언론정보학박사학위논문

An Algorithmic Approach to Personalized and Interactive News Generation

알고리즘에 기반한 개인화되고 상호작용적인 뉴스 생성에 관한 연구

2017년 2월

서울대학교 대학원
언론정보학과
김동환

Abstract

An Algorithmic Approach to Personalized and Interactive News Generation

Dongwhan Kim
Department of Communication
Seoul National University

Algorithms are increasingly playing an important role in the production of news content with growing computational capacity. Moreover, the use of the algorithm is taking up traditional human roles as increasing number of journalistic activities are mediated by software. For instance, the Los Angele Times runs software called Quakebot, which makes automated decisions on publishing news articles on abnormal seismic events. The Associated Press and Forbes have long been publishing algorithm-generated news content in collaboration with narrative-generation algorithm developers since 2014. The Washington Post also joined the trend by developing news reporting software for 2016 Rio Olympics.

We were motivated by the advent of various algorithm-generated news products. We reviewed current practices of algorithm-generated news and classified common algorithmic attributes to derive insights on how to maximize the capacity of the algorithm for more engaging and appealing news content

generation. The key opportunity areas we found were 1) broadening depth and breadth of input data enriches algorithmic computation, 2) personalizing the narrative in the context of news readers raises interest, 3) presenting interactive user interface components helps to engage news readers and make them more active news consumers.

We designed an algorithmic framework based on the proposed key concepts and implemented a news generation system called PINGS, which is capable of generating more personalized and interactive news stories. In this thesis, we describe the design process and implementation details that shaped the PINGS. We present a study on how news readers perceive the news values of the content generated by PINGS as well as the comments and opinions on its potential influence in the field and usability and usefulness of the system by recruiting experts for qualitative review. This thesis includes discussions on our approach to design and implement personalization and interactivity functions into a news system, and contributions it makes to the fields of journalism and HCI.

Keywords : Algorithm Journalism, Robot Journalism, News, Framework, Consolidated Database, Personalization, Interactivity, News System, HCI, Interface, Interaction, Design

Student Number : 2012-30846

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recently, there has been a growing number of discussions about utilizing computational capacity in the practices of journalism (Hamilton and Turner, 2009; Cohen et al., 2011; Gynnild, 2014). The use of a computer as a tool or technique in news reporting is not a novel idea, but it is changing how we access, organize, and make sense of information as well as the ways we interact with information (Daniel and Flew, 2010; Flew et al., 2012; Diakopoulos, 2011). This added capacity comes from the power of algorithms, which not only help us to find information but also provide a means to know the scope of what there is to know and how to know it (Gillespie, 2014). It means that finding more relevant, valuable, and reliable news in this information age is a step closer with the help of algorithms (van Dalen, 2012; Carlson, 2015).

Moreover, algorithms increasingly take traditionally human roles in the production of journalism as they are capable of making autonomous decisions and creating content that is indiscernible from that of human writers (Diakopoulos, 2014; Clerwall, 2014; Dörr, 2016). For instance, the software running on the servers of the Los Angeles Times, called Quakebot, detects abnormal seismic activities, writes a brief earthquake report, and publishes the news on an online platform without any human intervention. Quakebot encapsulates real-world activities through thousands of lines of computer

codes and acts like a journalist in designated circumstances. With increased software capacity, traditional media companies such as the Associated Press, Forbes, The New York Times, Los Angeles Time, and ProPublica adopted algorithmic approaches to news content creation in some degree (Graefe, 2016).

The leading narrative-generation algorithm developers, such as Narrative Science and Automated Insights, are also breaking down traditional boundaries of news industry by generating more targeted and relevant news stories from personal or public data. Their algorithm engines are known to count a wide variety of datasets that do not even have a direct relationship to the event, such as historical or sensor-based data, and are capable of rendering the narrative with varying tone and difficulty level to meet the desired reading preferences (Allen, 2013; Birnbaum et al., 2013). As the adoption of algorithm in content generation is becoming more of an inevitable trend, more thorough academic analysis of both the technology itself and its influence are required.

We reviewed many news products that are generated by algorithmic activities. These news products include newspaper-style articles from news media companies, summative and predictive reports for sporting events, personalized self-reports based on sensors, or even mobile app usage data. We intentionally covered a broad range of news contents that are believed to involve algorithmic computation in their production processes (even when the product may not have been generated entirely by algorithms). We also analyzed patent documents from various algorithm developers to uncover the underlying mechanisms of algorithm-driven narrative generation. Then, we identified common attributes and categorized them into a handful of

algorithmic dimensions that define what it means to have an algorithm for news generation. After reviewing the current practices, we were able to find the limitations commonly found among the existing algorithm-generated news products:

- Current news generation algorithms tend to make use of data from limited sources. To maximize computational capacity, algorithms must be given with data that are exhaustive and complementary. Building a consolidated database is the key to the success of algorithmic news generation.

- Even when the computational capacity of algorithms is often greater than any human can handle, we found that algorithms lack in counting individual context into account in generating news content. A potential system should communicate with news readers to become more personal storytelling agent.

- We do not need to be bounded with text-based news articles, but should aim to build more interactive and adaptive news service. However, we do not see notable cases of algorithmic news generation that provides more explorative news reading experience.

As more algorithms are mediating news generation processes, more studies were required to answer questions arose in the process: What are the types of news products that algorithms currently generate, and what are the strength and limitations found in them? What is missing from the current practices? How would news readers respond to the algorithm-generated news?

How would a news service be designed and implemented that maximized the potential benefit of such news-making process?

In this thesis, we tried to answer the series of questions that matter in how to design and implement algorithmic news generation system for more personalized and interactive news reading experience. We set the research questions to examine how algorithm framework can be implemented to overcome the limitations found from the current algorithmic news generation practices: develop a working system based on the algorithm framework, explore the design space of the algorithmic news generation system, and evaluate the system with the perspectives of general and expert news readers. To accomplish these goals, we conducted the following research activities in this thesis:

Theorized and Proposed a New Algorithm Framework

From the insights derived from the review of current practices, we proposed a new algorithm framework that is designed and implemented to overcome the limitations found in the process. The major objective of algorithm framework was to set to establish computational approach towards traditional journalistic processes, and to even extend the news making process into the next level: maximizing the capacity of algorithmic computation in collecting, processing, and presenting news information to readers. We found three opportunity areas where a new algorithm framework can be designed to be a more compelling creator of news content using algorithmic activities.

The first opportunity area was to construct a consolidated database

which would include not only the event-related data but also diachronous and synchronous data. Diachronous data enhances computational capacity by providing historical data to supplement the incoming data feed on an event. Synchronous data, on the other hand, includes other types of data that widen the options of the data analysis. Both diachronous and synchronous data empower algorithmic news generation by providing the context and give breadth and depth to the computation by enhancing the algorithmic decision-making process.

The second opportunity area was to frame narrative with a personalized angle for each news reader. The use of algorithm allows more personalized content generation for news readers at scale. Personalized news content could be generated by using profiles and preferences stated by news readers, historical logs of system usage patterns, and direct manipulation on the system interface if provided.

The third opportunity area was to present news content as a visualized and interactive news service so news readers could manipulate various user interface elements and generate their news stories rather than following the static and pre-made news. The design of algorithm framework was illustrated in figure 14 in the framework development chapter.

Designed an Interactive System for News Generation

Using the algorithm framework as the backbone, we implemented a fully-working prototype that changes the content and tone of narrative as participants interact with the system. We presented PINGS, an algorithm-driven news creation system that allows users to generate their version of

news stories by dynamically interacting with user interface elements. The design space of the system was explored upon the implications learned from the review of current practices and research framework on algorithmic news generation.

The user interface elements were composed of three dimensions: which were (1) timeline-based visualization of events, (2) list of data points to extend data-driven narratives, and (3) automated narrative displaying pane that can change the tone of narrative between objective and story modes. With the proposed interface elements, a user could freely add specific moments of the match or favorite players' performance records, and choose to see the news in straight and recap style or emotionally engaging and lopsided tone. The design of PINGS was illustrated in figure 23 in the system design chapter.

Conducted Evaluation Studies on the News System

We conducted a user evaluation of the system regarding the perceived quality of content and the perceived value of the system for personalized news story generation. There had not been many types of research conducted to examine personalized news generation system, so we reviewed some research to set up the criteria to evaluate both the quality of content produced by the system and the experience of using the system for personalized news content generation.

The results of this research indicated that this personalized and interactive news generation system run by algorithmic activities in the background was capable of generating significantly more interesting and pleasant to read

news stories than that of news articles generated by human journalists. The design of the user interface of the system was usable and useful in generating news stories that meet their expectations.

We concluded this thesis with the general discussions on the lessons learned from the research activities we conducted throughout this thesis, and the list of contributions to both journalism and HCI research fields.

Thesis Outline

In Chapter 2, we provided a theoretical background of this thesis. We highlighted aspects of increasing computational approaches found in the practices of journalism. The process of algorithmic news generation proposed in the research from 김동환 and 이준환 (2015) were introduced to give an overview of how algorithmic activities mediate traditional journalistic activities.

In Chapter 3, we reviewed many existing algorithm-generated news products and classified commonly found attributes in them. While each attribute is not distinguishable from each other in a mutually exclusive manner, the review helped to reveal commonly found characteristics as well as the shortfalls in utilizing algorithmic capacity in the current news products.

In Chapter 4, we stated research questions that we examined throughout this research. The first research question is to explore how algorithm framework can be designed and implemented. The research questions were set to study perceived quality of news content generated by the system, and perceived value of the system that generates personalized news stories.

In Chapter 5, we proposed all-new algorithm framework that is capable

of generating much more capable news products in the following ways: expanding the depth and breadth of news coverage with the construction of a consolidated database, framing news narrative with personalized angles, and presenting news stories in interactive and visual storytelling format.

In Chapter 6, the results of this research were reported. We depicted how proposed algorithm framework is implemented into a fully-functional algorithmic solution in the production of news content. We illustrated the design of algorithmic news generation system with the framework as the backbone and evaluation study of algorithmic news generation system.

In Chapter 7 and 8, we discussed the overall implications and insights learned from this research. We concluded this thesis with the summary of work and contributions as well as a discussion on opportunities for future work.

Chapter 2

Theoretical Background: The Algorithmic Turn in Journalism

In the modern world, computational processes and techniques are increasingly taking up traditional human roles as growing number of journalistic activities are mediated by software (Flew et al., 2012; Napoli, 2014; Carlson, 2015). While technology has always influenced journalism (Pavlik, 2000), the latest changes in the news production process are largely affected by the use of algorithms (van Dalen, 2012). As algorithms are taking responsibilities that were traditionally performed by human journalists, a need for socio-technological discussion on their influence in the industry also rises. In this chapter, we summarized the theoretical background on the effect of computation and algorithm in the functions and practices of the journalism industry.

2.1 The Computational Turn in Media

Computers and computation methods are changing many aspects of news media industry. The growing influence of computation capacity is related to more widely available datasets, the increasingly sophisticated and ubiquitous nature of software, and the developing digital economy (Flew et al., 2012). Also, the growth in data necessitated the increase in the computing power, which in turn caused advances in the statistical methods for analyzing

large datasets (Lantz, 2013, p. 6–7). In sum, the computational turn in news media can be said to be happening due to the following factors: increasingly available data, the growing digital economy, advance in statistical methods for analysis, and the general trend of softwarization of media.

The Growing Digital Economy

The burst of smartphone and social network services is a contributing factor in the computational turn happening in the media industry. As people spend more time in the digital world, more quantifiable traces of human activities are being made. The growth in the number of data sources resulted from the expanded diversity of available data formats for analysis. An interesting fact about big data is that the most of them are human-generated: 70 percent of the digital universe is generated by all of us through email and social media such as Facebook, Twitter, LinkedIn, Flickr, and YouTube (Craig and Ludloff, 2011, p. 4).

The Pew Research Center's biennial media attitudes survey found that 50% of the public now cites the Internet as the primary source for national and international news[1]. Among various social network services, Twitter and Facebook were the most significant players in providing news information. Twitter has been acting as a news media platform that is capable of real-time broadcasting in events such as the plane landing on Hudson River or the presidential election in 2008 (Kwak et al., 2010; Hermida, 2010). Also, Facebook made rapid progress in the number of unique visitors and the total

[1] http://www.people-press.org/2013/08/08/amid-criticism-support-for-medias-watchdog-role-stands-out/

time spent by its users, and it became the prominent news powerhouse than any other social media. Roughly two-thirds of U.S. adults use this service, and half of those users get news there, amounting to 30% of the entire population in U.S. (Fox and Duggan, 2013).

Increasingly Available Data

In recent years, enormous quantities of digital data are being produced as a byproduct of the revolution in computer technology such as the transformation of analog data to digital, competition among governments to share data, new technology-enhanced ways that people interact, and creation and monetization of data by many commercial entities (King, 2011). Twitter, a social networking service for posting short messages, offers open access to streams of user-generated data, so like Facebook, Instagram, and many other social network services through legitimate data transferring channels.

The sheer volume of data is exploding, and this phenomenon is called big data. The term refers to datasets whose size is beyond the ability of typical database software to capture, store, manage, and analyze (Manyika et al., 2011). In the case of Twitter, the complexity of data processing and the computational power required to make data analysis grow as the amount of data reaches to hundreds of billions of tweets. Therefore, the definition of big data inevitably involves its capacity to collect and analyze data with an unprecedented breadth and depth in scale (Lazer et al., 2009). In other words, the power of utilizing big data in analysis comes from the ability to perform a large-scale search, aggregation, and cross-referencing related datasets (Boyd and Crawford, 2012).

In research point of view, analyzing a large pile of aggregated data can tell more on many patterns involved in human activities. Now we can measure and therefore manage more precisely with data, which means we can make better predictions and smarter decisions than ever before (McAfee et al., 2012). This higher form of intelligence and knowledge induced by big data can be led to the generation of new insights that were previously impossible; with the aura of truth, objectivity, and accuracy (Boyd and Crawford, 2012).

Advances in Computational Methods for Analysis

With the growth in the size and complexity of data, computing power and statistical methods also advance to handle the given sets of data on request. More sophisticated machine learning algorithms that are capable of automating computational procedures are heavily under research due to this digital trend. Lately, it is reported that important advances are found in the study of deep learning algorithms that Google can now successfully classified cats among millions of YouTube videos automatically with significant accuracy (Le, 2013). It means the technology changes what we can learn from video processing at large scale that trillions of YouTube videos can now be identified without employing human labors.

Moreover, technological advances in algorithms introduced sophisticated data analyzing techniques that enabled real-time analysis or prediction using extensive data at scale. For example, the accumulation of tens of years of sports data allowed fairly accurate predictions based on the statistical analysis. FanGraphs and FiveThirtyEight, popular sports blogs, use domain-

specific statistical analysis methods such as sabermetrics and Elo-ratings in making autonomous predictions on the result of sports events even before matches begin. The founder of FiveThirtyEight, Nate Silver, a statistician and sports analyst, became famous after 2008 U.S. presidential election where he called almost complete outcomes by applying his statistical method for projection. As the volume and variety of data expand, the need for more reliable and scalable statistical analysis methods also increase.

Softwarization of Media

Today, our media activities are increasingly being mediated by various software running behind the scenes. Understanding what role do software play in media production and consumption is becoming more important to understand the changes happening in the media landscape. The software is a layer that permeates all areas of contemporary societies (Manovich, 2013, p. 15). Softwarization affects cultural activities such as sending and receiving messages, expressing ideas to others, gathering information, and distributing and accessing media contents on the web (Manovich, 2013, p. 24–25). In other words, software is a key actant in creating societies of control, as it makes possible a fundamental shift in how information is gathered, by whom, for what purposes, and how it is applied to anticipate individuals' future lives (Kitchin and Dodge, 2011, p. 86).

Softwarization also enables quantifying our media activities for further computational analysis. The search engine from Google uses a sophisticated algorithm called PageRank to calculate the relative importance of web pages to provide search results upon user's keywords. What empowers its search

ability is that Google uses the history of search keywords from the entire population of search engine users to provide tailored search results (오세욱 and 이재현, 2012). Google News, Facebook's NewsFeed, and even New York Time's online article recommendation system utilize the digital traces of user activities to provide more contextually appropriate information[2].

2.2 Computational Journalism

There has been a growing number of discussions about utilizing computational capacity in the field of journalism. The use of a computer, as a tool and technique, in news reporting is not a newly rising trend, but is increasingly changing how news contents are being produced and consumed. The definition of computational journalism means more than just applying computing power or technologies to journalism. Hamilton and Turner define it as "the combination of algorithms, data, and knowledge from the social science to supplement the accountability function of journalism" (Hamilton and Turner, 2009).

Applying the knowledge and methods of social science to make journalism more accountable dates back to Philip Meyer, who coined the term 'precision journalism' (Meyer, 1973). In his book, Meyer explains the use of computer tools allow journalists to manage, process, and analyze data to perform journalistic tasks. Later by Cox and other researchers, the idea of using computers in news gathering process succeeded to the practices of 'computer-assisted reporting (CAR)' (Cox, 2000). The introduction of database tools

[2]http://open.blogs.nytimes.com/2015/08/11/building-the-next-new-york-times-recommendation-engine

such as spreadsheet software and database management systems led journalists to work with various datasets and to become more accountable and scientific in reasoning and to argue their opinions (Garrison, 1998; Cox, 2000).

Utilizing a database as a tool and a method for further investigation is already claimed as the contributing factor in precision journalism and CAR, but the term 'database journalism' is also used to distinguish the effort of database-driven approach to the news production. Holovaty, in 2006, introduced an online news service called EveryBlock, which turned individual pieces of data elements (such as persons, locations, events) into news stories by letting the readers explore data [3]. More trending online websites such as Politifacts[4] are turning news into a more granular and structured information that can be recontextualized for multiple contexts (Chua, 2010). The latest trend of combining spreadsheets with widely available data, and tools to statistically analyze and visualize data is also referred to 'data journalism' (Rogers, 2013, p. 29). By using computational tools and techniques, journalists are expected to increase the depth of reporting quality, differentiate themselves from competitors, accelerate the process of journalism from news source to delivery, and provide an accurate analysis on behalf of supporting data (Daniel and Flew, 2010). It is also found that increased use of visualization tools helped journalists to integrate complex data into easy-to-understand and engaging stories with data (Segel and Heer, 2010).

While the definition of computational journalism varies from one research

[3] http://www.holovaty.com/writing/fundamental-change/
[4] http://www.politifact.com/

to another, it can be said to be the general tendency of using computer software to engage with techniques for the large-scale manipulation of data to enable new ways to access, organize, and present information (Flew et al., 2012). Diakopoulos offered a process-oriented definition of computational journalism as an inclusive term for the activities of journalism such as to apply computing in "information gathering, organization and sensemaking, communication and presentation, and dissemination and public interaction with news information, all while upholding values of journalism such as balance, accuracy, and objectivity" (Diakopoulos, 2011).

The computational capacity supports journalism not only in the production side of news content but also in the ways news readers consume them. Through an algorithm, how a story is told can be altered by offering different layers of stories that change upon the selections news readers make or by learning preferences and interest of news consumers. It can attract readers that are highly interested in such topics and helps to transform journalism into a more blended practice of reporting and social organizing (Hamilton and Turner, 2009). This rising trend of application of algorithm in the news making process, such as to collect and process tons of data, find patterns from large datasets, or even to generate information is widely found under the name of 'algorithm journalism' or 'robot journalism' (Dörr, 2016). The characteristics of algorithm journalism are reviewed in detail in the following section (also in Table 1).

