creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

Abstract

A Study on Topic Visualization Techniques for Story Telling with News Articles

Kun Koh (Kyle)
Department of Computer Science and Engineering
College of Engineering
The Graduate School
Seoul National University

With increasing amount of data that are published online, it becomes increasingly challenging to analyze and build narratives from the data, as well as producing graphics that support the narratives. Journalists express difficulties of performing such tasks during their news article production, because it often involves interacting with data experts and artists, and repeating the process multiple times. In order to alleviate this problem, it is necessary for journalists to collect their own data without the help from other experts and perform quick analysis to yield preliminary results, in order to see if the data potentially contain newsworthy information and determine if they need to be further analyzed either by themselves or by the professionals. Also, when the analysis is complete and newsworthy insights have been found, they need

to effectively deliver the findings to artists who produce graphics for the story to avoid producing irrelevant graphics or misrepresenting the underlying data. In addition, since the format of news articles is often limited by the publishing platforms, they are often restricted to producing visualizations that do not involve any interactive exploration of the data.

In this thesis, we present a design study of a tool that we have designed and implemented with the help from professional journalists to support the various stages in the news article production pipeline. Especially, we provide ways to collect, analyze, and visualize the information that are extracted from past news articles for storytelling using Wordle, a word cloud visualization method known for its attractiveness and aesthetic qualities [99]. We describe the interviews with the journalists about the challenges that they encounter in writing news articles such as the stress of interaction with experts or artists and the limitations that prevent them from producing interactive visualizations using the techniques developed over time in InfoVis community. Then, we describe the design process of NewsWordle, a tool that is designed to improve the overall news production. We evaluate the tool with journalists on the field with the case studies to see how NewsWordle can be utilized in their work-flow. We found that, by aligning the visualization techniques used for both analysis and creation of the visual prototypes for artists, the journalists can reduce time and effort involved in performing each task. We conclude the thesis with discussions for designing the tools that alleviate such challenges.

Keywords : Word Cloud, Text Visualization, Exploratory Search, Online Journalism

Student Number : 2009-23084

Contents

Abstract **1**

Contents **5**

List of Figures **9**

List of Tables **13**

Chapter 1 Introduction **15**

1.1 Background 15

1.2 Problem Definitions 17

1.3 Contributions 19

1.4 Organization of the Dissertation 20

Chapter 2 Related Work **21**

2.1 Producing Infographics and Data Visualization in Journalism. 21

2.2 Text Processing and Visual Analytics 24

2.3 Text/Topic Visualization 28

Chapter 3 Analysis on News Article Production **33**

3.1 News Article Production Pipeline Overview. 34

3.2 Challenges in the Current News Production Pipeline. . . . 36
3.2.1 A massive amount of information that needs to be explored 36
3.2.2 Increasing demand for news articles based on investigative journalism 38
3.2.3 Costly round trip between journalists and artists . . . 40
3.2.4 Production Cost of Interactive News Articles 42
3.2.5 Low visual literacy of readers 43
3.3 Limitations for Publishing News Articles. 44
3.3.1 News supply chain dominated by platform providers . 44
3.3.2 Mobile dominated online news consumption 46
3.4 Adoption of card-style news articles in Korea 47
3.5 Use of Wordle as Graphics 49
Chapter 4 ManiWordle for providing flexible control over Wordle 51
4.1 Analysis on the Original Wordle 52
4.2 Design Rationale. 54
4.2.1 Providing a compelling starting point. 54
4.2.2 Presentation words' importance with their size . . . 55
4.2.3 Reflecting users' intention as much as possible . . . 55
4.2.4 Provide fluent animation so users can follow changes . 56
4.3 Interactions 57
4.4 Implementation and optimization techniques 58
4.4.1 Adjusted rate of growth in spiral radius 59
4.4.2 Multi-thread optimization. 59
4.4.3 Collision detection with reduced-resolution texture . . 61

4.5 User Study Design 61
4.6 Controlled Experiment 62
4.6.1 Datasets and the Task 62
4.6.2 Participants 63
4.6.3 Hypothesis 64
4.6.4 Study Design and Procedures 64
4.6.5 Testing apparatus and setup 65
4.6.6 Results 66
4.6.7 Observations on the final layouts 69
4.7 Discussions and Implications 69
Chapter 5 NewsWordle **73**
5.1 Design considerations 73
5.1.1 The tool needs to fit into the production pipeline and enhance the process 74
5.1.2 The visualization needs to be of the consumable graphics while supporting analytics 74
5.1.3 The graphics has to be non-interactive and static . . . 75
5.1.4 The visualization prototype must be in high fidelity . . 76
5.1.5 Using news articles as data type. 77
5.2 Implementation 77
5.3 Database collection and management 79
5.3.1 Crawling data 79
5.3.2 Merging and Splitting Data 82
5.3.3 Processing Text 83
5.3.4 Exporting to Excel Formats 86

5.4 Summary View 87
5.4.1 Histogram 87
5.4.2 Range Slider 89
5.4.3 Gantt View 89
5.4.4 Binning 90
5.5 Cards 91
5.5.1 (Mani)Wordle 91
5.5.2 Words Tab 93
5.5.3 Style Tab 94
5.6 Publishing 94
5.7 Case Study 95
5.7.1 Crawling and Preprocessing 95
5.7.2 Filtering 96
Chapter 6 Discussions **101**
6.1 How NewsWordle Improves News Article Production. . . 101
6.2 Generalizations 104
6.3 Limitations 105
Chapter 7 Conclusion **107**
Bibliography **109**
Abstract in Korean **125**

List of Figures

Figure 1. An example of a news article production pipeline 17

Figure 2. NStreamAware [39] uses slices to show different date-time range, whose interval can be customized based on the data. Each slide contains simple statistics value and visual elements that can help users understand the underlying data. Juxtaposed layout of the slices help comparing adjacent slices. Especially, common visualization techniques across slices in the same position within the slices serves as small multiples that allows comparing multiple slices on the screen. NewsWordle borrows the concepts and allow configuring parameters for Wordles, not just individually, but across different cards. . 27

Figure 3. Naver [2] news article search on the keyword “Olympic” results in 2,880,723 articles, which makes it problematic for journalists to review individual news articles for excavating valuable articles and fact-checking. 37

Figure 4. Google Trend provides comparison of trends among various search keywords. Although it is possible see other related keywords associated with the search keywords, it is difficult to see how other associated keywords evolve over time. Data accessed on Nov 1st, 2016, http://trends.google.com 39

Figure 5. An example if an interactive news article from New York Times’

"Snow Fall". The mountain climbing map to the right updates automatically as the readers scroll through the story to show relevant information. 43

Figure 6. An example of a hosted news article by platform providers. The original news article was authored by the New York Times and is hosted on MSN, the platform provider's website. Such hosted articles are limited to providing interactive content within the platform providers policies and guidelines. 46

Figure 7. An example of a Card-Style news article that adapted the concept of *Card-Style UI* to enhance mobile news consumption experiences. Two cards are shown here and the third card can be accessed by swiping the screen to the left. Depending on the size, one or more card may fit into a screen. It also provides a book-style narrative as the story progresses by moving into the next card. 48

Figure 8. Main overview of ManiWordle showing a word cloud for the title and abstracts of All InfoVis Papers published since 1995. The word 'data' is selected and being rotated by dragging the circular handle on top. 52

Figure 9. All unpinned words fade away to the background when a mouse cursor hovers over the re-layout button (Top). After clicking on the re-layout button, unpinned words are re-arranged to form a packed cloud (Bottom). 58

Figure 10. The radius of spiral and the interval of collision checks are larger when dealing with bigger words compared to smaller words. . 59

Figure 11. (A) The original layout. (B) A user drags WordA to the top of other words. (C) The placer thread determines where WordB should go on a spiral. (D) WordB moves. 61

Figure 12. Task completion time (in average) for three text datasets. Error bars represent standard error. People spent significantly more time

creating a Wordle for their paper than the Wordle paper. . . 68

Figure 13. The final layouts produced using ManiWordle (Top) and Wordle (Bottom) by a user. The text was a Wikipedia entry on Yu-Na Kim. 71

Figure 14. Words from a participant's paper clustered based on their semantic meanings by the person using ManiWordle (Top). The layout by the same person using Wordle (Bottom). 72

Figure 15. An overview of NewsWordle application showing all news articles published published by Yonhap News Agency between Jan. 1st 2016 and Dec. 20th, 2016, on the term 프로야구(Professional Baseball). 79

Figure 16. A news office (news media) selection screen. Selecting multiple news offices may diversify the database, but the amount of time for downloading and processing the data increase. 81

Figure 17. The first screen for a database builder wizard. Users may specify the file, a keyword for search, and the date range for the news articles. 81

Figure 18. A windows showing the crawling process. Users may see the progress by identifying the number of the news articles currently being processed (to the left) and the processbar at the bottom with the estimated remaining time, which may depend on the network bandwidth and the processing power of the users' computer. . 83

Figure 19. Database merge wizard helps the users merge two databases into one. The first database (with the keyword, Hilary Clinton) is already loaded in the main window. The user can merge the database with the keyword, Donald Trump, and make a new database labeled 'Presidential Election.' 84

Figure 20. A word filter page. The users may create a global filter, or database specific filters and normalization rules. Also some commonly applied

filtering rules are provided for convenience. A drop-down menu for a word (2) shows that two words (‘대통령’, and ‘대통령의’) are normalized into ‘대통령’. 86

Figure 21. A basic statistics can be views by placing a moue cursor on to a historgram. 87

Figure 22. The tick marks can be dragged, added, and removed in order to adjust the interval of date-time range. 88

Figure 23. The histogram view can be configured to either stacked graph or grouped bar chart mode. 88

Figure 24. When the mouse cursor enters the colored boundary of the ManiWordle control, the card is enlarged in order to help manipulate the words’ positions. 92

Figure 25. When the mouse cursor is placed on top of each word, the same words in the other cards are highlighted. Also, the users can see the number of occurrence for the words in the news articles within the card’s date-time range. 92

Figure 26. The style tab allows users to change attributes of the ManiWordle. 93

Figure 27. An initial screen upon loading the database crawled with the keyword *Oh Seung Hwan(오승환)*. 96

Figure 28. Most common words are added to the global filter list. Also, any words that are shorter than 2 letters are removed. The only numeric word that were prominent were ‘34’ which is the age of Oh Seung Hwan. It shows up because of the convention of news articles that states someone’s name followed by their age. However, it is not newsworthy. 97

Figure 29. As the result of filtering and adjusting date-time range, 6 Wordles were generated. 98

List of Tables

Table 1. Subjective Responses to Six Questions (Average Ratings). The questions with significantly different ratings are marked with an asterisk (*). 66

Chapter 1

Introduction

1.1 Background

Since the introduction of the World Wide Web, news articles have been one of the most important contents. Especially, with the help of advances in data analysis techniques and tools using computers, many news articles report the information or insights extracted from the data. However, journalists often rely on other data experts to collect, analyze, and extract insights from the data for them, due to the difficulty and the lack of knowledge. Such process costs great effort and time due to the iterative nature of interacting with the others and even cause monetary expenses. Thus, journalists may not acquire necessary and newsworthy insights for news articles and are often left with the raw data that are yet to be thoroughly explored. Another challenge for journalists is producing appropriate infographics for news articles.

Although a teaser image or two may satisfy the news articles that do not have any statistical contents, any articles that quote statistical data or analysis require charts or visualization in order to convey the message more clearly and reduce the cognitive burden for readers. This is especially challenging in mobile environment, where any interactive visualizations that require elaborate manipulation of the object on the screen are not appropriate. Although journalists often rely on artists (or graphic designers) in order to offload the burden of producing good graphics, the current pipeline where the journalists request for figures may introduce another difficulty due to the reasons such as miscommunications in delivering journalists' intention or the difference in philosophy of using excessive visual embellishment, because it often comes down to choosing from tradeoffs between usefulness of char junks and integrity of the chart [7][38][44]. Therefore, journalists want to minimize the number of roundtrips requests for figures between artists and themselves. One way to achieve such goal is for journalists to produce high-fidelity prototyped graphics that the artists can refer to, so that the journalists' intentions are well delivered.

In addition to that, journalists express their fear that the news outlets are not getting enough 'clicks' on their websites, which directly correlate to their revenue, because large web portals, search engines, newsfeed apps, and social media have become primary channel for news consumption for many people, rather than visiting individual news media's website. This also exerts another problem that, if such platforms do not allow certain types of content, for example having many interactive elements, due to various reasons like security or overall performance of their services,

the journalists are only left with an option of producing old *long-text-with-photos* style articles. Therefore, all visual information has to be coded into rather simple static graphics to deliver messages.

1.2 Problem Definitions

Many individuals and organizations have developed their own version of news article production pipelines. Figure 1 shows one example of a news article production pipeline, derived from the interview with participating journalists. This is in line with the pipelines introduced in other literatures such as [101]. Through this thesis, we identify the problems that occur throughout the pipeline and try to alleviate the issues by providing a tool that can help streamlining the process.

First, we identified that the journalists are often blind to the data during the second

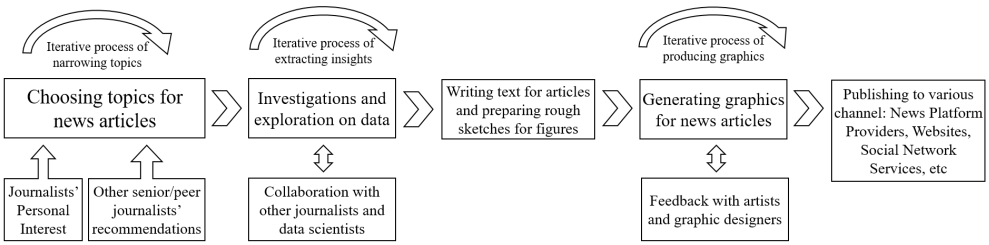

Figure 1. An example of a news article production pipeline. The first stage involves brainstorming phase where journalists choose topics for news articles. On the second stage, they collaborate with other journalists or data scientists to iteratively investigate and explore the data in order to extract newsworthy insights. On the third stage, they compose textual content of an article along with prototype sketches of the graphics to be inserted. On the fourth stage, journalists collaborate with artists and graphic designers to generate the figures. The figure may be purely aesthetic, purely functional (like plain charts), or both. On the last stage, they publish their articles through various publication channels.

stage of the pipeline in Figure 1, because journalists may lack in skills or resources to collect and analyze the data. Instead of providing extensive analytic capabilities, it is more important for journalists to quickly set up hypotheses about the data, collect, process, and analyze the data quickly and lightly in order to determine if the data may contain newsworthy insights. If the data only seem to contain noise, they can analyze set another hypotheses or try analysis on another sets of data. By doing so can they reduce the number of requests made to other journalists or data experts.

Second, we identified that the graphics generated by artists do not always reflect the journalists' original intention, as the graphics are often "created by the artists themselves after reading the textual content of the news article," or "from rough prototype sketches drawn by the journalists." Such miscommunications may result in multiple iterations of revising the graphics. In order to deliver the journalists' intention more effectively, the prototype graphics must be in high-enough-fidelity to convey all information. However, the limited resource and time for journalists for creating the prototypes lead to possible misinterpretation by the artists.

Third, because of the limited resources and time, the transition from the data analysis results in the second stage to the prototype graphics in the fourth stage of the pipeline needs to be streamlined. If the analysis results are in a meta-format (showing statistics in a tabular form or with the complex charts that are inappropriate for the news article figures), journalists then have to create a new prototype from a scratch, hence reintroducing the second problem described above (i.e. the journalists are required to make a prototype in another simpler form based on the analysis results).

Lastly, the graphics generated in the fourth stage of the pipeline are often limited to non-interactive forms because the news articles are often consumed through the publishing platform providers' websites such as Facebook [1] or Naver [2]. Also, the increasing use of mobile phones introduced a new challenge to graphics, because the interactivity and the visual complexity of the graphics have to be reduced compared to those of the desktop. Therefore, journalists prefer more static graphics to highly interactive visualizations that allow readers' own exploration of the data.

1.3 Contributions

In order to address these challenges, we setup an iterative design study that we have conducted with professional journalists. We first conducted interviews with them in order to understand how online news articles are produced and consumed. From there we extracted design requirements that reflect the problems that rises from the current production pipelines and consumer trend. Then, we designed a tool called NewsWordle to facilitate their process by providing efficient way to explore the digital news archive. Because data exploration strategies may greatly differ depending on the type of news article contents, we decided, after discussions with the journalists, to focus on producing the articles that reflect on other news articles published in the past to understand changes in macro topic trends. Also, we limited ourselves to production of text and photo based news articles, for they are the most frequently used format. NewsWordle helps journalists collect their own data in order to allow them to quickly evaluate if the data can produce newsworthy stories. If not, they can quickly collect different dataset without the cost of communication with data

experts. Also, by incorporating the data analysis methods into the final or near final graphics (which was a form of Wordle in our case, hence the term, NewsWordle) to be embedded in news articles, we alleviate the problems of producing separate prototype graphics based on the results of analysis from other tools. The professional journalists who participated in our designed expressed positive opinions towards the tool for allowing them to get involved more into both data analysis and graphic productions, resulting in smoother communications with their respective experts, while adding only little burden for themselves.

1.4 Organization of the Dissertation

The rest of the thesis is organized as follows. Chapter 2 presents literature review on text analysis, studies on infographics used by journalists, and text/topic visualizations techniques relevant to our study. Chapter 3 describes the challenges that journalists have to face when producing and publishing news articles due to current trend in online news consumption. In Chapter 4, we present ManiWordle, a manipulable word cloud generation tool that provides foundation for visualization changing topic trends in NewsWordle. In Chapter 5 describes design and functionalities of NewsWordle, a tool that we have built to meet the requirements. We also demonstrate how NewsWordle can help journalists explore and process news archives with real life example. In Chapter 6, we discuss design guidelines for building a tool that supports journalist produce text visualization for news articles with current limitations to their platforms. Finally, we finish the thesis with conclusion and possible future directions for the research.

Chapter 2

Related Work

In this section, we review previous work on infographics/journalism, visual analytics for text data, and text/topic visualizations.

2.1 Producing Infographics and Data Visualization in Journalism

The long attempt to overcome limitations of one-way communication channel from traditional mass media has been extensively studied. Although there are many non-digital approaches explored in the past, we focus on the use of infographics and text visualization techniques since the introduction of the web and mobile devices. Especially, we explore on how infographics and text visualizations are generated by the creators and how they are consumed.

Although graphics and charts always served a crucial role of conveying information or messages and helping readers understand the textual content of the news articles in the past, the infographics in the current generation of digital journalism are supported by computers and became interactive in order to amplifying cognition [17]. While infographics borrow many elements and techniques from the field of information visualizations for scientists and scholar to analyze complex data, the main purpose of the infographics is to tell a story through various media [10][91]. In order to understand how the viewers interpret data via visualizations, researchers have investigate how they capture the viewers' attention, how they guide the viewers into reading the visualizations in certain way, and the psychology behind using the visual narrative. For example, using visually salient features [51][91] may attract viewers' attention at a glance and may guide their eyes to read the visualization in the way they were intended. Chan et al. [19] adopted such strategy for making information dashboard to help viewers follow narrative flow of different visual elements and charts. When journalists massage data, NewsWordle uses transitional animations of words as salient visual features help them understand the changes reflected on Wordle. The production process of such graphics include, but is not limited: brainstorming and creating ideas for graphics; making low fidelity sketches; investigating and collecting relevant data and information; developing more high fidelity sketches; iterate over the design; and decorating and visual embellishment and styles [46][59][95]. It should be noted that these processes are not always performed by one person. People may take different role in the production pipeline in order to maximize

output and increase efficiency. Weber et al. [101] conductive extensive interview with various experts across different roles such as graphic designers/editors, art and design directors, programmers, and journalists currently working in the field. They have found that many media companies' production pipelines are still split into different roles. That is, the journalists (or the reporters) draw a big picture of the news article and coordinate the artists and the programmers to produce appropriate graphics. However, New York Times showed the shift in the traditional paradigm because everyone considered themselves as journalists, as the boundary of their roles is not as clear. Also, they would spurn from creating visualizations that might compromise the integrity and accuracy of the data [7][38][44], while still pursing for aesthetically pleasing images.

