Visual Analytics Dashboard Design, Development and Assessment of a Big Data

Challa sneha
Assistant professor in Department of CSE
Geethanjali College of Engineering and Technology

Chandra Kala Anneda
Assistant professor in Department of CSE
Geethanjali College of Engineering and Technology

Abstract: Design, Development and Evaluation of a novel Visual Analytics Dashboard for Big Data Analytics. The presented dashboard connects social activity from Facebook with a thorough event timeline of the factory disasters in the Bangladesh garment industry. Bangladesh depicts one of the largest garment industries in the world and their mostly female workers only receive a low wage. The goal of this thesis is to present a thorough understanding of the design and development processes needed to implement a Big Data Visual Analytics tool based on freely available open-source components in a robust, extensible manner. Moreover, an evaluation of the developed dashboard is performed based on a task-based user study in conjunction with software and database performance optimization. The user study concludes that the dashboard is easy to use in a productive manner without prior training and experience in using visual analytics tools. By using the presented dashboard, even novice users can gain profound understanding of the tragedies in Bangladesh, their background, and the resulting social media impact. Furthermore, the linked social media activity from eleven international companies in the garment industry can be interactively explored through different visualizations depicting actor mobility, conversation content, language distribution, and overall activity levels.

***Keywords: Big data, Social Data, Visual Analytics,* Acquisition, Sodato**

1. INTRODUCTION

In recent years, visual analytics tools have steadily been improved and adapted in order to work with large data sets, so-called Big Data, while providing accessibility to a growing audience. Although many of these data sets have historically been of proprietary nature, the growth of social media also spurred availability of huge collections of social media activity. Various social networks, such as Twitter and Facebook, provide extensive data while being increasingly used for research and business purposes.

In light of current research on social media activity in regard to major events, such as the 2013 German Bundestag elections using Facebook and Twitter data (Kaczmirek et al.; 2013), further analysis of distinctive events and their social media impact needs to be performed using state of the art visual analytics tools.

Under these preconditions, the 2012/2013 Bangladesh textile industry disasters which prompted massive exclamation from consumers and media outlets all over the world present a series of events worth studying on a global scale. Both the viewpoints of consumers and high-profile textile industry brands alike are captured in the social media dataset, with the latter publicly perceived as an adversary to workers' rights in the textile industry.

Given the opportunity to source social media data from Facebook for a perfectly sufficient timeframe for analysis of the Bangladesh textile industry event timeline, an interactive visual analytics dashboard based on the available data is designed and developed in context of this thesis.

To assess the quality of the developed dashboard, its accessibility and interactive components are evaluated by means of user and software testing. The evaluation is performed in order to gain a thorough quantitative and qualitative reasoning in regard to the value created for end users when visually analyzing complex event timelines such as the above-mentioned factory disasters in the textile industry. Additional insight on the impact of these events on social media activity and waves of sentiment against specific social media actors can be obtained through this analysis.

2. BIG DATA

Big Data is a modern terminology for large data sets which are hard to process with traditional tools due to their sheer size. Although the concept encompasses new technological developments in database and storage technology, it remains mainly a marketing term for interdisciplinary processing of available information. According to Buhl et al. (2013, p.66), "above all, [Big Data] is a multidisciplinary and evolutionary fusion of new technologies in combination with new dimensions in data storage and processing (volume and velocity), a new era of data source variety and the challenge of managing data quality adequately (veracity)".

Big Social Data is a term for Big Data which is obtained from the social media world. With Facebook and Twitter being some of the the most popular social networks on the internet, the data they provide to third parties over public and private APIs (such as the Twitter Firehose) can be called Big Social Data.

2.1. Visual Analytics of Big Social Data

As this thesis presents an IT artifact visualizing the Bangladesh factory disaster event timeline in light of social media activity from various companies' virtual

2.2. Theoretical Background

Facebook presences, a particular blend of technologies in Visual Analytics and Big Social Data is crucial.