Table 1: Comparing focus and key skills of different modes of journalistic practices

Type	Focus	Key Skills
Precision Journalism	To make journalism more accountable and scientific by applying methods from the social sciences (Meyer, 1973)	Applying social science methods to journalistic investigation
Computer-Assisted Reporting	To use computers to gather and analyze data for journalistic activities (Garrison, 1998; Cox, 2000)	Advanced use of computer tools for news reporting
Database Journalism	To work with datasets and database tools to become more accountable and scientific (e.g. EveryBlock, The Homicide Report (Young and Hermida, 2015))	Advanced use of spreadsheet/database software
Structured Journalism	To store elements of news into database and allow reuse for recontextualization (Chua, 2010)	Extracting and storing existing news article into smaller pieces of reusable elements (objects)
Data Journalism	To find, search and explore data for data-driven storytelling (Rogers, 2013)	Sensemaking on a large set of data to derive insights
Computational Journalism	To enable new ways to access, organize and present information using computing software (Hamilton and Turner, 2009; Flew et al., 2012)	Applying computing to the activities of journalism at large
Algorithm Journalism	To use algorithms in the process of journalistic activities, especially in the creation of news content (Diakopoulos, 2015; Dörr, 2016)	Applying content creation algorithms such as NLG to the production of information

How Computation Changes the Production of News

Through the added capacity of computational journalism, every step in the news production is expected to face changes at some level. For the discovery of and access to data, it allows collecting data from unprecedented news sources in various format. While some government and institutions open their data for public usage, some of them require heavy loads on cleaning and making sense of data. Also, many commercial data need more sophisticated computational techniques and computing power to set up proper data communication channels or to process tons of data that are generated in real-time. As more unstructured data (such as text phrases, graphics, and images, multimedia files) are becoming available on the web, the need for advancements in computing skills would also rise for richer data analysis.

Faster access to data enhances the competitiveness of journalists in a number of ways. It helps them to write reports in more scientific and investigative ways by enabling more sophisticated statistical analysis. Also, it helps them to find newsworthy events and derive insights from what have previously been unconnected and sporadic events. A collection of data not only becomes the source of news reporting, but visualization of the collected data set can also become interesting news report as well. For example, The Homicide report[5] from Los Angeles Times, or micro-local news from EveryBlock[6] present their large set of data in significantly visualized ways that news readers find the visualization itself convey news storytelling.

More advanced content management system (CMS) tools help journalists

[5] http://homicide.latimes.com
[6] http://everyblock.com

to tag and add metadata automatically when reports are written. The system does not add metadata to a report as a whole, but it adds them to individual elements inside the content of report such as names, places, other events as journalists are writing it in real-time. Chorus is one of the most popular CMS engines from Vox Media, and it is known to generate reports for multiple services under Vox including The Verge and SB Nation. It provides various useful functions such as to automatically tag elements for search retrieval, organize the report layout depending on the type and multimedia usage, and allows to add and update the content from the reviews made by the news readers. When one big chunk of a news article is broken down into pieces of reusable elements, many other kinds and types of news can be generated by recontextualizing stored elements in the database by adding and remixing different pieces of them (Chua, 2010).

2.3 The Algorithmic Turn in Journalism

The ability of computer software to support the activities of journalism by collecting and compiling enormous quantities of data comes from the power of the algorithm. An algorithm is regarded as a combination of logic and control component, where the logic component specifies knowledge to be used, and the control component determines the procedure it takes (Kowalski, 1979). The algorithmic turn happens when an algorithm is designed to produce proper output and derive insights even when the input data is beyond the capacity of any human computation. More precisely, the innovation of big data and its derived products are only possible with an

algorithm to transform an input through specified computational procedures into the output (Cormen, 2009).

The power of algorithm comes from its ability to make automated decisions by taking procedural steps (Diakopoulos, 2014). In journalism, the turn has come with algorithms performing various journalistic tasks such as editing, aggregating, publishing, and distributing of news contents automatically (Mager, 2012; Napoli, 2014; Gillespie, 2014). Recent progress in the development in natural language generation (NLG) enabled an algorithm to produce text from the computational representation of information (Reiter et al., 2000), which is natural enough to such an extent as to be perceived as text written by a human writer (Dörr, 2016).

As the volume and the variety of data grow, notable advances in the efficiency in algorithmic computation is also made. It means finding more relevant, valuable, and reliable news in the flood of news information is a step closer with the help of an algorithm. Moreover, an algorithm not only helps us to find information, but it provides a means to know what there is to know and how to know it (Gillespie, 2014). Algorithms are taking responsibilities that were traditionally performed by human journalists such as reporters, editors, photographers, and many other roles related to the production and consumption of news contents.

In general, algorithms are functioning in the following three realms of journalism sector: the creation of news and news-related content, filtering and curation of data and content for news readers, and data-driven approach to the functions of journalism.

Creation of News and News-Related Contents

With its autonomous power to collect, analyze and derive insights from data, algorithms are becoming a more active player in the realm of content production (Napoli, 2014). Algorithms have been developed and employed to perform what are traditionally known human roles such as poetry and music composition (Steiner and Dixon, 2012, p. 94–95), writing tweet messages in online services (Chu et al., 2010), and even in journalistic activities such as writing and publishing of news contents.

The key ability of algorithms in this realm is to generate content in natural human language based on the statistical analysis (and computational representation) of incoming data (Reiter et al., 2000). These algorithms are designed to mimic the writing style of a good journalist to bridge the gap between a machine and human. However, the patent document from Narrative Science stated that the tone and manner of text generated from an algorithm could be designed to meet specific needs or reading level of targeted news readers (Birnbaum et al., 2013). It means the form and type of news content do not need to follow that of a newspaper-style article written by a human journalist.

The type of content may also go beyond a plain text and include graphics, photos, audio and video, and any other forms of media elements. To utilize these media elements, it requires the construction of news content to invite more algorithmic approaches to let news readers interact with and contribute to the presentation of content (Royal, 2010). Moreover, leading technology companies are investing their resources in the development of modality-

changing algorithms such as to summarize text-based news article into image-based content (Ha et al., 2015), or to create text description on images (Vinyals et al., 2014). In sum, the capacity of algorithm expanded the realm of news production beyond renderings of a human thinking process. The algorithmic turn, therefore, changes what is news and what becomes news in this rapidly changing media environment.

Filtering and Curation of Contents for News Readers

One of the key functions that algorithms perform in the consumption of media contents is to assist its audiences in the process of navigating an increasingly complex and fragmented media environment. Central to this navigation process are the typically algorithmically driven recommendation, search and content aggregation systems that facilitate searching for and selecting content in an environment of such extreme content abundance (Napoli, 2014). Technologically unaided forms of search and navigation may no longer be practical or useful (Anderson, 2006).

The power of algorithm in finding and recommending the right content comes from the sophisticated logic to make automated decisions on how to prioritize, classify, associate, and filter the information (Diakopoulos, 2015). Google's PageRank uses the number of references a website has as a relative importance of the page and normalizing the weight of page importance for better search results (Brin and Page, 1998). Collaborative filtering, a technology that aims to learn user preferences and make recommendations based on user and community data (Das et al., 2007), is a type of algorithms that is widely adapted in many data-heavy commercial services such as

Amazon and Facebook. As personal and community-level activity data are readily available for analysis, algorithms are given enough power to customize its predictions and recommendations tailored to the wants and needs of the targeted audiences.

Data-driven Approach to Support The Functions of Journalism

As algorithm increasingly plays multiple roles in the production of news, it is required to uphold the core values of journalism such as being an objective and responsible agent in the system (Cohen et al., 2011; Diakopoulos, 2011). Hamilton and Turner (2009) write of the new opportunities by an algorithmic approach to journalism is the watchdog function of journalism. With its power to deal with large at scale and with speed, an algorithm can help journalists to explore and discover newsworthy events unless otherwise be unnoticed. Furthermore, the user of the algorithm may potentially benefit the community by producing commodity news on routine tasks (van Dalen, 2012).

To perform watchdog reporting by a human journalist, stated in Hamilton and Turner's research again, he or she would need "two kinds of assistance in doing so: first, ways to extract and integrate structured information from a variety of data sources such as text, video, and the web; second, they need tools with which to make visible and exploit patterns in data" (Hamilton and Turner, 2009). These are the kinds of computational capacity that do not yet meet the level of a human agent due to the technical complexity. However, remarkable progress is made in developing computation methods to deal with unstructured data such as deep learning and other related algorithms,

which in turn will make a computer software as a useful agent in the practices of journalism.

2.4 Algorithmic News Generation Process

For an algorithm to generate a news story from raw data, it is required to take a series of steps to process and analyze the data. A conventional algorithm engine should be able to gather data from various sources, extract meaningful and relevant information from data, find relatively more important moments in the event, determine the narrative frame to be applied in writing and to write and publish news article through various channels (김동환 and 이준환, 2015). Various research on algorithm journalism defines the process in their terms. One of the first papers in the field of algorithm journalism describes the process as to apply statistical analysis to interpret data, determine critical moments in the event, and produce stories (Allen et al., 2010). In the research published by 김동환 and 이준환 (2015), a five-step algorithmic news generation framework has been proposed (Figure 1).

The first step in the framework is to crawl data from various data sources. The crawling algorithm is designed to handle different specifications that are set by the service provider. The data might come through legitimate data transfer channels such as API (Application Programming Interface), or an algorithm may take the whole underlying code of a website. In a case of a baseball, an algorithm may collect box scores, play-by-play records, a batting average of batters, historical records, or player demographics (Allen

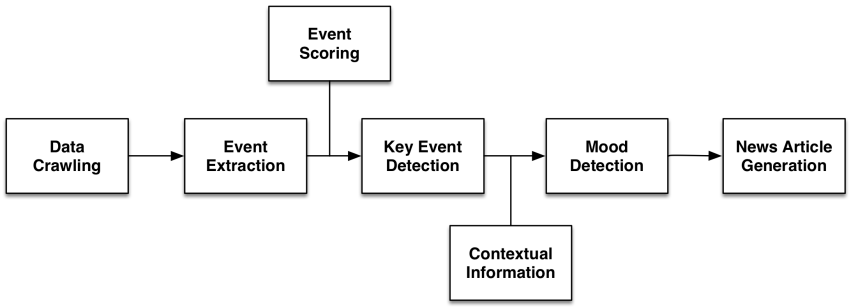


Figure 1: Algorithm News Framework

et al., 2010; 김동환 and 이준환, 2015).

The second step is to extract meaningful events from the entire set of data. When data is collected from various online sources, we cannot expect the data to be immediately usable. It often requires data cleaning processes, such as to correct format, split or merge entries, and clean unnecessary data (Gray et al., 2012). However, algorithms are still not capable of making a decision close to that of a human journalist. Therefore, a set of rules are given as an instruction to interpret the data, or a dictionary of domain-specific terms and metaphors are also needed to be prescribed for errorless data interpretation.

The third step is to determine the key event out of all events by comparing the event scores that are assigned to each event. The advance in statistical techniques made algorithms to perform more sophisticated computation than a simple calculation. In a case of a baseball again, sabermetrics (baseball statistics) provides many ways to identify key moments in a baseball game based on tens of years of historic major league baseball data. For news topics other than baseball, there exist domain-specific statistical methods that can

be applied to support algorithmic computation.

Mood detection is the fourth step in the process. When key events are determined in the previous step, an algorithm can set an angle to interpret the event just like human journalists would do to frame their perspectives on such incidents. For example, if some events are found to be happening more than average, then it can be said to be 'happening excessively' lately. Likewise, an event might surprise people when it is unexpected to happen due to its rareness.

Setting a frame on how to interpret an event can follow the quantifiable process by applying dynamic weighting system, which is a concept proposed in the paper from 김동환 and 이준환 (2015). Dynamic weighting system enables the algorithm to take contextual information into account in detecting the mood of the event. For example, a home run in a baseball game is usually included in the top highlight of the match. The home run is treated more important as the scores made by the home run increases. In such a case, a grand slam (scoring four runs) is four times more valuable than a single home run. However, the system does not put this grand slam event on top of other play records just because it earned more scores. A single in the last inning when the game is on a tie is more important than a grand slam made in the first inning if it is just one of many scores made by two teams. The algorithm needs to know that even a single run in the 9th inning is more important, and it is how the dynamic weighting system is designed to perform.

The last step of the framework is to write an article based on the mood detected from the previous stage. Recently, there has been remarkable advances

made in natural language generation algorithm, and the quality of algorithm-generated article also improved with this technological advancement. However, training an algorithm to write like a good journalist in every news domain has still many issues even with the current computational capacity.

The leading narrative generation algorithm developers such as Narrative Science and Automated Insights circumvent these problems by employing template sentences in generating naturally-read text phrases (Birnbaum et al., 2013; Allen, 2013; Davison and Guiro, 2014). In cases of languages other than English, the template-driven approach to narrative generation would be particularly useful since there are not enough programming libraries and packages to build the system. It raises the difficulty in programming software that generates text due to the limited resources to utilize. For the Korean language, more natural language toolkit is becoming available through Python community lately, but the output quality is still not at the level of a human journalist, and template-driven approach is often used.

Figure 2 is an example of news products that are generated using the algorithmic solution our team developed. With the algorithm engine we developed called 'Korean Baseball News Bot', we have been generating and publishing baseball news stories since the start of 2015 Korean Baseball League in March 2015. Since then, we produced more than 1,400 news and gathered more than 800 followers who liked our Facebook page for baseball news postings[7]. Also, we started to publish a finance report to an online news media called 'Financial News' from January 2016. The algorithm engine that publishes a financial news collects daily stock market

[7] https://www.facebook.com/kbaseballbot/

updates such as the rise or fall of KOSPI and KOSDAQ index, orders from institutions and individuals, and details on big name companies.

프로야구 뉴스로봇
Published by Jinsu Eun [?] · October 14 at 2:52pm ·

두산, LG에게 숨쉴틈도 안주고 득점 몰아부쳐...
(2016-10-08, 두산 11 對 1 LG, 잠실)

8일 잠실구장에서 열린 2016 타이어뱅크 KBO리그 두산과 LG의 경기에서 두산이 포문을 연 김재호의 2루타 이후 타선이 폭발하면서 LG를 상대로 크게 승리했다. 두산은 16만타 2홈런을 뽑아내며 타월한 경기력을 보여줬다. 두산은 0:0으로 경기 중이던 2회초, 김재호가 2루타를 때리며 2점을 얻었다. 그 이후에도 박건우, 오재일, 김재환 또한 승리하는데 큰 힘을 발휘했다.

2회 초 두산은 2사 1, 2루 상황에서 정수빈의 볼넷으로 2사 만루 상황을 만들고 김재호의 2타점 2루타로 선취 2득점했다. 그 후 3회 초에는 무사 2루 상황에서 오재일의 볼넷으로 1사 1, 3루 상황을 만들고 김재환의 1타점 2루타와 양의지의 1타점 적시타로 점수 차를 벌렸다.

3회 말 LG는 1사 2루 상황에서 김용의의 볼넷으로 1사 1, 2루 상황을 만들었으나 이후 이천웅의 삼진과 정성훈의 유격수 땅볼로 추격에 실패했다.

4회 초 두산은 김재호의 1점 홈런으로 1점을 달아났다.

4회 말 LG는 2사 1, 2루 득점찬스를 맞이하였으나 이병규가 안타 후 진루하던 이병규가 홈승부 실패로 아웃되어, 아웃카운트를 헌납해 추격에 실패했다.

5회 초 두산은 1사 1, 3루 상황에서 상대 투수 이준형의 실책으로 1점을 달아났다. 그 후 6회 초에는 1사 1, 3루 상황에서 김재환의 1타점 적시타로 1점을 달아났다.

6회 말 LG는 무사 1, 2루 상황에서 이병규의 희생플라이로 1점을 따라잡았다.

8회 초 두산은 무사 2루 상황에서 박건우의 1타점 적시타로 점수 차를 벌렸다. 오재일의 2점 홈런으로 2점을 달아났다. 그 후 9회 초에는 무사 1, 2루 상황에서 폭투로 1사 2, 3루 상황을 만들고 박건우의 유격수 땅볼로 1득점하며 경기결과를 확정지었다.

경기 결과 두산은 LG에게 대패를 안겨줬다. 오늘 경기의 결과 두산은 현재 1위(승률 0.650)이다. 한편 LG는 현재 4위(승률 0.500)이고 3안타 4볼넷 1타점 1득점으로 다소 아쉬운 플레이를 보여줬다.

(이 기사는 서울대학교 이준환 교수 연구팀이 개발한 기사 작성 알고리즘 '야알봇'이 작성한 기사입니다.)

경제>증권>증권일반

코스피 4.92포인트 하락, 1840.53포인트 거래 마감

[로봇저널리즘]

파이낸셜뉴스 입력 : 2016.01.21 15:27 | 수정 : 2016.01.21 18:07

관련종목 삼성물산(000830), CJ(001040), 현대차(005380), 삼성전자(005930), 삼성전자우(005935), 한국전력(015760), 삼성물산(028260), 카카오(035720), 셀트리온(068270), 컴투스(078340), 바이로메드(084990), 엠스리얼티1(094800), CJ E&M(130960)

코스피가 전날보다 4.92포인트(-0.27%) 하락한 1840.53포인트로 거래를 마쳤다. 이날 개인과 외국인이 각각 287억원, 2971억원어치를 동반 순매도하며 지수 하락을 이끌었으며, 기관은 3120억원을 순매수했다. 시가총액 상위 종목 중에는 오른 종목이 더 많았는데, 삼성전자(-0.62%), 한국전력(-1.94%) 삼성전자우(-1.14%)가 하락한 반면, 현대차(0.37%), 엠스리얼티1(0.49%) 삼성물산(0.68%) 등은 상승세를 보였다. 업종별로는 음식료업이 0.06%, 화학이 0.42%, 기계가 2.03% 상승했으며, 섬유의복이 -0.22%, 종이목재가 -0.57%, 의약품이 -1.49% 하락했다.

코스닥은 전날보다 3.84포인트(-0.57%) 하락한 665.84포인트로 거래를 마쳤다. 이날 개인과 외국인이 각각 249억원, 37억원어치를 동반 순매수하고, 기관은 259억원을 순매도했다. 시가총액 상위 종목 중에는 내린 종목이 더 많았는데, 셀트리온(0.09%), 바이로메드(3.69%) 컴투스(4.55%)가 상승한 반면, 카카오(-0.98%), CJ E&M(-1.74%) 동서(-3.16%) 등은 하락세를 보였다. 업종별로는 하락한 업종이 더 많았는데, 속박.음식이 0.0%, 운송이 0.1%, 섬유/의류가 13.59% 상승했으며, 건설이 -6.19%, 유통이 -1.32%, 금융이 -1.01% 하락했다.

"이 기사는 파이낸셜뉴스와 협업으로 서울대학교 이준환/서봉원 교수 연구팀이 개발한 기사 작성 알고리즘 로봇이 실시간으로 작성한 기사입니다."

Figure 2: News articles our research team generated using the algorithm solution developed for baseball and finance news

Chapter 3

Practices of Algorithmic News Generation

With the algorithmic turn in journalism, more media companies are adopting various algorithmic solutions in generating news and news-related content. An algorithm is known to be efficient in computing structured data into meaningful information and is capable of generating news more quickly, cheaply, and with fewer errors (Graefe, 2016). As technological advances continuously made, more news generation algorithms are expected to play significant roles in the production of news. From a simple message posting bot on Twitter to a sophisticated narrative generation engine, many algorithms are in use in the realm of news and news-related content generation. In this chapter, we analyzed a wide variety of existing algorithm-driven news products and classified most frequently found types of algorithm-generated news. Also, we were able to classify common algorithmic attributes by analyzing each news type and aligned them to the overall journalistic activities. We concluded the chapter with a discussion of the limitations found from the technological review.