In the process making visualizations, many graphic designers and artists do not benefit from the plethora of tools that were built to support creations of visualizations. Bigelow et al. [10] explained that they rely on low-level illustrative tools that give them maximum flexibility, because such flexibility gives designers more freedom in terms of forms of graphics they can produce. However, manually encoding large number of data without the help of these tools may lead to inaccurate portrait of the data, and because they have already invested much effort in plotting data points, it is difficult to go back once the process goes beyond the 'point of no return,' paradoxically reducing the flexibility. Also some more flexible visualization toolkits may require non-visual skills like programming like Processing by Reas et al. [85], D3 by Bostock et al. [11]. In designing NewsWordle, we wanted to make sure that

the journalists actually understand and utilize the tool by inviting them into a designing phase of the tool with a series of interviews and pilot sessions to test the tool in the news article production.

Lastly, the data cleaning process done by the journalists were often dominate by commonly available tools such as Openrefine by Ham et al. [47], Wrangler by Kandel et al. [54]. However, our own interview with a journalist suggested that the primary tool for data processing still remains to be Microsoft Excel.

2.2 Text Processing and Visual Analytics

While there are many text data analytics techniques developed in various field of research such as topic extractions, natural language processing, machine learning, and artificial intelligence, we review those that were related to visual analytics for text data found on the web, such as feed archives from social network services or news articles, since they are related to changes in topics over both short and long period of times and the data is consistently streamed, the type of data used by NewsWordle.

Hierarchicaltopics by Dou et al. [33] proposed topic hierarchical modeling methods using Topic-Rose-Tree, a method for performing hierarchical clustering, and used it for their visual analytics tool. SentiView by Wang et al. [100] performed sentiment analysis (also referred as polarity extraction in various research) on text data by combining uncertainty modeling and model-driven adjustment and use them to visualize the changes of attributes and relationships among the users of social networks. Sentiment analysis are largely used for opinion mining for product reviews or political views as in Twittercrowds [5], Access [8], Textwheel [28], Liu et al,'s [67],

Nguyen et al,'s [80], and Opinionflow [103]. Zhao et al. [106] proposed a method for analyzing sentiment based on emoticons. The effective techniques shine when analyzing data from the Internet forums and social network feeds, but not for news articles written by journalists for publications since they do not extensively use emoticons in their contents. Determining sentiment of both news articles and readers' comment will provide deeper insight on the data.

In order to extract keywords that consist of more than one word, Named Entity Recognition(NER) [21] technique is used. Because NER can be achieved in many different forms, many research adopted similar techniques, while often augmenting one or more rule that satisfies the characteristics of the data that they use. For example, Cui et al. [28] performed NER by applying a series of different rules to the original text data, such as Part of Speech (PoS) tagging, syntactic tagging, proper noun phrase classification, rules processing, alias expansion, and geographic normalization. Heimerl et al. [49] used Stanford's Core NLP [74]'s NER features that tokenizes, splits sentences, tags PoS, lemmatizes, and applies other miscellaneous rules. Also, although many research do not specifically mention the term NER, their text processing methods often involves different mixtures of NER techniques as in the example of NStreamAware by Fischer et al. [40] (Figure 2). Because NER can be influenced by the pre-defined set of words(dictionary) to get accurate data in specific domain, we decided to use the entire set of Korean Wikipedia's page titles as named entity for NER in NewsWordle, assuming that Wikipedia's page titles contain named entities and proper nouns such as organizations, people, countries. More details on

the processing is described in Chapter 5.

Categorizing text data can be done contextual clustering based on topics as well as temporal clustering. EvoRiver by Sun et al. [97] used manually labeled Twitter feeds to train Support Vector Machine and labeled the rest of tweets. Also by cross referencing the tweets with the authors and followers, they were able to how followers are influenced and attracted/distracted followers from certain topics. TwitInfo by Marcus et al. [75] focused on event detection by allowing users to specify the keywords to watch, and detect when tweets containing keywords or relative words spike in volume by comparing it with historically weighted running average, to compute data entropy from which they can be considered events. Also, Chuang et al. [24] proposed evaluating keyphrase extraction using statistical and linguistic model and evaluation method based on crowdsourced ratings to choose more relevant keywords and keyphrases that reflect the original text documents.

In order to facilitate comparison tasks supported by visualization, different strategies such as juxta-positioning, super-positioning, and encoding the values with other attributes are used [45]. The use of juxtaposition of visualizations is demonstrated by many previous works including Cui et al.'s [30], Compare clouds by Diakopoulos et al. [31], and VCloud by Lira et al. [66] juxtaposed word clouds from two different corpora to compare how each corpus represents the keyword by highlighting sentences from which the corpus was built, as well as frequently co-occurring words within the sentences. Lohmann et al. [71] used the concentric circle metaphor by placing most frequently co-occurring words from multiple documents in the center,

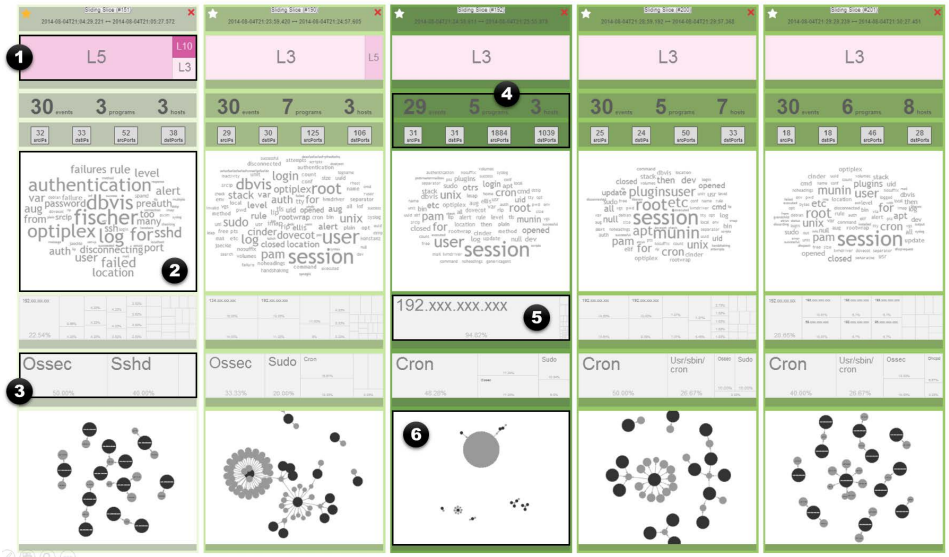

Figure 2. NStreamAware [39] uses slices to show different date-time range, whose interval can be customized based on the data. Each slide contains simple statistics value and visual elements that can help users understand the underlying data. Juxtaposed layout of the slices help comparing adjacent slices. Especially, common visualization techniques across slices in the same position within the slices serves as small multiples that allows comparing multiple slices on the screen. NewsWordle borrows the concepts and allow configuring parameters for Wordles, not just individually, but across different cards.

and each circular layer showing only the text shared by fewer documents, and the outer-most layer showing only the words that are prominent in each individual document. Perhaps the most relevant layout strategy for juxtaposing visualizations for each date-time range like NewsWordle is NStreamAware by Fischer et al. [40]. By representing each date-time range into a slice (Figure 2) (similar concept is called ‘Card’ in NewsWordle in order to match the concept of making of card-style news described in Chapter 3.4) with statistics and visualizations in order for users to

compare the slices. Since it can process stream of information in real time, new slices can be continuously added. Also, it supports merging slices in order to cluster them if adjacent slices have similar attributes. Because NewsWordle's primary users are journalists and it tries to seamlessly combine both analytics task and creation of animated Wordle that can be embedded into an actual news articles, the Wordle visualization tries to maintain aesthetical quality for readers pleasure while still supporting analytics task with effective animations [89]. Other example includes supporting comparison task is by embedding other visualization elements such as line chart or histogram into the word text's background as in Lohmann et al's [70] and Nguyen et al.'s [79].

2.3 Text/Topic Visualization

There are large number of text/topic visualization techniques that are designed to help users analyze the text data for various purposes. Many research used the word cloud or word embedding and encode data with one or more parameters like the font size, font families, colors, or position [13].

For example, ManiWordle [58] uses font sizes to represent the word's frequency, but not colors and positions. Also since it is designed to summarize a set of documents as a whole, rather than based on different categories or clusters, it does not visualize how topics were changed over the passage of time, unlike Chi et al. [22]. In order to provide analysis on time varying text/topic data, many research used stream/flow-themed visualizations with the concept of the stacked graphs/histograms that are connected among timelines in order to show evolution of topics over time in

ThemeRiver [48] style. For example, Byron et al. [15] explained different strategies and guidelines for generating stacked graphs to achieve scalability and aesthetic design. Dork et al. [32], Leskovec et al. [61] utilized stacked graph that shows stacked layers for each topic and their thicknesses to represent how important or frequent the topic was discussed. Luo et al. [72] used thickness over vertical accessed, but shoed topics in separate 'bubbles' instead of stacking them. Tiara by Wei et al. [102] also borrowed the concept of stacked graph to visualize topic clusters as layers and embed small word clouds within the layer that shows related words. This helps increase the space utilization. However, there can only be little room for the words on the narrow layers, if there are too many topics stacked at once, or the topic is relatively less prominent in the data.

TextFlow by Cui et al. [29], Rose et al. [90], Riehmann et al. [87], OpinionFlow by Wu et al. [103] and Xu et al. [105] utilized stacked graph, but allow layers to be merged into one (making a unified topic layer) and split as the topics get separated over time. While they can provide good overview of how different topic keywords evolve over time, it can still suffer from the lack of space for related words to be positioned. In contrast, NewsWordle provides uniform size canvas for each word cloud generated, while supplementing it with summary histograms for documents to show how actively such topics are discussed by the news articles. Archambault et al. [4][5] also took similar approach with ThemeCrowds/TwitterCrowds. They also split documents into different date-time ranges and provide visualizations for each into a treemap with words embedded that has been built by performing hierarchical

clustering. While NewsWordle does not provide such hierarchical clustering within a date-time range, it supports merging sets of documents collected by multiple topic keywords (used for article search) into one data, providing more flexibility for combining multiple search results, the feature often required by the journalists. Also NewsWordle helps understand the changes of topics using transitional animations.

In terms of the size of the word fonts, Collins et al. used word frequencies because it would begin to approximate the distribution of such words as the number of documents used for analysis increases [27]. Especially, domain specific terminologies are often repeated in the text and not substituted with synonyms, making the terms appear more frequent than other general vocabulary, making the frequency suitable for representing the topics embedded in the text. Bernstein et al. [9], Cui et al. [30], Dubinko et al. [36], Fisher et al. [39], Jo et al. [53], Koh et al. [58], Lee et al. [63], Malik et al. [73], and Wu et al. [103][104] also adopted similar strategies. Especially, the frequency of the words was often measured within subsets of documents based on clustered results. Buchin et al. [14] and Paulovich et al. [84] used both frequency and the spatial coding of the word in order to maximize the coverage of the geometric figures (maps), while still maintaining the size encoded data represented by the font size of the word. Also, Chi et al. [22] changed the size according to the contour of the word clouds that morphs into other form over time in smooth animations while still maintaining the relative sizes(importance) of the words. In NewsWordle, the sizes of the words were determined by the frequency of the nouns and adjectives words within a subset of the news article texts filtered by the date-time range in order to represent

the time varying topics of the news articles based on certain events (such as sporting events, plane crash, election, or wars and terrors) or a passage of time on longer periods like how topics on the cost of electricity changes based on the weather condition. Also, the distribution of the frequencies is not consistent for different document sets, we provided adjustable relative sizes of the words between the minimum size and the maximum size to make sure that the resulting word cloud is not dominated by a very few much larger (more frequent) words. While the absolute size of the same topic does not carry across different time periods, a secondary chart shows absolute differences in frequency of the words.

In this regards, the similar techniques are used when there are multiple word clouds generated simultaneously. Cui et al. [30], Fischer et al. [40], Ng et al. [78], Wu et al. [103][104] used multiple word clouds in order to facilitate the comparison of topic among different date-time ranges, while Kling et al. [57] allowed the comparison of different word clouds generated by the documents associated with different keywords. NewsWordle takes both approach by allowing journalists crawl news articles based on keywords they provide and generate multiple word clouds that represent each time period. There are also other metrics for measuring the importance of the words such as word weight calculated by LDA in Tiara [102].

Previous research also invastigated how word-cloud layouts are generated. Began as an attempt to visualize tags from the web [36][88], word cloud has been used in both scientific analysis and casual uses. Seifert et al. [92] suggested more packed word cloud design by scaling font and bounding box to make more aesthetically pleasing

word-cloud, which were later enhanced by Wordle by Viegas et al. [99] by allowing words' bounding boxes to be overlapped, as long as the actual contour of the letters (inked region) are not overlapped. As beautiful as they were in order to breakapart and go beyond the traditional goal of information visualization to support data analysis through visual means, the loss of semantic meanings among words hurt the robustness of the tool for analyzing data. Especially, besides using the font size to represent frequency, no other attributes are encoded by the parameters such as absolute/relative locations of the words, colors(hue), intensity, and rotation angle, as well as typefaces such as like decorations (underline, strike through), weight(boldness), style (normal, italic, oblique), and font families. There have been attempts to give semantic meaning to the locations of the words such as Barth et al [6], Cao et al. [16], Cui et al. [30], Diakopoulos et al. [31], Liu et al [68], Liu et al [69], Wu et al. [104], based on various weighting models that defines distances among words for relative positions or other encoded values such as timeline and clusters for absolute positions. Also, Kim et al. [56] took determistic approaches to avoid the issues of randomness of current word cloud layout algorithms. Chi et al. [22] tried to maintain the location of each words as the word cloud evolve over time in order to help viewers maintain the context and Gansner et al. [43] used graph theories to cluster words and produce packed layout. Also, Strobelt et al. [96] and took similar layout strategies as Wordle, but enhanced performance using various heuristics and modifying greedy approach of selecting initial positions of words before being placed on a canvas.

Chapter 3

Analysis on News Article Production

In this chapter, we describe the problems and challenges that journalists face while producing and publishing news articles in online space. We conducted interviews with two professional journalists who have been working in the field for 9 and 20 years respectively. They work for one of the major news agencies in Korea and are frequently engaged in dealing with data analysis, infographics design, and authoring news articles. As previously mentioned, we mainly discuss production of text and photo based news articles, the most common format. There are other types of news articles based on web 2.0 technologies, such as HTML5 and online streaming videos, whose use among editors is limited due to the difficulties described below.

3.1 News Article Production Pipeline Overview

Although the process may differ depending on culture within the community or company and types of news articles being produced [101], we describe one common example of news article production pipeline (Figure 1) based on the interview with the participating journalists and the literatures [46][59][95]. In the first stage, journalists are in constant review of possible topics for news articles. When there are particular trending events such as elections, sporting events, natural disasters, or accidents, they often get their ideas from news room meetings and other sources like internet, or even from the field. However, this stage serves as a brainstorming stage and the journalists may get inspirations from their personal interests. For example, if the cost of electricity is under discussion among the public during summer due to air conditioning cost, s/he may be interested in how hot summers in the past few years have been, which can turn into a narrative for a news article or a column. This process does not always have to happen on a meeting table. They might even receive recommendations for idea from senior journalists or colleagues within a company. Or, if s/he cannot find time to investigate the story, does not have capabilities like authorization for access to certain information, or simply does not have technical skills, s/he may recommend the idea to other people or seek for help from other peers and even interns. In addition, idea may come from reading other news articles from various sources, like the internet or other printed news media and television.

Once the idea is selected, journalist perform preliminary examinations on what they need to know for writing the news article. The sources for investigation range from peers to previous news articles on the same topics and simple internet searches, and

to the indexing company or the organizations' archive. If the idea is found to be feasible for a story, they launch full investigations. The idea may also be reviewed by the news room to see if it will make newsworthy stories or if it can conflict with the interest of the group.

At this stage, journalists may massage the data and information they acquired either by themselves or with other experts and professionals. For example, if the data analysis requires domain knowledge on specific topic, they may contact people that are familiar with the matter. Or if the data needs to be analyzed by more advanced techniques than the journalists can process, data scientist may provide advises. Many news media companies have contract with data analysis and mining firms that offer such services, or have an in-house data experts group that may assist other journalists for processing data.

Once enough insights and meta-stories have been collected, journalists write the news article's text and prepare for rough sketches or descriptions on the figures that need to be inserted into the article. Although journalists may create some figures, they often prepare outlines and descriptions for them and pass it onto artists (graphics designers), who will then produce images. If the figures contain data, rough sketches and the type of charts may be listed in the description. Sometimes, if the charts may become complicated, the raw data in a tabular form may be attached. The miscommunications can increase the cost of time and effort to reiterate the creation process for the graphics.

Once all materials are prepared, the news article is published through various

channels like online news platform providers, the company's own website, other news agencies or media, social network services, printed media, or even a news bot for messenger services. We designed NewsWordle to blend into the current production pipeline in order to enhance various stages and facilitate their process.

3.2 Challenges in the Current News Production Pipeline

3.2.1 A massive amount of information that needs to be explored and communications with data experts

One problem that journalists face in the news article production is that the sheer amount of information that needs to be process also increased compared to the past. Although advances in digital archives and search engines allowed journalists to easily search previously published news articles and contents based on keywords, the number of articles to review increase over time, leaving little time for them to examine individual articles (Figure 3). In addition, such massive amount of data returned from search result can result in low signal-to-noise ratio. Also, traditional search results do not reveal how theme around the keywords evolved over time. Although Google Trends (Figure 4) shows comparison among keywords, their theme, often associated with the keyword cannot be compared. For example, the search results on a keyword 'Olympic' may reveal that the events covered by the news articles change hour-to-hour as the time progress. Also, news articles may begin to cover specific athlete at certain point as s/he wins the event with a medal. At current stage, these types of trend analysis from news archive using statistics are difficult to achieve and may take long time to find meaningful insight.

Figure 3. Naver [2] news article search on the keyword “Olympic” results in 2,880,723 articles, which makes it problematic for journalists to review individual news articles for excavating valuable articles and fact-checking.

Because the data analysis can be costly with massive amount of data, journalists often rely on other people that expert in data analysis. Often times, the data experts are not hired in-house by the news agencies, but on a contractual base. The participating journalists argued that this introduces additional burden of communicating with the experts. For example, the journalists do not always have resources to check which data may contain the newsworthy information. Also, it is not always clear where or how the data may be obtained, if it requires programming skills like crawling that involves using APIs or pure web-crawling techniques, if not extracting from the

database systems. Also, the data cleaning and the filtering may not be always trivial. Therefore, journalists often make wild “guess” or hypotheses on what they may find from the data and discuss with the experts on how to collect and process. After a set amount of time, the data experts then report back to the journalists who make a request and explain any insights that they were able to extract, if at all. If the data do not contain any newsworthy insights, then the journalists have to either give up, or discuss on another round of iteration. This may significantly increase the time of production, a very important factor for many news articles.