Wong et al. (2012) underline major challenges that arise in Visual Analytics of data sets classified as Big Data, in particular scalability, summarization, difficulties in achieving smooth interactive UIs, and the development of effective methods for user-driven data reduction.

The analysis of social media actors, their actions and the artifacts they create is actively performed by researchers and businesses with focus on many different topics. For one, researchers use advanced methods of sentiment analysis to gain insight into the "reaction of people to events, topics and entities" based on Twitter data (Bravo-Marquez et al.; 2014, p.2). Additional commercial, web-based tools such as SAS exist which have a specialization on exploratory Visual Analytics of Big Data (Abousalh-Neto and Kazgan; 2012); but they lack the complexity needed for working with combined Big Social and event timeline data.

To conclude the theoretical background, previous research presents a healthy assessment of opportunities that arise through the use of Visual Analytics on large data sets, but on the other hand underlines several problems in handling of Big Social Data through traditional means of Visual Analytics software which need to be accounted for during implementation of the IT artifact.

3. DATA SOURCES AND ACQUISITION

This chapter introduces the two main data sources which are used in development of the Big Data Analytics dashboard. Data from these two equally important data sources is used during the realization of the Visual Analytics dashboard depicted in the first introductory chapter:
The first data source consists of social media activity from the virtual online presences on Facebook - so-called "Facebook walls" in domain language - of a wide selection of retail companies in the garment industry.
The second data source consists of a timeline of various events in regard to the publication of the series of Bangladesh factory disasters in years 2009 to 2014 via consumer-facing news outlets.

3.1. Social Media Activity from Facebook

Social media activity from Facebook is acquired using the Social Data Analytics Tool (SODATO) presented by Hussain and Vatrapu (2014) for eleven major companies in the clothing retail industry.
Only companies that present retail stores in Europe and/or the United States are selected in order to gain a representative set of social media activity that contains consumers' reaction to the Bangladesh disasters from above mentioned geographical regions.

The final selection of companies from the retail clothing industry is as follows:

- Benetton
- Calvin Klein
- Carrefour
- El Corte Ingles
- H&M
- JC Penny
- Mango
- Primark
- PVH
- Walmart
- Zara

After successful data extraction using the Facebook API, the SODATO tool supplies the Facebook activity for each company as a .CSV file which follows a specific data format convention.

SODATO CSV File Format	
dbitemid	int
dbpostid	int
timestamp	timestamp
lastupdated	timestamp
eventname	text
actorid	bigint
actorname	text
facebookpostid	text
typeofpost	text
link	text
commentlikecount	int

Figure 1. Structure of SODATO-supplied Facebook data sets

Figure 1 showcases the data format of the SODATO output. Each piece of Facebook activity is uniquely identified by the tuple of (dbitemid, dbpostid). Both timestamp and last updated contain timestamp information.
The most important distinction between different social media activity types depicts the eventname field, which holds any one of the values [POST, COMMENT, LIKE]. The eventname field thereby depicts the Action performed by the social media Actor who is specified by actorid and actorname.
The field face book post id is the unique identifier of each Facebook conversation, whereas type of post specifies the exact type of the Artifact as one of [status, SWF1, photo, offer, music, link, video, question]. The field link is optional and in some cases contains a hyperlink that will be shown in the Facebook UI, but it is of no relevance for this thesis. Finally, the comment like count field depicts the number of likes which were performed by other Actors on the Artifact at hand.