3.1 Overview

Generally speaking, journalistic processes consist of three distinct steps (Figure 3), which are to gather, process, and present (distribute) news information (Seib, 2002; Karlsson, 2011). The gathering stage refers to the act of collecting

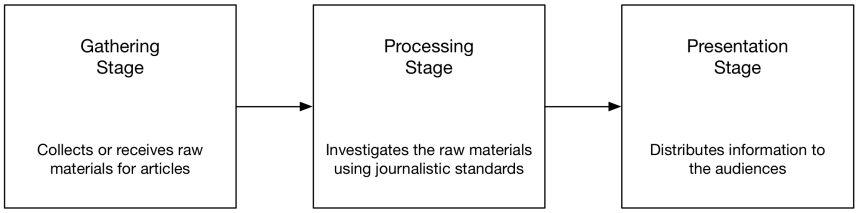


Figure 3: Three stages of journalistic processes

or receiving news-related raw materials. In the digital era, data collection expands to include user-generated content found on social media or automated data collecting channels. The processing stage is where the gathered data become refined information. In the traditional journalistic perspectives, this stage is where the raw data become journalistic information. In the digital world, keeping up with journalistic standards also matter and therefore various methodologies and procedures are applied to scrutinize the source data.

The output from this second stage might vary from a simple message to an investigative report. The structure of news information from this stage is increasingly being determined by the content format set by distribution channels in the modern digital environment. The third and final stage of the process is to present or distribute news information to the news readers. Again, the growing digital economy has expanded the distribution channels of such information and the delivery of news information is becoming much more complicated than ever before.

The current development trend of news creation algorithms is mainly focused on generating stories that are close to narratives written by humans. The stories can be as simple as filling in data points in sentences applying

pre-defined rules, such as the algorithms in Quakebot and The Homicide Report from LA Times. Quakebot is an automated system that lives on the Los Angeles Times' server, which receives emails from US Geological Survey and generates a report using pre-determined rules set by Schwencke, the programmer-journalist of LA Times. When the report is generated, the algorithm checks on the location and the magnitude of the earthquake. "If it happened near LA or the magnitude is over 6.0 scale, then the algorithm automatically set a post live. Anything smaller than that goes to the copy edit desk for editorial decision.[1]"

Algorithms in The Homicide Report track homicides by reading and analyzing cases from the coroner of LA County, and automatically generate simple reports by parsing data points (such as name, age, race, gender, date of death, neighborhood, etc.), and put them into pre-defined sentence templates (Young and Hermida, 2015). A typical report begins with this kind of sentence: "Jordi Astudillo, a 22-year-old Latino, was shot and killed Wednesday, Nov. 18 in the 900 blocks of West 41st Street in Vermont Square, according to Los Angeles County coroner's records." Regardless of the number of incidents and the release time of coroner's record, LA Times can publish every homicide reports with no additional human labor.

The sophistication in the structure and details in a message can be much more accurate as its algorithm matures. The leading algorithm solution developers such as Narrative Science and Automated Insights are known to be capable of generating more detailed, and data interpreted stories based on their algorithm solutions.

[1] http://www.businessinsider.com/quakebot-robot-la-times-2014-3

Narrative Science has started their business from a class project called StatsMonkey, an automated system that writes baseball stories using raw game data (Allen et al., 2010). For an autonomous story generation, the system determines what narratives apply to the description of the given baseball game. It makes decisions based on the box score, event logs, and baseball performance indicators such as the win probability or leverage index, which are widely used baseball statistics methods called sabermetrics. They allow the system to identify relatively more critical moments in the game, and the system generates angles based on the statistical comparison of data. An angle is a term borrowed from journalism, which determines an interpretation of a collection of events. Through this process, the system can make editorial decisions on the newsworthiness among a series of events.

After a spin-off, they have partnered with news media companies such as Forbes, Big Ten Network, and ProPublica in the algorithmic generation of news stories. The power of Narrative Science's reporting ability comes from its algorithm engine, Quill[2]. It is a natural language generation platform, which identifies the relevant and exciting data and automatically generates data-driven stories. To look into more detail, Quill collects data from various sources and extracts insights using the rules that are specified by their customer and the business domain the problem belongs. Then it applies natural language processing algorithm to generate human-like narratives. According to the series of patents registered by Narrative Science, their narrative-generation algorithm is capable of deriving features from ingested data, proposing an angle for narrative construction, retrieving situational and historical elements

[2]http://www.narrativescience.com/quill

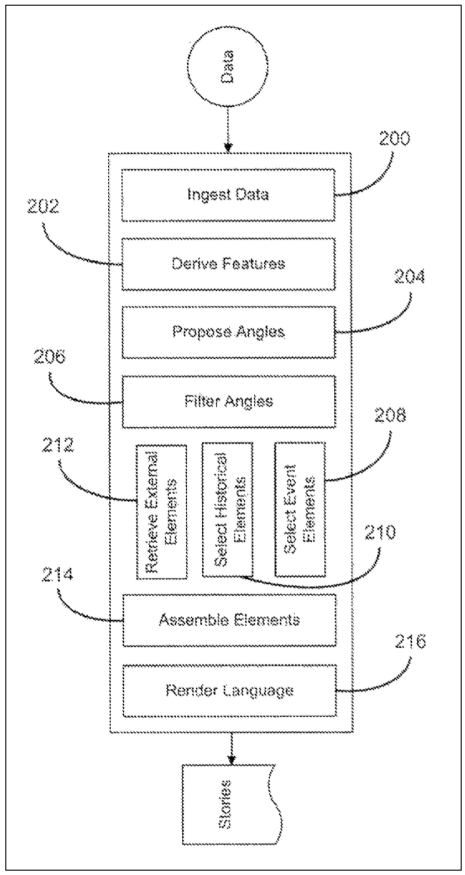


Figure 4: A flowchart on how Narrative Science process data into stories (Birnbaum et al., 2013)

to spice up the narrative, and rendering the story in a language at the required reading level (Birnbaum et al., 2013). Throughout the process, Narrative Science's algorithm can create stories that are close to those created by journalists (Figure 4).

Automated Insights, which started from a web service called StatSheet by Robbie Allen in 2007, is another big player in the market, which has

a narrative generation engine called WordSmith[3]. Automated Insights has been providing financial news to the Associated Press (AP) since 2013, and is known to produce more than four thousand news stories per quarter for AP alone[4]. Moreover, their portfolio expands beyond the realm of traditional media companies that they provide not only newspaper-style articles but also customized reports based on personal or community data. For instance, Automated Insights partnered with the website Bodybuilding.com to provide narrative summaries of each user's workout data. In this case, they collect data through sensors and gym records to provide information that matters to each particular person. Automated Insights also registered patents related to autonomous narrative content generation using templates (Allen, 2013). Their algorithm engine is known to be capable of generating more than three thousand news per quarter, according to their cooperate website.

The patent document from CBS Interactive contains even more interesting descriptions of how one of their algorithms can generate personalized game stories. Their algorithm may include personal traits and desires, such as a name, face, voice, video avatar, title, gender, preferred video game application or franchise, preferred gaming system, the level of expertise with a video game application or gaming console, league affiliation, and so forth (Davison and Guiro, 2014). Also, the description of the content type indicates that their algorithm is capable of generating information in many formats such as news ticker blurbs, radio interviews, or quick summaries of events. It means the product of algorithmic computation does not need to stay in the

[3]http://automatedinsights.com/product/

[4]https://automatedinsights.com/associated-press-leaps-forward/

form of plain English. Over time, more news types are expected to be tested, and algorithm-driven news products will be diversified regarding the content and model of news service.

We investigated 30 existing news products to review state-of-the-art practices of algorithmic news generation and classified most frequently found types of algorithm-generated news. Also, we were able to classify algorithmic attributes by analyzing each news type and aligned them to the overall journalistic activities. In the cases studied, algorithmic activities might have been applied to all three stages of journalistic processes. For example, some news products may have focused on a heavy analysis of raw data, but the report might be as simple as a data table, while others may have focused on generating human-sounding sentences for more natural news reading. Throughout the analysis of existing algorithm news products using the journalistic process model, we have classified seven different types of algorithm-generated news products and six dimensions of algorithmic attributes found among the selected news products.

3.2 Types of Algorithm-generated News

As mentioned above, the product of algorithmic news generation may be as simple as a tweet by a news bot on Twitter, but it can also be an investigative report written with an algorithm-detected narrative angle. To categorize the practices of current algorithm-generated news, we collected news products that are dynamically generated and distributed by using various algorithmic solutions. Some of them were specifically mentioned that they

are the result of an algorithmic calculation (such as Los Angeles' earthquake reports from Quakebot or financial earnings report on Forbes). Some of them were put on the list because algorithmic solution helped to pull contents from their internal database and distributed news information based on news readers' direct or indirect requests. Quartz and Aftershock from Oregon Public Broadcasting are such a case.

By analyzing 30 exemplary practices, we were able to identify and categorize the seven major types of algorithm-generated news. They are the data-centric report, news bot, news article, interactive news service, messaging service, immersive storytelling, and personalized report.

3.2.1 Data-centric Report

A data-centric report is the type of news product that is mostly focused on conveying data-centric information by analyzing and interpreting the given dataset. The major focus is to provide numbers or stats, and surrounding text are placed to give supplementary information to the figures.

4th Down Bot is a data-driven new service on the live analysis of 4th down events in football games, hosted by the New York Times (Figure 5). The algorithm behind this service is written in Python programming language, and their source code is openly available through code sharing website called GitHub[5]. To provide an automatic evaluation on football coaches' calls on 4th down plays, NYT analyzed over 15 years of NFL historical game data to determine whether a team would have been better off if the coach has called to punt or attempt a field goal. The key ability

[5]https://github.com/TheUpshot/4thdownbot-model

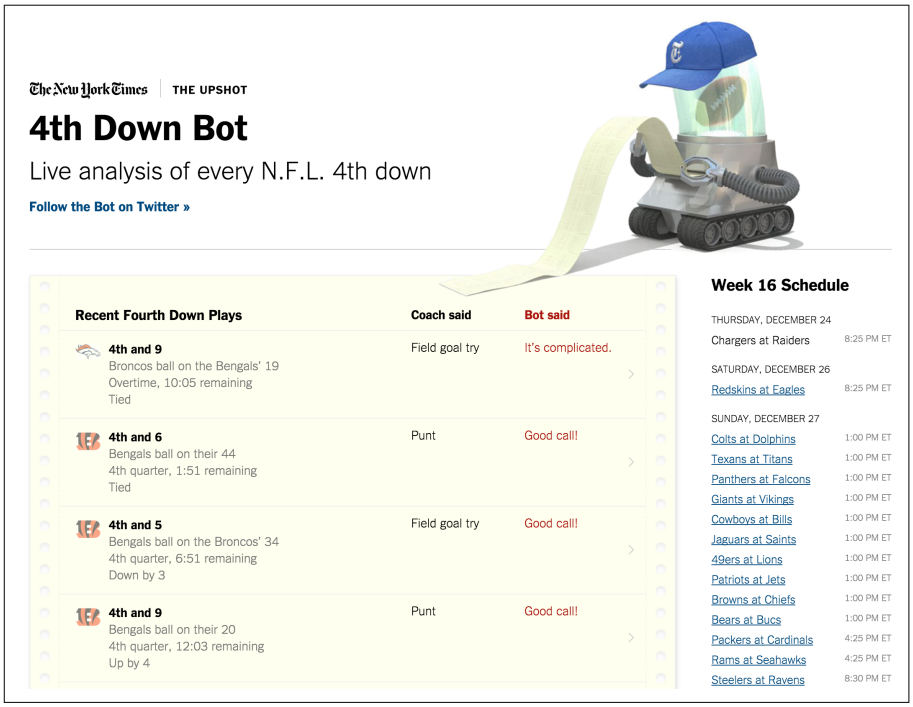


Figure 5: 4th Down Bot (NYT) provides live analysis of the calls made by NFL coaches

of algorithm in this news product is that it counts years of historical sports data to make automated judgments. In other words, it expands the set of data required for algorithmic computation to include not only the current game event but also other types of data.

The algorithm gathers text broadcasting data from sports media and classifies 4th down plays from all other games by applying rule- based text processing method. Then it analyzes each call made by football coaches by calling it “Good call,” “I would have gone for it,” or “It’s complicated.” The bot says the call made by the coach is a good call when the chance of winning percentage is higher than other coaching options. The algorithm

makes the analysis based on a statistical analysis on the result of thousands of similar plays in the historical database. The algorithm takes many variables in the computation, which includes the difference in score, the time remaining in the game, the number of timeouts each team has left, and the likelihood that a team makes a field goal or a first down.

3.2.2 News Bot

A news bot is defined as "automated accounts that participate in news and information dissemination on social networks" (Lokot and Diakopoulos, 2016). In particular, it is a software designed to follow behaviors of human in performing certain tasks such as to gather, organize, and publish information. News bots are observed across many social network services like Twitter, Facebook, Reddit, and even on Wikipedia (Lokot and Diakopoulos, 2016).

LA QuakeBot is one of many news bots found on Twitter, which delivers any abnormal seismic activities near Los Angeles (Figure 6). On Twitter, more than tens of accounts are actively sending tweets on earthquake events, especially in the name of cities on the Pacific rim. Along with Quakebot from the Los Angeles Times, the algorithm behind this news bot also collects data from the government source called USGS (United States Geological Survey). Since USGS sends earthquake reports in a structured format, the algorithm behind this news bot can process the data and extracts necessary text information for compositing a tweet on the event. The complexity of the text is at the minimum level: a tweet is often a single sentence long and has almost same sentence structure except for the number of magnitudes and the location of the earthquake.

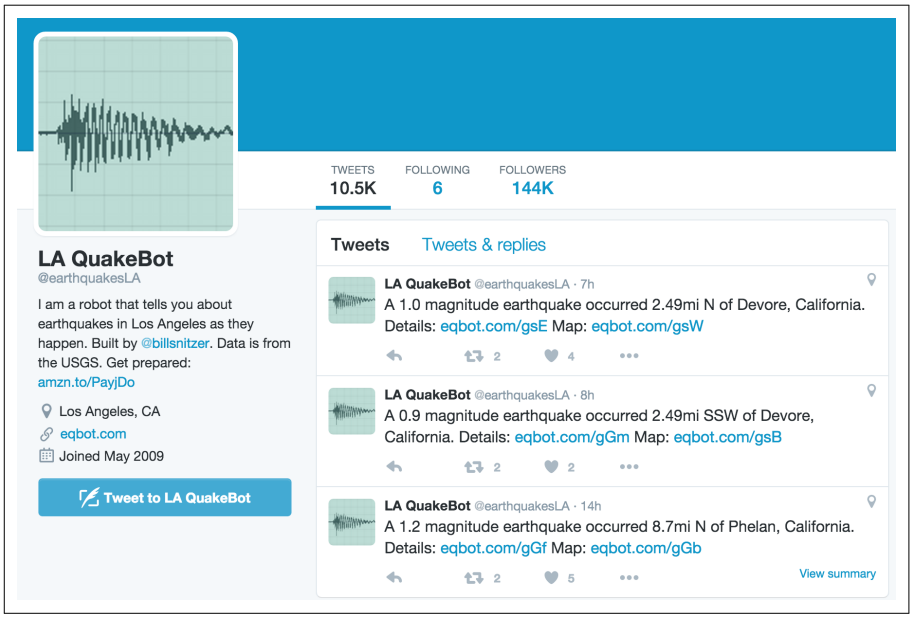


Figure 6: LA QuakeBot is a news bot on Twitter that tweets on any seismic activities near Los Angeles

According to the study from Lokot and Diakopoulos (2016), news bots are counted up to 8.5 percent of entire accounts on Twitter and 7 percent of Facebook accounts. The source of input data for these news bots may include not only government or institutional sources, but to include other websites, social media or blog, news media, a database of stories, and the list goes on. The output of news bots is also extensive, which includes disaster, sports, finance, business, politics, and weather.

3.2.3 News Article

The most common type of algorithm news would be a newspaper-style news article that mimics the writing style of a human journalist. Narrative

CUPERTINO, Calif. (AP) _ Apple Inc. (AAPL) on Tuesday reported fiscal first-quarter net income of $18.02 billion.

The Cupertino, California-based company said it had profit of $3.06 per share.

The results surpassed Wall Street expectations. The average estimate of analysts surveyed by Zacks Investment Research was for earnings of $2.60 per share.

The maker of iPhones, iPads and other products posted revenue of $74.6 billion in the period, also exceeding Street forecasts. Analysts expected $67.38 billion, according to Zacks.

For the current quarter ending in March, Apple said it expects revenue in the range of $52 billion to $55 billion. Analysts surveyed by Zacks had expected revenue of $53.65 billion.

Apple shares have declined 1 percent since the beginning of the year, while the Standard & Poor's 500 index has declined slightly more than 1 percent. In the final minutes of trading on Tuesday, shares hit $109.14, an increase of 39 percent in the last 12 months.

_____

This story was generated by Automated Insights (http://automatedinsights.com/ap) using data from Zacks Investment Research. Access a Zacks stock report on AAPL at http://www.zacks.com/ap/AAPL

Figure 7: Forbes publishes machine-generated finance news in collaboration with Automated Insights

generation algorithm developers, such as Narrative Science and Automated Insights, both registered patents on their particular technology on making more human-sounding narratives (Birnbaum et al., 2013; Allen, 2013).

The Associated Press started to publish automatically generated news on corporate earnings in July 2014, which is powered by a company called Automated Insights. They use an algorithm to crawl new earnings information from the servers of Zacks Investment, and publishes machine-generated

stories on them in less than a few seconds. The reports generated by Automated Insights are fast, scalable, and also error-free when the input data is correct. Their algorithm is known to be designed to mimic the writing style of a human journalist. However, it can also be said that the way the AP has been writing these earnings reports like a computer program, which copy-and-pasting numbers to the predetermined templates.

Narrative Science, another big player in this algorithm news market, partnered with Forbes in generating financial news. Similar to Automated Insights, they are also capable of generating automated summarization of companies' quarterly performance reports. As their patent describes, Narrative Science's algorithm determines the tone of narrative using various datasets other than the earnings report from the fiscal year. For example, it collects the analysis from human analysts, compares the performance of the targeted company with other similar companies, and the information about upcoming earnings release data of companies in the same sector.

3.2.4 Interactive News Service

Some of the interesting cases of algorithm-driven interactive news are found to be served in collaboration with traditional news media. In the cases below, the new stories were initiated by journalists and the technology helped them to generate different versions of stories that are customized to the needs of news readers on a digital platform.