3.2.2 Increasing demand for news articles based on investigative journalism

Although they began as simple replacement from the traditional paper based media such as newspapers or magazines, the digitalization of the article text opened a new way of exploring the archive with much less effort than it was required before. Searchable digital news archive provides various benefits to both consumers and producers. Such benefits come from being able to efficiently search for keywords and filter less relevant information for researching or fact-checking purposes. In order to facilitate this process, many methods were proposed to improve various stages of the entire process of news article creation. For example, natural language processing helps organizing texts structured form as well as extracting other metadata associated with the text such as sentiment analysis. Especially, with the introduction of Web 2.0 [3] where readers write comments and share the news articles through social network services with their acquaintances, network analysis on such comments and social

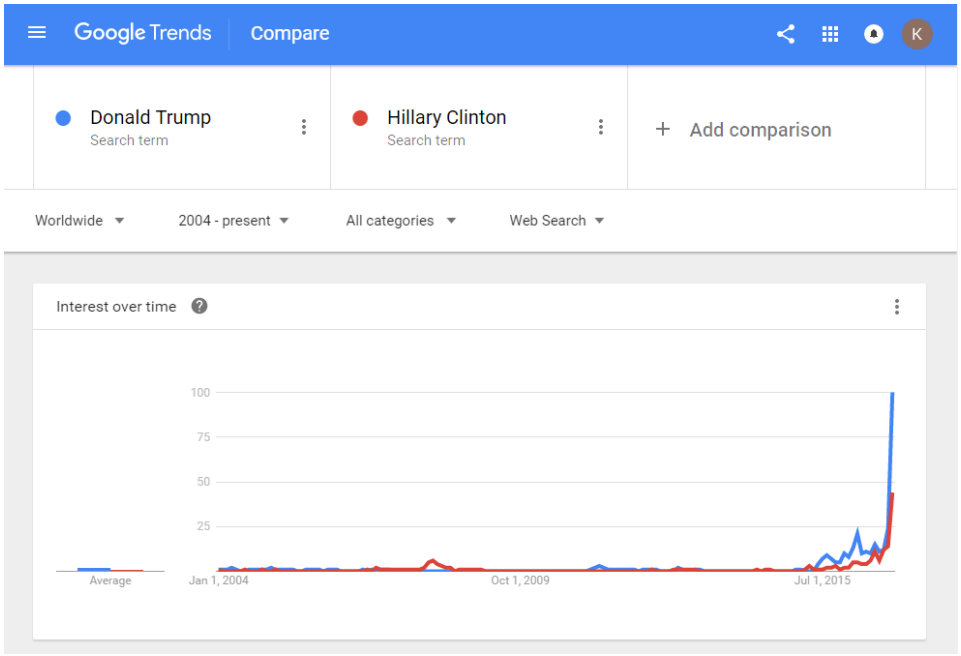

Figure 4. Google Trend provides comparison of trends among various search keywords. Although it is possible see other related keywords associated with the search keywords, it is difficult to see how other associated keywords evolve over time. Data accessed on Nov 1st, 2016, http://trends.google.com

shares may help find indication for repercussion that the articles may have. Major news outlets often have more than a thousand employees in editorial positions and the number of articles produced exceeds several hundred per day because online news articles are not limited by the spaces on the print paper [77]. Therefore, with increasing amount of online news articles, topic extraction techniques have become important to help summarize massive amount of text data into consumable sizes.

One participating journalist anticipated such investigative journalism based on

exploratory data will become more attractive to journalists, especially when the robot journalism [26][60], already began producing many fact-base news articles such as results from sporting events or hourly or daily stock market trend analysis. For example, Associated Press begin publishing most of their news articles for Minor League Baseball games using metadata extracted from the games [76]. This change results in covering more Minor League games than it was possible before with human labors. A participating journalist who have originally come from baseball statistics background felt the human journalists' shifting focus to other narratives that can only extracted with significant amount of investigation in the field or on the data. Therefore, the human journalists will be more suitable for producing investigative journalism with narrative talent, experience in the form of the editorials or special columns, based on domain knowledge and human insight/intuition. In order to achieve this goal, a concept of computation journalism [41] using data mining algorithms and artificial intelligence, along with tools designed to facilitate these process performed by the experts is essential to enhance the production of such news articles.

3.2.3 Costly round trip between journalists and artists

The participating journalists have expressed the difficulty of maintaining iterative design process of figures that supplement news articles. Although it is possible for the journalists to create their own figures, major news media has a separate department that specializes in creating figures and charts along with other visual embellishment for news article figures. The journalists make rough sketches for

highlights and charts, along with summary of the news article for the artists (or graphic designers) to generate appropriate images and graphics. This division of labor can be resourceful and cost effective, because few talent artists can serve multiple journalists. However, the problem arises when there are communications gap between the journalists and the artists. Perhaps the best way to reduce the gap may be for the journalists and the artists to pair up in a physical location and co-produce the content. However, their communication channel is often restricted by technical limitations such as using e-mails or instant messengers, and time it takes for journalists to make the rough sketch that is "crude, but good enough" to convey their intention to artists. Also, artists do not just produce images based on the requests. Because the concept arts they receive are usually rough and leave a room for artists to exercise their creative instinct to improve upon the concept, the end result sometimes may deviate away from the original request. More experienced journalists develop their own strategy to alleviate such problem. For example, one participating journalist said when s/he receives the figures and the graphics s/he requested from the artists, s/he quickly reviews them to see every idea is reflected. If not, s/he has to assess the error and make a decision. If the 'error' merely comes down to minor preference issue, but does not severely damage the integrity of the article such as using colors that journalists do not personally like for certain parts of the graphics, the journalist might not make any further changes and accept the result. If the error is minor and can be edited by the journalists her/himself, s/he now has to estimate the time it takes to make another request to the artists for fix or time and effort it takes to fix it by

her/himself. Since many news stories are time-sensitive and lose their values after certain times, journalists are in constant dilemma of choosing expedition over perfection.

3.2.4 Production Cost of Interactive News Articles

With the Internet boom of late 90's, people have discussed the possible usage of web technologies that enable (such as Java Applet or Macromedia Flash then, and now HTML5) that allow journalists produce a new form of news articles that involves interactive storytelling. However, despite numerous excellent example of interactive narratives, vast majority of the news articles published online remained the traditional style of text and photos. Although interactive news contents engage more audience [64], and news media use them as ways to 'show off' their ability to exploit new technology, the ROI(Return on Investment) is not always on par with the traditional news articles. This is also related to the limitations in publishing the articles, which we will explain in the following section. Also, the amount of labors that goes into creating such articles is often much higher, because they require participation from domain experts, data scientists, cognitive scientists, and graphics designer as well. For example, although a Pulitzer-winning interactive news article "Snow Fall" by New York Times [12] was well received and analyzed for its use of digital animation and interactivity [42][52], the production involved sixteen people including 11 graphics designers and the data was collected over six-month period. Because of the required time and labor, such interactive column does not comprise of the large proportion of the news articles produced every day.

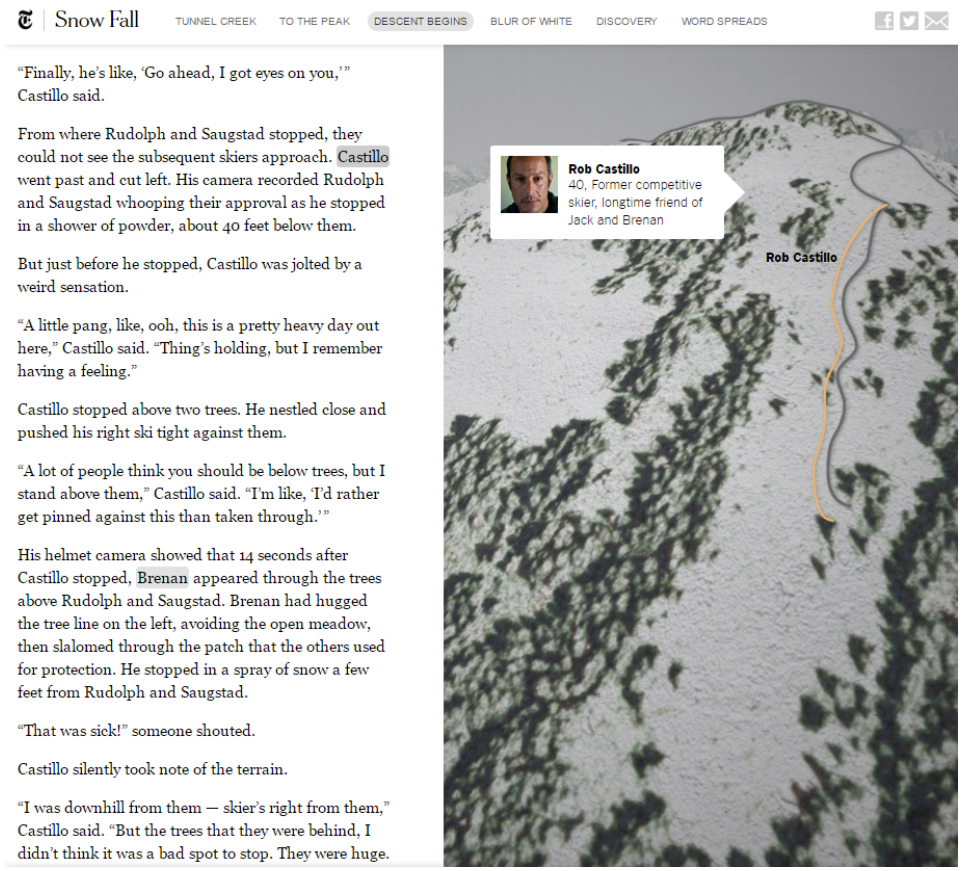

Figure 5. An example if an interactive news article from New York Times' "Snow Fall". The mountain climbing map to the right updates automatically as the readers scroll through the story to show relevant information.

3.2.5 Low visual literacy of readers

The participating journalists complained that fancy visualizations often create wall because the sophisticated nature of the graphics may be difficult to understand. While the problem can be partly attributed to lack of education in schools for newer types of representations of graphics, it if difficult to engage the readers to click the news article if the teaser image cannot deliver compelling graphics, or at least provide

abstract information at a quick glance. Therefore, journalists prefer sticking into primitive types of charts such as bar charts, donut charts, or any of "what we expect to see from Microsoft's Excel." Also, because the level of interactivity is limited by the platform providers as we'll discuss in the following section, journalist hesitate to spend effort in designing creative graphics that can only be understood by drilling down to detail. The teaser image has to provide "overview first", so that readers understand the message from the article, then "zoom and filter" by clicking on it to see more information. And other figures in the article can provide "details" on readers' "demand", coping with the visualization-information seeking mantra [93].

3.3 Limitations for Publishing News Articles

3.3.1 News supply chain dominated by platform providers

One of the problems that many online news media share is that few selected media take vast majority of internet traffics. It is difficult for a smaller news media to get attention from the public to generate revenue, especially when it becomes increasingly difficult to make the readers pay for news [25]. There are news aggregating services [62] such as web portals like Yahoo (https://www.yahoo.com) or MSN (http://www.msn.com) and social network services like Facebook (http://www.facebook.com) that became important channels for news media to distribute their news articles, where news consumption happens. There are also mobile/web services that provides news feeds based on the users' interest such as Pinterest (https://www.pinterest.com) or Flipboard (https://flipboard.com). Although these distribution channel may provide smaller news media companies

access to more and wider audiences, they have their own policies and guidelines for various aspects of news articles that they host including, but not limited to: kind of content, type of media, cover image, type of language used, allowing interactive contents, allowing 'outlinks' (a type of hyperlink that moves readers away from the platform to the other platforms such as news media's own website or services).

Participating journalists point out that the platform providers may use these rules at their own benefits, because "they need to make users stay within their own services as long as possible" (Figure 6). And by doing so, they want to "control their user experiences to satisfy their visitors." For example, once the news providers allow interactive contents to be hosted on their platform, it may introduce performance and security vulnerability. For example, many interactive contents used to require plug-in based technologies like Macromedia Flash in the past, and now requires HTML5 and Javascript, which may impact the web browser's performance. Since the platform providers "cannot review and censor all contents" hosted on their services, they limit the type of media to only texts and photos, and recently streaming videos.

3.3.2 Mobile dominated online news consumption

Because the platform providers restrict the use of interactive data exploration, news media need to provide an out-link from the platform providers to their own website where all interactivity can take place. However, since online news consumption market shifts to mobile devices among younger adults who are building news consumption habit [20], it becomes challenging to provide interactive content that

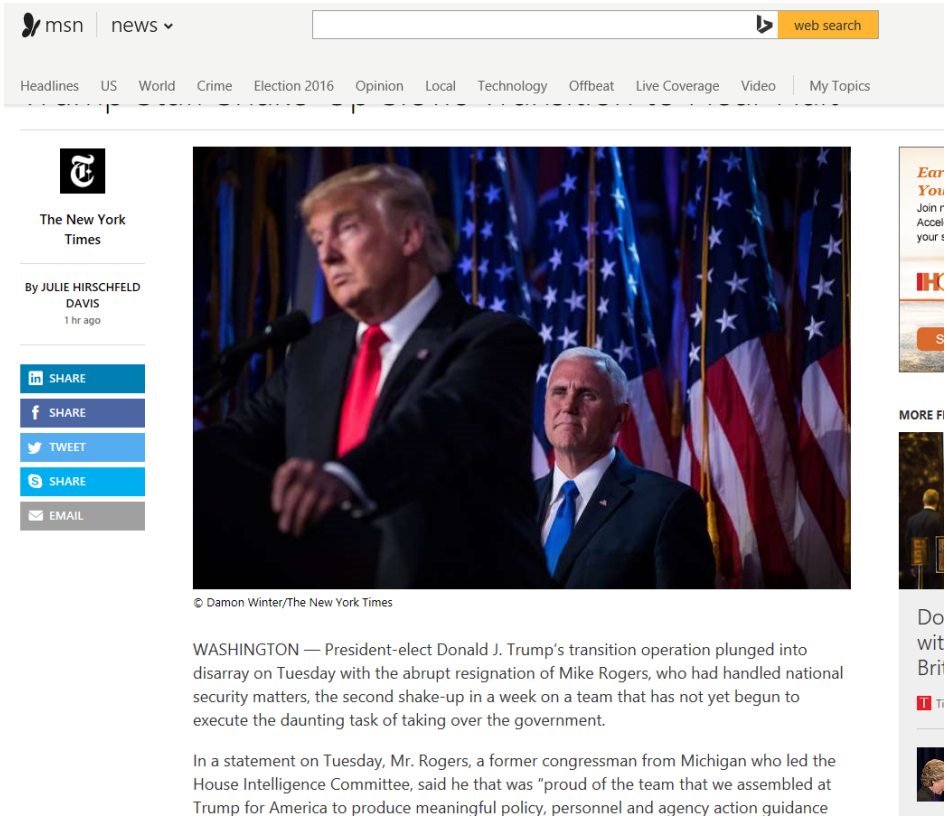

Figure 6. An example of a hosted news article by platform providers. The original news article was authored by the New York Times and is hosted on MSN, the platform provider's website. Such hosted articles are limited to providing interactive content within the platform providers policies and guidelines.

requires constant users' input as well as enticing them visit the news media's own websites, who may be "covered with online advertisements that consumes users' mobile data bandwidth", says one participating journalist.

In addition, although early criticisms of using web browsers' plug-in technologies such as Java Applets or Macromedia Flash have faded with the introduction of HTML5, the screen size and limited computational power makes interactive data exploration unsuitable.

3.4 Adoption of card-style news articles in Korea

In an effort to overcome some of limitations, news media companies and journalists began to seek for new types of news article formats that can adopt to current trend in online/mobile news consumption where interactivity can be limited. Although video streaming and live video streaming with the help of mobile devices are on the rise, producing good video content can be costly and time consuming. However, there were other attempts to exploit the interaction modals used on mobile devices. One notable example is the use of Card-Style UI (also known as Container-Style UI) design pattern found in mobile applications (Figure 7). Although there were news articles that adopted the form of slide-shows in the past, Card-Style news are produced in a square, or close to square format in order to satisfy both portrait and landscape use of the mobile display.

Figure 7. An example of a Card-Style news article that adapted the concept of *Card-Style UI* to enhance mobile news consumption experiences. Two cards are shown here and the third card can be accessed by swiping the screen to the left. Depending on the size, one or more card may fit into a screen. It also provides a book-style narrative as the story progresses by moving into the next card.

Card-Style news articles provide a number of benefits to both journalists and readers. For journalists, it is easy to provide impactful graphics for every page to engage readers. Also because the format resembles small presentation slides, it can be hosted on the platforms that support the image gallery such as Facebook or Instagram. Also, the slide-show nature of the format allows easy conversion to video files that can be hosted by video streaming services. For readers, navigation of card-style news article is intuitive. Also each page only provides consumable amount of information at once, it requires lower the cognitive load. It is especially important on mobile environment where the readers may be walking or on a bus/train while viewing content on their mobile devices. Journalists expressed that they expect our tool to be designed to support exporting text visualizations that can be suitable for creating Card-Style news articles.

3.5 Use of Wordle as Graphics

As mentioned earlier, Wordle only uses one attribute to show the text data, the font-size for the words' frequency. Because many other useful coding methods such as positions and colors are lost, it may not be optimal for its analyzing data. However, the journalists insisted that its aesthetic element is important as a teaser image for attracting viewer's attention before the article is 'chosen' to be read. Although it is possible to analyze raw data with more complex analytics tools to mind insights and produce separate graphics to represent it in news articles, such process would break the analysis phase and the graphics generation phase that the journalists has to deal with them separately. Therefore, in order to capture both the production efficiency and the aesthetics of Wordle, journalists seek for ways to use Wordle in the analysis phase as well. NewsWordle provided basic data massaging techniques and means to compare textual data to understand changes in topics and to help build narratives from them. By doing so, NewsWordle reduces the gap between data analysis and graphic production. In order to facilitate such tasks, we have made Wordle in NewsWordle to incorporate the elements of ManiWordle, enhancing its capability for producing high-fidelity prototype graphics.

Chapter 4

ManiWordle [1] for providing flexible control over Wordle

In this chapter we describe text visualization tool for manipulable word cloud generation called ManiWordle. ManiWordle improves upon a popular type of word cloud called Wordle [99] and provides flexibility of changing various parameters and word formation to customize meet the creators' specific requirements. We have discussed the possible extension to the manipulability and customizability of ManiWordle with the professional journalists while designing NewsWordle. From there, ManiWordle's layout strategies provided foundations for text visualization in

[1] A preliminary version of ManiWordle was introduced in TVCG [15]

Figure 8. Main overview of ManiWordle showing a word cloud for the title and abstracts of All InfoVis Papers published since 1995. The word ‘data’ is selected and being rotated by dragging the circular handle on top.

NewsWordle as it required rapid generations of multiple word clouds at a time. Also, fluid animation of ManiWordle helps journalists understand the manipulation of the data.

To reflect users’ intention while generating Wordle-like visualization, ManiWordle allows users to manipulate various aspects of original Wordle design suggested by Viegas et. al. [99] by supporting editable typography, color, and composition of individual word in intuitive ways, enabling them to have better control over the layout result.

4.1 Analysis on the Original Wordle

There are several occasions the auto-generated layout needs to be changed. For example, when a user is somewhat satisfied with the auto-generated layout while “playing” with combinations of adjustable parameters, it would be necessary to keep

the current layout, because making any further changes to the parameter may produce the layouts that are drastically different due to the randomness of Wordle generation and similar layout may not be easily reproduced. Therefore, the random generation cycles tend to stop once a user reaches somewhat-close-to-wanted layout. However, s/he may still have minor complains such as few less relevant words being too prominent, affecting the visual significance of some other important topic words. Simple removal of such words generates holes in the middle of layout and may damage aesthetic quality of Wordle. On the other hand, total reconstruction of the entire layout without the words removed may deviate away from the layout that s/he found attractive and result in undesirable overall layout compared to the previous result.

Also, a user may want to make specific words more distinguishable from the rest. One such example is when few proper nouns need to be emphasized, rather than blending in the word cloud. One frequently used technique is to add more of the same word into the input text in order to increase the font size of the word, the only data coding used in the original Wordle design. However, such techniques may not always satisfy the user's intention, because s/he might want to detach the word from the cloud and place elsewhere to serve as a title or labels using a specifically chosen colors.

