A Case-Based Reasoning Approach to Visualise Conversations

Nikolay Burlutskiy, Miltos Petridis, Andrew Fish and Nour Ali

Abstract At present, the complexity and scale of modern digital conversations between people is at its highest level but there is a gap in how to represent these conversations to a user. As a result, it is often hard for a user to understand the flow of a conversation and make an informed decision over it. However, an aesthetic and efficient visualisation can mitigate this drawback of data representation. In this paper, a case-based approach was proposed for choosing an appropriate visualisation for user's conversations. A case was formulated as a visualisation of a conversa-tion which a user decided to use for his analysis of the conversation. When a user decides to visualise a new conversation, the most similar visualisation type from previous users' experiences is selected for the visualisation of the new conversation. In this paper, the cases of visualisations of conversations from the IBM Many Eyes platform were collected and a case-based reasoning approach for choosing a visualisation of user's conversation was designed. Finally, the work of the proposed approach was tested on a sample email conversation, and then four participants eval-uated the appropriateness of the chosen visualisation types in comparison with other eight possible visualisations for the email conversation.

Key words: Case-Based Reasoning, Conversation Visualisation, Many Eyes

N. Burlutskiy (✉)
The University of Brighton, Brighton, United Kingdom
e-mail: N.Burlutskiy@brighton.ac.uk

M. Petridis
e-mail: M.Petridis@brighton.ac.uk

A. Fish
e-mail: Andrew.Fish@brighton.ac.uk

N. Ali
e-mail: N.Ali2@brighton.ac.uk

1 Introduction

The digital world has become an important media for human communication since the development of the first computers. These days people communicate everyday via e-mail, social networks, chats, forums and instant messengers. However, the complexity and scale of digital communication motivates people to find alternative ways to visually represent conversations which are more efficient for specific reasoning tasks. The interest in visualisation of conversations is represented by research work in the development of different visualisations [18, 15, 11, 3]. In addition to the research work on visualisation, users have been trying to visualise their data with various commercial visualisation tools [8, 10, 17].

The field of information visualisation has seen a rapid growth due to advances in software and hardware development and, as a result, a number of software visualisation tools have appeared on the market. These tools allow users to create a complicated visualisation on the fly by a single user's click. Moreover, a user can interact with visualised data and share it with others [8, 10, 17]. For instance, data in IBM Many Eyes can be uploaded in a format of text or tab-delimited data. The combination of dataset and visualisation encodes an experience which is a case in the classical view of case-based reasoning. In [7] the authors proposed a case-based recommendation system that is capable of suggesting popular visualisations to users based on the characteristics of their datasets and users' preferable visualisations in Many Eyes [10, 7].

Development of the best single visualisation of a conversation between people seems to be an impossible task. However, several visualisations proved to be preferable over other possible visualisations for specific tasks and conversations. For example, visualisation of e-mail communication has been researched for several years [11, 20, 2]. Visualisation of a social network and communication between people as a dynamically changing graph is a common ground for researchers in this field [9]. An attempt to visualise conversations to uncover their social and temporal patterns was done in [15]. In this work, a few visualisations, namely a graph, tree map-like visualisation, and a bar chart were shown to be useful in presenting properties of newsgroups. In [18] the authors proposed a fusion of two visualisations for showing sequence and reply relationships among the messages of a conversation simultaneously. However, a single visualisation of conversations does not suit all users' needs which is one of the main motivations for our work.

The addressed problem in this paper is how to choose a visualisation type for a given human conversation in order to help people to analyse, explore and understand the conversation. The Case-Based Reasoning (CBR) approach was applied for visualisation of conversations such as e-mail, instant messages, and social networks. The cases were defined as experiences of the visualisations of conversations. Next, a similarity measure for choosing a visualisation type for a target conversation was constructed and applied. In this work, a web-based visualisation platform IBM Many Eyes was chosen for collecting historical data of conversations and their visualisations. At the end of this paper, the results of visualisation for a sample email conversation using the proposed CBR approach are demonstrated and evaluated.

