TweetViz: Following Twitter hashtags to support storytelling

Lorena Lucas Regattieri
University of Alberta
Edmonton, AB
55 27 99767590
regattie@ualberta.ca

Ryan Chartier
University of Alberta
Edmonton, AB
recharti@ualberta.ca

Jennifer Windsor
University of Alberta
Edmonton, AB
jjwindsor@gmail.com

Geoffrey Rockwell
University of Alberta
Edmonton, AB
grockwel@ualberta.ca

ABSTRACT

How can visualizations of massive amounts of information be made more useful for data journalists? The availability of large amounts of publicly available user generated content is opening new opportunities to study social, cultural, and communications phenomenon. Computer assisted analysis now makes it possible to explore the relationship between nodes and text without having to choose between data size and depth. To create a visualization technique that would allowed us to reveal the network of actors and the main themes hidden in a large dataset, we had to work in a method of inquiry for social sciences. Based on the actor-network theory (ANT) we explored a dataset extracted from Twitter in order to map relationships and indicate new possibilities for journalists by discovering main themes around a hashtag, this way we interpret a layer of text multiple times, analyzing the nodes in its many attributes. Beyond the boundaries of 140 characters, this approach can succeed as it reproduces and reveals the dynamic connections contained in a collective phenomenon. In the last section, we demonstrate a prototype visualization that reveals behaviors and discourses within the large sample datasets. . We use the D3 visualization library to overlap related links and nodes to produce a comprehensible interactive visualization. Our model is interactive and allows us to identify part and whole pattern relationships constant with the three principles of information visualization: overview first, zoom and filter, then details on demand. This paper analyses networks from the perspective of ANT in order to create a visualization ready to support users when telling a story with data.

Categories and Subject Descriptors

D.3.3 **[Programming Languages]**: Language Constructs and Features – *abstract data types, polymorphism, control structures.*

D.2 **[Software Engineering]**: Design Tools and Techniques - *Flow charts, Object-oriented design methods, User interfaces.*

General Terms

Algorithms, Documentation, Performance, Reliability, Experimentation, Security, Human Factors, Standardization, Languages, Theory, Design.

Keywords

Data journalism, Actor-Network Theory, design, social network analysis.

1. INTRODUCTION

A fair number of events and social phenomenon find themselves connected; they are caused by a range of parts of a complex puzzle interacting to each other. As a society [1], we came to recognize that nothing is isolated anymore. If not yet to consideration, the “global village” is even more a reality in the current state of living, where everything is linked. New understandings about society and community life are guided by a concept of “glocal” - something that translates the current sensation of being both, global and local [2]. The use of twitter data to interpreted human behavior is not news. Every day, more researchers are overcoming the issue of understanding social relations using text analysis and information visualizations tools. The availability of large amounts of publicly available user generated content is opening new opportunities to study social, cultural, and communications phenomenon. Computer assisted analysis now makes it possible to explore the relationship between nodes and text without having to choose between data size and depth [3]. As the volume of available information expands, it is becoming increasingly important for techniques to be developed that will allow for networks of information to be effectively summarized and navigated. The alternative—what has come to be known as the “hairball”—is becoming increasingly unwieldy and obfuscatory, no matter how many colour based filters are applied. To overcome the hairball we have developed a new visualization technique that allows us to reveal the network of actors and main themes hidden within traditional network visualizations of large datasets. In this paper we reveal this technique and our methods for producing it.

2. METHOD

The project began as a conversation about how to visualize large quantities of data and how this process could support data driven information. We made the decision to focus on hashtag and the Twitter conversations surrounding these hashtags. The conversation led a set of agreed upon features that are represented in the original sketch. The tool had to do two simple things: visualize the frequency of hashtags in a data set and allow the user to click on specific hashtags and read the tweets associated with them. Every other feature we incorporated into the visualization serves one of these two purposes.

