Creating Visualizations: A Case-Based Reasoning Perspective

Jill Freyne[1] and Barry Smyth[2⋆]

[1] CSIRO Tasmanian ICT Center
GPO Box 1538, Hobart, 7001, Australia
jill.freyne@csiro.au
[2] CLARITY: Centre for Sensor Web Technologies
School of Computer Science and Informatics
University College Dublin,
Dublin, Ireland.
barry.smyth@ucd.ie

Abstract. Visualization is among the most powerful of data analysis techniques and is readily available in standalone systems or components of everyday software packages. In recent years much work has been done to design and develop visualization systems with reduced entry and usage barriers in order to make visualization available to the masses. Here we describe a novel application of case-based reasoning techniques to help users visualize complex datasets. We exploit an online visualization service, Many Eyes and explore how case based representation of datasets including simple features such as size and content types can produce recommendations of visualization types to assist novice users in the selection of appropriate visualizations.

1 Introduction

Manipulating complex data is now a familiar part of our everyday lives. and to help us there are a wide range of data analysis tools, from general purpose spreadsheets to more complex statistical analysis packages. Visualization is among the most powerful of data analysis techniques and is readily available either as standalone systems or as key components of common software packages such as spreadsheets. Great strides have been made in bringing a wide range of visualization options to the masses. For example, Microsoft's Excel offers 11 different types of chart (bar, line, pie etc.) and a total of 73 basic variations on these charts. Apple's Numbers spreadsheet is similarly well equipped and even Google's free Spreadsheets programme offers access to about 25 different variations of 6 different chart types.

Surely all of this puts sophisticated visualization within reach of the average user? The problem, of course, is that the average user is not a visualization expert

⋆ This work was supported by the Australian Government through the Intelligent Island Program, CSIRO and Science Foundation Ireland through Grant No. 07/CE/I1147.

and producing the right sort of visualization for a given dataset is far from trivial. Previous work in the area of visualization recommendation includes research into articulated task-orientated systems [4], early data property based systems [9, 8], hybrid task and data based systems which examine both user intent and the data at hand [11, 2] and more recent work which aims to discover patterns in user behaviour in preparation of a dataset in order to predict visualization requirements [6]. This work returns to the early data property based research as we exploit case-based reasoning techniques to make visualization recommendations. We believe that case-based reasoning [1] is very well suited to providing useful assistance in this type of task and in this paper we describe a case-based recommender system that is designed to do just this.

The starting point for this work is a Web based "social" visualization platform called *Many Eyes* that was created by the Visual Communication Lab in IBM Research's Collaborative User Experience group [10]. In brief, Many Eyes is a web-based visualization platform that allows users to upload datasets, chose from a wide variety of visualizations, and make the results available to others. To date over 33,000 datasets have been uploaded by nearly 8,000 users, creating 24,000 different visualizations. These "visualization experiences" encode important visualization knowledge in terms of the decisions taken by a user about how to visually represent a given dataset. In this way each visualization can be viewed as a *case*, with features of the dataset providing the *case specification* and the resulting visualization configuration providing the *case solution*. In this paper we propose that these visualization cases can be reused in the context of a new dataset, to make suggestions about appropriate visualizations.

2 Many Eyes

Many Eyes (http://manyeyes.alphaworks.ibm.com)is an online browser based visualization tool designed specifically to make sophisticated visualization easily accessible to web users but also to make the process of visualization a social one, where people can come together to discover and share what they see in publicly visualized data [10]. Many Eyes differs from other visualization software in its privacy constraints. All human contributed data is visible to the public, all datasets, visualizations and comments are publicly accessible. As Many Eyes is an experimental system the visualization options vary from the ordinary (histograms and pie charts) to experimental (word trees and matrix charts) and users have little assistance other than small graphics and a short textual description when choosing a visualization for their dataset.