However, because Wordle does not provide direct control over neither the location nor the color of an individual word. Changing locations of words are only supported through re-layouting the entire Wordle, instead of allowing them to be moved individually. Also, changing color palettes only allow users to change the entire color-

theme of Wordle with no control over which color is applied to each word. Although Wordle provides recoloring the words without changing the layout, it may take arbitrarily long time until the desirable outcome is shown.

Also, the angle of word is not customizable. A user may change proportion of each orientation applied to words by selecting them from pre-defined options (i.e., mostly horizontal, half and half, mostly vertical, etc.). Again, unless s/he wants all words to be placed horizontally or vertically, it may still take arbitrarily long time to set all words in the desirable orientation. For example, s/he may want to put all words horizontally and tilt few words in order to emphasize them (Figure 8).

In summary, users need the ability to change the parameters for the individual word as well as the entire layout.

4.2 Design Rationale

4.2.1 Providing a compelling starting point

Users may not have much idea about the end result of Wordle from the text. Also it may take long time to place and configure individual word. We decided to provide starting point as the original Wordle did, since Wordle layouts can be aesthetically pleasing to people. It takes a plain input text from a user, processes the content based on the Wordle parameters set by the user, and provide generated layout. Unless users choose specific parameters, it will choose random configurations and generate Wordles. Users can repeatedly generate random layouts until they are satisfied, and then begin customizing individual word.

4.2.2 Presentation words' importance with their size

Previous research shows that bigger words capture more attention than smaller ones [99]. We speculated that the creators of Wordles will be more interested where they place bigger words, meaning that they may spend more time customizing bigger words. Smaller words attract less attention, and can serve as 'space-filler' that fit into gaps among bigger words. Therefore, we prioritize placement of bigger words over small ones, because sporadically placed smaller words eliminate spaces that could otherwise be occupied by the bigger words. However, the smaller words still play important role: One of Wordle's aesthetic qualities come from compact layout among words, because smaller words can help set a shape to a cloud and provide the holistic view of the layout.

Since we allow users to *drag and drop* words to their desired location, we decided that bigger words will always have priority over the place if there is a conflict. When a smaller word is dropped onto a bigger word, the smaller one will automatically try to find the nearest available spot. If a bigger word is being dragged, smaller words move away from their current location to find other nearest available spot upon collision and leave room for the bigger word.

4.2.3 Reflecting users' intention as much as possible

We introduced the ability for users to "pin" words to the canvas to override other changes. This is to better support users' intention. As mentioned before, users of ManiWordle use randomly generated layouts as a starting point and begin modifying individual word. Therefore, if a word is manipulated by the user, it is likely that users

meant to keep the configuration, because the state in which the word is the result of users' customization. Therefore, we pin words that are modified in any ways by users. The pinned words will not move away from their current locations even while bigger words are being dragged over them. Only when the bigger words are dropped onto them do they move to other nearest available spots. This is because, although the smaller words were pinned, now users have shown their intention that bigger words are also manipulated. In this case, we go by our rule: *bigger words win*.

4.2.4 Provide fluent animation so users can follow changes

As described in the previous section, when users drag-and-drop words to change their locations, some other words get displaced from their current locations. Therefore, it may be necessary for users to keep track of which words being affected by their actions. We provide fluent animation between any changes user make. Even in the event of total rearrangement such as changing global angle settings or even randomizing the entire parameters, we compute minimum distance between the current state and the future state and animate all words accordingly. For example, if the randomization caused the word to have different colors, different angle/locations, and different font (whose word size is generally larger even with the same font size), we apply smooth transition to colors and angle/locations, while the font has to change immediately. The same method is applied for un-do and re-do, so that users can go back and forth between two states to compare them. All collision detect among words are done in real time, even when users are dragging the words. We later adopt the concept of traceable animation to show changing topic trend in NewsWordle.

4.3 Interactions

In the initial layout, no word is pinned. Once users click on a word, a framed rectangle appears to indicate that is it selected and pinned (Figure 9). Once pinned words remain pinned until the users explicitly unpin them. The extended leg on a framed rectangle can be dragged to rotate words.

Also, right clicking a word brings context menu, from which users can change colors and font of individual word. After users make adjustments to words and pin all necessary words that need to keep their status, they can click “re-layout.” This will initiate re-layout of all unpinned words in order to fill gaps that might have been made during drag-and-drop manipulations. When the mouse cursor is hovered on “re-layout’ button, all unpinned words fade to the background and so that users can confirm which words are pinned (Figure 9 - Top). Also, pressing shift a keyboard will also fade unpinned words. Keyboard shortcuts like Ctrl+Z and Ctrl+Y was mapped to undo and redo respectively.

4.4 Implementation and optimization techniques

ManiWordle was written in C# and build on XNA Framework 4.0 along with .NET Framework 4.6 on Windows 7. Rendering of the word-cloud itself was handled by XNA to exploit graphical capability of GPU, while leaving collision-detection to CPU. But basic layout algorithm follows the original Wordle generation pseudo code introduced in [99]. We further optimized the logic to support real time collision detection.

Figure 9. All unpinned words fade away to the background when a mouse cursor hovers over the re-layout button (Top). After clicking on the re-layout button, unpinned words are re-arranged to form a packed cloud (Bottom).

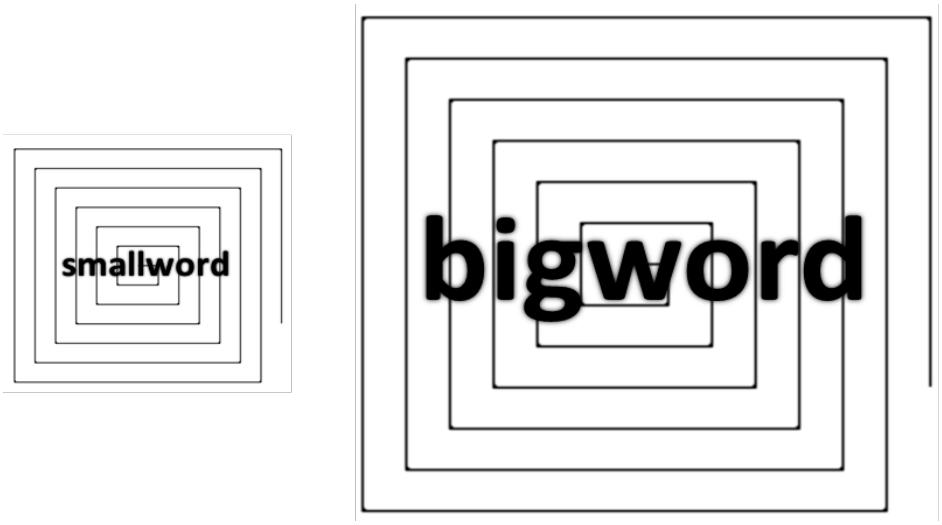

Figure 10. The radius of spiral and the interval of collision checks are larger when dealing with bigger words compared to smaller words.

4.4.1 Adjusted rate of growth in spiral radius

Compared to the original Wordle, ManiWordle uses adjustable spiral radius for finding appropriate spots for words (Figure 10). This is based on the observation that bigger words are more likely to collide with other words than smaller words. Therefore, collision check is done on a larger interval for bigger words. Also, we set the interval between two consecutive points on a spiral to be larger, based on the same observation.

4.4.2 Multi-thread optimization

We setup a thread running constantly and separately from the main rendering thread. The *placer thread* watches queue of the words that need to be placed on a canvas, and process them by making candidate locations and check for collision with any other words that are already placed on the canvas. Also the placer thread also has four child

threads that specializes in collision detections. The place thread picks four candidate location at a time from a spiral for the new word on a queue, and if one thread reports the word can be place on that location, it will put the word there. If all four report that the locations are not available, then the placer thread picks another four candidate locations from the spiral and repeat the process. Also when a word is being dragged to be relocated by users, any words that already on the canvas are checked for collision with the current word being manipulated. Once an existing word was determined to collide with the word being manipulated, the word is removed from canvas and put on a queue for the place thread (Figure 11). However, the removal and the collision detection is not visible to users. Only after finding appropriate spot on a canvas can the word appear to move to the location with smooth transition. It is possible to increase the number of collision detector threads for further speed boost. However, after four threads, we achieve real time interactivity and began to see diminishing return.

We used a six threaded version for collision detection in our user experiment to leave enough room for delay. However, in real life usage, using too many threads may quickly saturate CPUs' computing capability. On our experiment, we limited use of other resource intensive processes like web browsers or productivity applications.

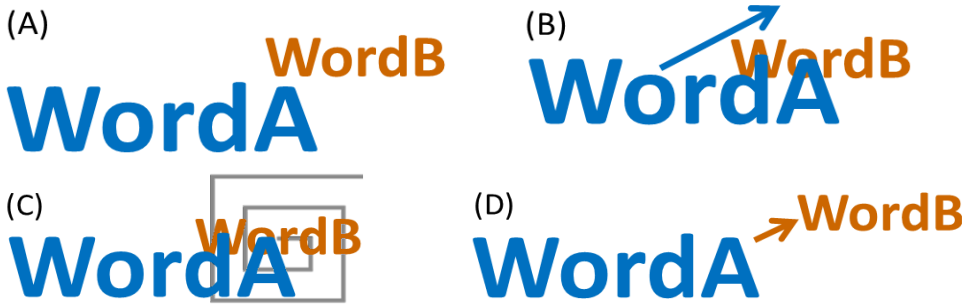

Figure 11. (A) The original layout. (B) A user drags WordA to the top of other words. (C) The placer thread determines where WordB should go on a spiral. (D) WordB moves.

4.4.3 Collision detection with reduced-resolution texture

Because vector-based collision detecting using fonts' embed spline curves may impact performance depending on the complexity of the font, it can result in unpredictable runtime. In order to overcome these problems, ManiWordle uses a 1-bit mask image that only has either inked or black value. Then 32 pixels (32 bits) are packed into one 32-bit integer value. This increases cache hit and reduced memory loads, compared to using the standard RGBA pixel values (32 bits per pixel). Also the size of pixel masks is only one third of the original in each dimension, resulting in one ninth of the number of pixels to be checked for collision detection. It is shown that to produce good results for the purpose of ManiWordle. Also we applied dilation [35] to the mask image, in order to prevent words being place too close to one another

4.5 User Study Design

We conducted two user studies to evaluate ManiWordle. The first study was a

preliminary usability study to identify ManiWordle's usability issues and users' general opinion on the tool. After making improvements upon the result, we ran a controlled experiment to if ManiWordle's enhanced interactivity and flexibility provides better subjective satisfaction for creating World-like word clouds.

For the preliminary user study, we recruited six participants (3 females) from the university's graduate school. No participants had any previous experience with using Wordle. After 15 minutes of training, they were asked to manipulate position, orientation, and typographical properties of several words to match a target layout presented in a printed form. The big words that are important were highlighted with an arrow mark in order to let them practice drag-and-dropping big words to change their locations and rotate to make new Wordle as similar to the provided one as possible.

They repeated the task three times with three different data: 1) a Wikipedia entry on Yu-Na Kim, a gold medal-winning figure skater; 2) a Wikipedia entry on StarCraft, a popular strategy video game; and 3) the academic paper that introduced the original Wordle [99].

All participants finished each task within 5 minutes without any difficulties. They enjoyed using ManiWordle, and said that it was "fun" and "intuitive". We did not identify any major usability issues.

4.6 Controlled Experiment

4.6.1 Datasets and the Task

We prepared three different text datasets that have varying emotional attachment

between the text and the participants. This is from the finding that 57 percent of Wordle users use their own text to generate the word cloud [99]. The first text with the least emotional attachment was the paper that introduced the original Wordle. No participants have reported to have read the paper before. The second text with moderate emotional attachment was a Wikipedia entry on Yu-Na Kim. This article was selected to reflect people's great interests and her popularity and reputation by her performance during the Olympic games at the time of the experiment. The last text with the greatest emotional attachment was the participants' own academic research paper that were published in the past. We had them remove the bibliographic information from each paper to remove any possible noise when counting words' frequency. For the practice session that preceded the actual experiment, a Wikipedia article on Beatles was used. Each participant was asked to make a Wordle that is as aesthetically appealing and satisfactory as possible for each of the three text datasets.

4.6.2 Participants

We recruited 12 participants (11 males and 1 female) among the graduate students at the university. We screen the participants to those who have a published conference or journal paper in English to get the third text data we need for the experiment. We asked them to email us the paper before coming to the lab, or had them bring the document file in a USB thumb-drive. They were given about 15,000 Korean Won for their participants. No participants reported to have seen Wordle before, although three of them have seen other tag clouding techniques on the web.

4.6.3 Hypothesis

We hypothesized that the enhanced interactivity and the ability for flexible control of individual word in ManiWordle will result in higher user satisfaction and let people feel more creative while creating Wordles. We also expected that the more the participants are emotionally attached to the text, the more effort they are going to put in creating Wordles.

4.6.4 Study Design and Procedures

We ran the study as a 2 (Visualization: Wordle, ManiWordle) × 3 (Text: the least, moderate, and the most attachment) within-subject design. Each participant performed the task (i.e., making her/his Wordle as aesthetically appealing as possible) for all the three text datasets using both visualizations. We shuffled the the order of visualizations to eliminate the learning effect. Also, the three text datasets (for different emotional attachment level) were also shuffled among participants.

Basic tutorial and practice session was given before conducting the experiment for both Wordle and ManiWordle to get participants familiarized with the interfaces and visualization. For each task, participants were told that they have five minutes to complete. However, we did not strictly enforce the time, since our goal was not to measure time sensitive productivities. Five-minute time limit was given as a guideline in order for them to complete the experiment in an hour without being exhausted. After 5 minutes, they were told to try to wrap up in another minute. However, they were still allowed to spend extra time if they were still not satisfied with the result. After each session with one visualization with all three text datasets, participants were

asked to fill out the post-session questionnaires for subjective evaluation. The same procedure was repeated for the other visualization. Upon completion of the entire experiment, they fill out a demographic survey along the interview discussing their preference between the two visualization tools and the subjective reasons. The experiment took about an hour.

Because the goal of ManiWordle was to enhance Wordle, ManiWordle was a complete superset of Wordle. If a participant was exposed to Wordle in the first session, we only explained the enhanced features for ManiWordle for the second session. For participants who used ManiWordle first, we were very careful not to use terms that might have negative connotations like “restriction,” “removal,” “restrain,” or “limit” when explaining Wordle for the second session. We did not want participants to feel they are given the additional functionalities of ManiWordle first and got them taken away for Wordle.

4.6.5 Testing apparatus and setup

Participants used a quad-core PC with a 27” LCD widescreen display running at a 1920x1200 pixel resolution. The system had NVIDIA 9800 GX2 GPU with 512 MB of memory. During the experiment, the program was maximized to fir the entire screen. All events and inputs were logged by the tool, so that they can be used for statistical analysis. The task completion time was manually measured by a conductor using a stop watch.

4.6.6 Results

We performed statistical analysis on participants' subjective responses from the questions (Table 1). We applied Friedman's Chi-Square test. Participants were significantly more satisfied with the result layout of ManiWordle than that of Wordle ($\chi 2(1) = 9$, $p = .039$). We did not find any other statistically significance between two visualization tools.

During the post-study interview, 10 out of 12 participants said that they liked ManiWordle more than Wordle. Especially, ManiWordle's ability to fine-tune the layout allowed them to make more satisfactory results. One participants specified that pinning big words firsts at his desirable locations and filling out the rest of the

QUESTIONS	WORDLE	MANIWORDLE
Q1: IT WAS EASY TO LEARN THIS VISUALIZATION.	5.54	5.23
Q2: IT WAS EASY TO USE THIS VISUALIZATION.	5.31	5.38
Q3: I LIKED TO USE THIS VISUALIZATION.	5.38	5.62
Q4: IT WAS FUN TO USE THIS VISUALIZATION.	5.46	5.77
Q5: I FELT CREATIVE WHILE USING THIS VISUALIZATION.	5.08	5.54
Q6: OVERALL, I AM SATISFIED WITH THE RESULT LAYOUT. *	5.31	5.77

Table 1. Subjective Responses to Six Questions (Average Ratings). The questions with significantly different ratings are marked with an asterisk (*).

word cloud via automated process was the easiest tactics for using ManiWordle. Another participant said that he never fully understood the rules of 'bigger words win' for conflict resolution because words font size is often difficult to judge, especially when two contesting words are greatly different in length. However, it was still better than Wordle that did not allow such fine tuning. Also, two participants said that fluent animation made ManiWordle less boring and more fun.

Two participants who preferred the original Wordle to ManiWordle said that they did not enjoy adjusting too many things. One participant said that he feared that he "couldn't get it back after making changes." The input log reveal that he did not utilize the *undo* feature which could have alleviated the problem. He said he "forgot" that it was there. The other participant said fine tuning the layout required too much effort and labor. He said Wordle is much simpler to use and the resulting layout is often as good as that of ManiWordle.

We also analyzed the effect of the emotional attachment level to the visualizations based on the amount of time users were willing to spend.

We also investigated the effect of the visualization and the emotional attachment level on the amount of time to complete a task. We ran a 2 (Visualization: Wordle, ManiWordle) x 3 (Text: the least, moderate, and the most attachment) analysis of variance and Tukey's HSD post-hoc test. We found a significant main effect of Text ($F_{2,66} = 3.42$, $p = .039$) with post-hoc tests showing that participants spent more time to create a Wordle for their own research paper than for the Wordle paper ($p = .03$) (Figure 12).

We also analyzed usage log for using both visualizations. 2 (Visualization) x 3 (Text) ANOVA test with the number of total interactions as the dependent variable revealed a significant main effect of Visualization ($F_{1,66} = 5.01$, $p = .03$). This result shows that participants initiated significantly more user interactions with ManiWordle than Wordle. Another 2 (Visualization) x 3 (Text) ANOVA test with the number of only the interactions that affected the global layout as the dependent variable. Because ManiWordle tend to get purer number interactions due to its ability to customize each word, we only counted interaction that would result in to global reconstruction of the word cloud. We found a significant main effect of Visualization ($F_{1,66} = 5.40$, p

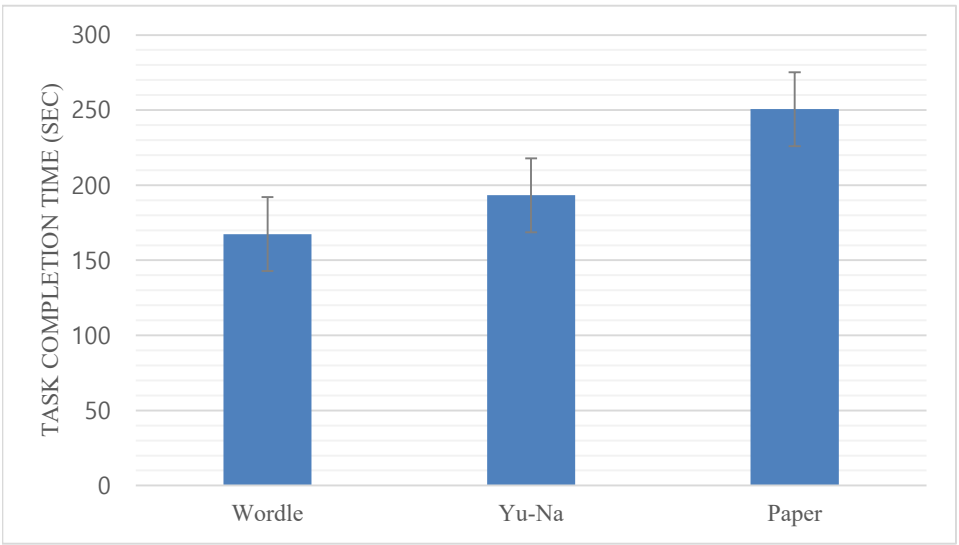

Figure 12. Task completion time (in average) for three text datasets. Error bars represent standard error. People spent significantly more time creating a Wordle for their paper than the Wordle paper.