The tweets themselves were extracted using the Twarc[1] tweet scrapper. Twarc is a command line tool that takes a single search term (in this case the string 'rob ford'), queries the twitter API (Application Programming Interface), and the downloads all of the metadata associated with whatever tweets it finds. However, Twarc alone produces a large amount of unnecessary data. For every 140-character tweet that Twarc downloads, approximately five thousand characters worth of metadata is received. All told, we collected about twenty gigabytes of twitter data. The next step was to filter this data, for that it was built another scrapper, also in python, that would search this data and return in csv format all of the information needed. In this case hashtags, but many other attributes such as: geolocation, mentions, and url are also available. This dataset returned approximately one gigabyte of data. In order to filter the data further, we used an R script to split the csv files, format character codes and time stamps, as well as filter out every tweet that does not contain a hashtag. This reduced the dataset to two hundred megabytes. We then uploaded the entire remaining dataset to a MySQL database through a PHP script. The final step was to query this database for visualization in a JavaScript library: D3. For reaching out a visualization dashboard that could provide interactive information, D3 proved to be extremely useful. All told we employed seven different programs across six different programming languages in order to pre-process the data.

3. DISCUSSION

This paper situates the debate and challenges posed by the large amount information available online. In this matter, we begin with a context of critical questions on Big Data. Mathematicians, philosophers, sociologists, and many scholars from different fields of study are claiming "for access to the massive quantities of information produced by and about people, things, and their interactions."[4] Big Data is a term use for a large combination of datasets together. Following Manovich[3] observations on the issue, which puts Big Data near a researcher using a simple desktop, "we want to combine human ability to understand and interpret - which computers can't completely match yet - and computers' ability to analyze massive data sets using algorithms we create."

Data driven journalism is a field that brings together the interdisciplinary studies involving the provocations in big data and information visualization. According to Paul Bradshaw, data can be both, used in the production and distribution of information in the digital era and a tool with which the story is told. In journalism, like any source, data can be treated with skepticism; and like any tool, it "should be use with conscious of how to shape and restrict the stories that are created with it." [5]. Just to have an idea, the graphics department at The New York Times, has a group of about 30 people responsible for the information graphics and multimedia presentations, such as: reporting and writing copy, processing datasets, web development, drawing schematics, designing print pieces, and developing and creating the interface of multimedia projects. When selecting subjects to research, data analysis, and reporting, people from many backgrounds are doing data driven journalism, the fact that now the abundance of data has increased exponentially is a major challenge for the ones working in the area of visual storytelling.

Social network sites like Facebook, Instagram, and Twitter became a central component of sociability in our contemporary society. User generated content is a way to measure qualitative data, from the metrics on the success of a product inside the market to tracing the news about a natural disaster, social media delivers a massive amount of information everyday. In studies of network analysis, Twitter has become a broad database for quantitative and qualitative scholarly analysis. With user generated content and the flow of information, the microblog is the virtual space for peoples perspectives online [6]. Twitter is a rich environment for data scientists looking to investigate the issues of Big Data, social relation, and data visualizations. While sharing financial results in February, 2014, Twitter announced that its number of users has passed 241 million monthly active users. From the 215 million monthly active users, there is around 100 million daily active users, generating 500 million tweets per day. For qualitative research, Twitter offers a great strategy to segment a topic of interest, which is the hashtag (#). A topic is indicated through the composition of a hashtag and a keyword. This is the average practice in the use of "tags" when categorizing web content, anyone familiar with bookmarking will rapidly understand the importance of labeling certain tweets. A hashtag gain importance when the text has a high rate of retweets, meaning that a message is republished many times. This specific word will then reach Twitter's trending topics and achieve a level of importance. This will end up creating from time to time, specific topics of conversation between users. In qualitative research and for the purpose of this research, we will track the hashtags in order to examine its parts in the course of a news event.

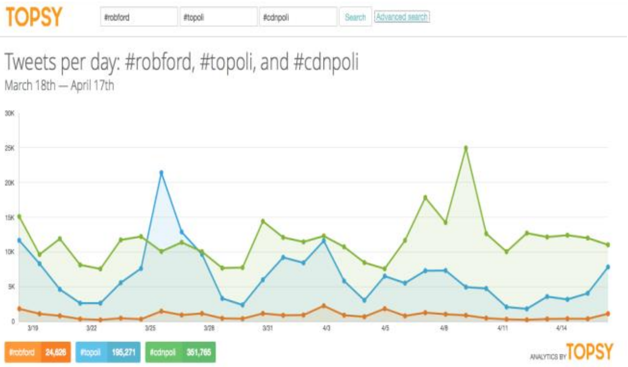

Figure 1. Frequency of hashtags overtime provides insights about topics: #robford #topoli #cdnpoli

Important and new questions emerge as we develop technical skills to overcome the "provocations" in Big Data, with computer assisted analysis it is possible to trace millions of opinions, ideas, feelings, and monitor those flux of information. Language, time, space, gain new features on the new method of information management. Thus, we need to think in new linguistic production associated with fast conversations on Twitter, for example, what would be the vocabulary during the course of an event, like a bomb explosion or a flooding? We can make these and many reflections analyzing the data extracted with the assistance of a

[1] Twarc was originally created to save tweets related to Aaron Swartz https://github.com/edsu/twarc

computer. In consequence, to tell stories based on these data visualizations.

