrink, A. T. Pang, and S. K. Lodha, Glyphs for Visualizing Uncertainty in Vector Fields, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 2, No. 3, pp266-279, 1996.

[3] C. M. Wittenbrink, A. T. Pang, and S. Lodha, Verity Visualization: Visual Mappings, *Computer Engineering & Information Sciences Technical Report*, University of California, Santa Cruz, 1995.

[4] E. R. Tufte, *Visual Explanations: Images and Quantities*, Evidence and Narrative, Graphics Press, February 1997.

[5] J. C. Roberts, Multiple-View and Multiform Visualization, *Visual Data Exploration and Analysis VII, San Jose, USA, in Proceedings of SPIE*, Vo. 3960, pp176-185, January 2000.

[6] E. R. Tufte, *The Visual Display of Quantitative Information*, Graphics Press, February 1992 Reprint.