= .02). This result shows that the participants performed significantly less inputs to globally change the layout with ManiWordle than with Wordle.

4.6.7 Observations on the final layouts

Many ManiWordle's resulting Wordles showed the layouts that are not easily reproducible using the original Wordle, if at all. For example, one participant used a color-change feature in ManiWordle (Figure 13 Top) in order to emphasize some words that are related to Yu-Na Kim's name, job, and victory in the world championship to be siginificant keywords. Extracting these true keywords, instead of just using the frequency, is efficient for human creators. Figure 14-Top shows a layout from a ManiWordle where the participant clustered the words based on the semantic meanings of each word. He showed a consistent pattern of clustering words for all three input texts and was disappointed that he could not do the same using the original Wordle. This type of clustering is a challenging problem in automated process and require massive computation power to perform natural language processing and machine learning techniques.

4.7 Discussions and Implications

The resulting layouts showed users are in the need of flexible controls provided by ManiWordle. First, ManiWordle yielded higher user satisfaction than Wordle did and was prefered. Also, participants made significantly less global changes using ManiWordle than using Wordle. In fact, among all inputs on ManiWordle, 46 percent were for configuring individual word.

The usability of interactive systems decreases as the system begin to offer its flexibility with more functionality [65]. Thus, we were encouraged to see ManiWordle was not any more difficult to learn to use than to learn Wordle, after seeing no significant differences from Q1 and Q2 in Table 1. Also, we were able to confirm our hypothesis that participants would spend more time in creating a Wordle when using the texts that are more emotionally attachable. However, we did not see evidence that the visualization may affect the time spent for creating a Wordle.

We initially anticipated that enhanced functionality of ManiWordle would make users feel more creative about their work. However, we did not find any significant difference between Wordle and ManiWordle on that regard. All our participants had a STEM background, meaning that they may be less engaged in graphics designs. We wonder if the result would be different if the participants were from different backgrounds, like arts or graphics design. However the participants still expressed it was fun to create Wordles with the tools, leaving possibility of applying different techniques to ManiWordle to improve its functionality in the future. One notable request from people was lack of ability to change the font sizes. Because the font size is the only coding scheme that reflected the data in Wordle, we did implement any tricks for the users to damage the integrity of the original data. However, since more recent studies on Wordle creation showed ability to manipulate underlying data along with multimodal inputs [53], it may be possible to support higher customization options. At the end, Wordle is a form of casual infovis [86], where the integrity of the data represented by the visualization may not be as import as when they are used for

scientific research. It is important to remember that the artists symbolize their strong point of view by breaking the conventional rules of visual analytics [98].

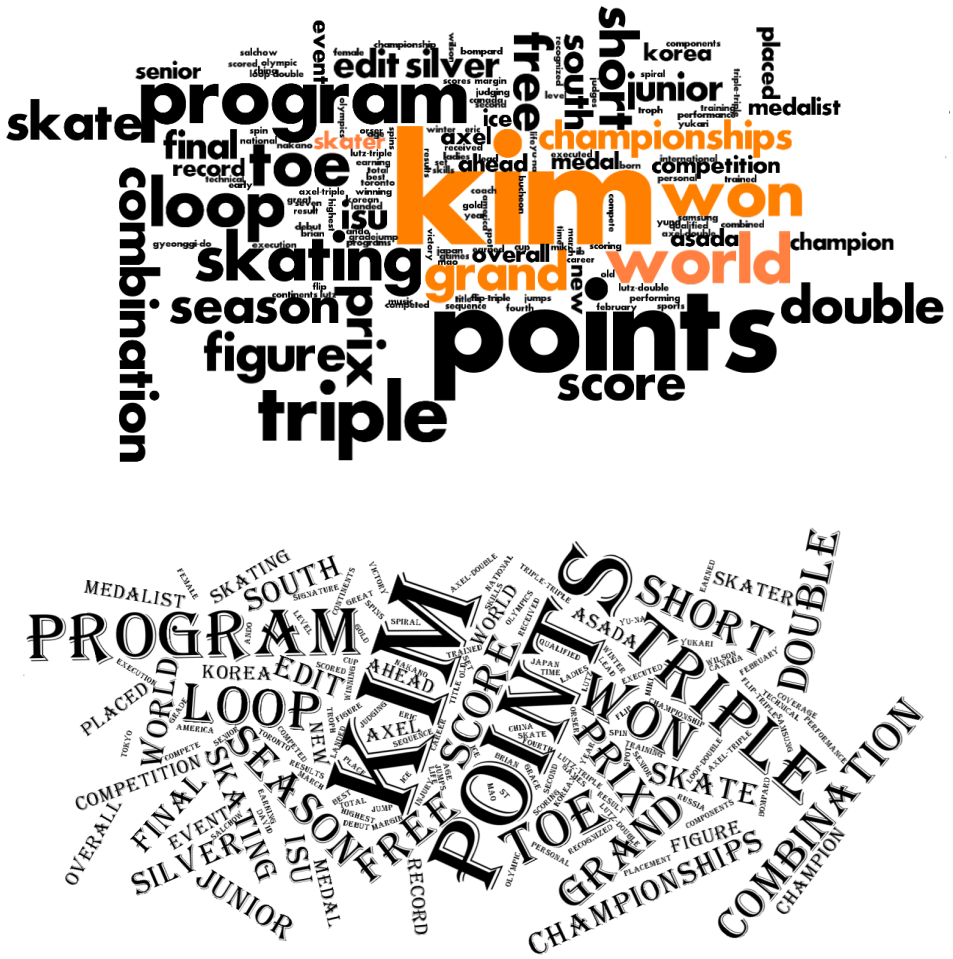

Figure 13. The final layouts produced using ManiWordle (Top) and Wordle (Bottom) by a user. The text was a Wikipedia entry on Yu-Na Kim.

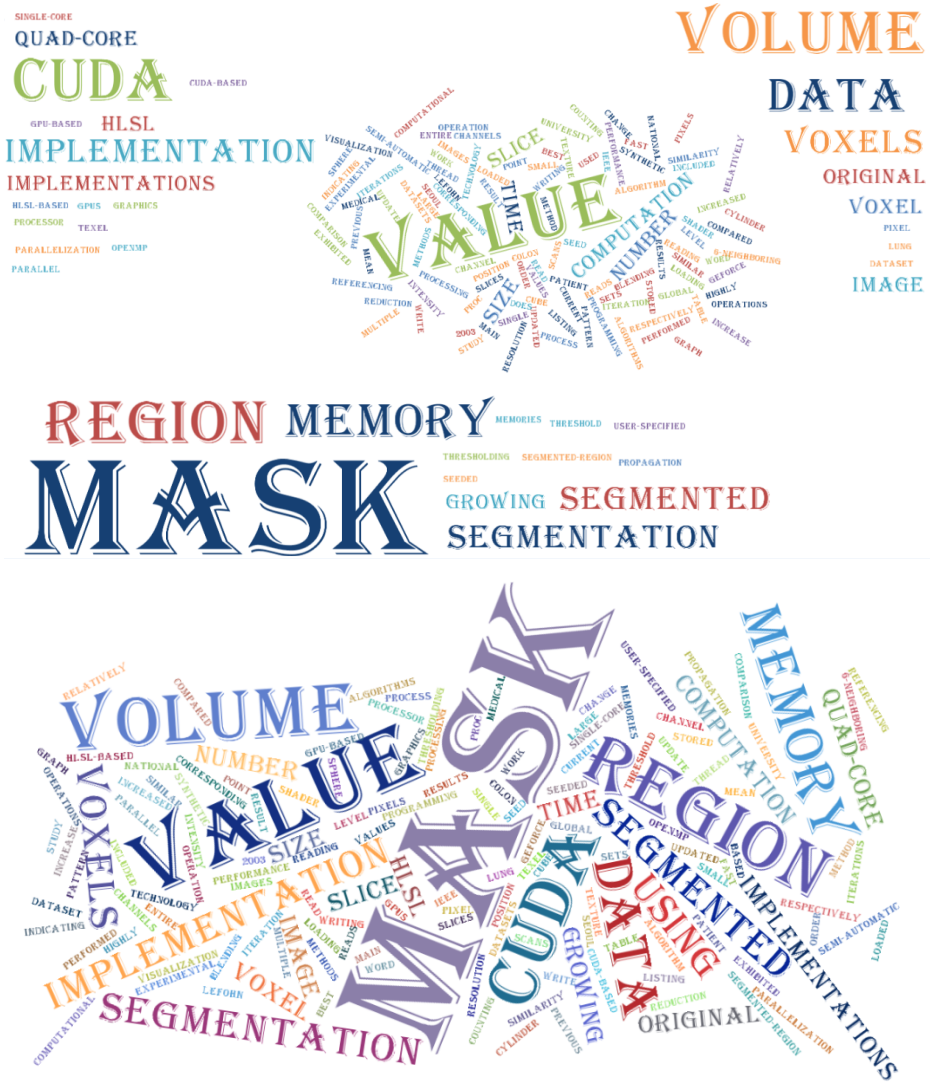

Figure 14. Words from a participant's paper clustered based on their semantic meanings by the person using ManiWordle (Top). The layout by the same person using Wordle (Bottom).

Chapter 5

NewsWordle

We designed and developed a text visualization analytics tool called NewsWordle. NewsWordle produces Wordle-like word clouds [99], that can be custom tailored by journalists to show graphics that show topics and trends over time and that can be directly used for publications in news articles. The text visualizations aspect of NewsWordle incorporates some of principles from ManiWordle to provide journalists a way to custom tailored Wordle that can be used in news particle productions. In order to facilitate news article production and publishing for journalists based on text data analysis, we designed NewsWordle with two professional journalists.

5.1 Design considerations

A number of design considerations were extracted both from interviews and design sessions while investigating the production pipeline of current news article productions based on text data analysis. The following sections describe the design considerations we have made with the participating journalists to alleviate the problems described in Chapter 3.

5.1.1 The tool needs to fit into the production pipeline and enhance the process

The main goal of the tool is to provide journalists better capabilities in various stages in the news article production pipelines to reduce reliance on the experts for both analyzing data and producing graphics. This improves the bottlenecks because the journalists can communicate with the experts in more details, thus reducing redundant iterations. However, our tool is not intended to completely replace the roles of the experts. For example, because the basic data processing and filtering can be done based on the journalists' knowledge on the issue, journalists can massage the data to contain more relevant information before handing over to the data experts, who may have necessary skills for more sophisticated data analysis, but without contextual knowledge on the data itself. Also, the visual representation of the data (in a Wordle form in NewsWordle) can also deliver a summary to the experts for the data to be explored. Furthermore, NewsWordle is not intended to provide just better templates for the graphics for journalists to produce all figures by themselves. It is designed to align the basic data processing and analysis, directly with the visual prototypes that describe the massages that the journalists intend to deliver to reviewers, to graphics designers. In order to meet these goals, NewsWordle does not manipulate the current news production pipeline, but rather integrate into it.

5.1.2 The visualization needs to be of the consumable graphics while supporting analytics

The main theme we found in text analytics was to show how trend changes over time. Previous research used the techniques for opinion mining and sentiment analysis [37]

[83] for product and content reviews as well as opinion flows on election. However, many analytic tools are designed to facilitate exploratory tasks to extract insight from underlying data, like in the example of [23][29][33][34][94] and often is not appropriate to be used for figures in news articles, due to their complexity. Although the concept of stacked graph to show topic changes as flows [15][32] may be intuitive at first glance, it is not scalable to using them in news articles due to its reliance on user interactions. Other trend visualization embeds more textual information [61] into flow visualizations like TIARA [82][102] which provided tighter integration of textual summary information to the visualization. Although such techniques can present more summary of the topics at once than the methods that heavily rely on user interactions, they utilize large canvas on a screen, only making them appropriate for readers with larger screens. However, the participating journalists argued that the visualization is more scalable for consumption if they can be useful regardless of the devices form factor or platforms. After a series of discussion with journalists, we decided to utilize (Mani)Wordle for both data analysis and production in order to streamline the process, because it still can intuitive for readers to see the main topics, especially the ones that are in larger fonts.

5.1.3 The graphics has to be non-interactive and static

We investigated other visualization frequently used in the news articles. The main theme that occurred is the used of infographics. An infographic is “an effective way to present complex data in a visual format that is compelling, provides rapidly available information, and is directly useful for decision-making purposes” [81].

Although the early days of visualizations for news media are done by graphics designers and statisticians, infographics have deviated away from explorative visualization in order to make it quickly consumable by the readers. Therefore, the exploratory visualization is utilized by the journalists in order to extract insights from data and the insights are then turned into narrative stories by the journalists and infographics by the artists.

Infographics designers are not without their dilemma for producing right figures for the articles. The main themes include, but are not limited to: Manual encoding of the data, placing data on existing graphics, relaxing the sequence of processing, and creating an effect data abstraction [10]. Technical limitations of current visualization tools may limit the designers' creativity, resulting in the charts that do not outstand from the others. In order to ensure creativity or the artist beyond the capabilities of the visualization tools, they prefer manually encoding data points onto a graphic. However, because manually encoding all data point may take significant amount of production times and effort it is difficult to explore various design options once the process begins. And if an error or missing data is found later from the resulting graphics, it can be very difficult to rework the entire manual encoding.

5.1.4 The visualization prototype must be in high fidelity

In order to ensure that the artists do not generate graphics that are irreverent or untruthful, it is important for journalists to provide visualization prototype with high-enough fidelity. This does not mean the visualization sketch/descriptions or prototypes have to be at the publish-ready level. However, they must contain all

essential information for the artists to work with [7][38][44]. That is, it may be important to make sure that the artists have to analyze and interpret the data again by themselves for producing graphics. Although it is possible for the artists to make inquiries to the journalists, it can often be burdensome when there are other requests coming in from other journalists. In order to minimize the roundtrip between the artists and journalists the fidelity of the prototype needs to be high enough. From there, the artists can then exercise their creativity.

5.1.5 Using news articles as data type

While there is a plethora of data that are readily available through various places like government websites, professional statistics media, we decided to crawl news articles contents as textual data. After investigating various sources with the professional journalists, we agreed that news articles are often the reflection of the contemporary trend, therefore mining the data from them would reveal important topics and underlying trends that would be newsworthy of a story.

5.2 Implementation

NewsWordle is built using C# and .NET framework 4.6.2 using Windows Presentation Foundation (WPF) (Figure 15). We have explored various options for platform including web, desktop platform, and mobile devices. Because journalists tend to work with the laptop or desktop devices provided by the company, NewsWordle was designed to be used with a mouse and a keyboard. Also, because of amount of time necessary to process data, and there can be multiple users running the

analysis at the same time, it would not be feasible to process all data on a centralized server. Therefore, all data processing modules had to be modified and packaged to run on the journalists' desktop running Windows 7 Operating System. Other relevant implementation information regarding the elements of the tool is described in detail in their respective sections.

The individual modules are built as separate project with appropriate frameworks and have been included in NewsWordle as a package. For crawling links from a news portal, Python 3.5 and Beautiful Soap[1] library were used. For part of speech (PoS) tagging and morphological analyzing, the process important in parsing Korean text, HanNanum[2] library was used with Java. Because analyzing news content requires named entity recognition (NER), NewsWordle automatically downloads the latest list of titles for Korean Wikipedia entries[3] upon crawling new news articles or re-analyzing the articles with the latest analysis methods. The database was managed by Sqlite3[4] library.

[1] Available at https://www.crummy.com/software/BeautifulSoup/

[2] Available at http://kldp.net/hannanum/

[3] Available at https://dumps.wikimedia.org/kowiki/latest/

[4] Available at http://sqlite.org

5.3 Database collection and management

In order to quickly evaluate if the topic is news worthy and if any insights can be drawn from the previous data, NewsWordle provides a tool for building databases by crawling news articles and support merging different databases.

5.3.1 Crawling data

There is not an easy way to search crawl news articles for analysis for journalists. While they can make requests to get full news contents of the past to an archive, it often takes few hours and days for the requests to be complied. Also, the archive may only record the news articles that are published by the same newspaper media. However, the content of news media maybe biased towards certain perspectives. For example, news media specializing in sporting news may contain fewer political news.

Figure 15. An overview of NewsWordle application showing all news articles published published by Yonhap News Agency between Jan. 1st 2016 and Dec. 20th, 2016, on the term 프로야구(Professional Baseball).

Therefore, we set a goal for crawling news articles from multiple sources as the data may reflect the trend more accurately [27] as the number of documents increase from sources.

The crawling of the articles can be achieved from a wizard page. At first, users(journalists) choose from which news media they want to crawl news articles (Figure 16). Sometimes, selecting the news media that specialize in certain theme may reduce the noise which may come from other less relevant news articles published by other news media. Also, since the crawling speed can be influenced by the network bandwidth, choosing more news media may result in longer crawling time. Once news media are selected, the users can specify the search keyword and date range (Figure 17) [1]. As mentioned earlier, choosing multiple news media and wider range of time may result into longer processing time. For example, search query 올림픽(Olympic) shows more than 3 million news articles which may take several hours, if not days to be accessed. During crawling the estimated remaining time is shown to the users, so that they can decide if they want to complete the process (Figure 18).

Another important thing to consider while crawling news articles' content is the possible copyright infringement. Because only the derived facts such as the result of

[1] The earliest news articles searchable by Naver [2] is currently Jan 1st, 1990 as of December 2016, although not all news media have searchable news article archives dating back to the point.

statistical analysis on the raw article content can be allowed to be republished, the news contents are processed as they are crawled, whose process is described in 5.3.3.

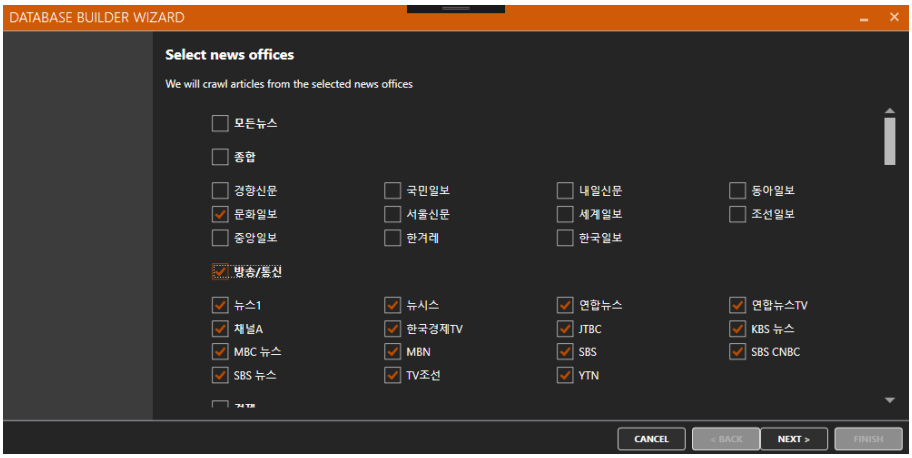

Figure 16. A news office (news media) selection screen. Selecting multiple news offices may diversify the database, but the amount of time for downloading and processing the data increase.

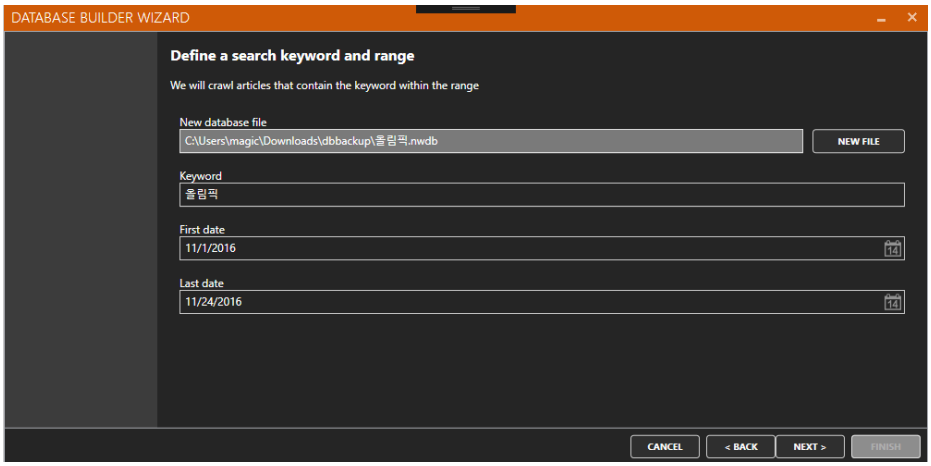

Figure 17. The first screen for a database builder wizard. Users may specify the file, a keyword for search, and the date range for the news articles.