The mapping controversies technique is a successful method to trace digital data. Cartography of controversies is a method created by Bruno Latour [7] and is broadly used in the communications field to map the debates around an specific object, subject, or event. This technique hinge on the idea that 'things' generate contested spaces, this way something new is produced following a large amount of material and subjective considerations. An Actor-Network-Theory (ANT) comprehension of events will move beyond the traditional dimensional image, between two or three common implications, extending to the meaning of the human factors, thus reducing the necessary to differ subject and object: "In a few words, when you look for controversies, search where collective life gets most complex: where the largest and most diverse assortment of actors is involved; where alliances and opposition transform recklessly; where nothing is simple as it seems; where everyone is shouting and quarrelling; where conflicts grow harshest. There, you will find the object of the cartography of controversies" [8].Considering Venturini's instructions to reach out for the controversy, we were lead to an investigation of an event that would be both, complex and big. A theme that would lead us to question the possibilities in the process of producing new visualizations, especially for data driven journalism. Knowing that we chose to pursue an empirical investigation within the course of news involving the Toronto mayor Rob Ford.

3.1 The story on the Rob Ford Controversy

A brief background about the case that explains the choice for data: starting in May 16, 2013, a series of reports about a video supposedly showing the Toronto mayor smoking from a glass pipe ends up circulating on the U.S media. Subsequently, media outlet Toronto Star also spread the news about a man their reporters claim in a video smoking crack. This is enough for the long controversy to begin. Since May, from denying allegations to new videos emerging from time to time in several news media, Rob Ford is an ongoing conversation on Twitter.

Building up from the theoretical references exposed above, we needed a dataset big enough to challenge us within the limits of back end and front end work with Big Data. With different themes underlying the discussion on Canadian and Toronto politics, the dataset extracted from Twitter around Rob Ford elaborates on how citizens are expressing their concerns on social, economics, and political issues in the society. The Rob Ford tweets set us up with long tail of conversations to follow, presenting us with a scenario demanding of critical thinking about information visualization. Moretti[9], Manovich[10], and Ruecker et al.[11] have drawn the attention of the literary research community to the value of visualization within the research process. Telling stories with data is about discussing theories of visual thinking and analytical design [12], however, it is also about engaging in a scholarly debate over the uses of a visual interface to investigate social data. We aim to bring together in our tool, an innovative method where anyone can quickly analyze, visualize and share information.

3.2 TweetViz: a tool to explore data[2]

In this section we demonstrate a prototype visualization that reveals behaviours and discourses within the large sample datasets. Our model is interactive and allows us to identify part and whole pattern relationships constant with the principles of Shneiderman's[13] visual information-seeking mantra: overview first, zoom and filter, then details on demand. We use the D3[3] visualization library to overlap related links and nodes to produce a comprehensible interactive visualization. In developing this technique we are untangling what would otherwise be "hairballs," aligning relevant information from the inside out, displaying clusters, outliers, patterns and trends, making visible to users "differences that make a difference". [14] An overview provides the gist of the data — the substance or salient aspects of the information and a perceptual shortcut. It is the 'macro' referred to when discussing micro/macro readings of information graphics: the texture of detail that we don't immediately need to direct our full attention to that cumulates into larger, coherent structures. Gist provides a summary of the data at a low cognitive cost for the viewer in terms of time and mental energy. The image (2) shows an early sketch of the concept, it was designed to allow comparisons to be made within an eye span and provides a general context for the entire dataset. The user then has a basis to draw on for further drill-down decisions.