2 Case-based Reasoning for Visualisation of Conversations

The classical CBR approach for problem solving has four steps: 1) retrieving one of previously experienced cases, 2) reusing the retrieved case, 3) revising the chosen solution, and 4) retaining the new experience for further usage [1]. Adapting the classical CBR approach for choosing a visualisation for user's conversation requires the definition of what a case is for this particular problem domain.

2.1 Definitions and Problem Statement

In this paper, a case c_i is a pair consisting of a dataset D_i of a conversation and a visualisation V_i of the dataset D_i:

$$c_i = (D_i, V_i) \tag{1}$$

The cases of the visualisation of conversations were collected from the Many Eyes platform and then a set of appropriate cases was formed by filtering the datasets. Then a similarity measure was constructed as a rule for comparison of a new conversation with the previously experienced cases.

To start with, the problem of visualisation for human conversations was formalised. A graph-based model of conversations defined below was proposed:

Definition 1. Let A be a set of actors and M the set of messages involved in a conversation C. Each message $m \in M$ has a source, $s(m)$ and a target, $t(m)$, actor in A. The conversation graph $G(C)$ is the directed multigraph $G = (V, E)$ with $V = A$ and $E = M$.

Problem Statement. Find an appropriate visualisation type V_i for user's conversation C using knowledge of chosen visualisation types for previously visualised conversations.

The choice of an appropriate visualisation type for a human conversation was done using the CBR approach. Prior to applying the CBR paradigm for visualisation of conversations, the most common problems people encounter when exploring a conversation were analysed. For this purpose, the literature on visualisation of conversations was studied and then the Many Eyes platform was used for understanding what kind of problems people are trying to solve by visualising the conversations [21, 4, 6, 13, 19, 5]. The list of these problems is below:

- *Finding a sequence of messages:* each message is considered to be an event of an atomic communication between people. To explore the sequence, a multi-modal visualisation can enhance user's experience in exploration of his conversation.
- *Discovering patterns in communication:* the user can be directed toward an answer regarding the way in which he has communicated with people in a particular conversation by providing him with visualisations where cyclic or sequential patterns of communication are represented explicitly.

- *Aggregate the information on a conversation:* if user's conversation become very long then he might want to see only an aggregated view of his conversation.
- *Understanding the granularity of time:* the user can look at his conversations at different time granularity. For example, the user may want to see the intensity of his day-to-day conversations or look at his month-to-month messages. By looking at different granularity, the user can have an insight into an aggregated picture of his communication rather than being lost in his daily messages.
- *Supporting temporal questions:* the problem of finding a particular message or a conversation of interest is common when exploring user's conversations. The visualisation must allow a user to answer his temporal queries easily.

The goal of visualisations of conversations between people is to get a deeper insight of the nature and properties of communication. A well-chosen visualisation of conversations can help people to find answers for their questions.

As a source for the cases, visualisations of different conversations available on the IBM Many Eyes website were chosen.

2.2 Many Eyes as a Source for Visualisation Cases

IBM Many Eyes is a visualisation platform where a user can upload his dataset in a format of comma-separated values and then he can choose a visualisation. The platform provides twenty one different visualisations including bar charts, bubble charts, tree map, graph, scatter plots, etc. A user can choose a visualisation for his dataset and then he can manipulate his visualisation. For instance, he can flip axes, zoom in or zoom out, and interact with the visualisation.

On the 20th of October, 2014 there were 490,982 datasets and only 186,314 visualisations. This counts for less than 38% of all the uploaded datasets. In this work, only the visualisations of conversations are of interest. Thus, the datasets were filtered using keywords such as 'conversation', 'email', 'chat', 'twitter', 'facebook'. In total, there were 5819 datasets and only 2322 visualisations for this set of search words. Then meaningless and non-related datasets and visualisations were eliminated. For this purpose, all 2322 visualisations were analysed in order to map the datasets to the proposed model of a conversation through recording the number of actors and messages for each conversation. The number of actors and events in a case were defined as follows:

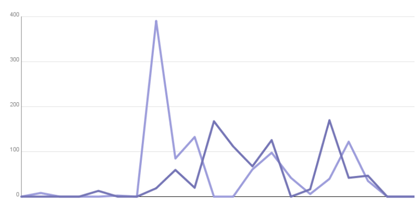
- *Number of actors:* $|V|$ the number of people involved in a conversation which is the number of nodes $|V|$ of the graph $|G|$.
- *Number of events:* $|E|$ the number of messages sent and received by the actors. For example, if an actor sent a message to another actor and received a reply, the number of events will be two.