5.3.2 Merging and Splitting Data

Crawling results from a keyword may not represent the theme that a user(journalist) may wish to find. For example, s/he may want to see the topic changes on seven Korean baseball players who play in the Major League Baseball. However, a search result from a keyword 'Major League Baseball' contain many other MLB related news that are not relevant to the Korea players. Although a search keyword does not have to be strictly one word (for example, a keyword 'Major League Baseball' consists of three words), it is difficult to construct a search keyword that can retrieve the news articles that cover all seven players. Therefore, NewsWordle allows databases built from multiple search results to be merge into one, so that the trend analysis can be performed on all articles. It is also useful when there are multiple ways to spell a word by different journalists and media. It should be noted that, in many cases, one news article can contain more than one search keywords. In other words, different search queries can result in many news articles that overlap with each other. We make sure that the news articles are not counted more than once during the merge process, while keeping the info on which keyword the news articles contain. Figure 19 shows the wizard windows that helps merging two databases. There are two different options for merge. "Merging DBs into one DB" means the database will still maintain the dataset from each database, so that the users can see the separate statistics and compare how many articles are duplicate in each DB, whereas "Merge DBs and Consolidate" means that two databases are essentially the same, and the separate statistics are not necessary. The latter is useful when the two search keywords

to make each database represent the same entity. For example, the search results from the keywords ‘South Korea’ and ‘The Republic of Korea’ can be consolidated.

5.3.3 Processing Text

The news contents are processed with both automatic algorithms and manually defined filters.

The automatic algorithm is applied during the crawling phase, and can be performed again if the news articles need to be reprocessed with a new list of Proper Nouns (NER) or when the processing algorithms are updated. In order to support upgrading the processing module separately, it is built as a discrete project. Once a news article’s content has been crawled, the text is process with HanNanum library which analyzes

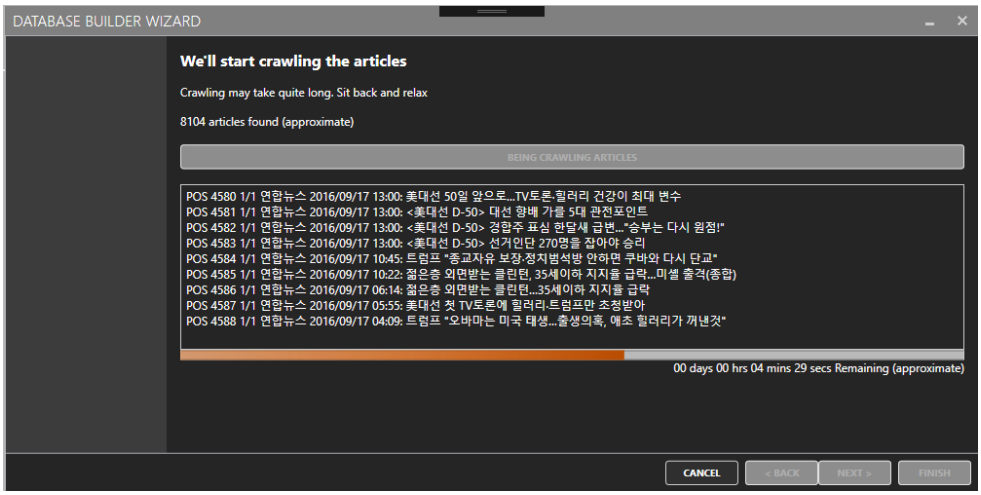

Figure 18. A windows showing the crawling process. Users may see the progress by identifying the number of the news articles currently being processed (to the left) and the processbar at the bottom with the estimated remaining time, which may depend on the network bandwidth and the processing power of the users’ computer.

morphemes and tag Part of Speech (PoS). In order to use the most up-to-date list of Proper Nouns, NewsWordle automatically downloads the list of titles for Korean Wikipedia pages. Although we do not automatically extract named entities from the data using n-gram analysis, the Wikipedia has entries for variety of topics including names of famous people, tv shows, movies, sport teams, abstract concept, diseases, accidents, and etc. Since the morpheme analysis and PoS tagging are CPU intensive works, the processing cannot keep up with the crawling the news content.

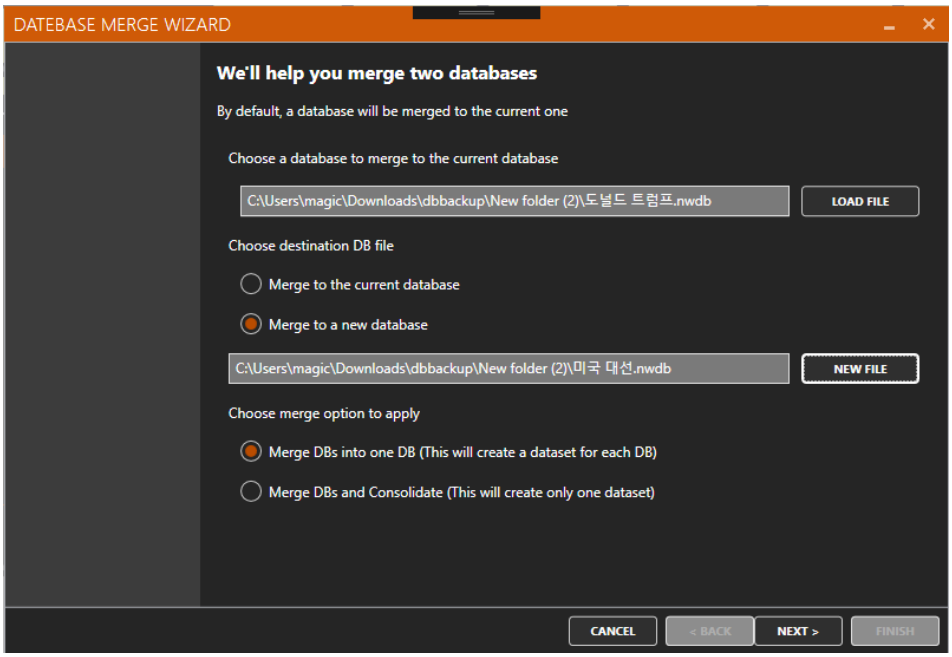

Figure 19. Database merge wizard helps the users merge two databases into one. The first database (with the keyword, Hilary Clinton) is already loaded in the main window. The user can merge the database with the keyword, Donald Trump, and make a new database labeled ‘Presidential Election.’

In order to efficiently utilize the resources to minimize the crawling and processing, we crawl the news articles in parallel threads, while feeding the contents to the processing module pipeline in non-sequential order, which are then re-ordered by the processing module. By doing so, the crawling process and text processing can be done in parallel, reducing the overall time for building the databases.

The users can also add manual filters and rules. There are three different levels of manual processing. The first 2 levels are done in a separate word processing page (Figure 20). The first filter is a global filter (Figure 20-4) that is applied to all databases. The filtering rules are not bound to any specific database and are kept separately from the them. Usually, the name of the news media, meaningless proper nouns, the words like 'copyrights' that show up many news articles for a copyright notice, part of website domains and protocols like 'http'. The second filter is a database specific filter (Figure 20-1,3) This filter is applied only to a database currently loaded into a main window, and is only bound to the database. Figure 20-1 shows the entire list of the words currently in the database, except for the ones listed in the global filter. The users can apply filters by selecting the words that need to be filtered and moving them to Figure 20-3. Our evaluation of prototype with the journalists has revealed that, in many cases, short words or the words that contain numeric values are often manually removed. We provide the buttons (Figure 20-5) that apply these heuristics. Also, the words can be normalized into other words. For example, 'Major League Baseball' and 'MLB' may both be normalized into 'MLB'. Also, 'Donald Trump' and 'Trump' may both be normalized into 'Trump'. By using

this, the users can also merge different words that are spelled differently. The third and last filter is a filter that is applied only to one Wordle at a time. Since the filter screen shows statistics for the entire database, the third filter can be defined in the card-views in the main window. We discuss more details in 5.5.1.

5.3.4 Exporting to Excel Formats

NewsWordle is designed and evaluated with the professional journalists in order to enhance the entire news article production pipeline, it can support making Wordle visualizations that can be directly inserted into a news article as figures. However, they often expressed the need for viewing the raw statistics of the words and attributes

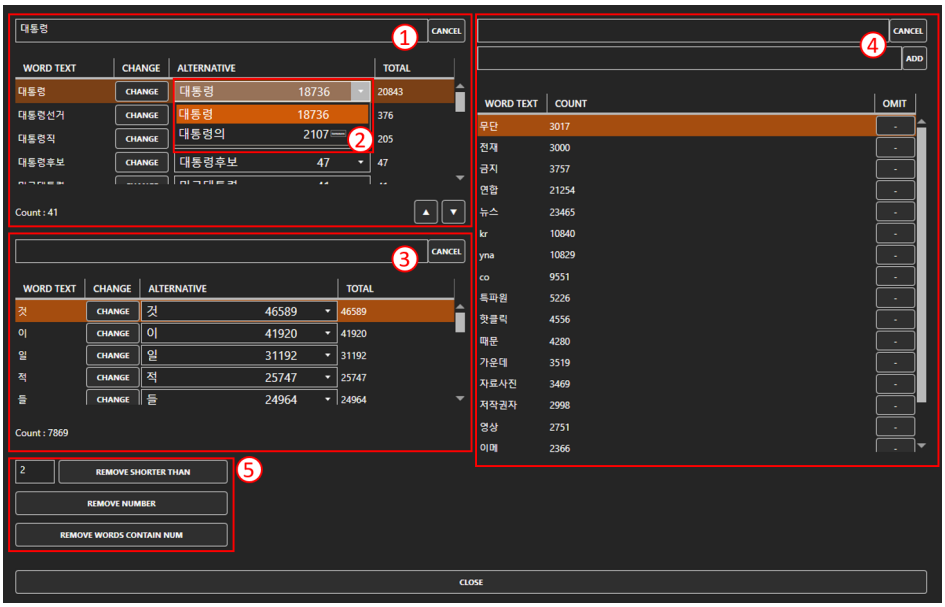

Figure 20. A word filter page. The users may create a global filter, or database specific filters and normalization rules. Also some commonly applied filtering rules are provided for convenience. A drop-down menu for a word (2) shows that two words (‘대통령’, and ‘대통령의’) are normalized into ‘대통령’

that went into making Wordles. Therefore, NewsWordle helps export the database into an Excel file that can be reviewed by other data analysis tools. The dataset includes the words that are currently shown in a Wordle, the words that were present in the previous Wordle, but has faded, and the words that were not present in the previous Wordle, but became prominent. This helps understand how the topics have changed within the given date-time range, compare to the preceding range and succeeding range.

5.4 Summary View

The main window has a summary view (Figure 15-1,2) that represents basic statistics of the dataset.

5.4.1 Histogram

The histogram (Figure 15-1) shows how many news articles are preset in each bin. (More details on binning is discussed in 5.4.4) The height of each histogram is always normalized by setting the bin with the most number of news articles to be the tallest. The specific number of news articles and their keywords can be seen by hovering mouse over their values (Figure 21).

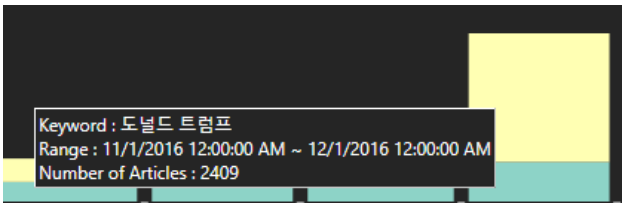

Figure 21. A basic statistics can be views by placing a moue cursor on to a historgram.

Figure 22. The tick marks can be dragged, added, and removed in order to adjust the interval of date-time range.

As mentioned earlier, many news articles may contain more than one keywords. Therefore, simply stacking them in a histogram may not accurately portrait the nature of the data (i.e. many articles overlap across multiple keywords.) In order to facilitate avoid the skew, users can also change them to grouped bar charts (Figure 23 Bottom).

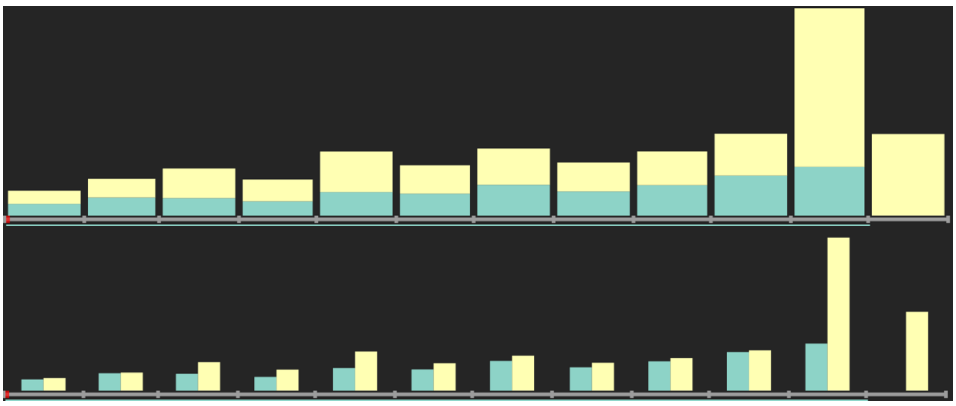

Figure 23. The histogram view can be configured to either stacked graph or grouped bar chart mode.

5.4.2 Range Sliders

The range sliders are located below the histogram view. The range slider (Figure 15-2) can be used to define the interval (date-time range) of each histogram, which corresponds to one Wordle card. The users can drag tick marks to left or right to adjust the interval. As the mouse cursor is placed along the slider or on top of each ticks, a popup shows up to show the specific date and time. They can add a new tick mark by right-clicking on a region where the ticks are not present to split intervals. Also, by right clicking on existing tick marks, they can remove ticks (effectively merging two adjacent intervals) or go into detail settings to specify the time in finer granularity. The latter is important when journalists already have pre-existing knowledge on the specific time of an event, from which the topics are likely to change (such as accidents or political speech, etc.) In order to give perspective of their interaction, when there is a change in the range sliders, their corresponding Wordles and cards are automatically updated in real time.

5.4.3 Gantt View

The Gantt view located below the range represents the earliest and the latest time specified the users when the database was crawled (see Figure 17). By hovering a mouse cursor on top, they can see total number of news articles are crawled for the keyword (rather than seeing the number of news articles within the date-time interval as in the Histogram view). Although the intention of the Gantt view is to allow journalists to merge database collected over different period of times, this turned out to be a rather rarer scenario. They preferred to crawl news articles on the same date-

time range using multiple keywords in order to directly compare how much each keyword is mentioned during the same period. The only time they would compare datasets collected from different time period is when the comparison needs to be made based on the beginning of an event to certain length of time. For example, journalists may be interested in how general opinions have evolved over a week for two different (but comparable) movies that are released in different time. (e.g. comparing general sentiment of Star Wars Episode 7 and Star Wars Episode 3) However, such scenario can also be achieved by analyzing them separately.

5.4.4 Binning

Upon loading a database, NewsWordle automatically detects the earliest and the latest date of published news articles in the database and uniformly divided tick marks are pre-populated. The possible binning intervals are 5 yearly, yearly, quarterly, monthly, weekly, daily, hourly, ten minutely, and minutely. It automatically chooses the interval that the resulting number of bins is as close to ten as possible. ten is the target number of Cards that the journalists try to produce when making card-style news. Also, instead of just beginning from the earliest time in the database, we try to set the interval to begin at 'nice numbers.' For example, if "weekly" is chosen, the starting day is always the first Monday before the earliest time in the database. Once users begin to manipulate the range sliders, the mode automatically changes to 'custom'.

5.5 Cards and ManiWordle

The cards (Figure 15-4, 5, 6) corresponds to each of date-time range defined in the range sliders.

5.5.1 (Mani)Wordle

By default, a (Mani)Wordle (Figure 15-5) is generated for each card. Although not all features of ManiWordle had to be imported, it allows crucial interaction mechanism, the ability to move words and pin them. Because multiple cards must be displayed simultaneously, the Wordle canvas can be too small to interact with. However, when the mouse cursor moves into a colored border (Figure 15-4), the card is enlarged to the users manipulate the layout (Figure 24). Also, placing the cursor on top of each word will highlight the same words in other cards. It should be noted that, while the highlighting can be achieved by change the words to more saturated colors, it is less suitable for Wordle-like word clouds because the colors are already used extensively. Therefore, we decided to dimmed the other words as well as applying glow effect.

Just like in ManiWordle, the location of each word can be manipulated by a simple dragging gesture. Also, since it shows the intention of setting the location of the particular words, they are ‘pinned’ (shown with the underline in highlighted mode). Also, just like ManiWordle, the users can re-layout and re-color the words without affect the positions and the colors of the pinned words. They can also unpin on the canvas. Lastly, users can change the color of individual words by selecting the color shown in the context menu by right clicking the word. Because each ManiWordle is drawn on a smaller canvas on a screen, it always zooms in to fit the layout, the feature

not implemented in ManiWordle before.

Figure 24. When the mouse cursor enters the colored boundary of the ManiWordle control, the card is enlarged in order to help manipulate the words' positions.

Figure 25. When the mouse cursor is placed on top of each word, the same words in the other cards are highlighted. Also, the users can see the number of occurrence for the words in the news articles within the card's date-time range.

5.5.2 Words Tab

In 5.3.3, we explained how global filters and database-bound filters are defined by users. The last filter is a Wordle-specific filter that is only applied to the current ManiWordle. There are two ways to a generate filter: By right-clicking on a word on the Wordle control directly and choose ‘remove’; or to choose it from a ‘Words’ tab and remove it the table (Figure 15-6). Also, in order to allow users to remove the words that were not seen in the global filter window, but in Wordle control, we allow them to remove the words globally from the Wordle control also.

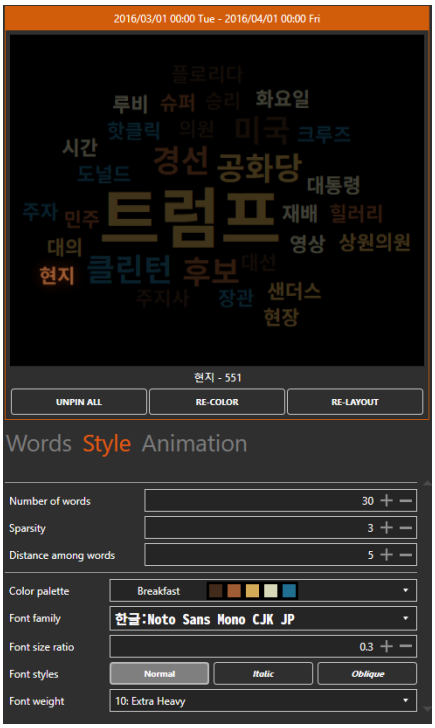

Figure 26. The style tab allows users to change attributes of the ManiWordle.

5.5.3 Style Tab

The style tab (Figure 26) provides options to manipulate the attributes of the ManiWordle control. A participating Journalists has expressed that the style of each card does not always have to be unique. In fact, they preferred to have the style that is consistent across all cards in order to give uniformed feelings. Therefore, by default, any changes to one card's style is synchronized with other cards. The users can also specify the sparsity, the value that defines how scattered the words are on a canvas. Also distance among words define minimum distance that the actual inked regions of the words are set apart. If too small values are specified, the words maybe placed too close to one another. Color palettes provide both predetermined and custom palettes that the users can define. Also, various typeface attributes like font families, font styles, and font weight can be configured.