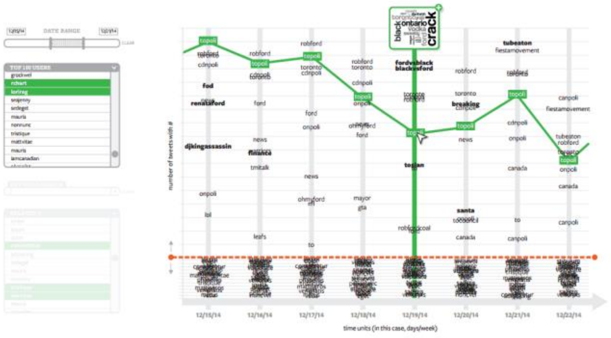

Figure 2. First sketch of the tool would display a word cloud for each day

Data visualizations excel at expressing comparative or relational aspects of data in order to highlight significant connections and identify patterns or trends. In the same way that mapmakers often focus on certain predetermined features of a landscape rather than depict an exact replica of an area from above, our first task in creating an overview of more than a million tweets was to consider which features were most likely to reveal relevant structures within, and context for, the data. When choosing a temporal framework for the visualization, patterns and trends (as evidenced by changes in the dataset such as new hashtag appearances, spikes in frequency and emergent word occurrence patterns) were revealed. It became possible to compare and contextualize data changes with real-world events. We chose hashtag frequency for the y-axis reasoning that it offered the broadest indication of tweet topic, and other means of drilling-

[2] TweetViz Prototype is available at http://analytics.artsrn.ualberta.ca/viz/hashtag.html

[3] D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG and CSS http://d3js.org/

down such as username and keyword search would then provide the viewer greater detail after. Highlighted hashtag occurrence over time, in the context of how often it appears, provides a macro view of a conversation arc over a given period. We also chose to highlight outliers — hashtags that only appear once in the data set — reasoning that they might provide a unique perspective from outside of occurring trends and patterns. After the broad strokes of the overview, the user can explore the data more closely. The ‘zoom’ Schneiderman referred to typically means changes in the scale of magnification — in TweetViz, it is semantic in nature. The user can move from a macro reading of the data to closer examinations of the text. In the original sketches, this is accomplished by either a small word cloud generated for a given hashtag each day, or in the tweets themselves in a second panel. Filtering is achieved with a date-range selector and a username and keyword search.

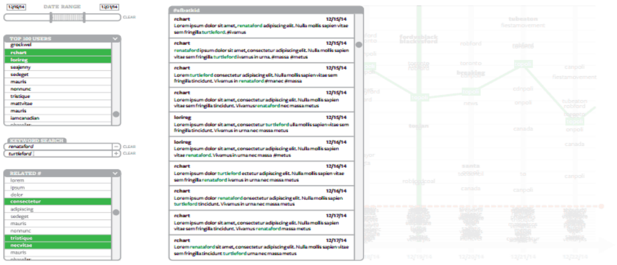

Figure 3. User can explore tweets by user or hashtag

A significant design concern for large data sets is dealing with occlusion: ensuring that the design inhibits visual elements overlapping as much as possible. In the early sketches, we designed a division between the 10 most commonly occurring hashtags and the rest of the hashtags in order to minimize overlap: when the slider bar is raised, the user can see all but the top 10 occurring hashtags in their relative (and often occluded) arrangement; when the bar is lowered, greater vertical space lessens overlap for the top 10.

The current visualization offers a toggle between relative and absolute views of the top 10 hashtags, and uses jitter — the slight, irregular movement of overlapping hashtags — to reveal overlapped elements at minute intervals. In the next paragraphs, we engage in the process of untangling the "Hairball" by building our own tool. The visualization dashboard consists of two screens. The first is a visualization of the relative frequency of each hashtag in the data. The larger the percentage of tweets that that hashtag gets used in the higher it appears on in the chart. Secondly, we also wanted to visualize the contents of these tweets. This is done in two ways. Firstly, the original design had a word cloud associated with each node, this word cloud is designed to offer an ‘at a glance’ insight into the content of the tweets represented by a hashtag. Secondly the user can click on a node and transfer the tweets in that node to the second screen. The viewer is simply a widget that allows the user to sort, filter, and read individual tweets.