Many Eyes has three core processes, data upload, visualization creation and social discovery and discussion. Due to the system's open access policy each process can be undertaken independently, for example any user can create a visualization on any data set contributed or comment on any data set or visualization created. This platform creates an ideal online environment for collaboration, cooperation and communication around a set of data and its visualizations.

2.1 Data Upload

Raw data uploaded to ManyEyes an be freeform text or tab-delimited data. In an effort to keep entry barriers to using Many Eyes low the system has the capability to recognise and process tab-delimited data, allowing users to copy data directly from Microsoft Excel or Open Office. Many Eyes carries out initial analysis on all tabular data at upload time and makes assumptions as to the type of data, textual or numeric, contained in each column. All datasets are accompanied by metadata. Uploaders are required to provide textual information such as a title for the dataset and encouraged to provide other relevant information such as the source and description of the data. The system appends further metadata including the creation date and creator's details before making it public.

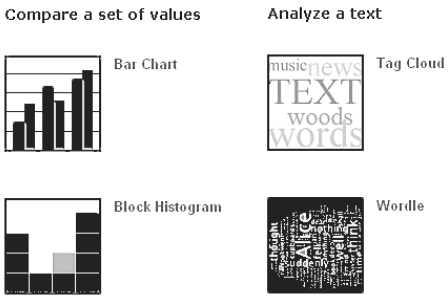

Fig. 1. Selection of category and visualization types in Many Eyes

2.2 Visualization Creation

Many Eyes has 6 categories of visualizations, containing a total of 16 visualization types some of which can be further sub-categorized. Maps for example can be broken down into country specific maps. Users are shown a predefined list of visualization categories and types (see Fig. 1). Sample category titles include "track rises and falls over time", "analyze text" and "seeing the world" amongst others. Each subcategory or chart type in the option list is accompanied by 1-2 explanatory sentences to guide the user in their decisions. Further information relating to each chart type describing its strengths and weaknesses and its appropriateness for varying data types is available but users must navigate away from their current process in order to locate this information [5]. Understandably not all of the visualization types in Many Eyes are suitable for displaying both unstructured text and tabular data. Six of the 33 visualization types have been used for text data visualization and 31 of the visualization types have been used to chart tabular data. On selection of a chart type Many Eyes automatically generates a visualization, assigning chart parameters etc when only one suitable option is available and asking for user conformation when multiple options exist.

2.3 Sharing & Discovery

Many Eyes was designed to enable a new kind of social data analysis. Its development team believe that visualization is a catalyst for discussion and collective insight about data and as such they designed Many Eyes as a collaborative visualization tool, providing users with a platform for discovery, sharing and discussion around people, datasets and visualizations. Each registered member has a public profile page which contains personal details, watchlists, topic hubs and details of activity on the site. Users may communicate directly with each other via profile pages or communicate around a specific dataset or visualization.

In the context of the"knowledge worker", the availability of datasets and associated visualizations provides a rich environment from which non expert visualizers can learn. Novice or inexperienced users may discover datasets similar to theirs in order to decide how to effectively uncover the messages contained in their raw data. Many Eyes provides various methods for browsing and searching its repository of data and visualization pairs. We believe that case-based reasoning techniques could automate the process of discovering suitable visualizations for contributed datasets. By creating cases which represent simple dataset features such as the presence of numeric and textual content as well as the size of the dataset we aim to capture the expertise demonstrated by expert visualizers to assist users in selecting the best chart for their data.