The Opportunity Gap is a data-driven news on which states are providing advanced courses for low-income high school students (Figure 8). Earlier studies have shown that taking advanced classes is a critical factor for success

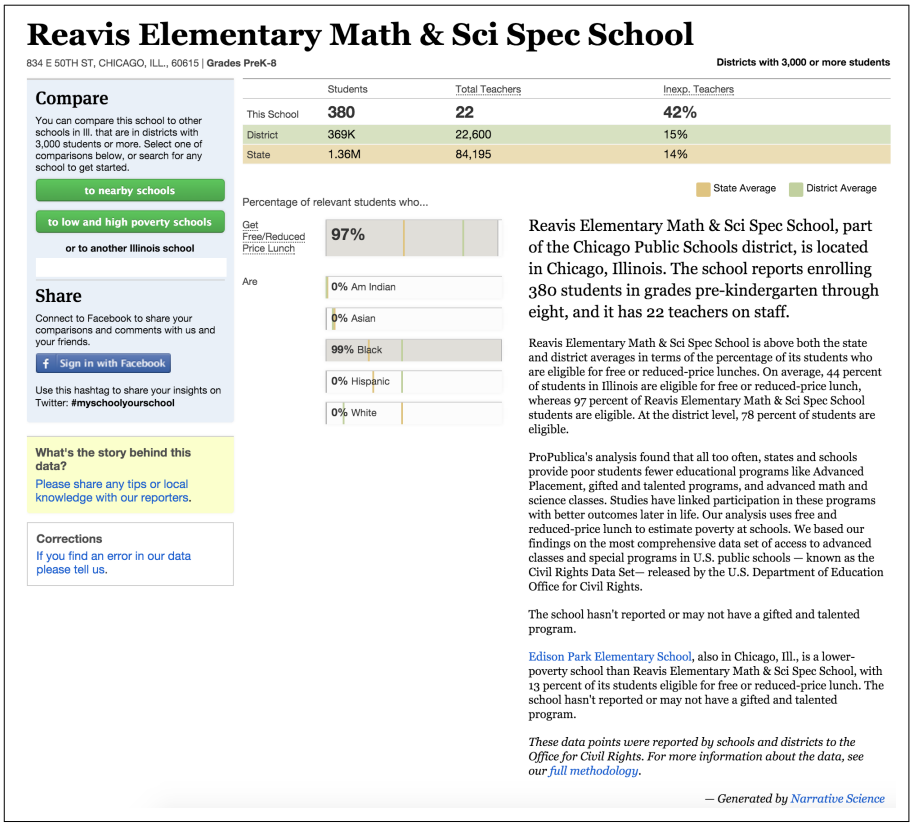


Figure 8: Narrative Science's algorithm creates news content for an investigative news report from ProPublica

later in life (Cole et al., 2009). ProPublica used a dataset from the U.S. Department of Education to investigate whether low-income students have equal access to advanced courses in high school, and created an interactive news service where a reader can enter a zip code to find and compare the details of schools' performance. With technological help from Narrative Science, they generated summarized reports on each of 52,000 schools, which would have been an overwhelming task for human labor. The crux of the algorithm-driven description is to include not only a summary of the

data for an individual school but also text comparing it to other schools in the state with various poverty levels. Their algorithm successfully takes different environmental conditions into account for better analogies and provides journalistic insight in generating an investigative report. Also, their system creates different sentences even for schools with the same kind of data, so that reading is more natural and not repetitive.

Another exemplary case of an interactive news service is found with Oregon Public Broadcasting (OPB), which rolled out a news app to accompany earthquake preparedness in the state of Oregon. Aftershock is created to provide a personalized news report about the estimated impact of earthquakes on a 9.0 scale where users enter search queries. Aftershock uses 384 possible combinations of data including shaking, soil liquefaction, landslide risk, and tsunamis. Users see a version of the story that is relevant to the location they have selected. In this way, a person in the center of Portland finds a different warning message than that of the person in Eugene.

3.2.5 Messaging Service

Much like asking questions by sending a text message to a friend who knows everything about a certain issue, a message news service is responding differently to the interactions its news readers make. As more messaging services are catching popularity, more news services make the use of existing infrastructure from popular messenger services or at least uses the metaphors such as sending and receiving text messages in delivering news information to service users.

Current news products do not support and make use of the interaction

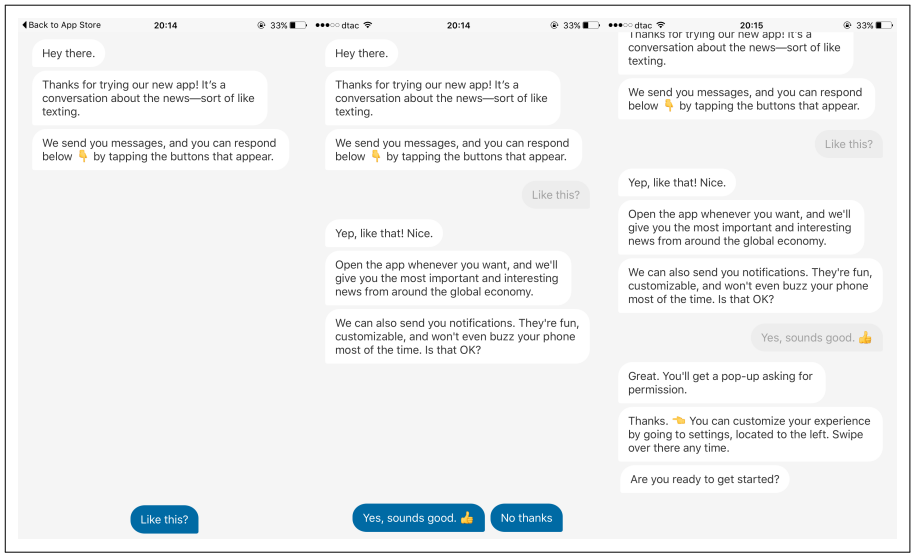


Figure 9: Quartz news app sends news blurbs in a form of text messages

that the reader makes while reading the news. Quartz's mobile news app has set an interesting remark by providing new in the style of chat software. This new kind of service allows for a more conversational consumption of news, which opens up the chance to provide news stories based on direct user inputs as opposed to showing news articles in a linear and fixed sequence. Quartz first sends a short text blurb and gives one or two choices that its users can make. For the most cases, they can show their interest in the topic by selecting the option for more detailed information or skip the issue by selecting the next button. If users want to see more details, Quartz's algorithm sends a series of more detailed information including text, graphics, or even movie clips. If users want to skip the topic, the app shows news blurbs about other subjects. Again, the users can choose to see more of the topic when they find an interesting news information.

Currently, Quartz news app does not take user profile or any other user analytics into consideration in selecting the news blurbs. Nor the service allows users to specify questions they would like to ask to the system. It means the depth of information is still at the level of a brief news report with Quartz. It only provides information that is prepared in advance. A good storytelling news service will take this a step further and generate stories on-demand based on requests. As this kind of application can handle more ad hoc responses by learning the patterns of news consumption alongside the profile details and knowledge level of each user, a more dynamic and contextual knowledge gaining process can be realized.

3.2.6 Immersive Storytelling

News products in this category provide highly interactive and immersive news-reading experience. The products reconstruct what is happening in the real world into an interactive and immersive computing environment, so the users can learn by experience or playing with the interface in gaining the news knowledge.

Cutthroat Capitalism is a web-based game hosted by Wired, which is introduced as an exemplary case of a news game that operates under the same economics logic of the Somali pirates in the book Newsgames (Bogost et al., 2012). The original article discusses the economics mechanics that run behind the scenes of Somali pirates[6]. Wired converted the article into an interactive game that runs on a web browser. By letting users control a pirate ship and take a role of a hostage-taker and negotiator for ransom. The

[6]http://www.wired.com/2009/07/ff-somali-pirates/

Figure 10: A news game from Wired that makes users to learn economical situations of the Somali pirates by playing the game

gameplay depicts economic conditions similar to that under which Somali pirates live. The game has successfully captured the reality into a fun and enjoyable game that may enhance a better understanding of the situations happening in the waters off Somalia. This news product was selected as one of the major categories of algorithm-generated news since this news game has great opportunity to tell different and more engaging stories to its players. Algorithms are heavily used in constructing the computing environment and set up the logic to deliver different messages and interactions upon players' choices.

1000 Days at Syria is another web-based game that brings the reality into a window on a desktop[7]. Created by Mitch Swenson, it is a hypertext-based news game that changes its story plot as a user makes choices while following the narrative structure. In the form of an interactive storybook, it

[7]http://onethousanddaysofsyria.squarespace.com/

vividly describes the first thousand days of the Syrian conflict. The algorithm used in this case is simple: it gives multiple choice options at certain points of the story, takes back what news readers selected, and gives a pre-determined set of text phrases in response to their choices. The complexity of software is set to minimum, but the power of storytelling is considerable. News readers get a chance to think about what are the available options for the characters in the story to survive, and discover the consequence of their choices. It is more efficient way of conveying messages compared to reading plain text article.

3.2.7 Personalized Report

Automated Insights and Narrative Science both developed narrative gener- ation algorithms and their line of products went beyond traditional newspaper-style articles. Both companies use their algorithm solution to generate reports or stories that reflect personal preferences and customized datasets for each user.

Automated Insights partnered with Yahoo! Fantasy Football in generating personalized football reports for the subscribed users. Automated Insights used statistical method and natural language generation algorithms to transform raw football data into match recap, draft reports, and match preview. Its algorithm, therefore, is applied to generate text content at varying length. With more than 7 million unique visitors per month[8], it is impossible for Yahoo! to create a report that meets wants and needs of every subscriber using human labor. However, Yahoo! can generate various versions of reports

[8]https://siteanalytics.compete.com/football.fantasysports.yahoo.com

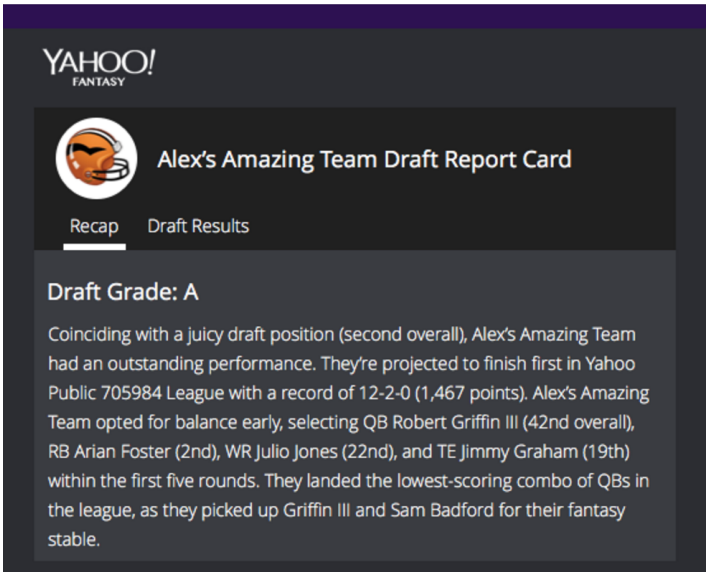


Figure 11: Yahoo! Fantasy Football provides personalized football reports in collaboration with Automated Insights

at scale, and more naturally sounding game reports by collaborating with Automated Insights. Yahoo! reported that they specifically asked Automated Insights to generate reports in more engaging tone than factual reports, such as to criticize disappointing plays and to write in humorous tone for more enjoyment out of football games recap.

Similarly, BodySpace, GreatCall, Edmunds are not traditional news media companies at all, but Automated Insights expands their line of business to include the ones who need to generate stories based on customized data sets. Automated Insights takes customized data sets as input, and generates reports in the form of a newspaper article. The audience of such report is just the user himself, which means each report is pinpointed to meet the needs of its users.

3.3 Analysis of Algorithmic Attributes

After analyzing algorithmic attributes commonly found among these news types, we were able to classify six dimensions of algorithmic attributes that define what it means to have an algorithm to generate news. Furthermore, each dimension of attributes aligns with the overall journalistic processes. For example, traditional news production requires a reporter to collect, check, and interpret data from various sources to write a news report. Similarly, an algorithm would also need to crawl, clean, and build a database to perform algorithmic tasks.

Each dimension, therefore, maps to one of the three journalistic processes to operate as an active creator of news information. Also, any algorithm-generated news can be said to be made by pairing different dimensions of attributes. For example, an automated sports news may use text broadcasting messages from sports media as an input, and create a brief report by performing statistical analysis in an objective viewpoint. Another algorithm news may take government data as an input and can be delivered as an intermediary news that makes heavy use of images and videos in conveying news knowledge.

The dimensions we found from the analysis are input source and data domain (which belongs to gathering stage in terms of journalistic processes), computation method and narrative framing (processing stage), and output type and interactivity level (presentation stage). The list of news products and corresponding algorithm attributes are shown in table 2.

Table 2: Algorithmic attributes of the current practices of algorithmic news generation

	Gathering Stage				Processing Stage					Presentation Stage				
	Input Source				News Domain				Computation Method				Narrative Angle				Output Type				Level of Interactivity			
	Web/API	Government Report	Local/Priv. Database	Sensor-generated	General	Sports	Finance	Local	Statistical analysis	Rule-based Processing	Natural Language Gen.	Other Methods	Data Only	Objective	Mass Customized	Personalized	Report (number+form)	Short Text	Long Story	Interactive Content	Static	Exploratory	Selective Navigation	Participatory
FanGraphs	o					o			o				o				o					o		
NFL Predictions (538)	o					o			o				o				o					o		
4th Down Bot (NYT)	o					o			o				o				o					o		
Earthquake Bots (Twitter)		o					o			o			o					o			o			
StatsMonkey (NS)	o				o					o				o				o			o			
Finance News on Forbes (NS)	o						o				o			o					o					
BigTenNetwork (NS)	o				o						o			o					o					
GameChanger (NS)	o					o					o			o					o					
Finance News on AP (AD)	o						o				o			o					o					
GreatCall (Link App, AI)				o	o						o					o		o			o			
BodySpace (App, AI)				o	o						o					o		o			o			
Yahoo! FantasyFootball (AI)			o			o					o					o		o			o			
Thomson Financial Report	o						o				o			o					o		o			
e-Sports News (CBS)	o					o					o			o					o		o			
Quakebot (LAT)		o						o			o			o				o			o			
Homicide Report (LAT)		o			o			o			o			o				o			o			
Yahoo! News Digest	o				o					o						o		o				o		
Flipboard	o				o					o					o			o				o		
Google Trends		o			o					o				o					o			o		
Techbot Report (Techholic)		o			o					o				o				o			o			
The best and worst places...(NYT)	o							o		o				o				o				o		
Opportunity Gap (ProPublica)	o							o			o			o				o				o		
Aftershock (OPB)	o							o		o				o				o				o		
Quartz App				o	o					o					o				o					o
NYT Election Bot/w Slack			o		o						o				o			o						o
News2Images (Naver)	o				o						o		o								o			
Wibbitz	o				o						o			o							o			
Cutthroat Capitalism (Wired)			o					o			o					o						o		
1000 Days at Syria			o					o			o					o						o		
Harvest of Change	o				o						o													o

3.3.1 Gathering Stage

For an algorithm to generate a news content, a series of steps are required to collect, clean, and organize the raw data. In recent years, huge quantities of digital data are being produced as a byproduct of the revolution in computer technology, new technology-enhanced ways that people interact, and the creation and monetization of data by many commercial entities (King, 2011). As the volume and variety of data exponentially increase, knowing the source of input data is becoming more important as those points can profoundly influence the output quality of algorithm-driven news products.

Input Source

The most common sources of data for algorithmic news creation are either websites or APIs (Application Programming Interface) from various media and commercial companies. A crawling algorithm must be designed to handle different specifications set by data providers. For example, an algorithm may take the entire code of a website or collect specified data formats through legitimate data communication channels. The data providers, in this case, are mostly other news media and commercial companies for recontextualizing existing content (Chua, 2010), but may also include publicly available data from government institutions. Some regularly send email or smartphone messages triggered by designated events. Quakebot and The Homicide Reports from LA Times make proper use of such data sources. Also, sensors are becoming popular as they create constant streams of data, and it is possible to collect personally meaningful data with the rise of

smartphone and other sensor-embedded wearable devices.

Data Domain

The products by news creation algorithms are heavily positioned on sports and finance domain. Data in these domains are consist of numbers such as game scores and stock prices, which require less complex computation technique in generating an informative content. By widening the idea of what is news, a brief report on individual's daily activities is increasingly recognized as a legitimate source of news since the report comes into one's newsfeed as if someone wrote an article on one's health information. The trend can also be identified by a growing number of services in this domain. BodySpace and GreatCall made the partnership with Automated Insights to generate personal reports using sensor-generated data. The personal trails of life are logged and reported in the form of a short news article with the help of algorithms.

3.3.2 Processing Stage

The dimensions of the algorithm in the processing stage represent the characteristics of news products on how they are generated using which computation method, and how the perspective and tone of the narrative are determined to frame the narrative. Mostly presented as finalized news content, the analysis of the underlying logic of commercial news products available on the market may not be entirely accurate in evaluating computation methods used in making such products. Diakopoulos (2014) suggested a useful methodological approach for research oriented investigation: applying

the concept of reverse engineering (Diakopoulos, 2014). Reasoning on how the output is made from the input gives hints on the understanding of the underlying components even when they are not so visible. We expect such analysis on algorithmic attributes provides a deeper understanding of elements found in algorithm-generated news products.

Computation Method

Broadly speaking, news creation algorithms either collect structured or unstructured datasets for data processing. Structured data are relatively easier to make rules for data modeling. Sports-related data are often stored and processed in a structured way: mostly in numbers and in fixed fields. Also, there exist domain specific statistical methods in sports: sabermetrics[9] and Elo-ratings[10]. Both methods help to find and quantify critical moments in sporting events that can be automatically detected by algorithms. 2015 NFL predictions (serviced by FiveThirtyEight[11]) uses Elo-ratings to provide a data-driven report on the forecast of each football team's chance of winning the Super Bowl. The percentage output comes from the software-driven analysis of alphanumeric data.

For unstructured data, a text processing algorithm may include cleaning the noise or parsing the text data from the collection of crawled data. The process may be as simple as to finding frequently used terms in the text

[9] Sabermetrics provides statistical methods for analyzing baseball events, which uses historical baseball records to find undervalued players, changes in the probability of winning the game at each at-bat situation, and many other records (Tango et al., 2007).

[10] Elo-ratings is the measure of teams' and players' skill levels based on the historical game data, which counts over 30,000 ratings in total.

[11] http://projects.fivethirtyeight.com/2015-nfl-predictions/

document, but it can also be a much more complicated data mingling process, such as finding hidden patterns or summarizing text phrases. Recent advances in natural language processing algorithm enhanced the capability of generating more complex and longer sentences. Computation methods also include other algorithms that are developed to perform designated actions, such as to convert text into other forms or to construct the virtual reality environments for enhanced news experiences.

Narrative Angle

Yahoo! Fantasy Football games generate a personalized report for each user by using the algorithm solution from Automated Insights. The algorithm uses personal game settings and records in the creation process, which makes each report different from other reports. While the text in the report is written based on user's personal game settings, the tone of the text itself follows the form of a sports news article from a typical news media outlet. Fantasy Football users read their report as if their team in the gaming world is real, which include information on their team and players' performances which are different for every one of them. More tailored customization in the tone of information is found on a report generated by BodySpace app (from Bodybuilding.com). The sample text of the app includes "Your new high in a number of bicep exercises is seven," "You did a total of 28 bicep sets as well, composing 25% of your week's workouts." By using data from various sensors, this app provides a tailored report in a tone that changes based on the user's performance.

3.3.3 Presentation Stage

Algorithmic attributes in the final stage are to shape and deliver news product to optimize news reading experience of its readers. In this stage, the form of the final output of news product, whether it is a short text or long story, or an interactive service, is determined. Also, the delivered news product can be as simple as static text phrases, but it can also be a messaging service where the choice of news topic is made by newsreaders rather than the editor on the media side.

Output Type

FanGraphs and NFL predictions provide interpretation of game data and statistical analysis in the form of a table. Usually, these reports are filled with numbers, which would be a good starting point for sports fans to fill in the gaps by themselves. Algorithm bots can be easily programmed for various reasons, such as to make surveillance reports, auto-upload repetitive posts, or transform alphanumeric data into other forms of representing data. We can often find automated tweets on Twitter that are designed to serve a single function. Narrative generation algorithms from Narrative Science and Automated Insights are the advanced versions of such news bots, and they are capable of generating longer stories by dynamically computing and determining the perspectives of the story. Most of their news products are, therefore, in this long story format.