One notable feature of NewsWordle is to define the font size ratio of the Wordle control. Since some words can be much more frequent than the others, the size of less frequent words can be too small. This heavily depends on the data and there is no universal scaling curve that fits all. Therefore, we allowed users to set the size ratio between the smallest (the least frequent) word and the largest (the most frequent) word.

5.6 Publishing

When the users are ready to publish the Wordles, they are exported into image files. This allows news reporters to produce an output that can be embedded into a news article without getting extensive help from the artists. Also, even if the resulting files

are not directly published, it can be used as a guide for the artists when they create graphics using the tools they are familiar with due to high fidelity nature of the ManiWordle; i.e. the word layout has been hand-crafted by the journalists in order to reflect their intentions.

5.7 Case Study

We have conducted a case study that simulates the entire process of making graphics for news articles that involve news article analysis. The sample news articles were collected using NewsWordle between January 1st, 2016 and November 30th, 2016 on *Oh SeungHwan*, (a Korean Pitcher Major League Baseball). For the case study, we chose Yonhap News Agency as the source, due to the participants' familiarity with the media.

5.7.1 Crawling and Preprocessing

An initial search query has revealed that there are 809 news articles published between the time period for the keyword. The crawling, morpheme analysis, and Part of Speech tagging took about 50 seconds to finish. Upon completion, NewsWordle showed an initial screen with Wordless using random attributes (Figure 27).

Figure 27. An initial screen upon loading the database crawled with the keyword *Oh Seung Hwan(오승환)*.

5.7.2 Filtering

Initially, the monthly binning has been automatically selected since the date covers from January to November. It would be more appropriate to choose monthly over yearly, since choosing the latter would result in just one histogram and a Wordle for the entire database. By observing that view has been histogram view, there have been 2 or 3 news articles on Oh except for August. We can also see that the initial Wordles have too many noises such as the name of the news media (연합), domain names for the journalists' email (yna), copyright notices (무단 전재 금지, 저작권자) and etc. In order to apply general filters, we open a filter windows and add them to the global filter lists

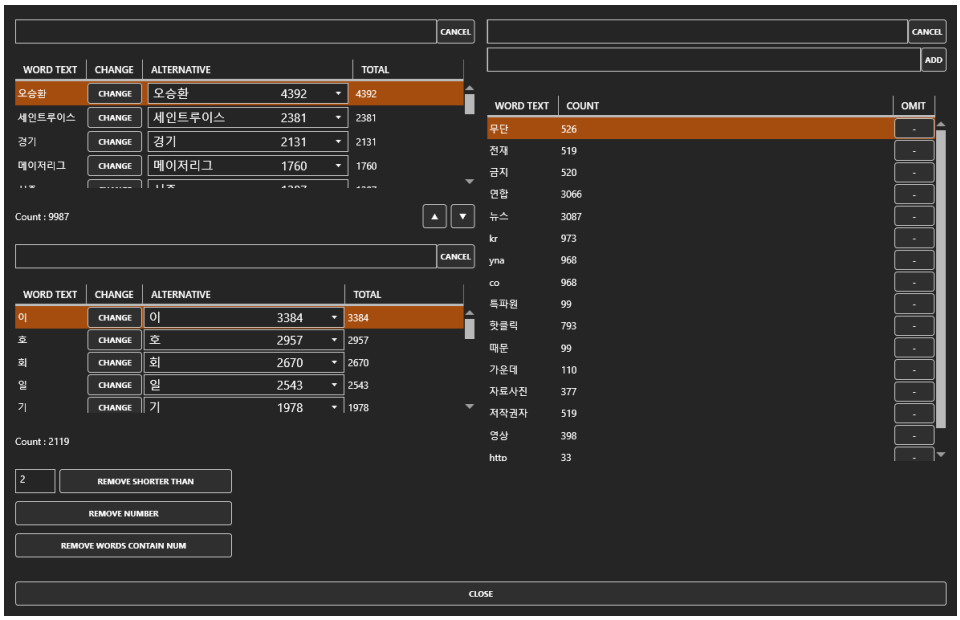

Figure 28. Most common words are added to the global filter list. Also, any words that are shorter than 2 letters are removed. The only numeric word that were prominent were '34' which is the age of Oh Seung Hwan. It shows up because of the convention of news articles that states someone's name followed by their age. However, it is not newsworthy.

After filters were applied, we still see that the Wordles are largely dominated by the player's name, which is excepted, but not significant findings. We also removed few other words that are expected in the dataset, but not significant, such as 'major league', 'St. Louis Cardinals' the team *Oh* has joined. After applying series of filters, Distinct patter began to emerge (Figure 29).

First, January had distinct keywords which are not found in other time period; *Gambling* and *Lim Chanyong*. This is when Oh was reported to have visited Macau

Figure 29. As the result of filtering and adjusting date-time range, 6 Wordles were generated.

(1) Jan 2016 (2) Feb 2016 (3) Mar~Apr 2016

(4) May~June 2016 (5) July~Aug 2016 (6) Sep~Nov 2016

for gambling with *Lim*. Still, the most important keyword is *contract*, since *Oh* made a contract with *St. Louis Cardinals*. Also the word *KBO* (*Korean Baseball Organization*) is prominent, because the news articles reported that *KBO* is investigating the gambling hubbub (Figure 29-1).

When the *spring camp* season came, the news articles began to mention *Jupiter*, *Florida* (Figure 29-2). Oh has begun his spring training as well. Also, the word *Park Byoungho* appeared as he signed up with *Minnesota Twins*. During March (Figure 29-3), *Trial* match, so called Cactus and Grapefruit Leagues were the most discussed topics. That is also when *Park Byoungho* began to show potential to become a star player.

When the season began (Figure 29-4), news articles primarily focused on Park Byoungho who hit many homeruns, while *Oh* was just mentioned shortly for appearing in the game as a bullpen pitcher. Therefore, only few words that are related to *Oh* during this period. As the season progressed, *Park's* record faded and Kang Jungho made a come-back from his injury. Also *Oh* began to appear for a closer role, as the word '*마무리*' can be seen at the time (Figure 29-5) During the last couple month, *Oh* began to appear regularly as the Cardinals' closer, whose decision was made by the manager (감독). Throughout the whole seasons, the news articles have mentioned the word *ballspeed* and *fastball*, for which *Oh* is known.

Chapter 6

Discussions

The participating journalists has shared their opinion on how NewsWordle helps their works to improve their production pipelines.

6.1 How NewsWordle Improves News Article Production

Supporting data collection and data analysis was important. These steps were often not achieved by the journalists due to limited resources. Before making any requests to data experts for data collection and analysis, the journalists had to make careful assumption that the data would reveal meaningful and newsworthy trend. Also, because they may not be familiar with what the current data collection/analysis technologies are capable of and what their limitations are, the journalists have to discuss such issues with data experts before the requests can be made. Since this process involves collaborating with the others, the cost of abandoning the data and

reiterating become significant. With NewsWordle the journalists can now quickly collect the data and analyze them without the help of other data experts. Because they can specify the attributes for crawling data defined by the news media, search keywords, and date rage, they can collect all data as they become necessary during analysis and production. They can also massage the dataset by merging/splitting and filtering noises to extract meaning insights. If they do not acquire any significant or newsworthy findings from the data, they can perform another search. This reduces the costs of communications and reiterations with data experts.

Second, using the visualization that supports both analysis and generations for article contents reduced the production time. Since Wordle is a form of word-clouds that emphasizes particularly on the aesthetic qualities, it may not be as robust as other visualizations that are curated for visual analytics. However, using separate visualization techniques for data analysis and generating graphics suitable for article figures turned out to be problematic. If the visualization used for analytic task cannot be directly transformed into a publication-ready graphics, the journalists have to make separate requests to graphic designers/artists. And since making such requests often involves: explaining the underlying data; screenshots of the interactive visual analytics tool that lead them to such insights; and rough sketches on the content of the graphics for what and how the data is portrayed, the process can also become costly, as in the previous example with data experts. Therefore, the journalists requested that discrepancy between the visualization used in visual analytics and the final graphics to be minimum. Also, they argued that the visualization to be scalable

for consumptions regardless of the devices form factors or the platform.

We reviewed various text visualization techniques with the journalists, many of which involved interactivity or had non-aesthetic looks. At the end, Wordle was selected as the visualization to achieve such goals in NewsWordle. The exported images can be directly embedded into the news articles that are consumable on any type of devices. In addition, when the journalists feel necessary to add other miscellaneous visual embellishment to the originally exported image by NewsWordle, the high fidelity nature of Wordle give the artists good directions for their work. Lastly, they are aesthetically pleasing as agreed by the journalists and shown in the previous literatures [98][99]. Especially, even in increasing uses of mobile devices for news consumption, the journalists argued that aesthetic images made of large fonts would be beneficial in grabbing the readers' attention.

Nevertheless, some adjustments had to be made to the visualization in NewsWordle to support the journalists' goals. Because both the strengths and the limitations of the original Wordle come from the randomness of the form and layout, we enhanced the Wordle in NewsWordle with interaction techniques applied to ManiWordle. By doing so could the journalists make a visualization as close to the final graphics for news articles as possible. Also, since the original ManiWordle design provides only limited way to massaging the data, we introduced the filters and the rules to help journalists cleanup the dataset. The journalists have identified that three levels of filters were especially useful for because it reduces time for creating the same filtering rules for each database for filtering noises. The global filter would remove the noises that are

present in any databases like copyright notices and certain proper nouns. The database-bound filter was useful for filtering the words that are not particularly relevant to the current database, but maybe so in other databases. For example, the named entity ‘major league baseball’ would be the most frequently used words in the news articles crawled with the same keyword. The word itself would provide little insight on what the articles talked about ‘major league baseball.’ However, if the same word begins to appear in news articles crawled in different categories, such as political news, it may be noteworthy. Therefore, the separation of the global filter and the database-bound filter provided the journalists greater flexibility. Lastly the card-bound filter allowed finer tuning on generating a Wordle image.

6.2 Generalizations

Although the scope of NewsWordle was to use news articles that are published in the past as data, there are general strategies for making tools for journalist.

Because it can be costly to rely on other experts or contracts to analyze the data and produce graphics, it is important to minimize the iterations between journalist and the data scientists/artists. To achieve such goals, the tool used by journalists need to allow fast iteration of data explorations using their own resources only. Heavy reliance on external resources like custom servers cannot be utilized every day. For many part, rather simpler analysis of the text data may suffice their uses of detecting topics changes and the trends.

Also, the tool needs to produce the figures that are in high-enough fidelity. Once the artists feel the need of exploring and interpreting the data by themselves, the end

result may deviate away from the journalists' intention, resulting in re-work of the entire process. To avoid getting this duplicate process, the figures generated by the tool can be supplemented with the metadata that were used to generate them.

6.3 Limitations

We have identified some limitations to current NewsWordle design.

First, the histogram view can be improved by adopting more visual elements that can present both the number of articles in each database and how many of them overlap with each other. It can be used to see how cohesive the keywords. If many news articles mention multiple keywords simultaneously, we can assume that those keywords are well coupled in terms of their themes. In such case, it may be sufficient to crawl data only on one keyword, whereas if there are only few news articles the keywords are discrete and have to be collected separately. Cleverer visualization may help journalists identify such characteristics of the datasets. For example, the concept behind Pixel bar charts by Keim et al. [55] may be applied. Also, the date-time range of each card does not have to be uniform, thicker, but shorter bars may be improved by embedding other visualizations onto it, as in the examples of TreemapBar by Huang et al. [50]

Also, the use of Gantt view can be improved. Although the intention of Gantt char was to show the range from which the articles were collected from, it was not effective in showing how the articles were actually distributed. For example, the journalists may request to crawl news articles on a keyword 'Miley Cyrus' starting from 1991. However, since the singer debuted in 2001, there will be only a small number of news

articles between 1991 and 2001, if at all. The current design of Gantt view in NewsWordle does not reflect this. While the journalists can still refer to the histogram view to get the same insight, it would save screen space if Gantt and histogram views can be represented in more compact way without sacrificing their readability.

Also, the journalists argued that it would be useful if the system can automatically generate textual description of Wordles, especially in order to avoid journalists' mistakes. This has led us to an interesting future research area, a semi-guided robot journalism. Because the unstructured text of news articles' content has been transformed into the structure data, it can be used to generate textual descriptions. Especially, such automation will reduce the overall production time for journalists, because they can simply edit the generated textual description instead writing sentences without a starting point. Also, it will help reduce missing significant features that were not visual salient during the journalists' own visual exploration. For example, the topics that have not been discussed for a while and have reemerged later can be detected by machine learning.

Chapter 7

Conclusion

For this thesis, we presented a design study of topic visualization techniques on news article data for story telling by the journalists working in the field. We have recognized the problems of applying current state-of-arts InfoVis tools to journalists' own production pipeline. We identified that the journalists are often limited by the lack of resources for data collection and analysis and difficulty of communications with the experts. Focusing on analyzing old news archives, we designed a new tool named NewsWordle for journalists to facilitate: news article crawling, topic analysis based on date-time ranges, Wordles that show topics and trends, and the ability to generate the figures that can be embedded directly into a news article, or used as high fidelity prototype for artists to improve upon. For creating Wordle, we adapted techniques from ManiWordle and allowed journalists generate a semi-guided layout

of the Wordle. We found that using NewsWordle, the journalists were able to collect and cleanup multiple datasets that consist of news articles from multiple sources until they were able to find meaningful and newsworthy stories. Also, they were able to generate images that can be used in a news article, or can deliver intention to artists/graphic designers in high fidelity. Through a case study, we have also identified the strength and weakness, as well as some generalization strategies.

Some possible future direction to consider is the ability to crawl, analyze, and visualize readers' comments. Although not selected as an initial requirement, it became to be seem more necessary as the journalists began to wonder how readers react to the news articles. We think sentiment analysis based on comments can be compared to those of the news article written by the journalists to see if there is any discrepancy between them. Because opinion mining and sentiment analysis is a topic that is widely explored in the past literature, we believe it will be a promising future venue for NewsWordle as well.

Bibliography

[1] Facebook, http://www.facebook.com

[2] Naver Corporation, http://www.naver.com

[3] Alexander, Bryzan. "Web 2.0." A New Wave of Innovation for Teachning and learning (2006): 32-44.

[4] Archambault, Daniel, Derek Greene, Pádraig Cunningham, and Neil Hurley. "ThemeCrowds: Multiresolution summaries of twitter usage." In Proceedings of the 3rd international workshop on Search and mining user-generated contents, pp. 77-84. ACM, 2011.

[5] Archambault, Daniel, Derek Greene, and Pádraig Cunningham. "Twittercrowds: Techniques for exploring topic and sentiment in microblogging data." arXiv preprint arXiv:1306.3839 (2013).

[6] Barth, Lukas, Stephen G. Kobourov, and Sergey Pupyrev. "Experimental comparison of semantic word clouds." In International Symposium on Experimental Algorithms, pp. 247-258. Springer International Publishing, 2014.

[7] Bateman, Scott, Regan L. Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. "Useful junk?: the effects of visual embellishment on comprehension and memorability of charts." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2573-2582. ACM, 2010.

[8] Bautin, Mikhail, Charles B. Ward, Akshay Patil, and Steven S. Skiena. "Access: news

and blog analysis for the social sciences." In Proceedings of the 19th international conference on World wide web, pp. 1229-1232. ACM, 2010.

[9] Bernstein, Michael S., Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairam, and Ed H. Chi. "Eddi: interactive topic-based browsing of social status streams." In Proceedings of the 23nd annual ACM symposium on User interface software and technology, pp. 303-312. ACM, 2010.

[10] Bigelow, Alex, Steven Drucker, Danyel Fisher, and Miriah Meyer. "Reflections on how designers design with data." In Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces, pp. 17-24. ACM, 2014.

[11] Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer. "D³ data-driven documents." IEEE transactions on visualization and computer graphics 17, no. 12 (2011): 2301-2309.

[12] Branch, John, Snow Fall, The Avalanche at Tunnel Creek http://www.nytimes.com/projects/2012/snow-fall New York Times, (2012)

[13] Brath, Richard, and Ebad Banissi. "Using Font Attributes in Knowledge Maps and Information Retrieval." Proceedings of Knowledge Maps and Information Retrieval (KMIR) at Digital Libraries (2014).

[14] Buchin, Kevin, Daan Creemers, Andrea Lazzarotto, Bettina Speckmann, and Jules Wulms. "Geo word clouds." In 2016 IEEE Pacific Visualization Symposium (PacificVis), pp. 144-151. IEEE, 2016.

[15] Byron, Lee, and Martin Wattenberg. "Stacked graphs–geometry & aesthetics." IEEE transactions on visualization and computer graphics 14, no. 6 (2008): 1245-1252.

[16] Cao, Nan, Jimeng Sun, Yu-Ru Lin, David Gotz, Shixia Liu, and Huamin Qu. "Facetatlas: Multifaceted visualization for rich text corpora." IEEE transactions on visualization and computer graphics 16, no. 6 (2010): 1172-1181.

[17] Card, Stuart K., Jock D. Mackinlay, and Ben Shneiderman. Readings in information visualization: using vision to think. Morgan Kaufmann, 1999.

[18] Chyi, Hsiang Iris, and Monica Chadha. "News on new devices: Is multi-platform news consumption a reality?." Journalism Practice 6, no. 4 (2012): 431-449.

[19] Chan, Yeuk-Yin, and Huamin Qu. "FinaVistory: Using Narrative Visualization to explain social and Economic relationships in financial news." In 2016 International Conference on Big Data and Smart Computing (BigComp), pp. 32-39. IEEE, 2016.

[20] Chan-Olmsted, Sylvia, Hyejoon Rim, and Amy Zerba. "Mobile news adoption among young adults examining the roles of perceptions, news consumption, and media usage." Journalism & Mass Communication Quarterly 90, no. 1 (2013): 126-147.

[21] Chieu, Hai Leong, and Hwee Tou Ng. "Named entity recognition: a maximum entropy approach using global information." In Proceedings of the 19th international conference on Computational linguistics-Volume 1, pp. 1-7. Association for Computational Linguistics, 2002.

[22] Chi, Ming-Te, Shih-Syun Lin, Shiang-Yi Chen, Chao-Hung Lin, and Tong-Yee Lee. "Morphable Word Clouds for Time-Varying Text Data Visualization." IEEE transactions on visualization and computer graphics 21, no. 12 (2015): 1415-1426.

[23] Choo, Jaegul, Changhyun Lee, Chandan K. Reddy, and Haesun Park. "Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization." IEEE

transactions on visualization and computer graphics 19, no. 12 (2013): 1992-2001.\

[24] Chuang, Jason, Christopher D. Manning, and Jeffrey Heer. "“Without the clutter of unimportant words”: Descriptive keyphrases for text visualization." ACM Transactions on Computer-Human Interaction (TOCHI) 19, no. 3 (2012): 19.

[25] Chyi, Hsiang Iris, and Angela M. Lee. "Online news consumption: A structural model linking preference, use, and paying intent." Digital Journalism 1, no. 2 (2013): 194-211.

[26] Cohen, Sarah, James T. Hamilton, and Fred Turner. "Computational journalism." Communications of the ACM 54, no. 10 (2011): 66-71.

[27] Collins, Christopher, Fernanda B. Viegas, and Martin Wattenberg. "Parallel tag clouds to explore and analyze faceted text corpora." In Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on, pp. 91-98. IEEE, 2009.

[28] Cui, Weiwei, Huamin Qu, Hong Zhou, Wenbin Zhang, and Steve Skiena. "Watch the story unfold with textwheel: Visualization of large-scale news streams." ACM Transactions on Intelligent Systems and Technology (TIST) 3, no. 2 (2012): 20.