3 The Many Eyes Dataset

The dataset used for this work represents approximately 21 months of usage of Many Eyes from January 2007 and covers 33,656 separate dataset uploads and 24,166 unique visualizations from 15,888 registered users. It is worth noting that only about 43% of uploaded datasets are actually successfully visualized. In other words 19,111 datasets are not associated with any visualizations. In turn, just over 60% of users who uploaded datasets went on to store a visualization. This is surely a telling comment on the challenges faced by users when it comes to choosing and configuring suitable visualizations of their data. It seems that in many cases users just did not have the visualization experience (or the time) to select from the many different charting options and configurations that are offered. In general there are two basic types of dataset in Many Eyes. *Text* datasets are a *bag-of-word* type datasets whereas *tabular* datasets are the more traditional column-based datasets, using a mixture of data types. Text datasets can be visualized using a limited set of specialized visualizations (e.g., matrix charts, tree-maps, or tag-clouds). In total 4702 text datasets were visualized resulting in 7090 visualizations (1.5 per dataset). The visualization of 9880 tabular datasets resulted in 16,848 different visualizations (1.7 per dataset). There are 16 core Many Eyes visualization options that are well suited for tabular data.

4 A Case-Based Recommender for Many Eyes

The Many Eyes repository of datasets and visualizations is more than a simple collection of raw dataset and charts. It is reasonable to assume that each combi-

nation of dataset and chart is the result of a deliberate visualization exercise. As such it encodes some latent decision making process by which the dataset 'owner' came to settle on a particular visualization option which addressed his/her particular objectives. Of course such objectives may extend beyond the simple need to visually summarise a particular dataset. In many cases it is reasonable to assume, for example, that the user will have considered the aesthetics of particular visualization choices, adding an extra dimension to their decision making.

In short then, the combination of dataset and visualization encodes an *experience*. It is a *case* in the classical view of case-based reasoning. And in this paper we propose to take advantage of this perspective in order to develop a case-based recommender system that is capable of suggesting good visualizations to users based on the characteristics of their particular dataset. This will be of particular interest and benefit to less experienced Many Eyes users, who, in the past, have failed to produce visualizations for their datasets. Of course the recommendations may also be of interest to more experienced users by highlighting alternative visualization options that they may be less familiar with.

To this end we propose to augment the existing Many Eyes system with a CBR component. The role of this component is as follows. When a new dataset is selected the CBR system converts the dataset into a suitable set of features and uses these features to find a set of similar cases from the visualization case base. The visualizations associated with these cases are ranked and returned to the user as a set of recommendations. In the following sections we will summarise the case representation, retrieval, and ranking techniques that are used.

4.1 Case Representation

We will begin by assuming each case represents a single visualization of a single dataset. Thus, each case, c_i is made up of a dataset component, d_i and a visualization component, v_i as shown in Equation 1. In fact there is also additional information that is also sometimes available such as the rating associated with a particular visualization, r_i. In case-based reasoning parlance the dataset component corresponds to the *specification* part of a case, the visualization component corresponds to the *solution* part of a case, and the rating component can be viewed as the *outcome* of the solution. In this paper we will focus on the specification and solution side of visualizations cases, largely because the Many Eyes dataset is very sparse when it comes to the availability of ratings data.

$$c_i = \{d_i, v_i\} \tag{1}$$

The representation of the visualization component is straightforward, at least for this paper, since each case solution is just the type of visualization used, $chart(v_i)$, because we are focusing at the moment on recommending a particular visualization type when faced with a new dataset. Going forward, one can envisage more complex solution features if we wish to reason about particular features of the visualization, such as the axis placement, label usage etc.

Each dataset is characterised by a set of simple features that relate to the type of data contained in the dataset. We distinguish between text and tabular datasets by extracting different features for each. For example, for text datasets we extract features that include the total number of terms (*terms*), the number of unique terms *unique*, and the terms themselves can be used as part of the specification $(t_1, ..., t_{terms})$; see Eq. 2. For tabular datasets we can extract features such as the number of textual columns, col_{txt}, the number of numeric columns, col_{num}, the number of data points (rows), *rows* and a bag-of-words textual description derived from any metadata associated with the dataset, *desc* (e.g., column headings, title etc). In the case of numeric columns we also extract features that reflect the maximum, minimum, average, and standard deviations of the column $(min_i, max_i, avg_i, dev_i)$ and for string columns we extract the number of unique strings, $unique_i$. In this way each case is represented as a feature-based dataset and solution as in Eq. 3.