Interactivity Level

While most of the newspaper-style text generation algorithms offer a static, text-only reading experience, data-centric reports often allow users to explore and find their meaning from the data. As more digital news services encourage wild interactions, some innovative news services have been born and are now on active duty. The New York Times announced a collaborative news service with Slack, which allows users to ask questions and receive live updates on the 2016 presidential election as if chatting with a friend[12]. Quartz also announced a news app that provides news information in a conversational manner (Figure 9). It first sends a short piece of information in the form of a text message. When a user shows an interest in the message by choosing one of the answers the app provides, it shows another chunk of text and waits for further user engagement.

3.4 Discussion

The machine has learned from and follows the news writing style of a human journalist, which is to write in an objective and descriptive tone. While the gap between machine-generated text and that of human may still exist, the perceived differences between the two versions are not substantial (Clerwall, 2014). The result from Clerwall's experiment shows that the text written by a human journalist was scored higher in the quality of writing (properties such as pleasant to read, well-written, clear, interesting, and coherent). However, the properties related to the credibility and representa-

[12]http://www.nytimes.com/interactive/2016/us/politics/election-bot.html

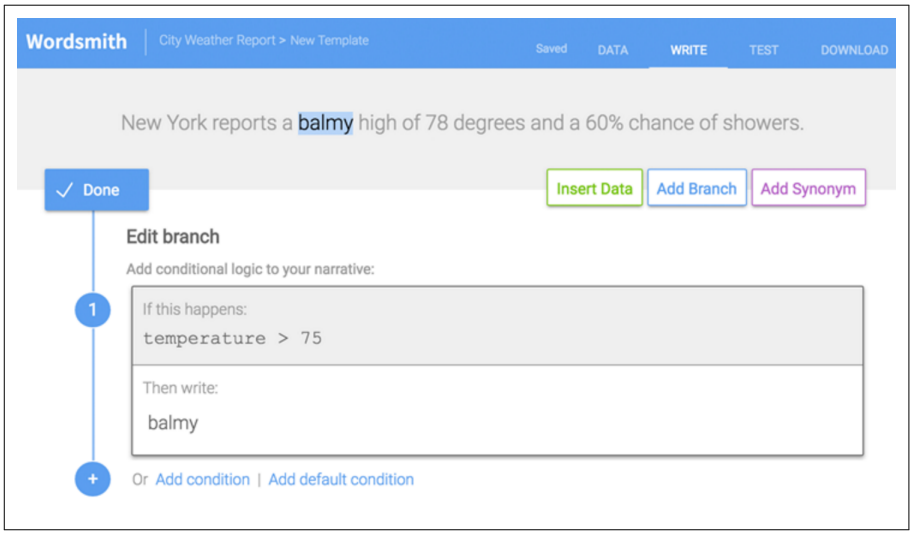


Figure 12: Wordsmith (algorithm solution from Automated Insights) offers algorithmic solution to convert structured dataset into plain English

tiveness nature of a news article were performed better by the software (such as descriptive, useable, objective, informative, and accurate). An algorithm is found to have strength in writing articles that are clear and objective. The direction of current technical development of algorithm as news content creator, therefore, has been set to deliver straight and summarized reports.

Automated Insights announced a new software called Wordsmith, which allows users to upload their dataset for the automatic generation of narrative content. It reads data off a spreadsheet uploaded by the user and generates narratives that change the tone of description based on the specified conditions (Figure 12). For instance, if the screen size of a product in the data is smaller than 25 inches, the system uses ‘good-sized’ to describe the product. If the screen size is between 25 and 50, then this adjective word changes to ‘huge.’ Screen sizes above 50 inches are described as a ‘mind-bogglingly

massive' product. Wordsmith is capable of generating different descriptions for each product automatically, and therefore, avoids using human labor for this straightforward and repetitive task.

Although Wordsmith is not dedicated content generation tools for news media, their ability to interpret any dataset users upload and to generate personalized narratives at scale shows the possibility of them becoming general purpose information-generation tool that even includes news stories. The next level of algorithmic activities in the practices of journalistic activities may already be here.

However, the review of existing algorithm-generated news products revealed limitations in utilizing data from various sources and how content are structured and presented to news readers. The limitations are listed below:

Data from Limited Sources

From the Quakebot of LA Times to machine-generated reports on Forbes, current narrative generation engines are still at the level of writing a simple and straight report in the few domains such as sports and finance. When these engines write about a summarized report on a baseball game event or the stock market of the day, they only count the broadcasted text data or JSON feeds that are directly related to the event. These engines, especially the ones from Narrative Science and Automated Insights, are known to be capable of generating text in a much more complicated way. However, we only see a news brief in news media platforms. By widening the range of datasets to collect and analyze, the algorithm news can be much more exciting. For a baseball game, analyzing external datasets as historical MLB

records on players and teams, and other baseball-related news databases allow the algorithm engine to produce text in more detailed and investigative manner. Also, if combined with personal data and sensor-based context data, more personally attaching news stories are expected to be generated.

Lack in Individual Context

As the algorithm follows the writing style of journalists, the output of algorithmic computation is still at the level of mass produced text. The algorithm can be designed to compute with the context of individual readers and produce the targeted output for each person. Personalizing the content is one of the most compelling advantages of the algorithmic approach to news content generation. The activities readers perform while reading an article, such as searching, clicking on links, can be accumulated by the system and analyzed for further analysis. Also, machine learning techniques can be applied to learn the patterns of readers or find correlation among news articles at large scale. Once the computer learns patterns, it can start to aggregate and remix stories from various sources and makes its modeling of topics. Then, the algorithm becomes capable of finding and writing about the currently rising trend and unseen patterns found in big data.

Linear Narrative Structure

On average, the quality of algorithm-generated narratives is found to be as good as human written articles (Clerwall, 2014). However, most of the machine-generated products are still at the level of static text articles. The focus of existing products is set to generate narratives by mimicking

the tone and style of a good human journalist. The collaborative project between ProPublica and Narrative Science, The Opportunity Gap, sets a good direction of how autonomous content generation algorithm should function: interactive and on-demand news generation. The limitation of the project is that Narrative Science had to compute all possible combination of outcome in advance and pre-generate all related text without knowing who would want which kind of information. It needs more sophisticated algorithms in generating the text on demand. The potential benefit of having more computation in journalism include topic detection and sensemaking of news information through visualization (Cohen et al., 2011).

Monotonous News Format

Current news products do not support nor make use of the interactions that news readers make while reading news on digital platforms. Traditional news media agencies continuously explore better ways to convey their news products using their platform, such as Quartz mobile app for interactive and participatory news information delivery. Quartz has set an interesting remark by providing news information through chat blurbs. A user reads the blurb sent from the app, then selects an answer from one or two options given by the system on whether to continue reading on the topic or to proceed with the next one. However, the depth of information is still at the level of a brief news report with Quartz. It only provides information that is prepared in advance. A good interactive news service would take a step further and compute the type and depth of information on real-time, and also take user profile into account in generating the content on the fly. As stated

above, Harvest of Change has also inspired many news readers by applying virtual reality technology to conduct an interactive news storytelling. This project has room for technological contributions that algorithms can make: it can be designed to convey different messages upon the behaviors and movements made while the headset is worn, and the storytelling will become much more personal and on-demand in this virtual environment.

Chapter 4

Research Questions

According to the chief product officer of Automated Insights, they are capable of generating two thousand articles per second and have generated 1.5 billion machine-generated articles in the year 2015 alone[1]. Since the computation time for an algorithm to transform a handful of data into a typical news article takes almost no time, counting the number of news articles their algorithm can create seems meaningless.

As the efficiency and capacity of algorithmic computation evolve, news media companies are aggressively adopting algorithmic solutions to generate simple and straight news articles via in-house or joint development efforts. Began with Quakebot from the LAT, traditional media companies are now generating thousands of machine-generated news articles per day, especially in sports and finance domains. One of the most recent cases is found with the Washington Post in generating algorithm news for Rio Olympic games[2]. According to their blog, their algorithm engine will automatically generate short updates on a daily schedule of events, medal events and rank, and 15 minutes alert before the start of a medal event.

Some media companies emerge from the traditional print model of

[1] Automated Insights' performance records were introduced by Joe Procopio, the CPO, in the remote presentation among United Nations International Telecommunications Union, Automated Insights, and the hci-d lab of Seoul National University on June 23rd, 2016.

[2] https://www.washingtonpost.com/pr/wp/2016/08/05/the-washington-post-experiments-with-automated-storytelling-to-help-power-2016-rio-olympics-coverage/

news delivery to more interactive and two-way communicative model of news services. Quartz and TechCrunch both initiated conversational news model, but their approaches were different: Quartz announced a stand-alone mobile application that requires installation from App Store, but TechCrunch created a news bot that runs on Telegram, which is a popular messaging service that opened up their platform for third party vendors for chat-bot development. Both methods have pros and cons, but the important thing is that new and interesting news dissemination models are given to news readers.

Moreover, algorithms are increasingly taking the role of news content creator in areas other than traditional news domains. Narrative-generation algorithm developers are often capable of handling data in diverse domains and construct a narrative in easily-readable format. Their algorithms can generate more targeted and relevant stories for individual users in the form of newspaper-style article, such as personalized health and workout report based on an individual's training records (in collaboration with the BodySpace) or automated vehicle descriptions that is generated by the profile of cars registered to their car-sales website (in collaboration with the Edmunds). They are not the kind of content that a news media company would deal with, but the growing influence of these algorithm developers allowed a new type of news services be born.

As illustrated above, more algorithmic computation is making substantial changes in how news products are made and shared among news readers. Meanwhile, we also found limitations commonly found among the listed products. The major drawbacks we derived from the review are (1) limited

data sources, (2) lack of individual context, and (3) linear and monotonous news structure. These limitations hinder the algorithmic computation from being differentiated and distinctive news creator. We set our first research question to overcome the limitations commonly found in the existing practices and maximize the capacity of algorithmic computation for enhanced news creator:

> **RQ1: How can the algorithm framework for news generation be designed to overcome the limitations of the existing algorithmic news generation process?**
>
> RQ1-1: How would the algorithm framework be designed to expand the breadth and depth of data gathering in algorithmic news generation?
>
> RQ1-2: How would the algorithm framework be designed to take individual context into account when generating news with algorithmic activities?
>
> RQ1-3: How would the algorithm framework be designed to present news information with user-driven interactivity?

After a new algorithm framework for news generation is proposed, we explored how the framework can be applied to real-world problems and be designed and developed with the underlying algorithm framework. The study on the second part of this thesis includes step-by-step procedure on how algorithm and algorithmic solutions are applied to gather, interpret, process, frame, and generate text phrases for news readers. The functions

of algorithm framework are aligned to the stages of journalistic processes, which was introduced in the previous chapter, in order to explicate the activities of the algorithm in terms of journalistic processes in the news making process.

With the algorithmic framework that functions, we built a working prototype of a news system that generates news content entirely by algorithmic approaches. The algorithm framework, which is implemented in the first study, becomes the backbone of the news system. The implications learned from the review of existing news products (Chapter 3) and research framework (Chapter 5) are used in deriving insights on what and how to build such a system. In other words, the user interface elements and the style of interaction are designed upon the implications in the way the system can overcome the limitations found in current practices. We explored and discovered design space for algorithmic news generation system, which is introduced in the second part of the research results. The second research question is to examine how news generation algorithm framework and the system can be designed and developed to derive implications for the software and user interface implementation.

RQ2: How an algorithmic news generation system be designed and developed with the underlying algorithm framework?

RQ2-1: How the conceptual modeling of algorithm framework be implemented and integrated into a news generation system?

RQ2-2: What is the design space of the user interface of algorithmic news generation system?

After the system development, we conducted an evaluation study on how would news readers use this system to generate their versions of news, and how much would they appreciate the quality of output content and enjoy participatory news making process. The evaluation was held to examine the quality of the proposed algorithmic news generation system in the following ways: which were (1) the comparison of the perceived news values between the content generated by the algorithm system and a human journalist, (2) the comparison of the perceived news values on the content generated by the proposed news-generation system, and (3) the evaluation study on their experience with this interactive and personalized news generation system from three distinct user groups: news readers, journalists (news experts), and HCI researchers and practitioners (system experts).

The first and second part of the evaluation, to compare the perceived news values between a human-written news and the system generated news, was conducted by recruiting participants for in-lab experiments. For the third part of the evaluation, which was to evaluate the quality of the system and its output content, we conducted interviews with three different groups of users to take multiple viewpoints into evaluating the news generation system. Since the system is only at the level of a prototype rather than a published news service, we had to assess the system regarding their perceived value of the system as a usable and useful tool for news information seeking device, and also the perceived news values of the content generated by the system. For the third research question, therefore, we recruited three different categories of users to derive the overall implications and discussions on the system development process.

RQ3: How is the algorithmic news generation system perceived by news readers and domain experts?

RQ3-1: How is the algorithm generated news content perceived by news readers and domain experts in terms of news values?

RQ3-1: How is the algorithm system perceived by news readers and domain experts in terms of the system quality?

Chapter 5

Developing Algorithm Framework for News Generation

The use of algorithm can potentially generate news different from that of a human journalist: its ability to take a broad set of data into account, tailoring information to the wants and needs of its readers, and adding interactions and visualizations to induce more engagement and help them to make better understanding of news information. In this chapter, a new algorithm framework is proposed, which can generate news content using various algorithmic activities. The procedure of how algorithm framework functions consist of three stages, which are (1) data synthesizing, (2) narrative framing, and (3) news presentation stages.

5.1 Opportunities for Algorithmic News Generation

As algorithms are playing increasingly significant roles in news generation, the need for more thorough analysis on how to apply algorithm solutions to various processes of news production stages. The review on the current practices of algorithmic news generation revealed an interesting trend in the delivery of news information to interested individuals. We propose a new framework for algorithmic news generation based on the analysis of current practices of how algorithm generates news products. The major implications found from the review are listed below:

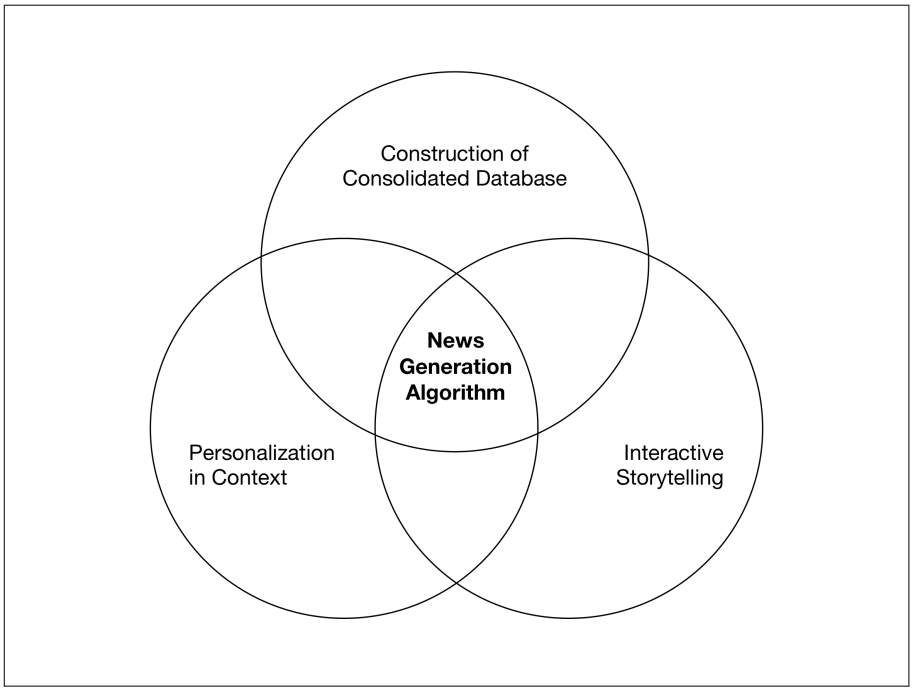


Figure 13: Three-circle diagram that illustrates the key concepts proposed in this thesis

- Existing algorithms collect data from limited sources that prevent the generation of more thorough data-driven stories.
- Existing algorithms lack in providing more personally tailored stories even when their biggest strength is the power to compute large scale data and generate differently rendered stories per individual reader.
- Existing algorithms generate news in learn and monotonous format, while it can be designed to generate content on-demand, which meets the requirements set by the channel or platform where news is distributed.

After reviewing existing algorithm-generated news products, we found three opportunity areas where the algorithm can have the greatest impact on the production of news information (Figure 13). The first area is to maximize the capacity of data handling. Algorithmic computation is especially powerful in pull data from multiple sources and integrates layers of information. The second area is to frame narrative generated by the algorithm with the angle of personalization. Personalizing the content itself is closely related to the use of consolidated database discussed in the previous section. The content becomes more personal when the system can make use of data that contains the context around the person. The third opportunity area is to construct interactive and visual storytelling in delivering news information.

5.1.1 Constructing Consolidated Database

In recent years, the volume of data is exploding where the size is beyond the ability of typical database software tools to capture, store, manage, and analyze (Manyika et al., 2011). The definition of big data inevitably involves the capacity to collect and analyze data with an unprecedented breadth and depth in scale (Lazer et al., 2009), and the power of big data comes from the ability to search, aggregate, and cross-reference large datasets (Boyd and Crawford, 2012). From a research point of view, analyzing a large pile of aggregated social data can tell more about the trends and patterns involved in human activities. Now we can measure, and therefore manage, more precisely than ever before, which means we can make better predictions and smarter decisions than ever before (McAfee et al., 2012).

In other words, the competitive advantage of algorithmic computation

over human capacity comes from its power to make use of unlimited volume and variety of data. To put it in another way, the quality of algorithm-driven products is enhanced when a large set of data is fused into the data processing system. In addition to event-related data, news generation algorithms can take two additional categories of external data that add depth and breadth to data gathering for enhanced algorithmic computation: diachronous and synchronous data.

Diachronous data enhances computation accuracy by providing historical data to supplement the incoming feed data on an event. The term diachronic is referred to an accumulated or historical data, which in turn, enables more accurate and efficient computation for predicting the future outcomes. The accuracy of algorithmic computation in FanGraphs and NFL Predictions from FiveThirtyEight increase when decades of accumulated game records are statistically analyzed for references. In other words, their algorithms can take 'diachronic' data into account to make better performances.

A news startup 'RecordedFuture' produces real-time threat reports based on the analysis of largely available data on the web[1]. Their intelligence algorithm first harvests data from various sources including private networks, IRC channels, forums, social media, and other feeds of data in real time, across all languages. Then the algorithm uses natural language processing and machine learning algorithms to structure information and organize threat information in the order of time, from the past to future events. The power of their intelligence algorithm comes from its ability to analyze a large pile of data in depth, and count diachronous data into account when making an

[1] http://www.recordedfuture.com

algorithm-based prediction on what has not happened yet.

Synchronous data, on the other hand, widens an algorithm's ability to consider the context and other perspectives when making autonomous decisions. When Narrative Science and Automated Insights generate sports news, their algorithms count not only real-time streaming game data, but also other related data such as players, teams, and other game events. It can be said to include "synchronic" data in the computational analysis. In Narrative Science's patent document, the input data for a sports news may include "event data in the form of a game box score, historical data about a team and/or player, more general game, player, team and/or league history data (such as records), and forward-looking data about games, players, teams, league schedules, etc" (Birnbaum et al., 2013). The importance of having diachronic and synchronic data as well as any derived data from remixing these datasets are highlighted in this patent document.