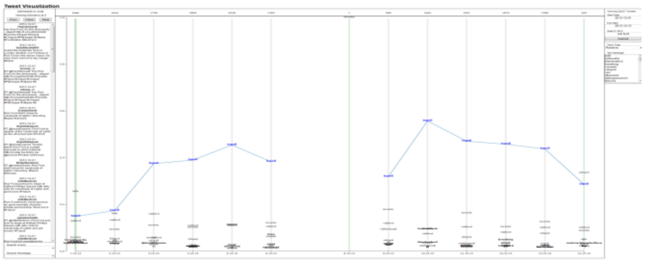

Figure 4. Visualization of the Relative Frequency of the hashtag #Topoli overtime

Due to the incredible quantity of tweets that twitter processes on a daily basis, the unique identification numbers assigned to each tweet was massive. Unfortunately, not every program handles large numbers in the same way, and due to the large assortment of programs in use, not all of this data was translated between languages perfectly. Another problem encountered is due to character encoding. Because twitter is an international platform, it is extremely lenient in which characters it allows. Unfortunately, due to the large amount programs and data formats used, not all of which allow by default the entire unicode character set, certain characters needed to be removed from the set (notably all newlines, carriage returns, and some foreign symbols I could not identify) and certain characters were lost in translation. An example of where this problem appears is in the ‘t’ hashtag in the rob ford data set. Unfortunately, ‘t’ is only a small part of the hashtag itself, but the rest does not render properly. Beyond the prototype stage a better solution to this project needs to be addressed. Request size also proved to be a problem. Javascript is a client side service, and in order for it to visualize properly the entire data set needs to be processed and transferred to the user computer. Unfortunately, due to the size of the project, these requests tended to overwhelm the earlier versions of the project. Early versions of the twitter viewer actually fetched the full text of every tweet it was analyzing. This was necessary because it was the easiest way to generate the word clouds dynamically. However, this soon proved to be too much for JavaScript to handle. Instead, we needed to preprocess all of the data on the server. Unfortunately, this meant that the word clouds needed to be generated outside of D3. Due to the difficulties to visualize, it was decided to cut the world clouds and only visualize the content through the tweet reader.

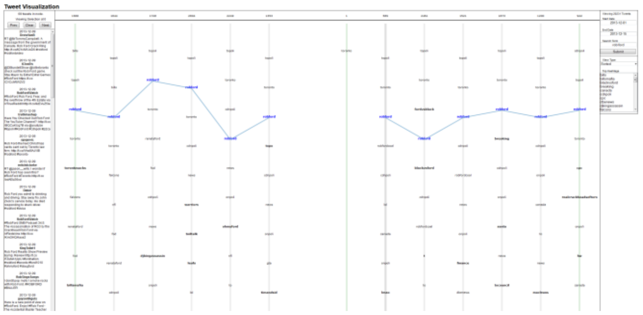

Figure 4. Visualization of the sorted Frequency of the hashtag #RobFord overtime

3.3 Reporting on issues and findings

Due to the incredible quantity of tweets that twitter processes on a daily basis, the unique identification numbers assigned to each tweet was massive. Unfortunately, not every program handles large numbers in the same way, and due to the large assortment of programs in use, not all of this data was translated between languages perfectly. Another problem encountered is due to character encoding. Because twitter is an international platform, it is extremely lenient in which characters it allows. Unfortunately, due to the large amount programs and data formats used, not all of which allow by default the entire unicode character set, certain characters needed to be removed from the set (notably all newlines, carriage returns, and some foreign symbols I could not identify) and certain characters were lost in translation. An example of where this problem appears is in the 't' hashtag in the rob ford data set. Unfortunately, 't' is only a small part of the hashtag itself, but the rest does not render properly. Beyond the prototype stage a better solution to this project needs to be addressed. Request size also proved to be a problem. JavaScript is a client side service, and in order for it to visualize properly the entire data set needs to be processed and transferred to the user computer. Unfortunately, due to the size of the project, these requests tended to overwhelm the earlier versions of the project. Early versions of the twitter viewer actually fetched the full text of every tweet it was analyzing. This was necessary because it was the easiest way to generate the word clouds dynamically. However, this soon proved to be too much for JavaScript to handle. Instead, we needed to preprocess all of the data on the server. Unfortunately, this meant that the word clouds needed to be generated outside of D3. Due to the difficulties to visualize, it was decided to cut the world clouds and only visualize the content through the tweet reader.

In terms of visualization, crowding turned out to be the biggest problem in the visualization itself. Once the initial prototype was built on a small subset of the data, it became immediately apparent that some of the assumptions made in the original design were not true. The first assumption was that spacing between the top few hashtags would be relatively even. We could visualize the top hashtags as a relative percentage and use a slider bar to 'squish' all of the lower hashtags down allowing us to push them out of the way and focus on the higher percentage hashtags. In the Rob Ford data set, this is false, and in fact, the opposite is true. The top hashtags are completely dominant, and only the top three or so are actually visible on a relative scale with everything else squishing into the bottom. Instead, of using a slider bar to push the lower less important hashtags out of the way, it became apparent that we needed a way to focus in on the lesser hashtags and push the dominant ones out of the way.