$$c_i = \{terms, unique, t_1, ..., t_{terms}\}, chart(v_i) \tag{2}$$

$$\begin{aligned} c_i = \{col_{txt_1}, col_{num_1} rows, desc, type, (min_1, max_1, avg_1, dev_1 | unique_1) \\ , ..., type_n, (min_n, max_n, avg_n, dev_n | unique_n), chart(v_i) \end{aligned} \tag{3}$$

4.2 Similarity and Retrieval

Given a new target case c_T (made up of a particular dataset) the task of the recommender system is to locate a set of similar cases that can be used as a source of visualizations. For the purpose of this paper we concentrate on some tried and tested similarity techniques using simplified versions of the above case representations. For example, to compute the similarity between tabular dataset cases we use the similarity metric shown in Eq. 4 which simply calculates the relative difference between the number of textual and numeric columns and rows between the target dataset and the case dataset; in this instance uniform weighting is used and so $w_f = 0.33$.

$$sim(c_T, c_i) = 1 - \sum_{f \epsilon \{col_{txt}, col_{num}, rows\}} w_f \bullet \frac{|c_T(f) - c_i(f)|}{max(c_T(f), c_i(f))} \tag{4}$$

A similar approach to similarity assessment is used for the text based dataset, by comparing the datasets by the total number of terms and total unique terms. While these similarity techniques are extremely simple, they provide a useful starting point for this work. In the evaluation section we will demonstrate that even these simple techniques work well when it comes to driving high quality recommendations, while at the same time leaving a number of options open for more sophisticated similarity techniques as part of future work. Thus, given a target case c_T we can use the above similarity techniques to produce a ranked list of n similar cases as the basis for recommendation.

4.3 Generating Recommendations

Each of the n cases retrieved will be associated with a single visualization. The same visualization type may occur in more than one case and so we can identify a set of k different visualization types from these n cases. We need a way to rank these visualizations so that those that are associated with more similar cases are preferred over those that are associated with fewer, less similar cases. To achieve this Eq. 5 scores each of the n visualizations, v_i, as the sum of the similarity scores associated with the retrieved parent cases; $chart(v_i, c_j) = 1$ if v_i is the chart used in c_j and is 0 otherwise. The result is a ranked list of visualization recommendations, $v_1, ..., v_k$ in descending order of their aggregate similarity scores as per Eq. 5

$$score(v_i, c_T, c_1, ..., c_n) = \sum_{\forall j=1...n} sim(c_T, c_j) \bullet chart(v_i, c_j) \tag{5}$$

5 Evaluation

This work is motivated by the fact that less than half (43%) of the datasets uploaded to Many Eyes are actually visualized. We believe that this is at least in part due to the confusion of choice that faced the novice first-time uploader. Our hypothesis is that even a simple form of case-based recommendation will help to improve the visualization rate by making proactive suggestions to the user about which visualization technique might best suit their dataset. In this section we will describe the results of a recent large-scale, off-line, leave-one-out style evaluation using the live Many Eyes dataset.

5.1 Set-up

The core Many Eyes dataset was transformed into a set of 22,935 visualization cases covering 14,582 different unique datasets and 33 visualization types. These cases included 6800 text cases and 16135 tabular cases. For the purpose of this evaluation we are interested in understanding the extent to which our simple CBR strategy can produce useful visualizations, compared with a number of benchmark strategies, which differ in terms of how cases are selected or recommendations are produced. The different techniques are summarised as:

1. *CBR* - the basic CBR approach described above is used to produce a ranked list of the top k visualizations from a set of n similar cases
2. *Popular* - this strategy simply recommends the k most popular visualizations (globally) in Many Eyes.
3. *Exact* - this is a hybrid recommender strategy which identifies a pool of similar cases like the CBR approach but only identifies cases which exactly match the target features. It then calculates the popularity of each visualization type for this pool. Specifically in the case of tabular data the set of similar cases identified match the number of text and numeric columns in the target case and in the text based datasets the word count of similar cases is within a close defined range of the word count of the target dataset.