Google Trends is a website that shows collective intelligence on the search terms put into Google's search engine[2]. In its graphical representation on a search term, the horizontal axis represents the time, and the vertical axis shows the total number of searches made by the entire Google users. A search made by a single person would not have meant much to other Google users, but putting all search queries together over time with the scale of the global users made search keywords relevant to many. The website also shows some trending stories that are gaining popularities in the last 24 hours or featured stories that are globally on trend. Much like Twitter's trending topics and Instagram's hashtag function, Google Trends also makes a good

[2]http://www.google.com/trends

use of synchronous data to curate the right information at the right time.

5.1.2 Personalization in Context

One of the major strength of algorithmic computation is its fast and scalable data analysis that count the context of information recipients. In other words, there is a significant opportunity for an algorithm in news generation that it enables the creation of news content for an individual news reader.

In communication research, the effect of tailoring in message creation has been widely studied. Tailoring a message means using data about the given individuals to determine what content to generate, and the contexts and frames surrounding the content (Hawkins et al., 2008). In these studies, the messages were found to be more effective at engaging and persuading the audiences when messages are tailored compared to generic mass messages (Rimer and Kreuter, 2006; Roberto et al., 2009). According to Hawkins et al. (2008), there are three distinct strategies in which to achieve the goals of tailoring messages which are personalization, feedback, and content matching. Among these strategies, personalization is the one that is used as a method to enhance message processing by increasing attention or motivation.

The definition of personalization varies in different fields of research. Fan and Poole (2006) performed a meta-analysis on varying definitions of personalization, and their study provides comprehensive findings on the definition of personalization in multiple fields. In cognitive science, it is "explicit user model that represents user knowledge, goals, interests, and other features that enable the system to distinguish between different users

(Brusilovsky and Maybury, 2002, p. 31).” In terms of building an interface, personalization means “the understanding of the user, the user’s tasks, and the context in which the user accomplishes tasks and goals” (Karat et al., 2000, p. 50). In computer science, a more technology oriented definition of personalization is applied: “a toolbox of technologies and application features used in the design of an end-user experience” (Kramer et al., 2000, p.44).

Collectively, the common thread in these research can be said to involve activities of tailoring information and services to consumers based on certain knowledge about them to achieve targeted goals in mind (Tuzhilin, 2009). The key quality of personalization is to provide ‘a means to know what there is to know and how to know’ the information that corresponds to an individual’s specific interests and needs (Gillespie, 2014; Lavie et al., 2010). Compiling a large set of personal data that involve users’ profiles and preferences are the key enabler for such personalization (Pariser, 2011; Napoli, 2014; Kizilcec, 2016).

Looking into the details on what can be personalized in an information system, Fan and Poole (2006) also proposed four major aspects that can be manipulated for personalization: the information itself (content), how the information is presented (user interface), the media through which information is delivered (channel), and what users can do with the system (functionality). By implementing an algorithm framework, we expect all of these aspects should be considered to achieve the goal of making a personalized news generation system.

Personalizing news content is closely related to how the system constructs

a database for data analysis. The content becomes more personal when more data with an individual context are taken into the consolidated database from multiple sources. There are two different types of personalization: an explicit method that uses direct user inputs, and implicit method that infers preferences from data (Thurman and Schifferes, 2012). These content-based approaches have been applied to various forms of news and news-related services, such as providing a personalized selection of news information for personal news agents, news readers for wireless devices, and web-based news aggregators (Liu et al., 2010).

A direct input from news readers is a more convenient way of collecting data. The types of explicit data include the logging of a user's click or touch on the interface, descriptions in a user profile, search logs, sensor-generated data, physiological measures of any kind, photos and videos the user watched and liked, and more. Due to the growing digital economy, more and more data are expected to be generated and at the same time, more services are expected to be launched that utilize these datasets. The election bot launched by New York Times in collaboration with Slack is an example of how news information can be provided on behalf of the direct input from its users. Depending on one's knowledge level and interests, different questions will be asked, and the bot will need to meet the needs of varying users. The downside of the NYT election bot is that the answering process is done manually: it requires human labor to process the questions and to answer questions instead of automatically composing the narratives. Nor it does not answer in consideration of any contextual information, such as the profile and preferences of users, in the communication process.

The recent developments in computation techniques are geared towards finding and learning patterns from the trail of digital activities. Users' social media activity is a good source of information about the person, which allows the personalization implicitly. While direct inputs into social media, such as uploading posts and tweets, are an explicit source of data (O'Banion et al., 2012), making a secondary analysis of an individual's data can also give interesting information about the user. For example, the pattern of news consumption is a good source of input for machine learning, which allows the system to make automated predictions on what users would like in real-time. Once the machine learns with enough training data, it is known to produce fairly accurate predictions or classifications.

In sum, the system-driven personalization by learning implicit data and the user-initiated personalization by making use of explicit user data are both closely related to generating personalized content for news readers. An algorithm framework needs to be designed in a way to automate system-driven personalization in the message processing and to offer a customizable user interface for active engagements from news readers.

The three most widely used personalization tactics are identification, raising expectation, and contextualization (Hawkins et al., 2008). Identification is a tactic that recognizes and integrate person's detail in communication, and raising expectation refers to include messages such as 'the following message has been created especially for you.' Contextualization is to frame the communication message in a context that is meaningful to the recipient. To frame the content is what journalists do to create an article with an angle. In this research, we set our focus in personalizing news content by

contextualization.

5.1.3 Interactive Storytelling

There also has been substantial research efforts in defining interactivity in many fields of research. Some scholars suggested communication models distinguish machine interactivity from person interactivity, and its definition is "the extent to which users can participate in modifying the form and content of a mediated environment in real time." (Steuer, 1992; Hoffman and Novak, 1996; Sohn, 2011).

According to Murray (1997), the properties of digital environments are suggested as follows: procedural, participatory, spatial, and encyclopedic. Since the birth of Eliza, an intellectual computer program that shares text-based conversation with its users, there has been a series of efforts in developing software solutions that can manipulate sharing of conversations with a system like communicating with a person (Murray, 1997, p.71–74). Murray defines the term interactivity the following way: "the primary representational property of the computer is the codified rendering of responsive behaviors, and this is most often meant when we say that computers are interactive." Some of the most recent interactive news services such as Quartz succeed the concept introduced by Eliza in which it allows human to human communication type of interactivity in conveying news information.

The other type of interactivity involves human to machine communication, where interactivity refers to the behavior of users on how they exert controls to manipulate the interface level elements that the computing system mediates (Bolter et al., 2000). In news media, interactive news refers to the type of

news products mostly focused on delivering objective and transparent news information to their readers. Backed up by the vastly increased amount of data and greater power of data-mining software, how we consume news information becomes more participatory and interactive (Jenkins, 2006; Flew et al., 2012; Lewis and Usher, 2013).

The most significant change in this era of news experience is that readers acquire news at any time by selectively exploring a wide variety of digital sources (Liu et al., 2010; Fox and Duggan, 2013). When a system supports rich interactions in modifying the components of the user interface for more personalized information seeking, we call it a 'customized' information system (Blom, 2000; Beam, 2014). What we often call as an interactive system belongs to this second type of interactivity. This kind of product often affords interactivity as a method to offer more control in the reception of such information, and the types of interactivity include features such as zoom in for detail, hover highlighting, annotation balloons, hyperlinks to the source of data, or any other methods that help to make sense of the given information.

Moreover, an active participation allows a news story to evolve as an ongoing stream of live information. The greater the responsiveness, the tighter the symbolic coupling between the actions of users and procedural representations are made from the sophisticated implementation of interactivity functions (Bogost, 2007). To achieve a significant level of participation, a system must consider greater sophistication in designing an interactive news information system. Algorithms can add different layers of stories and therefore generate different types of news content depending on readers'

interests and choices (Hamilton and Turner, 2009). In other words, news services can be designed to provide unique information that attracts a subset of readers highly interested in the selected topic and sustain their interest. Therefore, providing different news content based on algorithmic computation is the key design factor of a news system that meets the needs of news readers (Beam, 2014).

Segel and Heer left interesting, yet inspiring, remarks on the changes when interactivity is added to a traditional form of storytelling: "Stories in text and film typically present a set of events in a tightly controlled progression. While tours through visualized data similarly can be organized in a linear sequence, they can also be interactive, inviting verification, new questions, and alternative explanations" (Segel and Heer, 2010). Interactive and visual storytelling, in turn, is crucial to providing an intuitive and fast exploration of large data and accommodating storytelling (Wojtkowski and Wojtkowski, 2002).

In sum, Table 3 illustrates the three major opportunity areas for an algorithmic news generation with operationalized definition. In the following section, we propose an algorithm framework that deals with each of the opportunity areas as a procedural process for algorithmic activities involved in operationalizing the original definitions.

5.2 Algorithm Framework for News Generation

In this new algorithm framework, we defined how algorithm framework should function on every stage of news generation process. In order to maxi-

Table 3: Three opportunity areas for algorithmic news generation

	Definition	In the Framework	Algorithmic Activities
Constructing Consolidated Database	To collect and analyze data with unprecedented breadth and depth in scale (Lazer et al., 2009)	To construct database with diachronous and synchronous of data as an input to the algorithmic computation.	Crawl and learn data from various sources: diachronous (historical datasets) and synchronous data (other types of data that are relevant to the current event).
Personalization in Context	To determine what content to generate, and the contexts and frames surrounding the content (Hawkins et al., 2008).	To frame the narrative content in a context via explicit and implicit preferences.	Make use of explicit and implicit preferences from users in narrative construction: explicit (profile or direct input), implicit (behavioral patterns and log data).
Interactive Storytelling	To which users can participate in modifying the form and content of a mediated environment in real time (Steuer, 1992; Hoffman and Novak, 1996).	To present news information within a user interface for customized news reading and visualized storytelling.	Afford various interactivity functions for controling user interface elements, such as to zoom-in and out, navigating using hyperlinks, mouse hovering, etc.

mize the capacity of algorithmic computation and quality of content generated by the framework, news generation algorithm should be designed to tackle the opportunity areas suggested in the previous section, which were to construct consolidated database to maximize computational power, to apply personalized framing in data interpretation for narrative construction, and to present news content with added interactivity and visualized way for more engaging and self-driven storytelling.

These algorithmic activities are aligned to journalistic processing stages, which describe the processes required for traditional news production. In the following section, we propose an overall procedure of how algorithm

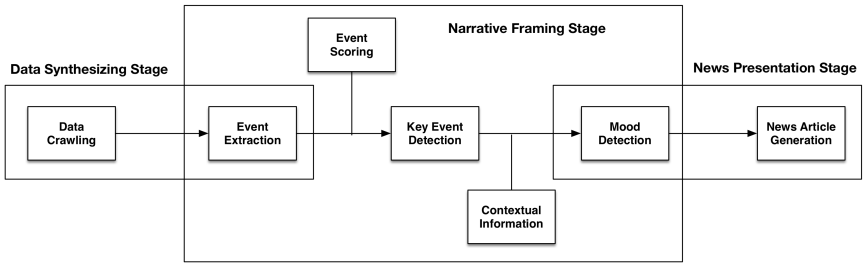


Figure 14: The overall structure of algorithm framework

framework can be built to achieve the goals set by the new framework development. The gathering stage is where an algorithm collects a large pile of data and synthesize them into a consolidated database. The processing stage is equivalent to applying different computation methods to prioritize and set perspective angles that emerge from data interpretation. In the final stage of news generation, an algorithm presents news content to the readers in various forms depending on how an interactivity layer is added and a visualization scheme is applied to the output content.

The concept of an algorithm framework for news generation is already introduced in our previous research (김동환 and 이준환, 2015), and we modified the methodological model from the research to illustrate how each stage of news generation process function with respect to the application of algorithmic activities (Figure 14). Each stage of the overall framework is explained in detail in the following sections.

5.2.1 Gathering: Data Synthesizing

Building a consolidated database is the most important goal in the first stage of algorithm framework. Similar to human-centered journalistic

processes, the quality of algorithm-generated news depends on securing exhaustive and exclusive data as the input to the algorithmic computation process. In order to build a consolidated database, the algorithm must utilize data from various other sets of data (Figure 15). In this thesis, we defined the process of constructing a consolidated database as the combination of three groups of databases: (1) event-related database, (2) diachronous database, and (3) synchronous database.

The first and the foremost database is the event-related database, which consists of data for the event that the framework is working on for news generation. It is the main database that all the algorithmic computation is based on, while the other two external databases extend the depth and breadth of algorithmic activities by providing historical and other types of data to make up the automated decision-making process.

Gathering data from various input sources can be performed in multiple ways. The most common method to collect data would be to create source-specific crawling code. There may need a crawling software designed for each source of data, since the format and configuration of data may vary depending on the source. For example, a sports news media may send (or internally exchange) text broadcasting messages that contain detailed activity records in JSON file format via API (Application Programming Interface) channels, while the changes in today's stock market information are often collected directly from HTML files.

Meanwhile, government agents and other institutions usually publish their annual reports PDF documents. It is often more difficult to convert text and figures in PDF into the structured data structure for computation and

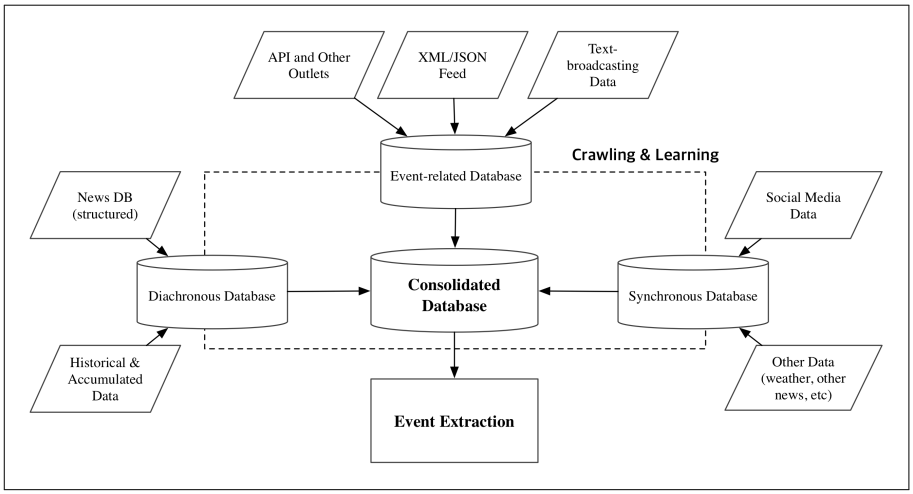


Figure 15: Data synthesizing stage of the framework

therefore interpreting these type of input data require a different strategy in data gathering. All of these cases require different specifications in setting up data communication channels, which means multiple algorithms are required for data crawling. Usually, a crawling software is written in programming languages such as Python or Ruby, which are gaining popularity due to their strength in handling text data.

Use of diachronous data and synchronous data also expand the ability of algorithmic computation for quality reporting. If only summative text phrases are all it is required for output, then the database may not need to include these data in the first place (event-related data would be more than enough). However, in any other cases, building a consolidated database enhances algorithm's decision-making process by providing excessive and complementary data sets that can be utilized in multiple ways. In the case of a finance news, the database may dynamically add historical records on

the market conditions when similar events took place, and also insert the audit result of event-related company's financial records. The machine can generate much deeper and clear analysis on one company when all of the data sets listed above are collected.

Diachronous data helps to find patterns and trend that might arise in the current event, and synchronous data helps to make rules for anomaly detection and change perspectives on interpreting the feed data. With the burst of social network services, the amount of digitally traceable data also grows exponentially. For example, entire tweets that Twitter users make can be reached and stored using Twitter's firehose service, and hashtags that Instagram users add on photos can be collected through their API after simple authentication process. Thus, social media data became an excellent source of data for extensive data analysis. The results of these analyses may not be directly applied to the interpretation of news events, but give enough sense on how the general public think and behave.

In other words, they help algorithms to make plausible inferences that enhance the automated decision-making process. Analysis of clickstream and behavioral data such as the time spent on one page to the other and search keywords entered at Google also strengthen computational capacity. The building of a consolidated database for news content generation should not miss the opportunity to include types of data that can tell unspoken wants and needs of news readers.

In addition to data mining, the ability of algorithm framework would be maximized if it can handle user-generated data or the system log data to learn for better analysis. The framework would need to collect the explicit

preferences stated by news readers, or it can learn from the patterns and behaviors on what news readers want to see in sports news. In order to collect log data, a news system or a news platform is required so news readers can navigate to find news contents they like. For instance, Google offers a snippet of codes that tracks the behaviors and click streams that users make on a website. Google uses the data to adjust and specify their target of advertising efforts. A similar process can be applied to an algorithm-driven news system, where the tracking data are analyzed and learned to provide more tailored and context-aware news stories.

5.2.2 Processing: Narrative Framing

In this second stage, the algorithm extracts events out of the consolidated database and makes them ready for narrative construction. The procedure includes putting raw data into a codified cleaning process and prioritizing meaningful data that are essential to the current event by scoring all events. Furthermore, detecting key events takes another computational process driven by statistical analysis (Figure 16). Which methods to apply may differ depending on the news domain.

From event scoring to context information, we developed computational process called 'complex weight matrix.' The complex weight matrix computes the weight on each event via both static and dynamic calculation of the importance of events to process data to informational content. Static weights are the pre-determined scores upon the rules set by the system developers. If an event is about scoring a home run in sports, a rapid increase in the stock market index or an earthquake happened right next to a city, and then

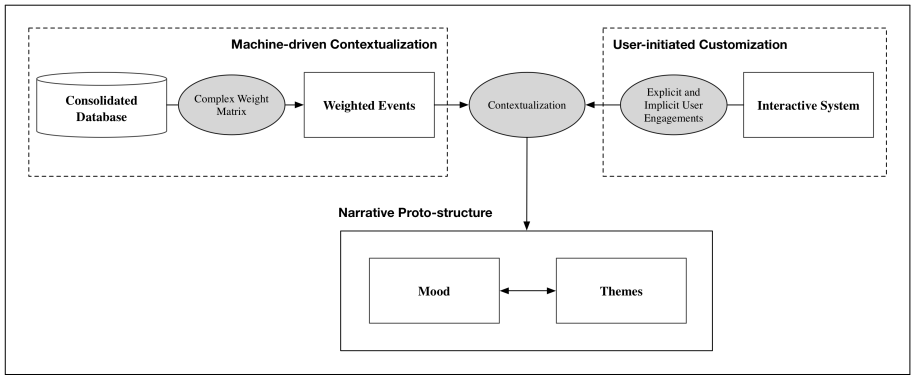


Figure 16: Narrative framing stage of the framework

such an event is usually regarded as newsworthy events that need immediate attention. A static weighting system is more reliable and fast at detecting events that are important for everyone. However, the content generated through static weights can only include obvious and predictable events.

Dynamic weight, on the other hand, adds variance to the computational approach to key event detection by computing relative importance of events. The winning score is more dramatic than a single scored in the first inning in a sports game, and a slight increase in the price of the stock I hold is worth to know even when the entire market is losing. Static and dynamic weighting mechanisms complement each other. It is always important to detect what is really important to everyone, and knowing what made the difference is also critical in news report. The complex weight matrix uses both methods for narrative framing, and therefore it enables the algorithm framework to generate news with the tone of an argument.

In this algorithm framework we propose, the complex weight matrix filters the key events and prioritize all other events to derive the angle of

narrative flow, which means the news content is framed to deliver a message with an angle in the next stage. Moreover, any explicit or implicit user engagements made with the news service can be used to refine the weighting system. Explicit user engagement includes the profile and preferences, IP address, and any other context-specific information. It might even include direct communication messages via chat-style interactive services. Implicit user engagement includes the habitual patterns, the profile of the user, and crowd's trending data, which are implicit and require learning process with sample training data for accurate algorithmic computation.