Unfortunately, it seems very hard to automatize the process of matching the datasets to the model of a conversation since the uploaded datasets were very diverse and in many cases they were obfuscated for more anonymity. However, 40 cases were classified for the preliminary study in this paper (Table 1).

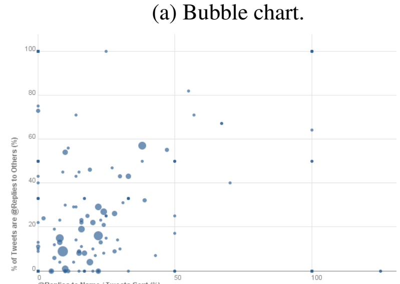
Examples of such visualisations of conversations are shown in Figure 1. In Figure 1a there is a bubble chart visualisation of a conversation between 18 actors, who are represented as circles. The size of the circles is proportional to the number of messages sent by each actor. Figure 1b represents a line graph where the number of email messages received by one actor over a fixed time is shown. In Figure 1c the scatter plot shows actors as circles with their sizes proportional to the number of messages sent to each of them. Figure 1d depicts the most common words in communication between two people over time.

All the 40 chosen conversations from Many Eyes were visualised by eight different types of visualisations such as Bar Chart, Matrix Chart, Line Graph, Bubble Chart, Treemap, Word Cloud, Block Histogram, and Network Diagram. The proposed CBR approach chooses one of these visualisation types for an input dataset.

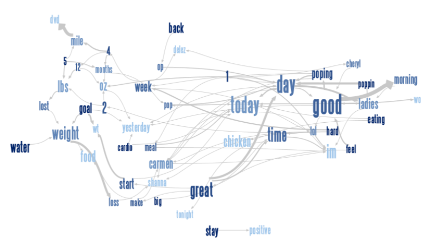
(a) Bubble chart.



(b) Line graph.



(c) Scatter plot.



(d) Graph-like visualisation.

Fig. 1: Examples of visualisations of conversations in Many Eyes.

Table 1: Conversation datasets and their visualisations.

Keywords	Datasets	Visualisations	Chosen as Cases
Conversation	459	71	11
Email	385	112	2
Chat	650	112	10
Twitter	1060	707	15
Facebook	3265	1320	2
Total	5819	2322	40

2.3 Choosing a Visualisation Type

Each dataset D_i of the collected cases was visualised in the past using a visualisation type V_i. In order to choose a visualisation type for an input dataset D, it is compared to all case datasets D_i using a similarity measure (Subsection 2.4). Then the visualisation type of the closest dataset is chosen to visualise the input dataset. It is important to mention that the choice of a visualisation for a particular dataset depends on the user's task. In this paper, it was assumed that the user's task is one of the following: finding a sequence of messages, discovering patterns in communication, aggregating the information on a conversation, understanding the granularity of time, and performing time filtering queries. For each of these tasks there is an appropriate visualisation type V_i which depends on the properties of the conversation such as the number of actors $|V|$ and the number of messages $|E|$ in the conversation.

The visualisation of the user's conversation must support interaction with the data behind his visualisation together with flexibility to change the visualisation for the user's needs. Many Eyes supports these requirements.

When a user is recommended a visualisation by the CBR approach, he can decide to use another visualisation for his conversation. In this scenario, a system built on the base of the proposed CBR algorithm can utilize that information for learning from the users' behaviour.

The area of visualisation of conversations and human communication has been researched thoroughly. A very brief list of conversation visualisations with their evaluations can be found in [21, 4, 6, 13, 19, 5].