4. *PopularContext* - similar to *Popular* but it treats textual and tabular visualization types as separately
5. *Random* - recommend a set of k random visualizations.

5.2 Methodology

Our evaluation takes the form of a standard leave-one-out test. For each target case, c_T, we use its specification features to represent a new dataset and generate a set of k visualizations using each of 5 recommendation strategies; note that the k is based on the number of unique visualizations retrieved by the CBR strategy. Our first measure of quality looks at how often the target visualization is present in the set of k recommendations; so an *accuracy* of 60% means that the target visualization is present in 60% of the recommendation sets of size k. The second looks at the average position of the target visualization in the recommendation lists. There are two ways to do this. One is to focus on those recommendation lists that do have the correct target visualization and then compute the average position of the target visualization in the final recommendation list. This so-called *average position* approach ignores recommendation lists that do not contain the correct visualization though, and therefore benefits the less accurate strategies. As an alternative we can compute a position value across all recommendation lists by assigning a $k + 1$ penalty to those lists that do not contain the target visualization *adjusted position*. This is a conservative penalty because it assumes that the correct visualization is actually in position $k + 1$, which may not be, but it serves to at least remove some of the bias associated with *average position*.

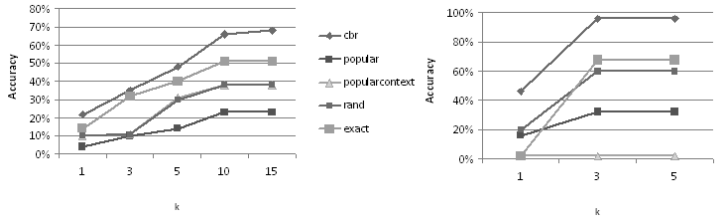

Fig. 2. Accuracy of predicted visualization types (a) tabular and (b) text.

5.3 Results

Recommendation Accuracy Fig. 2(a)-(b) show the accuracy results separately for the textual and tabular cases. These results clearly support the use of the CBR recommendation strategy. Overall CBR is seen to outperform all other techniques with particularly impressive results for the CBR technique in the easier textual case recommendation scenario. There is also a very consistent benefit associated with the similarity-based technique used by CBR compared with the simpler matching used by $Exact$, with the former delivering relative improvements of 25%-50% across a wide range of k values.

Recommendation Position The results of the positional analysis of the recommendation techniques are presented in Fig. 3 (a)-(b). In terms of the average position statistic the local recommendation techniques such as CBR and $Exact$ are delivering improved performance compared to the global benchmarks, although there are a number of anomalies. For example in Fig 3(a), at $k = 3$, we see that CBR delivers its correct recommendations with an average position of 1.5. However, the popularity-based techniques achieve a better average position of just over 1. But remember, at this setting CBR is recommending a correct visualization among its top 3 recommendations more than 20% of the time versus 5% of the time with popularity-based approaches. By introducing a positional penalty we find that the local techniques do consistently better than all other benchmarks; see Fig. 3(c)-(d).