A series of computation took place in the narrative framing stage, and the final deliverable of this second stage is the prototype of narrative structure, which we call a 'proto-structure' before generating the full narratives. The definition of narrative proto-structure in this research is what the algorithm internally creates as an interim deliverable before moving on to the final stage, where the elements in the proto-structure are assembled and contextualized for news story generation or service development.

The procedure for generating a proto-structure of the narrative is similar to how a human journalist sets a frame of the news content. Journalists set their perspective to interpret the mood of the happenings; then they write stories with their perspectives on the matter. In other words, the person has set a frame on how to lead the narrative by looking at the data. What a human journalist knows by heuristic needs to be taken into a procedural computational process for machines. The algorithm needs to mimic the framing process in order when creating a proto-structure of the narrative.

The proto-structure consists of two major components, which are mood

and themes. To set up a mood of the event, a human would know relatively more important event by years of practice or just from common sense, but the algorithm needs to follow instructions written in codes or iteratively be trained with extensive data for automated decision making. In this framework, the mood is determined by interpreting consolidated database using the complex weight matrix. In addition to automatic mood detection through algorithmic weight computation, narrative proto-structure also makes use of a pre-determined set of rules to interpret the events established by domain experts. The professional and domain-relevant opinions are built as a set of rules defined in themes, and the proto-structure is generated using both mood and themes detection for elaborated narrative generation in the following stage.

5.2.3 Presentation: News Presentation

In news presentation stage, the framework converts the narrative proto-structure to a natural and human-sounding narrative with two algorithmic approaches: natural language generation and template-based text generation (Figure 17). The basic model of natural language generation involves multiple stages, where what (knowledge) needs to said is determined (communicative goal) that meets the expectation of the person (user) in a certain tone of language (discourse) (Reiter et al., 2000; Manurung, 2004). In this framework, knowledge refers to the consolidated database that requires determination on how to structure and aggregate data to realize narrative to news readers. Recently, open source programming resources for natural language processing are increasingly available, which in turn lowers the barrier of computational

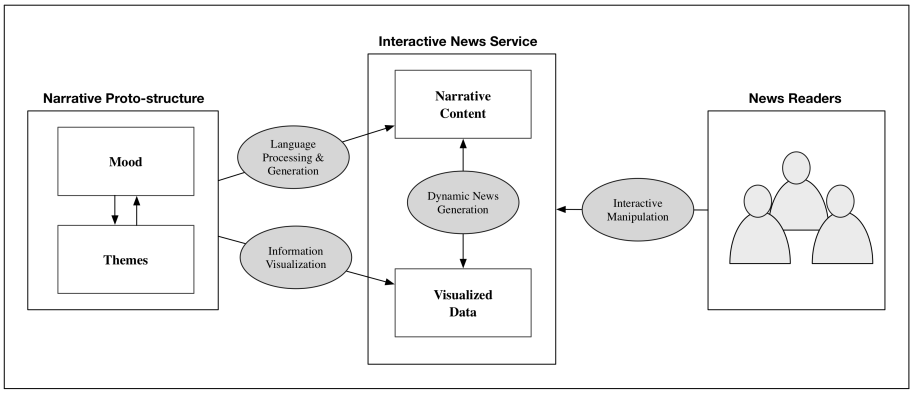


Figure 17: News presentation stage of the framework

attempts to narrative generation.

Template-based text generation, on the other hand, is good supplementary tactics that can be applied to narrative generation. It may be as simple as to plug-in various data elements into designated positions in a sentence. The generation of a sentence can be as simple as to insert today's date in an introductory greeting sentence, or it can be much more complicated that various descriptions of a victim in a local crime report. This method can sometimes be an even more efficient way to produce a narrative in news domains or languages that are currently difficult with natural language generation. In the Korean language, for example, concatenating morphemes or tagging part of speech require much more sophisticated techniques due to limited availability of Korean corpus and software packages.

After narrative generation, the framework present news information in a way it is optimized to the context of news readers. The objective of the framework development in this stage is to support personalized customization of news information for each user. Personalization and customization are

closely linked concepts and are frequently used interchangeably in many studies. Like personalization, the definition of customization also varies based on the field of study. Personalization occurs in an information system, while customization is defined as the amount of user involvement in the process of personalizing the system (Blom and Monk, 2003; Sundar and Marathe, 2010; Beam, 2014).

The definition of personalized customization is that an algorithm news framework is designed to take personal context into account when creating content, and it aims to provide a news service that is customized to the needs and wants of the news consumers. We expect personalized customization algorithms to be more scalable and responsive to real-time requests and be able to offer personalized news reading experiences for each and every news reader. It means the disruption of the current process of journalistic activities. Rather than generating news stories for everyone on a daily basis, news content can be created through a specific request and delivered to one person.

Much like Flipboard's algorithm called 'Duplo[3]', the algorithm in the framework would also need to consider the constraints of the system on the recipient side of news information and alters the layout of contents dynamically to meet up with the varying size and layout requirements from users. The more margin is given to the framework means more flexible and lengthy the content is. If only a subtle portion of margin is given to the framework, then it can only produce information that can be fitted to a device with a small screen, such as a smart watch.

[3]http://www.flipboard.com

Since the output of framing stage was only a prototype of narrative structure, the presentation stage can flexibly turn it into different types of news product. Depending on the domain of news, type of mood, marginal portion, and the amount of user engagement, the final product of this news algorithm framework is presented in a different form. For a report that needs to be delivered fast and summarized manner, a newspaper-style news brief will be the best fit. If the user wants to jump to the latest updates on the presidential election event, for instance, the framework may consider having a chat-style interface to quickly engage with users in real-time.

5.3 Discussion

In this chapter, we proposed a new algorithm framework for news generation that is designed to overcome the limitations of existing algorithm-generated news products (RQ1). Recently, algorithmic news generation gained popularity in generating news in limited domains such as sports, finance, environmental, and local news. These domains had a relatively low barrier for algorithmic news generation for multiple reasons: data came in structured form, data were openly available or easily accessible via crawling software, and a simple and straight style data-driven report was expected and acceptable for news readers.

However, the algorithm is capable of being more than a simple data conversion tool. Its ability to compute large data in a short time enabled more thorough and investigative news generation. Also, it is able to create personalized content based on the context and preferences of each news

reader, and it could even update the content based on the person's behavioral patterns, explicit and implicit preferences, previously made choices, and the person's friends or other community members' collaborative filtering and collective intelligence. There were numerous ways to enhance the production and presentation of news with algorithmic activities.

To answer the first research question, which was to examine how can the algorithm framework for news generation be developed that maximizes algorithmic capacity, we created the overall structure of how an algorithm framework must be designed to maximize the potential of algorithmic news generation (Figure 14). In this chapter, we found three key opportunities areas where algorithms can make changes in generating a different yet more engaging news story, and we proposed all-new algorithm framework where algorithmic activities are aligned to the process suggested in the traditional journalism research (Karlsson, 2011). .

The first and foremost step in the framework was to build a consolidated database. A human journalist may conduct an interview, search government record, create a poll to collect public's opinion, and look up previous or historical records to find similarities in interpreting the current news data to write a better news story. Similarly, the algorithm framework collected input data from various sources and constructed a consolidated database to maximize the capacity of algorithmic computation, and therefore better news could be made. The major objective of algorithmic activities in this stage was set to secure as many sources of data for input and program algorithm software to properly extract meaningful information from various types input data in non-unified format.

The expanded use of algorithms requires other sources of input data that the model does not include. However, the major focus of our suggestion on the framework construction was to take diachronous and synchronous data into account for generating news products. The use of diachronous and synchronous data would help algorithmic computation in generating quality news content. If a short and straightforward text is required, then the database might not need to include data in these groups, but for any other cases, a consolidated database would grow to include multiple layers of databases in a structured way.

Diachronous data help to find patterns and trends that might arise in a current event, and synchronous data can help to make rules for anomaly detection and changing perspectives of interpreting feed data. Any direct and indirect engagements with news readers help algorithm to refine the output with more context. It could be a direct input, such as the direct manipulation of the interface of the service and the IP address for location-specific information retrieval. Also, the habitual patterns, the profile of a user, and trending data are important indirect data to consider in tailoring the context for the construction of a consolidated dataset.

In the processing stage, an interim narrative structure was generated by applying statistical analysis and rule-based processing techniques to interpret the events with the classified narrative frame. We proposed a complex weight matrix to automate decision-making process by allowing the framework to mix and match various events for weight computation for more natural and relevant framing determination. Through this complex narrative framing process, news readers could be told and offered different layers of stories

depending on their levels of interest and their choices (Hamilton and Turner, 2009).

The narrative proto-structure was then turned into a news content or an interactive news service by generating narratives using natural language generation algorithm and adding an interactive layer if requested from the news services or media platforms of news readers. The algorithm framework was programmed to generate narrative content using either rule-based language generation or dynamic sentence generation method. While these methods had advantages and limitations in handling undefined situations or generating witty expressions, we used both methods that are applied different parts of narrative to maximize the ability of algorithmic narrative generation.

When presenting news, there could be multiple strategies to turn proto-structure news content into a more optimized news product. We defined proto-structure news as the interim content that is created by algorithmic computation. Since the final output could be presented in any format, a proto-structure needed an additional algorithmic process other than to generate news in plain text form. Much like Flipboard's layout algorithm called Duplo[4], which alters the layout of content dynamically to meet the varying size and layout requirements of the device of the user, a dynamic news presentation algorithm might take the content from the proto-structure to the next phase.

In computer science, the term 'budget' for an optimization algorithm refers to allocating computing resources efficiently for optimized computing results (Chen et al., 2000). If more budget is allocated, the result of the algorithmic computation (of optimization algorithms) might become much

[4]https://techcrunch.com/2014/03/23/layout-in-flipboard-for-web-and-windows/

greater. If small budget is allowed, then the available information space also diminish. In practice, a small device, such as a smart-watch, only allows the smallest budget, so only the most relevant information was selected through computation to fit on the screen. The algorithm for budget selection, therefore, might well determine how information would be shaped and presented to news readers, and it would also affect the kinds of interaction style and method the news product must incorporate.

Furthermore, the output of algorithmic computation does not need to be fixed in the current linear structure, nor even as an article on the web. As Lark and Quartz's news service has partially demonstrated, the information delivered by such systems can potentially be more useful by actively learning and adjusting the type and level of news content. Also, algorithms may utilize current location information, and the type of smart devices news readers may have for news generation. Baseball fans, for example, can freely ask and explore different parts of the game events and make personal highlights of the game by including their data points (such as favorite players or MVP of the match). The conversation can happen as a news clip or voice readout on their wrist using a smartwatch app. The information, therefore, can be shared in the way we casually share conversation with friends who have common interests and viewpoints, and the news evolves from newspaper-like articles to smart news services.

When all the suggested implications are considered, we believe the algorithm framework will be capable of providing more appealing and engaging news reading experiences. We expected the algorithm framework to handle real-time requests in scale, and able to offer personalized and

interactive news reading experiences for each news reader. In previous research, by allowing people to dive into data and directly manipulate data with their perspective, the content was perceived to be more credible and reliable (Dietvorst et al., 2015b). Data exploration and algorithmic transparency would enhance the accountability function of journalism. Transparency could also be a useful lever to bring to bear on algorithmic power when there is sufficient motive on the part of the algorithm's creator to disclose information and reduce information asymmetry (Diakopoulos, 2014).

In the following chapter, we examined the research questions by conducting two studies. The first part of the study was to implement the algorithm framework that fully functions in a selected domain. The much details on how we made the software code and the deliverables from each stage of the framework were included. The second part of the study was to explore the design space of algorithm-driven news generation system. We created a working prototype to test the concept of the algorithm framework and conducted a user evaluation to examine perceived news values of algorithm-generated news and news readers' experience of using the news system to generate their version of news stories.

Chapter 6

Design and Evaluation of the PINGS: Personalized and Interactive News Generation System

In Chapter 5, we suggested a new conceptual algorithm framework for news generation, which defines algorithmic activities happening at all stages of journalistic processes as well as the implications for designing an algorithm system.

In Chapter 6, we described the process of how a fully-functioning algorithm framework for baseball news generation is built. We presented PINGS, which stands for personalized and interactive news generation system. We designed and implemented PINGS that collects data from various sources, allows its users to interact with the system in generating the narrative, and changes the tone of narrative between an objective article and personalized story mode. We then reported on the evaluation study on the design and results of the evaluation. We concluded this chapter with the implications and discussions emerged from the results of the evaluation.

6.1 Overview

The results of this research consist of three parts, (1) the development of proposed algorithm framework for news generation, (2) the design and

```
# PSEUDOCODE
PROGRAM ConstructGameDatabase Algorithm:
    DEFINE data schema
    GET event data using crawl() method
        EXTRACT match-related data
        CREATE match event database
    UPDATE internal data structure
        COMPUTE performance records of teams and players in real-time
        UPDATE game, players, season, series, team and 5 other databases


PROGRAM ProtoStructureGenerator Algorithm:
    LIST all match-related events in the internal view
    SET weight on events using ComplexWeightMatrix()
        ASSIGN static weight on each event using pre-defined rules
        COMPUTE dynamically varying importance using statistical analysis
    LIST key events in the internal view
    GENERATE narrative proto-structure
        APPLY themes to assemble proto-structure
        DETECT mood for event interpretation


PROGRAM NewsPresentation Algorithm:
    GENERATE news stories
        GENERATE news title
            FILTER expressions based on mood and theme selection
            GET pre-defined template expressions for title
            ASSEMBLE expressions into title sentence using natural language generation rules
        GENERATE news body
            APPLY templates to select expressions
            GENERATE expressions using natural language generation rules
            ASSEMBLE expressions into text phrases by adding conjunctions, postpositional particles
    VISUALIZE narrative proto-structure
        VISUALIZE match event into timeline-based visualization
        GET match and historical data sets from databases
        UPDATE news story dynamically upon user interactions
        UPDATE databases and ComplexWeightMatrix() with the analysis of user selections and preferences
```

Figure 18: Top level pseudocode on how PINGS operates

implementation of algorithmic news generation system, and (3) the evaluation of the system in terms of perceived quality of news content and the experience of system usage. Figure 18 illustrates the overall structure of this interactive news generation system that we designed and implemented on top of the algorithm framework. The psuedocode depicts how algorithm takes data from various sources, converts data into a narrative form, and presents the news content to readers in an interactive and visualized format.

The algorithm framework is designed to handle data from various news sources. However, we are going to implement and design an algorithmic news generation system in sports domain, specifically baseball news. The reason we chose baseball as an exemplary case of algorithmic news generation is that there are a wide variety of data available for baseball including text

broadcasting messages from Internet media companies, diachronous data such as historical records of match results and accumulated player records, and synchronous data such as news from other media or other game events.

Also, baseball fans are used to tell their favorite teams and players explicitly, which makes it easier to tailor information to the wants and needs of news readers when developing an interactive prototype for the proof-of-concept. We will discuss the lessons learned and implications for designing an algorithmic news generation system in broader news domains in the discussion chapter. In this chapter, we will focus on demonstrating how we implemented the algorithm framework and interactive news system.

The goal of the first part of this research is to implement a fully-functioning algorithm framework. We designed the framework to construct a consolidated database that collects data from various sources. As stated in the previous chapter, collecting diachronous and synchronous data for computation allows the algorithm to maximize its potential to generate news products that count large scale of data into the computational process that is beyond the capacity of any human journalist.

Unlike a human journalist, an algorithm cannot conduct an interview directly with players or coaches, and it has limitations to inferring the whole out of a part. To overcome these limitations, the framework needs to be built to nurture algorithm's strengths that it is a fast, reliable, and customizable data gathering and analyzing agent. The data collection, therefore, is designed to collect as many data sources as possible to maximize the capacity of algorithmic computation. Since the proposed algorithm framework is primarily applied to generate baseball news, we started the study by learning what

attributes of algorithmic dimensions that can be applied to this specific domain and securing available data sources are.

The second part of the study is to explore the design space of PINGS, where the algorithm framework developed in the first part becomes the backbone of the entire system. The underlying algorithmic activities are intended to follow the journalistic processes for content generation. We designed the user interface to explore how consolidated database are applied as interactive interface elements, such as to offer historical match results between the two baseball teams or visualized the flow of gameplay that changes perspective upon the selection on which team to see. The important aspect of the second study is to examine how end users would use the system to generate their version of baseball stories and how much they would appreciate the elements of the interface in doing desired tasks.

For the third study, we conducted an evaluation study of the system to explore how would news readers proactively generate news articles when the proper tool is given and their experience of using the system. The implications and discussion on both studies are also reported in the latter part of this chapter.

6.2 Underlying Framework Development

The implementation of the news system follows the process of algorithm framework (Figure 14). Each stage needs to be implemented in the set order since the deliverable from each stage becomes the source of data for the algorithmic computation on the following stage. Consolidated database is

constructed as the output of the process depicted in data synthesizing stage, and narrative framing stage utilizes the database to build a proto-structure of the narrative after interpreting database with the mood determined by algorithmic computation. The proto-structure is iteratively updated as the system learns the behavior of users in using the system interface, and finally, it presents the news information in the form that best suits the context of the users. In the following section, we describe the details on how we implemented the algorithm framework for baseball news generation.

Table 4: The operationalization of key concept for PINGS

	Algorithm Framework	Operationalization
Data Synthesizing Stage	To construct database with diachronous and synchronous of data as an input to the algorithmic computation.	Diachronous data refers to historical datasets that add depth to the algorithmic computation, which includes the performance records of players for multiple seasons, match history between rival teams, and news database.
		Synchronous data refers to other types of data that add breadth to the algorithmic computation, and it includes other teams' match result, performance rankings for teams and players, and news articles from other media.
Narrative Framing Stage	To frame the narrative content in a context via explicit and implicit preferences.	Explicit preferences refer to direct user input made through the user interface such as click on the name of players, favorite moments, and the tone of narrative that is interchangable through the button.
		Implicit preferences refer to indirect data such as behaviroal patterns and log data collected via the news system. Since PINGS is only an interactive prototype at this stage, implicit preferences were not gathered and utilized in this research.
News Presentation Stage	To present news information within a user interface for customized news reading and visualized storytelling.	The types of user interface elements for PINGS include timeline-based visualization of batter events, list of players for each team, and narrative content that changes upon the mode selection.
		Interactivity functions implemented for PINGS include hover highlight, animated balloons for pop-up information, click on buttons, etc.

6.2.1 Data Synthesizing Stage

In general, a sports news about the result of a match deals with what has happened in the game between two teams. For a baseball game, a game begins with a pitcher throwing a ball to a catcher, and a batter tries to make a hit before the pitcher takes three strikes. If the batter makes a single, then the batter advances to the first base and the second batter walks into the batter box. A team scores when a runner on a base comes back to the home base. The team with more scores wins the game, and the game ends when the leading team takes 27 out counts from the other team. The algorithm framework needs to collect the series of data generated as these events are happening in time.

Table 4 depicts how we designed the whole consolidated database that refers to and brings data from various related data sets. When a match begins, a series of text broadcasting messages come in as events. Each event contains a short description of play-by-play information on whether a batter made a hit, scored, or strikeout by a pitcher. The framework keeps track of who is the batter and pitcher at every event, and therefore it can update the records of players and teams in real-time as well as to pull request to search for any abnormal records for newsworthy events. For instance, if a cleanup hitter made a home run, which made him take the first place in the number of home runs in this season, then the system looks up and saves the current status to other relational databases such as batter records, teams, and seasons databases, and send relevant data points to 'articles' database for writing a news on it. In designing the schema for the consolidated database,

we heavily made use of diachronous and synchronous databases such as match events in other baseball parks and the updated seasonal information regarding the overall team and player rankings.