2.4 Similarity Measure

In order to decide which visualisation types are the most appropriate to use for visualising the user's conversation, k-Nearest Neighbours (k-NN) algorithm was used. The idea of this algorithm is to choose k neighbours for the input dataset as the most similar to the input dataset. For this work, three ($k = 3$) neighbouring datasets were chosen and their types of visualisation were used for visualising user's dataset. For choosing a neighbouring dataset of the input dataset, a distance metrics between the input dataset and the datasets collected from Many Eyes was defined. This metric was decided to be an Euclidean distance:

$$D_j = \sqrt{\omega(x - x_j)^2} \tag{2}$$

where D_j is an Euclidean distance between a vector of an input dataset $x = (|V|, |E|)^T$ and $x_j = (|V_j|, |E_j|)^T$ is a vector of a case dataset collected from Many Eyes, $j \leq 40$ where 40 corresponds to the number of cases. The vectors x and x_j have a number of actors $|V|$ and events $|E|$ as their elements. The vector $\omega = (\omega_1, \omega_2)$ represents the importance of the features, namely the number of actors $|V|$ and the number of events $|E|$, for comparing two datasets. The weights in the vector ω are represented as numerical values between 0 and 1. In this work, the weights with equal stress on importance for both the number of actors $|V|$ and messages $|E|$ were assigned. Thus $\omega_1 = 0.5$ and $\omega_2 = 0.5$. The result of the k-NN algorithm was a choice of the closest visualisation types in the described metrics for an input dataset based on previous visualisations collected from the Many Eyes platform. Since the k-NN algorithm with k=3 was applied, the user was given the three closest visualisation types for his conversation.

In the next section, the application of the algorithm for a sample email conversation is shown.

2.4.1 An Example for Visualisation: An Email Conversation

In this section, the work of the proposed CBR approach for a sample e-mail conversation is shown. The conversation had four actors $V(G_1) = \{A, B, C, D\}$ and it had eighteen messages sent between the actors (Figure 2). The order of this directed multigraph of the conversation $|V(G_1)|$ is 4 and the size $|E(G_1)|$ equals to 18. Thus, the dataset has the following parameters: $(|V|, |E|)^T = (4, 18)^T$

First, the distance between the sample dataset and each of the 40 case datasets

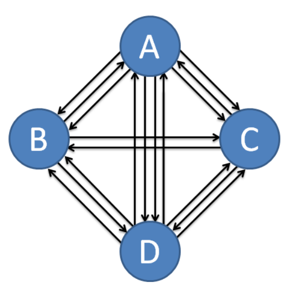


Fig. 2: A multigraph representing the email conversation between four actors.

was calculated, it resulted in 40 values. Since k=3 for the k-NN algorithm, the three nearest neighbours for this dataset were chosen by selecting the three smallest numbers from these 40 values calculated in accordance with the formulae 2.

$$D_1 = \sqrt{0.5(4-1)^2 + 0.5(18-18)^2} = 2.1$$
$$D_2 = \sqrt{0.5(4-1)^2 + 0.5(18-11)^2} = 5.4$$
$$D_3 = \sqrt{0.5(4-1)^2 + 0.5(18-10)^2} = 6.0$$

Thus, the distances $\{D_1, D_2, D_3\}$ for the three nearest neighbours were $\{2.1, 5.4, 6.0\}$. The visualisation types for these datasets chosen in the past were Line Graph, Bubble Chart, and Matrix Chart accordingly. Ideally, the three neighbouring datasets would have had the same visualisation type. However, it seems that for the datasets similar to the sample dataset users decided that these three visualisation types are the most appropriate. In order to understand why the proposed CBR approach gave these visualisations and how appropriate these visualisations are in facilitating the user's comprehension of his conversation visualisations, an evaluation described in the next section was designed and conducted.

3 Evaluation

The goal of the designed pilot evaluation was 1) to evaluate the choice of conversation visualisation from eight different visualisations and 2) to compare the manual selection of a visualisation to the choice made by the CBR algorithm described earlier in the paper. As the main part of the evaluation, four participants were asked to answer seven questions for each of the eight alternative visualisations of the sample email conversation described in Section 2.4.1.