3.2. Data Processing Steps

Various data processing steps need to be performed before the social data and the event timeline can be used for visual analytics purposes. These processing steps are a major component of the data acquisition pipeline.
Processing of Event Timeline Due to the manual collection of all information which is included in the event timeline, no further processing steps apart from conversion of the spreadsheet into .CSV file format is needed.
Processing of Facebook data The SODATO-provided Facebook activity data sets are generated as independent files for each company's Facebook wall, and need to be combined into one for using them as a whole data set that can be filtered or expanded on demand.
Figure illustrates the data acquisition process through which the author was able to obtain social media data from Facebook for a wide range of international clothing retailers.
The general concept follows the stages of the "Big Data Value Chain" introduced by Miller and Mork (2013), with steps of preparation, organization and integration of the data prior to visualization and analysis.
Data preparation tasks are performed in a pre-processing step which converts all .CSV files to from their character encoding UTF-16 to the more commonly used UTF-8 and handles edge cases in which the generated SODATO output lacks proper data type encapsulation.
Subsequently, a data normalization phase performs sanity checks on the input data and identifies malformed data or unneeded information.
Lastly, all distinct data sets are aggregated while conserving information regarding their original source in an additional variable. The aggregated data is then imported into a database management system (DBMS), from which it can be accessed for visual analytics purposes.

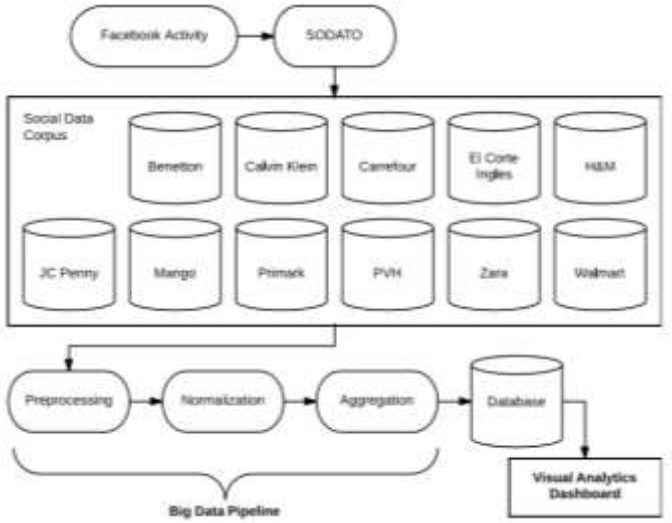

Figure 2: Big Data Acquisition Pipeline of the Social Data from Facebook used later on in the Visual Analytics Tool

3.3. Data Storage

The character of the two underlying data sets which are used in the future visual analytics dashboard has been outlined in previous sections. This section discusses the storage of the acquired data, which will be used later on for development of the dashboard.

Storage of Event Timeline The structured nature of the Bangladesh event timeline makes it easy to choose a fitting storage type for the events. Based on this, a relational database management system (RDBMS) has been chosen. The decision is further reinforced by the rather small amount of events (96) and the simple data structure without external dependencies.

Storage of Facebook Activity Data The semi-structured nature of the Facebook activity data which was acquired from the Facebook walls of eleven companies using the SODATO tool and processed according to the steps described in section 3.3 presents a more difficult decision.

With a total of 8GB, the combined file size of the raw Facebook activity data is quite large, but obviously not large enough to warrant the use of large-scale distributed database systems such as Apache Hadoop, based on which Google realized the first Map/Reduce-style processing of Big Data (White; 2009).

Choice of DBMS This presents us with the opportunity to employ broadly established open-source RDBMS like PostgreSQL instead of distributed Big Data ready databases like Apache Hadoop. This is due to the fact that multi-core systems with beefy hardware and plenty of working memory in the double-digit Gigabyte range are easily available nowadays.

For obvious reasons, a strict ordering of all events and social data by their timestamps and dates is needed for the creation for a visual analytics dashboard as outlined in previous chapters. Given the structured analyses which need to be performed asynchronously in an interactive, user-specified manner, the decision was made to use a RDBMS, with the availability of SQL's expressiveness in a RDBMS being a major contributing factor for this choice. Queries stated in SQL allow for easier modeling of requirements than NoSQL and Map/Reduce approaches of distributed database systems, and present no drawbacks in the use cases at hand.

To conclude this chapter, the data storage for the subsequent steps in creation of the visual analytics dashboard will be performed in a single-machine PostgreSQL database which can be queried using SQL. After data import, table optimization and index creation, the final size of all data in the database approaches 80 GB. This means the data can be kept in working memory to a good extent, thereby improving overall performance of the database.