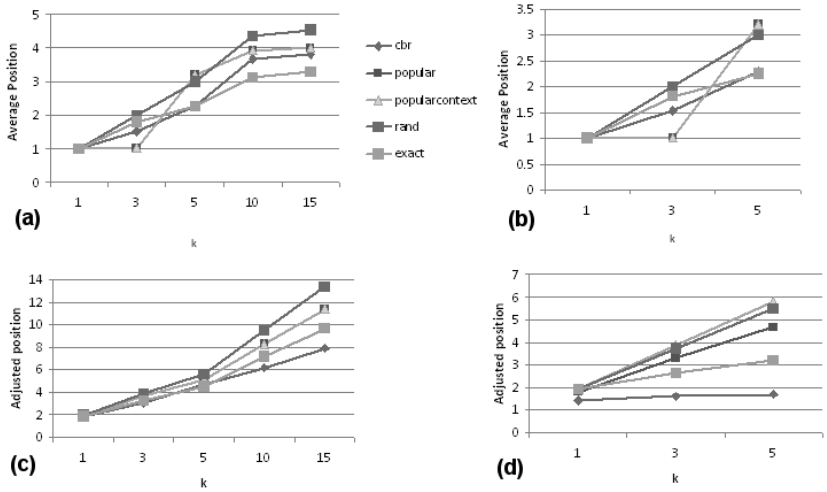

Fig. 3. Average position of the target visualizations (a) tabular, (b) textual and adjusted position of the target visualizations:(c) tabular, (d) textual

5.4 Summary

Even a relatively simple approach to case reuse has delivered useful results which may make a difference to Many Eyes users in practice. In each case we have found the case-based approach to outperform all of the other benchmarks that were tried, consistently producing more accurate recommendations nearer to the top of the recommendation list. Of course these findings need to be validated. They may be based on real-user data but they have not been tested on live users in the field. Nevertheless with these findings we can be optimistic about the prospect of success in such a future trial.

6 Conclusions

The objective is this work is to help users of a Web based visualization system to produce better visualizations by recommending visualizations that have been previously used for datasets that are similar to their own. To that end we have started with a very simple case recommendation technique, but this has performed very well in practice, significantly outperforming a number of benchmarks. However, there remains plenty of room for improvement and as future work a number of obvious next steps present themselves:

1. *More Sophisticated CBR.* Incorporating some notion of semantics into the representation and similarity computation should be possible.
2. *Introducing Adaptation.* Users will benefit greatly from configuration support when it comes to actually using a particular visualization. This includes deciding which fields are associated with which axes, scale settings, etc.
3. *Ratings & Provenance.* The Many Eyes visualization data contains rating information and information about the creator of the particular visualization that could be used to greatly improve the algorithms [3, 7].

References

1. A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39 – 59, 1994.
2. Elisabeth André and Thomas Rist. The design of illustrated documents as a planning task. pages 94–116, 1993.
3. Peter Briggs and Barry Smyth. Provenance, trust, and sharing in peer-to-peer case- based web search. In *ECCBR*, pages 89–103, 2008.
4. Stephen M. Casner. Task-analytic approach to the automated design of graphic presentations. *ACM Trans. Graph.*, 10(2):111–151, 1991.
5. Catalina M. Danis, Fernanda B. Viegas, Martin Wattenberg, and Jesse Kriss. Your place or mine?: visualization as a community component. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 275–284, New York, NY, USA, 2008. ACM.
6. David Gotz and Zhen Wen. Behavior-driven visualization recommendation. In *IUI '09: Proceedings of the 13th international conference on Intelligent user interfaces*, pages 315–324, New York, NY, USA, 2009. ACM.
7. David B. Leake and Matthew Whitehead. Case provenance: The value of remembering case sources. In *ICCBR*, pages 194–208, 2007.
8. Jock Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141, 1986.
9. Steven F. Roth, John Kolojejchick, Joe Mattis, and Jade Goldstein. Interactive graphic design using automatic presentation knowledge. In *CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 112–117, New York, NY, USA, 1994. ACM.
10. Fernanda B. Viégas, Martin Wattenberg, Frank van Ham, Jesse Kriss, and Matt McKeon. Manyeyes: A site for visualization at internet scale. In *IEEE Transactions on Visualization and Computer Graphics*, volume 13(6), pages 1121–1128, 2008.
11. Michelle X. Zhou and Min Chen. Automated generation of graphic sketches by example. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 65–74. Morgan Kaufmann, 2003.