At the beginning of an interview, each participant was introduced to the eight basic different types of visualisations used in Many Eyes: Bar Chart, Matrix Chart, Line Graph, Bubble Chart, Tree Map, Word Cloud, Block Histogram, and Network Diagram. Then the participant received an explanation for the meaning of graphical elements in each of the visualisations. A simple example of a visualisation for a conversation between two people was given in order to confirm that the participant understood the layout of all eight visualisation types. Finally, the eight visualisations designed using IBM Many Eyes for the sample email conversation (Section 2.4.1) were given to the participant (Figure 3) and the following seven questions were asked:

- How many people participated in the conversation?
- How many messages were sent?
- Who initiated the conversation?
- Who sent the second message?
- Who sent the third message?

- How many time steps there were in the conversation?
- How many messages did the person A sent to the person B?

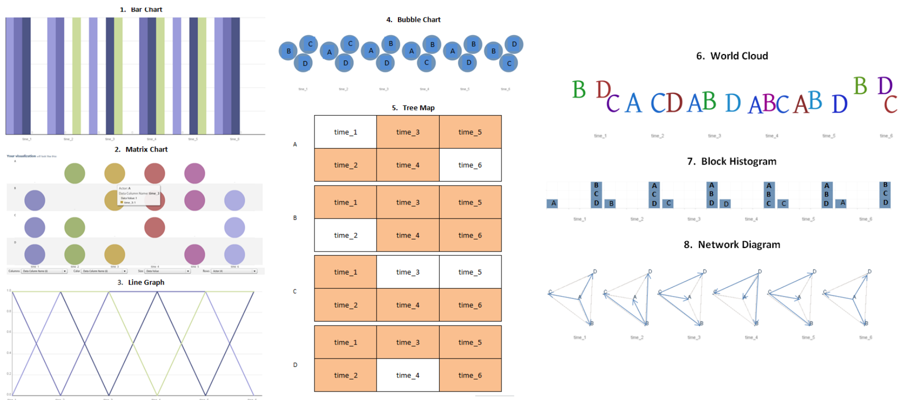


Fig. 3: The eight visualisations of the email conversation for the evaluation: 1) bar chart, 2) matrix chart, 3) line graph, 4) bubble chart, 5) treemap, 6) word cloud, 7) block histogram, and 8) network diagram.

Each question was ranked by the participant on a scale from zero to five. The higher the value, the easier the visualisation for answering questions over the visualised conversation. In other words, 0 meant that the visualisation was very confusing for answering the question and 5 meant that it was easy to answer the question using the particular visualisation. The results of the average scores for the answered questions are in Figure 4. The mean value with standard deviation of the scores for each of the visualisations are in Table 2. At the end of the interviews, the participants were requested to give their feedback on what they thought about the visualisations.

The evaluation showed that the Network Diagram, Matrix Chart and Block Histogram outperformed other visualisations with the average scores of 4.7, 4.6, 4.5 and standard deviations of 0.3, 0.5, 0.8. The relatively low values of the standard deviations meant the answers of the participants were quite consistent. The worst visualisations with regard to the participants' answers were Line Chart and Bar Chart with 2.8, 3.3 for the average scores and 1.1, 1.7 for the standard deviations.

In the next section, the results of the evaluation are discussed and then the results are compared to the choice of the CBR approach.

4 Discussion and Future Work

The results of the pilot evaluation and the automatic choice of the CBR approach showed a discrepancy, which can be explained by the design of the CBR approach