4. VISUAL ANALYTICS DASHBOARD DESIGN

In this chapter, the design process of the Visual Analytics dashboard is outlined. Furthermore, the fundamental components of the dashboard are chosen based on the available datasets and possible visualizations. Then, different ways of realizing the Visal Analytics dashboard are discussed and a decision is made. Finally, the architecture of the underlying software and database is presented.

4.1. Design Goals & Objectives

The design of a Visual Analytics Dashboard as depicted in the previous chapters needs clear goals in terms of visualization options, interactive components, target devices and many more. During subsequent steps, for example when implementing the dashboard, these previously formulated goals and objectives provide crucial orientation in many decisions and trade-offs that have to be made.

Multidimensionality A Visual Analytics Dashboard consists of a mash-up of multiple visualizations which can be utilized by the user in combination to maximize efficiency. The type and size of each visualization need to be carefully evaluated.

Accessibility The dashboard should be accessible as easily as possible for users. It should therefore have as few hard dependencies in terms of installed software, operating system or device type as possible.

Responsiveness The dashboard needs to be responsive to different device types and screen sizes. The realized IT artifact should be reasonably accessible using devices with different form factors and adapt accordingly. Both a 4K display used in a conference room and a normal tablet should be able to display the dashboard. Consequentially, the dashboard should also work without problems on a standard-sized workstation computer and modern laptops.

Performance Another key objective for the visual analytics dashboard displays performance of both user- and server-facing software components. The large-scale data set which is the foundation of the dashboard created in this thesis necessitates increased processing needs. This also impacts overall memory footprint and disk space consumption. In order to achieve smooth performance in productive use, some sort of duty-sharing between server and client software components needs to be established. Thereby, workload may be shifted as needed and user interface waiting times are reduced.

Ease of Use From the end user's perspective, ease of use depicts an important non-functional requirement which needs to be addressed. Factually, all software is made accessible to end users in combination with additional documentation, a user manual or even training lessons. Even though these measures present additional value for the end user, the visual analytics dashboard should be designed in a way that enables the user to work with the dashboard without any prior briefing or training on how to use it.

Extensibility Lastly, during realization of the visual analytics dashboard, an extensible framework should be used so that future changes can be implemented with only moderate effort and without unnecessary technical hindrances.

4.2. Design Principles

The design of a Visual Analytics dashboard needs to follow a set of core principles, through which the above stated goals can be achieved.

Key outtakes from a thorough investigation on design principles for visual analytics dashboards by Kang et al. (2011) are that dashboards need to facilitate clue-finding, have smooth transitions between different stages of analysis, support evidence marshalling, allow flexibility in organizing and support task resumption after pursuing alternative paths of analysis.

Furthermore, Keim et al. (2010, p.7) propose that the design of a Visual Analytics dashboard should enable the user to "provide

timely, defensible and understandable assessments" of the underlying data.

Additional principles applied to design and realization of the dashboard is:

- Detail on Demand: The detail on demand principle strives to first present an easily graspable overview to the user and not the full depth of the available data. This overview can be processed visually and intellectually in short time. Only subsequently, when the user decides to, the level of detail shown in the visual analytics tool can be increased.
- Ready-made Visualizations: The Visual Analytics dashboard presented in this thesis is based on social media data from Facebook. The dashboard consists of a combination of multiple visualizations. Therefore, each visualization needs to highlight unique features of the underlying social interactions. This allows the dashboard as a whole to be kept clean and organized, and prevent it from becoming too complex.
- User-centric Design (UCD): Abras et al. (2004, p.2) emphasize that in user-centric design, "the role of the designer is to facilitate the task for the user and to make sure that the user is able to make use of the product as intended and with a minimum effort to learn how to use it". When designing the interface, a focus is put on optimization of the user experience. This goes as far as directly consulting the end user to find out additional requirements. Hence, a strong accentuation is put on the user of the dashboard to seamlessly solve visual analytics tasks.