[29] Cui, Weiwei, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. "Textflow: Towards better understanding of evolving topics in text." IEEE transactions on visualization and computer graphics 17, no. 12 (2011): 2412-2421.

[30] Cui, Weiwei, Yingcai Wu, Shixia Liu, Furu Wei, Michelle X. Zhou, and Huamin Qu. "Context preserving dynamic word cloud visualization." In 2010 IEEE Pacific Visualization Symposium (PacificVis), pp. 121-128. IEEE, 2010.

[31] Diakopoulos, Nicholas, Dag Elgesem, Andrew Salway, Amy Zhang, and Knut Hofland. "Compare clouds: Visualizing text corpora to compare media frames." In Proc. of IUI Workshop on Visual Text Analytics. 2015.

[32] Dork, Marian, Daniel Gruen, Carey Williamson, and Sheelagh Carpendale. "A visual backchannel for large-scale events." IEEE transactions on visualization and computer graphics 16, no. 6 (2010): 1129-1138.

[33] Dou, Wenwen, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky. "Hierarchicaltopics: Visually exploring large text collections using topic hierarchies." IEEE Transactions on Visualization and Computer Graphics 19, no. 12 (2013): 2002-2011.

[34] Dou, Wenwen, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle X. Zhou. "Leadline: Interactive visual analysis of text data through event identification and exploration." In Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, pp. 93-102. IEEE, 2012.

[35] Dougherty, Edward R., Roberto A. Lotufo, and The International Society for Optical Engineering SPIE. Hands-on morphological image processing. Vol. 71. Washington: SPIE Optical Engineering Press, 2003.

[36] Dubinko, Micah, Ravi Kumar, Joseph Magnani, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. "Visualizing tags over time." ACM Transactions on the Web (TWEB) 1, no. 2 (2007): 7.

[37] Esuli, Andrea, and Fabrizio Sebastiani. "Sentiwordnet: A publicly available lexical resource for opinion mining." In Proceedings of LREC, vol. 6, pp. 417-422. 2006.

[38] Few, Stephen, and Perceptual Edge. "The chartjunk debate." Visual Business Intelligence Newsletter, no. June (2011): 1-11.

[39] Fisher, Danyel, Aaron Hoff, George Robertson, and Matthew Hurst. "Narratives: A visualization to track narrative events as they develop." In Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on, pp. 115-122. IEEE, 2008.

[40] Fischer, Fabian, and Daniel A. Keim. "NStreamAware: Real-time visual analytics for data streams to enhance situational awareness." In Proceedings of the Eleventh Workshop on Visualization for Cyber Security, pp. 65-72. ACM, 2014.

[41] Flew, Terry, Christina Spurgeon, Anna Daniel, and Adam Swift. "The promise of computational journalism." Journalism Practice 6, no. 2 (2012): 157-171.

[42] Freixa, Pere, Carles Sora, Joan Soler-Adillon, and J. Ignaci RIBAS. "Snow Fall and A Short History of the Highrise: two approaches to interactive communication design by The New York Times." Textual & Visual Media. Nº7 (2014): 63-84.

[43] Gansner, Emden R., Yifan Hu, and Stephen North. "Visualizing streaming text data with dynamic graphs and maps." In International Symposium on Graph Drawing, pp. 439-450. Springer Berlin Heidelberg, 2012.

[44] Gelman, Andrew, and Antony Unwin. "Infovis and statistical graphics: different goals, different looks." Journal of Computational and Graphical Statistics 22, no. 1 (2013): 2-28.

[45] Gleicher, Michael, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D. Hansen, and Jonathan C. Roberts. "Visual comparison for information visualization." Information Visualization 10, no. 4 (2011): 289-309.

[46] Golombisky, Kim, and Rebecca Hagen. White space is not your enemy: A beginner's guide to communicating visually through graphic, web & multimedia design. Taylor & Francis, 2013.

[47] Ham, Kelli. "OpenRefine (version 2.5). http://openrefine. org. Free, open-source tool for cleaning and transforming data." Journal of the Medical Library Association: JMLA 101, no. 3 (2013): 233.

[48] Havre, Susan, Beth Hetzler, and Lucy Nowell. "ThemeRiver: Visualizing theme changes over time." In Information Visualization, 2000. InfoVis 2000. IEEE Symposium on, pp. 115-123. IEEE, 2000.

[49] Heimerl, Florian, Steffen Lohmann, Simon Lange, and Thomas Ertl. "Word cloud explorer: Text analytics based on word clouds." In 2014 47th Hawaii International Conference on System Sciences, pp. 1833-1842. IEEE, 2014.

[50] Huang, Mao Lin, Tze-Haw Huang, and Jiawan Zhang. "TreemapBar: Visualizing additional dimensions of data in bar chart." In 2009 13th International Conference Information Visualisation, pp. 98-103. IEEE, 2009.

[51] Itti, Laurent, and Christof Koch. "Computational modelling of visual attention." Nature reviews neuroscience 2, no. 3 (2001): 194-203.

[52] Jacobson, Susan, Jacqueline Marino, and Robert E. Gutsche. "The digital animation of literary journalism." Journalism (2015): 1464884914568079.

[53] Jo, Jaemin, Bongshin Lee, and Jinwook Seo. "WordlePlus: Expanding Wordle's Use through Natural Interaction and Animation." IEEE computer graphics and applications 35, no. 6 (2015): 20-28.

[54] Kandel, Sean, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. "Wrangler: Interactive visual specification of data transformation scripts." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3363-3372. ACM, 2011.

[55] Keim, Daniel A., Ming C. Hao, Julian Ladisch, Meichun Hsu, and Umeshwar Dayal. "Pixel bar charts: A new technique for visualizing large multi-attribute data sets without aggregation." In IEEE Symposium on Information Visualization, 2001. INFOVIS 2001., pp. 113-120. 2001.

[56] Kim, KyungTae, Sungahn Ko, Niklas Elmqvist, and David S. Ebert. "WordBridge: Using composite tag clouds in node-link diagrams for visualizing content and relations in text corpora." In Sy stem Sciences (HICSS), 2011 44th Hawaii International Conference on, pp. 1-8. IEEE, 2011.

[57] Kling, Felix, and Alexei Pozdnoukhov. "When a city tells a story: urban topic analysis." In Proceedings of the 20th international conference on advances in geographic information systems, pp. 482-485. ACM, 2012.

[58] Koh, Kyle, Bongshin Lee, Bohyoung Kim, and Jinwook Seo. "Maniwordle: Providing flexible control over wordle." IEEE Transactions on Visualization and Computer Graphics 16, no. 6 (2010): 1190-1197.

[59] Krauss, J. "More than words can say: infographics." Learning and leading with technology 5, no. 39 (2012): 10-14.

[60] Latar, Noam Lemelshtrich. "The Robot Journalist in the Age of Social Physics: The End of Human Journalism?." In The New World of Transitioned Media, pp. 65-80.

Springer International Publishing, 2015.

[61] Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. "Meme-tracking and the dynamics of the news cycle." In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 497-506. ACM, 2009.

[62] Lee, Angela M., and Hsiang Iris Chyi. "The rise of online news aggregators: Consumption and competition." International Journal on Media Management 17, no. 1 (2015): 3-24.

[63] Lee, Bongshin, Nathalie Henry Riche, Amy K. Karlson, and Sheelash Carpendale. "Sparkclouds: Visualizing trends in tag clouds." IEEE transactions on visualization and computer graphics 16, no. 6 (2010): 1182-1189.

[64] Li, Xigen. Internet newspapers: The making of a mainstream medium. Routledge, 2013.

[65] Lidwell, William, Kritina Holden, and Jill Butler. Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design. Rockport Pub, 2010.

[66] Lira, Wallace, Fernando Gama, Hivana Barbosa, Ronnie Alves, and Cleidson de Souza. "VCloud: adding interactiveness to word clouds for knowledge exploration in large unstructured texts." In Proceedings of the 31st Annual ACM Symposium on Applied Computing, pp. 193-198. ACM, 2016.

[67] Liu, Shenghua, Wenjun Zhu, Ning Xu, Fangtao Li, Xue-qi Cheng, Yue Liu, and Yuanzhuo Wang. "Co-training and visualizing sentiment evolvement for tweet events." In Proceedings of the 22nd International Conference on World Wide Web,

pp. 105-106. ACM, 2013.

[68] Liu, Shixia, Xiting Wang, Jianfei Chen, Jim Zhu, and Baining Guo. "Topicpanorama: a full picture of relevant topics." In Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on, pp. 183-192. IEEE, 2014.

[69] Liu, Xiaotong, Han-Wei Shen, and Yifan Hu. "Supporting multifaceted viewing of word clouds with focus+ context display." Information Visualization (2014): 1473871614534095.

[70] Lohmann, Steffen, Michael Burch, Hansjörg Schmauder, and Daniel Weiskopf. "Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds." In Proceedings of the International Working Conference on Advanced Visual Interfaces, pp. 753-756. ACM, 2012.

[71] Lohmann, Steffen, Florian Heimerl, Fabian Bopp, Michael Burch, and Thomas Ertl. "Concentri Cloud: Word Cloud Visualization for Multiple Text Documents." In 2015 19th International Conference on Information Visualisation, pp. 114-120. IEEE, 2015.

[72] Luo, Dongning, Jing Yang, Milos Krstajic, William Ribarsky, and Daniel Keim. "Eventriver: Visually exploring text collections with temporal references." IEEE Transactions on Visualization and Computer Graphics 18, no. 1 (2012): 93-105.

[73] Malik, Sana, Alison Smith, Timothy Hawes, Panagis Papadatos, Jianyu Li, Cody Dunne, and Ben Shneiderman. "Topicflow: visualizing topic alignment of twitter data over time." In Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining, pp. 720-726. ACM, 2013.

[74] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven

Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit." In ACL (System Demonstrations), pp. 55-60. 2014.

[75] Marcus, Adam, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. "Twitinfo: aggregating and visualizing microblogs for event exploration." In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 227-236. ACM, 2011.

[76] McCormick, Rick. “AP's 'robot journalists' are writing about Minor League Baseball now”, The Verge, http://www.theverge.com/2016/7/4/12092768/ap-robot-journalists-automated-insights-minor-league-baseball Accessed on Oct 10th 2016

[77] Meyer, Robinson. “How Many Stories Do Newspapers Publish Per Day?” the Atlantic, http://www.theatlantic.com/technology/archive/2016/05/how-many-stories-do-newspapers-publish-per-day/483845/ Accessed on Oct 5th 2016

[78] Ng, Yik-Wai, and Huamin Qu. "TrendFocus: Visualization of trends in financial news with indicator sets." In 2014 International Conference on Big Data and Smart Computing (BIGCOMP), pp. 7-12. IEEE, 2014.

[79] Nguyen, Dinh Quyen, Christian Tominski, Heidrun Schumann, and Tuan Anh Ta. "Visualizing tags with spatiotemporal references." In 2011 15th International Conference on Information Visualisation, pp. 32-39. IEEE, 2011.

[80] Nguyen, Vu Dung, Blesson Varghese, and Adam Barker. "The royal birth of 2013: Analysing and visualising public sentiment in the uk using twitter." In Big Data, 2013 IEEE International Conference on, pp. 46-54. IEEE, 2013.

[81] Otten, Jennifer J., Karen Cheng, and Adam Drewnowski. "Infographics and public

policy: using data visualization to convey complex information." Health Affairs 34, no. 11 (2015): 1901-1907.

[82] Pan, Shimei, Michelle X. Zhou, Yangqiu Song, Weihong Qian, Fei Wang, and Shixia Liu. "Optimizing temporal topic segmentation for intelligent text visualization." In Proceedings of the 2013 international conference on Intelligent user interfaces, pp. 339-350. ACM, 2013.

[83] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2, no. 1-2 (2008): 1-135.

[84] Paulovich, Fernando V., Franklina Toledo, Guilherme P. Telles, Rosane Minghim, and Luis Gustavo Nonato. "Semantic wordification of document collections." In Computer Graphics Forum, vol. 31, no. 3pt3, pp. 1145-1153. Blackwell Publishing Ltd, 2012.

[85] Reas, Casey, and Benjamin Fry. "Processing: a learning environment for creating interactive Web graphics." In ACM SIGGRAPH 2003 Web Graphics, pp. 1-1. ACM, 2003.

[86] Pousman, Zachary, John Stasko, and Michael Mateas. "Casual information visualization: Depictions of data in everyday life." IEEE transactions on visualization and computer graphics 13, no. 6 (2007): 1145-1152.

[87] Riehmann, Patrick, Manfred Hanfler, and Bernd Froehlich. "Interactive sankey diagrams." In IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., pp. 233-240. IEEE, 2005.

[88] Rivadeneira, Anna W., Daniel M. Gruen, Michael J. Muller, and David R. Millen.

"Getting our head in the clouds: toward evaluation studies of tagclouds." In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 995-998. ACM, 2007.

[89] Robertson, George, Roland Fernandez, Danyel Fisher, Bongshin Lee, and John Stasko. "Effectiveness of animation in trend visualization." IEEE Transactions on Visualization and Computer Graphics 14, no. 6 (2008): 1325-1332.

[90] Rose, Stuart, Scott Butner, Wendy Cowley, Michelle Gregory, and Julia Walker. "Describing story evolution from dynamic information streams." In Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on, pp. 99-106. IEEE, 2009.

[91] Segel, Edward, and Jeffrey Heer. "Narrative visualization: Telling stories with data." IEEE transactions on visualization and computer graphics 16, no. 6 (2010): 1139-1148.

[92] Seifert, Christin, Barbara Kump, Wolfgang Kienreich, Gisela Granitzer, and Michael Granitzer. "On the beauty and usability of tag clouds." In 2008 12th International Conference Information Visualisation, pp. 17-25. IEEE, 2008.

[93] Shneiderman, Ben. "The eyes have it: A task by data type taxonomy for information visualizations." In Visual Languages, 1996. Proceedings., IEEE Symposium on, pp. 336-343. IEEE, 1996.

[94] Shi, Lei, Furu Wei, Shixia Liu, Li Tan, Xiaoxiao Lian, and Michelle X. Zhou. "Understanding text corpora with multiple facets." In Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on, pp. 99-106. IEEE, 2010.

[95] Siricharoen, Waralak V. "Infographics: The new communication tools in digital age." In The International Conference on E-Technologies and Business on the Web (EBW2013), pp. 169-174. The Society of Digital Information and Wireless Communication, 2013.

[96] Strobelt, Hendrik, Marc Spicker, Andreas Stoffel, Daniel Keim, and Oliver Deussen. "Rolled-out Wordles: A Heuristic Method for Overlap Removal of 2D Data Representatives." In Computer Graphics Forum, vol. 31, no. 3pt3, pp. 1135-1144. Blackwell Publishing Ltd, 2012.

[97] Sun, Guodao, Yingcai Wu, Shixia Liu, Tai-Quan Peng, Jonathan JH Zhu, and Ronghua Liang. "EvoRiver: Visual analysis of topic coopetition on social media." IEEE transactions on visualization and computer graphics 20, no. 12 (2014): 1753-1762.

[98] Viégas, Fernanda B., and Martin Wattenberg. "Artistic data visualization: Beyond visual analytics." In International Conference on Online Communities and Social Computing, pp. 182-191. Springer Berlin Heidelberg, 2007.

[99] Viegas, Fernanda B., Martin Wattenberg, and Jonathan Feinberg. "Participatory visualization with Wordle." IEEE transactions on visualization and computer graphics 15, no. 6 (2009): 1137-1144.

[100] Wang, Changbo, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang. "SentiView: Sentiment analysis and visualization for internet popular topics." IEEE transactions on human-machine systems 43, no. 6 (2013): 620-630.

[101] Weber, Wibke, and Hannes Rall. "Data visualization in online journalism and its

implications for the production process." In 2012 16th International Conference on Information Visualisation, pp. 349-356. IEEE, 2012.

[102] Wei, Furu, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. "Tiara: a visual exploratory text analytic system." In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 153-162. ACM, 2010.

[103] Wu, Yingcai, Shixia Liu, Kai Yan, Mengchen Liu, and Fangzhao Wu. "Opinionflow: Visual analysis of opinion diffusion on social media." IEEE transactions on visualization and computer graphics 20, no. 12 (2014): 1763-1772.

[104] Wu, Yingcai, Thomas Provan, Furu Wei, Shixia Liu, and Kwan-Liu Ma. "Semantic-preserving word clouds by seam carving." In Computer Graphics Forum, vol. 30, no. 3, pp. 741-750. Blackwell Publishing Ltd, 2011.

[105] Xu, Panpan, Yingcai Wu, Enxun Wei, Tai-Quan Peng, Shixia Liu, Jonathan JH Zhu, and Huamin Qu. "Visual analysis of topic competition on social media." IEEE Transactions on Visualization and Computer Graphics 19, no. 12 (2013): 2012-2021.

[106] Zhao, Jichang, Li Dong, Junjie Wu, and Ke Xu. "Moodlens: an emoticon-based sentiment analysis system for chinese tweets." In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1528-1531. ACM, 2012.

국문 초록

서울대학교 대학원

전기.컴퓨터공학부

고 건

온라인 텍스트 데이터의 양이 증가함에 따라 데이터를 분석하고 스토리텔링을 위한 내러티브를 구성하고, 이를 위한 그래픽을 생성하는 일이 어려워 지고 있다. 이에 언론인들은 뉴스 기사를 생성함에 있어 데이터 전문가나 아티스트들과의 상호작용이나 개선을 위한 반복 작업에 어려움을 느끼고 있다. 이러한 문제들을 줄이기 위해 언론인들은 직접 데이터를 수집하여 빠른 분석을 해 볼 수 있어야 하며, 초기의 결과에 따라 데이터에서 기삿거리의 유무 여부와 추가 상세 분석에 대한 필요성을 결정할 필요가 있다. 또한 분석이 완료되어 필요한 인사이트가 발견 되더라도 이를 아티스트에게 효과적으로 전달하여 스토리에 연관된 그래픽을 생성냄과 동시에 데이터를 잘못 나타내는 일을 방지하는 것이 필요하다. 또한 출판 플랫폼의 제한으로 인해 상호작용이 들어간 시각화 기법을 사용하는데 제약이 들어가기도 한다.

이 논문에서는 현직 언론인들의 도움을 얻어 기존의 기사 생성 파이프라인의 다양한 단계들을 개선하기 위한 도구인 ‘뉴스워들’의 디자

인 스터디를 제시한다. 특히 과거에 출판된 뉴스 기사들의 텍스트 데이터를 수집, 분석, 시각화한 후미적인 측면을 강조하는 단어 구름 기법인 ‘워들’을 통해 이를 표현하는 방법을 제시한다. 일단 현재 언론에 있는 기자들이 정보 시각화 커뮤니티에서 지속적으로 다루고 있는 인터렉티브 시각화 기법을 뉴스 기사에 도입하기 위해 겪어야 하는 도전과제와 어려움에 대해 설명한다. 그리고 실제 현장에 있는 기자들의 도움을 얻어 대량의 과거의 신문기사 텍스트를 탐색할 수 있는 도구의 요구조건과 디자인을 도출하는 과정과 기자들과 함께 원하는 요구조건 충족여부를 평가하고 케이스 스터디를 통해 해당 도구가 기사 송출에 활용되는 과정을 묘사한다. 이 과정을 통해 분석 작업과 아티스트들을 위한 시각 프로토타입에 활용되는 시각화를 하나로 활용하면 각 작업에 들어가는 시간과 노력을 줄일 수 있다는 것을 알아낼 수 있었다. 마지막으로 이같은 문제들을 해결하는 도구의 디자인에 대해 논의한다.

주요어 : 단어 구름, 텍스트 시각화, 탐험적 검색, 온라인 저널리즘

학 번 : 2009-23084