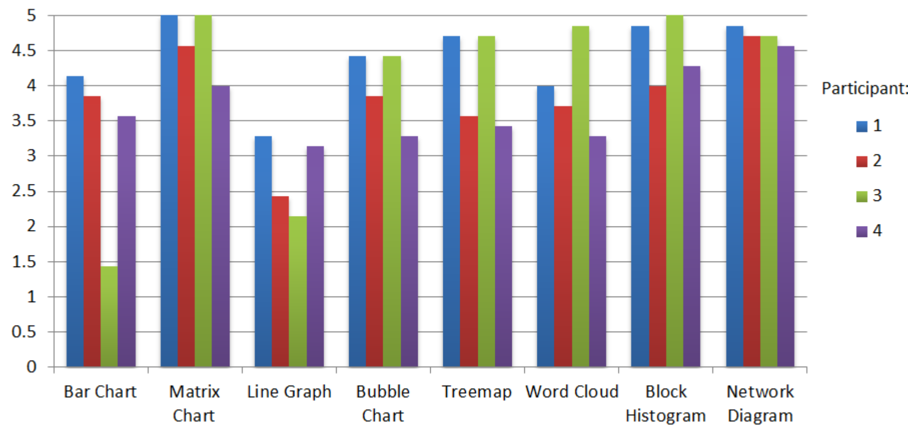


Fig. 4: The results of the evaluation for each of the four participants by visualisation.

Table 2: The mean value and the standard deviation for each visualisation.

Question / Visualisation Type	1	2	3	4	5	6	7	8
1. The Mean Value of the questions score	3.3	4.6	2.8	4.0	4.1	4.0	4.5	4.7
2. The Standard Deviation of the questions score	1.7	0.5	1.1	1.1	0.9	1.0	0.8	0.3

and the conducted evaluation. The three nearest visualisations for the sample email conversation were Bubble Chart, Line Graph and Matrix Chart. In these cases, the users visualised a) the number of his facebook and email messages received by days, b) the number of emails received over the time, and c) the number of facebook messages received during the day at different times. On the contrary, the evaluation showed that the Line Graph visualisation of the sample email conversation was not suitable for answering the questions over the conversation. Thus, the questions the user asked over his visualisation are very important and must be fixed or known in advance. In this paper, the asked questions were fixed to seven (Section 3).

Also, the application of the proposed CBR for choosing a visualisation type of a conversation showed a necessity in further elaboration on the following aspects: 1) refinement of the case model for visualisations of conversations, 2) what attributes other than number of actors and messages in a conversation are important for a case, 3) how many cases are required for statistically significant results of the CBR, and 4) how to evaluate the results of the CBR for visualisation of user's conversation.

Many Eyes lacks flexibility in terms of visualisation since the visualisation types and formats for uploading data are limited. From another side this simplicity makes Many Eyes easily accessible by a wide audience without any previous experience in visualisation. This fact motivates us to look into the visualisations offered by Many Eyes for visualising conversations and developing a recommendation system for advising visualisations of users' data and conversations in particular.

As a future work, we are planning to test the proposed CBR approach for other datasets in order to understand how to tune the approach for more accurate choices of visualisations. For this purpose, clarification of what a case is needed, temporal and context attributes of conversation must be introduced. Only a small set of visualisations crawled from Many Eyes was analysed, we are planning to extend the search beyond the keywords we used. Also, we are planning to use other sources for cases, for example, Google Fusion Tables [8].

In long term we are interested in the accessibility of the CBR choices of visualisation to an audience with different levels of expertise in using conversation visualisation for analysis. For this purpose, a formal evaluation of the visualisations using ”Physics of Notation” guideline [14] and its operationalization methods based on this guideline [16] has been planned.

5 Conclusion

Conversation visualisation is a promising area of research in facilitating the process of understanding digital conversations. Current practice of reasoning over these conversations is mostly in the form of queries filtered by keywords, amounts, date, and other parameters, and further analysing using spreadsheets. Previous experience of visualisation of conversations is helpful to assist people in visualizing their conversations. CBR is a powerful paradigm to utilize this previous experience in order to improve the appropriateness of conversation visualisations. In this paper, it has been shown that using visualised datasets uploaded to Many Eyes is feasible to make a choice of visualisation for a conversation easier by recommending to the user appropriate types of visualisations. The next step in the development of this approach is to collect more cases of conversation visualisations by extending the search by keywords and by using alternative sources such as Google Fusion Tables. Then, an evaluation for a focused group is planned to be conducted.

Acknowledgements The authors would like to thank the University of Brighton, especially the Doctoral College for funding the project on Temporal Reasoning and its Visualisation.