Important Considerations when designing dashboards for Big Data As previously discussed in section 2.4, important considerations arise when performing visual analytics due to very large data sets. Research by Wong et al. (2012) indicates that under these circumstances, various problems may occur which are hard to circumnavigate. One problem is that human cognitive capability might not be able to keep up with the growth of data available for visualization purposes, thereby posing a challenge to end users. Another problem depicts user-driven data reduction, in which the design of user interfaces for data reduction becomes increasingly difficult as new dimensions of data are integrated into visual analytics tools.

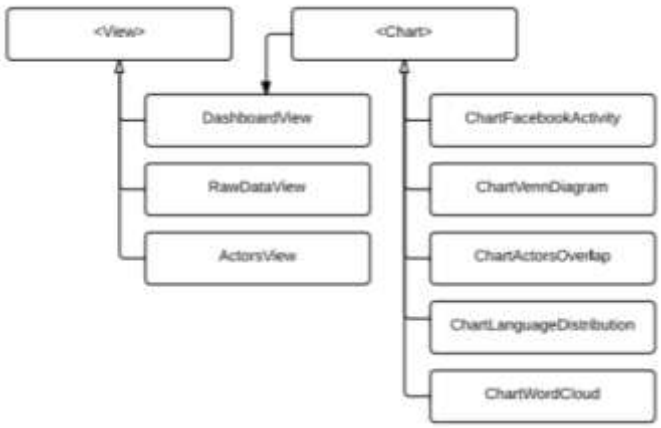

Figure 3.: Software Architecture

Views within the Dashboard Within the single-page web application, three different views are abstracted to a View metaclass which offers stateful navigation capabilities.

The implementations of the View meta class are:

- DashboardView: This is the main view of the web application which contains the Visual Analytics dashboard. It is initially shown to the user.
- RawdataView: This view presents a detailed search interface for the Facebook activity data. Many visualizations in DashboardView refer to RawdataView in order to provide the user with further information.
- ActorsView: This view presents a dedicated interface for analysis tasks related to Actor Mobility as depicted in section

The visualizations of actor mobility in Dashboard: View refers to ActorsView in order to provide the user with further details when requested. Furthermore, ActorsView presents a handy set of tools for analysis of actor mobility and cross-postings between different time frames and facebook walls.
Visualizations within the Dashboard The Visual Analytics dashboard depicted by Dashboard View includes five different visualizations.

5. CONCLUSION

In this thesis, a detailed examination of contemporary challenges in Big Data Analytics has been performed. Research has reached a point where social media activity is ubiquitous, yet hard to collect and analyze. In conjunction with complex event timelines as depicted by the Bangladesh garment factory disasters, the data at hand presents numerous opportunities for attaining deep insights. In this context, visual analytics present the means of reaching those insights to many users with different backgrounds, both experts and novices alike.
The novel implementation of a Big Data Visual Analytics Dashboard designed and developed in the course of this thesis showcases, that the creation of visual analytics software which meets the high requirements of present-day datasets is viable, and can be achieved by a single programmer with limited resources. Furthermore, the developed IT artifact leverages open-source visual analytics frameworks to a maximum extent in order to achieve a pure implementation of important concepts in visual analytics such as the detail on demand principle.
A thorough evaluation showcased the effectiveness of the tool's approach on visual analytics. Both the client- and server-side components of the Visual Analytics Dashboard present performance at par with commercial tools, and can seamlessly be used under many operational circumstances.
Additionally, the results of the user study performed during this thesis indicate, that the presented Visual Analytics dashboard combines a high ease of use with the ability of performing many different interactive analyses on a large dataset. Moreover, the Visual Analytics tool put forward may be utilized through any modern browser on a multitude of different devices and screen sizes, with visualization display times as low as in the hundreds of milliseconds.