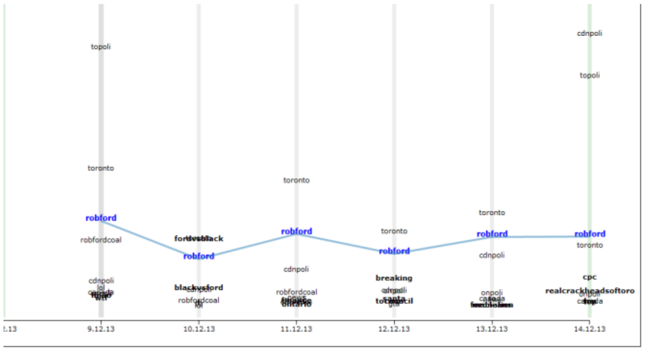

Figure 5. Issues when hashtags overlap each other

4. CONCLUSIONS

In short, our research builds up from a solid theoretical reference to visualize relationships in a network, the alliance between computing methods and the humanities consider quali-quantitative techniques that means more than overlapping statistical resources and ethnographic approach. We believe that information is only visible when the user can have the opportunity to click on, explore, discover, and share new findings. Data analysis can serve as technique to reveal the different structures of the same story and to provide new lens to see levels of information. When journalists use data to do their jobs they shift from being the first one to communicate to being the ones telling people what a certain progress of an event may actually mean. This tool can be appropriate by for journalists trying to visualize news and events, using data to transform something abstract into something everyone can understand and relate to the real events. With the curiosity to continue to think critically on how to display digital information and to explore data, for the future work we hope to overcome the issues with data encoding and crowding in our tool.

5. ACKNOWLEDGMENTS

Funding for the project generously supplied by Just What do They Do (JWDTD), Implementing New Knowledge Environments (INKE), and Social Science and Humanities Research Council of Canada (SSHRC).

6. REFERENCES

[1] Castells, M. 1996. *The Rise of the Network Society: The Information Age: Economy, Society and Culture, Volume 1*. Blackwell Publishers, Inc, Malden, MA.

[2] Wellman, B. 1999. *Networks in the Global Village: Life in Contemporary Communities*. Westview Press, Ed. Boulder, CO.

[3] Manovich, L. 2012. Trending: The Promises and the Challenges of Big Social Data. In *Debates in the Digital Humanities*. Minnesota, MI: The University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/15

[4] Boyd, D. and Crawford, Kate. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon.In *Information, Communication, & Society* 15:5, 662-679.

[5] Gray, J. Chambers, L. and Bounegru, L. *The Data Journalism Handbook How Journalists Can Use Data to Improve the News*. O'Reilly Media.

[6] Wu, S. Hofman, J. Mason, W. and Watts, D. 2011. Who says what to whom on Twitter. In *International Conference On World Wide Web*, WWW'11,New York., NY.

[7] Latour, B. 2005. *Reassembling the Social: an Introduction to Actor Network Theory*. Oxford University Press. Oxford.

[8] Venturini, T. 2010. Diving in magma: how to explore controversies with actor-network theory. In *Public Understanding of Science* 19 (2009): 1-16.

[9] Moretti, F. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History.* Verso, London.

[10] Manovich, L. and Douglas J. 2009. Cultural Analytics. In *Plenary address at the Digital Humanities 2009 conference*. University of Maryland. June 22-25.

[11] Ruecker, S. Radzikowska,M. and Sinclair S. 2011. *Visual Interface Design for Digital Cultural Heritage: A Guide to*

Rich-Prospect Browsing. Farnham, Surrey: Ashgate Publishing,

[12] Tufte, E. 2001. *The Visual Display of Quantitative Information*, 2nd. Ed. Graphics Press LLC, Cheshire, Conneticut.

[13] Shneiderman, B. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the IEEE Symposium on Visual Languages* (Washington. IEEE Computer Society Press, pages 336-343) IEEE'96.

[14] Tufte, E. 2006. *Beautiful Evidence. Graphics Press*. Chesire, Connecticut:CT.