References

1. Agnar Aamodt and Enric Plaza. Case-based reasoning; foundational issues, methodological variations, and system approaches. *AI COMMUNICATIONS*, 7(1):39–59, 1994.
2. Daniel Angus, Bernadette Watson, Andrew Smith, Cindy Gallois, and Janet Wiles. Visualising conversation structure across time: Insights into effective doctor-patient consultations. *PLoS ONE*, 7(6):e38014, 06 2012.
3. N. Burlutskiy, M. Petridis, A. Fish, and N. Ali. Enabling the visualization for reasoning about temporal data. In *Visual Languages and Human-Centric Computing (VL/HCC), 2014 IEEE Symposium on*, pages 179–180, July 2014.

4. Judith Donath, Alex Dragulescu, Aaron Zinman, Fernanda Viégas, and Rebecca Xiong. Data portraits. In *ACM SIGGRAPH 2010 Art Gallery*, SIGGRAPH '10, pages 375–383, New York, NY, USA, 2010. ACM.
5. M. Farrugia, N. Hurley, and A. Quigley*. Exploring temporal ego networks using small multiples and tree-ring layouts. In *4th International Conference on Advances in Human Computer Interfaces ACHI*, 2011.
6. Simone Frau, Jonathan C. Roberts, and Nadia Boukhelifa. Dynamic coordinated email visualization. In Vaclav Skala, editor, *WSCG05 - 13th International Conference on Computer Graphics, Visualization and Computer Vision'2005*, pages 182–196, Plzen, Czech Republic, January 2005. (Jan 31 - Feb 4).
7. Jill Freyne and Barry Smyth. Creating visualizations: A case-based reasoning perspective. In Lorcan Coyle and Jill Freyne, editors, *Artificial Intelligence and Cognitive Science*, volume 6206 of *Lecture Notes in Computer Science*, pages 82–91. Springer Berlin Heidelberg, 2010.
8. Google. Fusion tables, March 2014. http://tables.googlelabs.com/.
9. Steffen Hadlak, Heidrun Schumann, Clemens H. Cap, and Till Wollenberg. Supporting the visual analysis of dynamic networks by clustering associated temporal attributes. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2267–2276, 2013.
10. IBM. Many eyes, March 2014. http://www.manyeyes.com.
11. Sasha Jovicic. Role of memory in email management. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '00, pages 151–152, New York, NY, USA, 2000. ACM.
12. T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*, pages 3–10, Sept 2013.
13. J. Lee, S. Lin, and K. Karahalios. Visualizing patterns of social and communicative behavior in children using plexlines. VAHC 2013, 2013.
14. Daniel L. Moody. The physics of notations: Toward a scientific basis for constructing visual notations in software engineering. *IEEE Transactions on Software Engineering*, 35(6):756–779, 2009.
15. Marc A. Smith and Andrew T. Fiore. Visualization components for persistent conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 136–143, New York, NY, USA, 2001. ACM.
16. Harald Strrle and Andrew Fish. Towards an operationalization of the physics of notations for the analysis of visual languages. In *Model-Driven Engineering Languages and Systems*, volume 8107 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2013.
17. TIBCO. Spotfire, March 2014. http://spotfire.tibco.com.
18. Gina D. Venolia and Carman Neustaedter. Understanding sequence and reply relationships within email conversations: A mixed-model visualization. Technical Report MSR-TR-2002-102, Microsoft Research, January 2003.
19. Fernanda B. Viégas and Marc Smith. Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspaces. In *Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4 - Volume 4*, HICSS '04, pages 40109.2–, Washington, DC, USA, 2004. IEEE Computer Society.
20. Jae wook Ahn, Catherine Plaisant, and Ben Shneiderman. A task taxonomy for network evolution analysis. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):365–376, 2014.
21. Rebecca Xiong and Judith Donath. Peoplegarden: Creating data portraits for users. In *Proceedings of the 12th annual ACM symposium on User interface software and technology (UIST '99). ACM Press*, pages 37–44. ACM, 1999.