REFERENCES

[1]. Abousalh-Neto, N. A. and Kazgan, S. (2012). Big data exploration through visual analytics, Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, pp. 285–286.
[2]. Abras, C., Maloney-Krichmar, D. and Preece, J. (2004). User-centered design, Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications 37(4): 445–56.
[3]. Bajaj, V. (2012). Fatal fire in bangladesh highlights the dangers facing garment workers, New York Times 25.

[4]. Bhavnani, S. K., Dang, B. and Divekar, R. (2013). Accelerating translational insights through visual analytics, AMIA.

[5]. Bostock, M. (2012). D3. js, Data Driven Documents . URL: http://d3js.org/

[6]. Bravo-Marquez, F., Mendoza, M. and Poblete, B. (2014). Meta-level sentiment models for big social data analysis, Knowledge-Based Systems . URL: http://www.sciencedirect.com/science/article/pii/S0950705114002068

[7]. Buhl, H., Ro¨glinger, M., Moser, F. and Heidemann, J. (2013). Big data, WIRTSCHAFTSINFORMATIK 55(2): 63–68. URL: http://dx.doi.org/10.1007/s11576-013-0350-x

[8]. Burke, J. (2013). Bangladeshi factory collapse leaves trail of shattered lives, The Guardian .

[9]. Dahl, R. (2012). Node. js: Evented i/o for v8 javascript. URL: https://www.nodejs.org/

[10]. Diakopoulos, N., Naaman, M. and Kivran-Swaine, F. (2010). Diamonds in the rough: Social media visual analytics for journalistic inquiry, Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on, pp. 115–122.

[11]. Gerasch, A., Faber, D., Ku¨ntzer, J., Niermann, P., Kohlbacher, O., Lenhof, H.-P. and Kaufmann, M. (2014). Bina: A visual analytics tool for biological network data, PLoS ONE 9(2): e87397. URL: http://dx.doi.org/10.1371/journal.pone.0087397

[12]. Haggerty, J. and Haggerty, S. (2009). Visual analytics of an eighteenth-century business network, Enterprise and Society . URL: http://es.oxfordjournals.org/content/early/2009/09/21/es.khp051.short

[13]. Hearst, M. and Rosner, D. (2008). Tag clouds: Data analysis tool or social signaller?, Hawaii International Conference on System Sciences, Proceedings of the 41st Annual, pp. 160–160.

[14]. Himi, S. A. and Rahman, A. (2013). Workers unrest in garment industries in bangladesh: An exploratory study, Journal of Organization and Human Behaviour 2(3): 49–55.

[15]. Hussain, A. and Vatrapu, R. (2014). Social data analytics tool (sodato), Advancing the Impact of Design Science: Moving from Theory to Practice, Springer International Publishing, pp. 368–372.

[16]. Ihrig, C. J. (2013). The express framework, Pro Node. js for Developers, Springer, pp. 189–204.

[17]. Islam, M. A., Deegan, C. et al. (2014). Social audits and multinational company supply chain: A study of rituals of social audits in the bangladesh garment industry, Available at SSRN 2466129 .

[18]. Kaczmirek, L., Mayr, P., Vatrapu, R., Bleier, A., Blumenberg, M., Gummer, T., Hussain, A., Kinder-Kurlanda, K., Manshaei, K., Thamm, M. et al. (2013). Social media monitoring of the campaigns for the 2013 german bundestag elections on facebook and twitter, arXiv preprint arXiv:1312.4476 .

About the authors:

Challa sneha working as an Assistant professor in department of CSE at Geethanjali college of Engineering and Technology affiliated to JNTU Hyderabad. She has 2+ years teaching experience. Her Research interests include Big Data Analytics, cloud computing and Artificial Intelligence.

Chandra Kala Anneda working as an Assistant professor in department of CSE at Geethanjali college of Engineering and Technology affiliated to JNTU Hyderabad. She has 2+ years teaching experience. Her Research interests include Big Data Analytics, cloud computing and Artificial Intelligence.