Baseball game events can be collected by crawling text-broadcasting messages, which is often serviced by Internet sports media websites. The messages are sent in the form of JSON file format, which is commonly used when sending and receiving text messages on the web. The implementation of the algorithm framework began with designing programming codes that collected JSON file from an Internet sports portal (Figure 19). This JSON file contains play-by-play data, which are series of descriptions of every action made by players in the game and are recorded in real-time. Since all the significant events are recorded in this file, the algorithm framework can analyze the file and make automated judgments on constructing the match details. Figure 20 shows the list of events extracted from JSON data after cleaning and filtering cluttered data.

The framework not only collects the data for the match event but also gathers other related data such as information about batters and pitchers, team statistics, and seasonal information. In other words, building a software architecture that includes diachronous and synchronous data is the primary focus of the implementation. For instance, the batting average is a calculation of the number of hits divided by the number of appearances the batter made at the plate. It is based on the analysis of accumulated performance records over time. The batting average can also be calculated to show the batter's relative strength to a specific team or pitchers of the opponent team. It means the algorithm framework can utilize this performance indicator to interpret

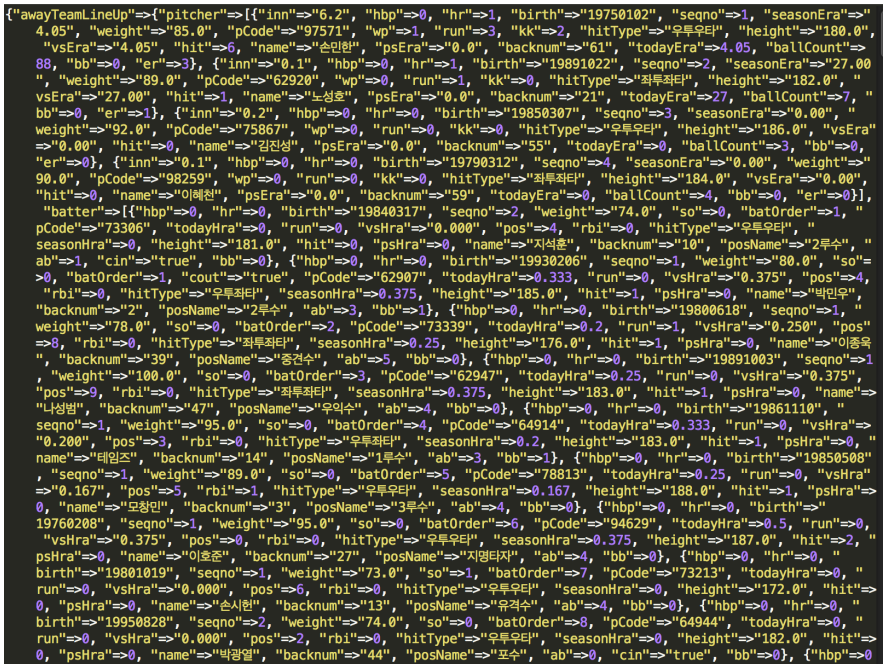


Figure 19: Snippet of text broadcasting messages crawled in JSON file format

\way:Home/Inn/Out/b1_/b2_/	b3_/WE/	WPA/	Pitcher/Batter/Hit___/	sq/Live text/
₹산 0 : 0 LG 1초 0死 - - -	0.500	0.000	임찬규 박건우 -	0{:etc=>"경기시작", :situation=>[3, 2]}
₹산 0 : 0 LG 1초 1死 - - -	0.522	▼0.022	임찬규 박건우 -	9{:etc=>"1회초 두산공격", :batter=>:gb_out, :out=>{:batter=>[:gb_out, "1루수"]},
₹산 0 : 0 LG 1초 2死 - - -	0.538	▼0.016	임찬규 오재원 -	17{:batter=>:gb_out, :out=>{:batter=>[:gb_out, "2루수"]}, :situation=>[1, 1]}
₹산 0 : 0 LG 1초 3死 - - -	0.548	▼0.010	임찬규 오재일 -	23{:batter=>:so, :out=>{:batter=>[:so]}, :situation=>[1, 2]}
₹산 0 : 0 LG 1말 1死 - - -	0.526	▲0.022	보우덴 김용의 -	30{:etc=>"1회말 LG공격", :batter=>:fly_out, :out=>{:batter=>[:fly_out, "좌익수"]},
₹산 0 : 0 LG 1말 2死 - - -	0.510	▲0.016	보우덴 이천웅 -	35{:batter=>:gb_out, :out=>{:batter=>[:gb_out, "2루수"]}, :situation=>[1, 1]}
₹산 0 : 0 LG 1말 3死 - - -	0.500	▲0.010	보우덴 정성훈 -	43{:batter=>:so, :out=>{:batter=>[:so]}, :situation=>[1, 2]}
₹산 0 : 0 LG 2초 1死 - - -	0.523	▼0.023	임찬규 김재환 -	49{:etc=>"2회초 두산공격", :batter=>:gb_out, :out=>{:batter=>[:gb_out, "2루수"]},
₹산 0 : 0 LG 2초 2死 - - -	0.540	▼0.017	임찬규 양의지 -	54{:batter=>:gb_out, :out=>{:batter=>[:gb_out, "3루수"]}, :situation=>[1, 1]}
₹산 0 : 0 LG 2초 2死 민병헌 - -	0.528	▲0.012	임찬규 민병헌 1루타(0)	58{:batter=>:hit, :B=>[[:hit]], :situation=>[1, 2]}
₹산 0 : 0 LG 2초 2死 허경민 민병헌 -	0.506	▲0.022	임찬규 허경민 1루타(0)	65{:batter=>:hit, :B=>[[:hit]], :runner_run=>[[1133, 1, 2]], :situation=>[2, 2]}
₹산 0 : 0 LG 2초 2死 정수빈 허경민 민병헌	0.475	▲0.031	임찬규 정수빈 -	74{:batter=>:bb, :B=>[[:bb]], :runner_run=>[[1131, 1, 2], [1133, 2, 3]], :situation=
₹산 2 : 0 LG 2초 3死 - - -	0.336	▲0.139	임찬규 김재호 2루타(2)	83{:batter=>:hit, :runner_run=>[[1132, 1, 3]], :runner_home=>[[1131, 2], [1132, 3
₹산 2 : 0 LG 2말 1死 - - -	0.311	▲0.025	보우덴 서상우 -	90{:etc=>"2회말 LG공격", :batter=>:so, :out=>{:batter=>[:so]}, :situation=>[1, 0]}
₹산 2 : 0 LG 2말 2死 - - -	0.294	▲0.017	보우덴 양석환 -	97{:batter=>:so, :out=>{:batter=>[:so]}, :situation=>[1, 1]}
₹산 2 : 0 LG 2말 2死 이병규 - -	0.307	▼0.013	보우덴 이병규 -	101{:batter=>:er, :B=>[[:er, "유격수", 1142]], :situation=>[1, 2]}
₹산 2 : 0 LG 2말 3死 - - -	0.284	▲0.023	보우덴 유강남 -	105{:batter=>:fly_out, :out=>{:batter=>[:fly_out, "좌익수"]}, :situation=>[2, 2]}
₹산 2 : 0 LG 3초 0死 - 박건우 -	0.236	▲0.048	임찬규 박건우 2루타(0)	109{:etc=>"3회초 두산공격", :batter=>:hit, :B=>[[:hit]], :situation=>[1, 0]}
₹산 2 : 0 LG 3초 1死 - - 박건우	0.244	▼0.008	임찬규 오재원 -	115{:batter=>:gb_out, :runner_run=>[[1138, 2, 3]], :out=>{:batter=>[:gb_out, "1루
₹산 2 : 0 LG 3초 1死 오재일 - 박건우	0.232	▲0.012	임찬규 오재일 -	121{:batter=>:bb, :B=>[[:bb]], :situation=>[5, 1]}
₹산 3 : 0 LG 3초 1死 - 김재환 오재일	0.151	▲0.081	임찬규 김재환 2루타(1)	128{:batter=>:hit, :runner_run=>[[1134, 1, 3]], :runner_home=>[[1138, 3]], :R=>[[:h
₹산 4 : 0 LG 3초 1死 양의지 - 김재환	0.109	▲0.042	임찬규 양의지 1루타(1)	135{:batter=>:hit, :runner_run=>[[1361, 2, 3]], :runner_home=>[[1134, 3]], :R=>[[:h
₹산 4 : 0 LG 3초 2死 양의지 - 김재환	0.137	▼0.028	임찬규 민병헌 -	140{:batter=>:fly_out, :out=>{:batter=>[:fly_out, "중견수"]}, :situation=>[2, 1]}
₹산 4 : 0 LG 3초 2死 허경민 양의지 김재환	0.129	▲0.008	임찬규 허경민 -	149{:batter=>:bb, :B=>[[:bb]], :runner_run=>[[1136, 1, 2]], :situation=>[2, 2]}
₹산 4 : 0 LG 3초 3死 - - -	0.158	▼0.029	임찬규 정수빈 -	155{:batter=>:fly_foul, :out=>{:batter=>[:fly_foul, "1루수"]}, :situation=>[2, 2]}

Figure 20: At-bat events extracted from JSON file

data with various diachronous perspectives.

In baseball, each team competes with other teams to be placed higher in the ranking, as well as players compete with others for individual performance rankings. Therefore, more thorough report on a game result may include information of how other teams and players performed, especially when one's favorite team's and players' records are at stake. For instance, if one's favorite player is placed in the first place of the number of home runs made in this year, then the news reader would appreciate if the batting records of other sluggers in other games are stated for comparison.

6.2.2 Narrative Framing Stage

After the construction of a consolidated database, the algorithm framework processes the database to frame the narrative. Framing a narrative means the algorithm can set an angle to interpret the whole event, which is equivalent to how a human journalist establishes the tone of an argument in writing news content.

In baseball, many sabermetrics-based statistical methods help algorithm to make an automated interpretation on game events. As stated in the previous chapter, sabermetrics uses historical game data to analyze and predict the incoming data in real-time. It enables the algorithm to make machine-driven predictions on the winning probability depending on the scores of each team and inning information, the relative importance of each at-bat event, etc. On the other hand, the interactions made by users throughout the system enhance the weighting mechanism with the explicit preference update, which is worth to mention in this section. When an interaction is made, the system

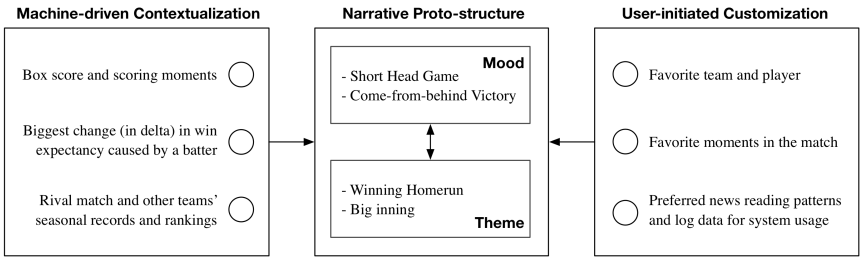


Figure 21: A use case of how a narrative proto-structure is made, when the match was determined by a winning home run after a close game

determines if the action is about directed preference update. The system calls the complex weight matrix to prioritize, re-select, and apply a new set of rules for a narrative proto-structure generation. Figure 21 depicts how PINGS generates a proto-structure of news content based on both machine-driven contextualization and user-initiated customization.

We used a method called 'win expectancy' that is used to calculate the probability for a team to win the match. The table 5 shows the chance for a team to win the game depending on the following conditions: (1) which inning the game is at, (2) the different in scores, (3) if any or all bases are loaded with runners, and (3) the out counts. For instance, the probability of a home team to win the game when the game just started is about 54%, since the history of baseball match results showed that home teams have slightly higher chance of winning the games at their ballparks. When a home team is leading the match by 4:3 in the 7th inning with no runners on base and no out counts are made yet, the probability of winning the game is 79%. If a batter makes a single run and the home team leads the match by 2 points (5:3), then the percentage jumps to 89%. The home team is now ready to

Table 5: Sample of win expectancy table that shows the probability of winning the match for the home team

Inn-Base-Out	-10	-5	-3	-1	0	1	3	5	10
7110	0.001	0.031	0.102	0.304	0.500	0.696	0.897	0.969	0.999
7120	0.001	0.027	0.089	0.267	0.441	0.626	0.856	0.953	0.998
7130	0.001	0.023	0.077	0.235	0.391	0.578	0.839	0.948	0.998
7140	0.001	0.021	0.070	0.212	0.353	0.520	0.783	0.923	0.996
7150	0.001	0.019	0.065	0.202	0.340	0.530	0.825	0.944	0.998
7160	0.000	0.016	0.057	0.177	0.298	0.469	0.766	0.917	0.996
7170	0.000	0.015	0.051	0.161	0.272	0.426	0.739	0.907	0.996
7180	0.000	0.014	0.047	0.147	0.249	0.391	0.683	0.874	0.993
7111	0.001	0.033	0.111	0.328	0.539	0.740	0.921	0.978	0.999
7121	0.001	0.030	0.102	0.301	0.496	0.690	0.894	0.968	0.999
7131	0.001	0.028	0.094	0.281	0.464	0.659	0.883	0.965	0.999
7141	0.001	0.026	0.089	0.265	0.439	0.622	0.850	0.952	0.998
7151	0.001	0.023	0.080	0.242	0.403	0.604	0.868	0.961	0.999
7161	0.001	0.022	0.075	0.230	0.382	0.569	0.833	0.947	0.998
7171	0.001	0.020	0.069	0.211	0.353	0.527	0.807	0.937	0.997
7181	0.001	0.020	0.068	0.206	0.344	0.511	0.776	0.920	0.996

win the game. However, if the away team makes a score in the very next inning, the percentage for the home team to win the game drops to 75%.

The percentage changes after each batting event, whether the batter makes a hit, home run, or strike-out. The most significant advance in using this performance metric is that it enables real-time calculation on the winning percentage. Using Tom Tango's win expectancy chart[1], we have adjusted the detailed statistics of historical records of Major League Baseball leagues to meet the status of Korean baseball league. In implementing the framework, we used this edited version of win expectancy for more accurate computation.

In addition to this statistical approach, we also created themes to determine

[1] ftp://ftp.baseballgraphs.com/wpa

Table 6: Sample themes for baseball news content generation

대분류	중분류	소분류	상세	WPA값을 이용 연산	Pr
키플레이	승리팀투수	퍼펙트게임		투수 WPA 합산 + 10	1
키이벤트	승리팀투수	노히트노런	선발투수의 노히트노런	투수 WPA 합산 + 10	1
키이벤트	공격팀타자	끝내기안타	끝내기안타	단일이벤트 WPA	2
키이벤트	공격팀타자	끝내기홈런	끝내기홈런	단일이벤트 WPA	2
키이벤트	공격팀타자	멀티안타	한 선수가 3안타 이상	안타를 친 타석들의 WPA 합산	3
키이벤트	공격팀타자	멀티홈런	한 선수가 2홈런 이상	홈런을 친 타석들의 WPA 합산	3
키플레이	승리팀타자	팀홈런	팀이 4홈런 이상 기록	홈런을 친 타석들의 WPA 합산	4
키플레이	승리팀타자	빅이닝	하프 이닝에서 4득점 이상을 기록	하프 이닝 내 모든 타석 WPA합산	5
키플레이	패배팀투수	빅이닝허용	하프 이닝에서 4실점 이상을 기록	이닝 내 투수들의 WPA 합산	6
키플레이	승리팀투수	팀완봉(영봉)	투수진이 경기를 무실점으로 마무리	투수들의 WPA 합산	6
키이벤트	수비팀야수	결승실책	야수의 실책으로 인한 실점이 결승점	단일이벤트 WPA	7

the mood of the match event. For example, if a team wins the game by preventing the opposing team from achieving a single hit, then the game is called ‘No-hitter’ game. For baseball fans, knowing the game ended with no-hit and no-run is the single most important thing to know. Also, if three or four batters make home runs in raw, then these rare batter events are worth to report. Table 6 is some of the sample themes we developed to determine the most important event of the game. We analyzed news articles from various news media to classify these themes they use in reporting baseball news.

In order to improve the ability of algorithm framework to generate news that is more appropriate and intelligent, we applied the complex weight matrix that uses both machine-detectable weights from sabermetrics and

rule-based theme classification method as a way to compute dynamic narrative proto-structure. In baseball news, a score by a single home run is usually worth to mention when reporting the game summary. However, the win expectancy metrics helps to find relatively more important scores or home runs such as the winning run of a tight match.

Moreover, the narrative frame becomes more comprehensive when the pre-made rule detects this winning run, and the algorithm framework can be set to highlight the plays made by the person who made this winning score. Complex weight matrix, therefore, enhances the quality of decisions made by the algorithm framework and helps to write more natural and engaging news stories as a writing from a specialist in baseball.

After a proto-structure is made, the system determines if the algorithm needs to generate text for news stories or to render a visualization of data components onto the user interface. It is what happens in the presentation stage, and we are going to introduce how the algorithm framework presents news content back to users in the following section.

6.2.3 News Presentation Stage

Once the narrative proto-structure is generated, the algorithm framework at news presentation stage converts the proto-structure into news content that meets the form and layout of the news outlet, which can be in the form of traditional newspaper-style news article or an interactive news service. First and foremost, the primary focus of algorithmic computation in this stage is to generate stories that meet the expectation of news readers.

The algorithm framework generates narratives in two ways: template-

Table 7: Templates for generating randomized sentences by inserting metadata

대	중	소	상세	문장	meta1	meta2
T1	승리팀	역전승	선취점 내주고 승리	#(승리팀은/는) #(패배팀과/와)의 경기에서 역전승을 거두었다.	#(승리팀)	#(패배팀)
T1	승리팀	끝내기	경기 종료 직전 기록	#(승리팀은/는) #(끝내기로/으로) 승리했다.	#(승리팀)	#(끝내기)
T1	승리팀	공멸	실책이 전체 실점의 50% 이상	#(승리팀은/는) #(패배팀과/와) 경기에서 #(양팀실책)을 합작하며 수준 이하의 경기력을 보였다	#(승리팀)	#(패배팀)
T1	승리팀	엑스트라이닝	9회까지 득점이 없거나 동점	#(승리팀과/와) #(패배팀)의 경기는 연장전으로 넘어갔다.	#(패배팀)	#(승리팀)
T1	패배팀	석패	단순점수차 1점	#(패배팀은/는) #(승리팀과/와) 경기에서 #(점수) 아깝게 패배했다.	#(패배팀)	#(승리팀)
T1	홈팀	무승부	점수차이 없음	#(홈팀은/는) #(어웨이팀과/와) 경기에서 #(점수) 무승부를 했다.	#(홈팀)	#(어웨이팀)

based processing and dynamic sentence generation. The first method, template-based processing, is to construct a narrative using multiple pre-made templates. A template, in this case, is a full sentence with some vocabularies marked as data points that can change the theme selected by using complex weight matrix. For example, a sentence might be "#(name_of_batter) hit a #(type_of_hit) in the match between #(home_team_name) and #(away_team_name), which made #(win_team) win the match." Multiple different sentence can be generated using this template (refer to figure 7 for more templates). For the World Series match between Chicago Cubs and Cleveland Indians, the catcher David Ross made the winning run by hitting a home run. In this case, the template can generate "David Ross hit a single home run in the match between Chicago Cubs and Cleveland Indians, which made Cubs win the

game."

This template can be used repetitively as long as the winning run is made by a home run, and is the most important thing to report in the news about this match. The news about the match between Cardinals and Pirates, for instance, can have the following sentence: "Jedd Gyorko hit a double in the match between St. Louis Cardinals and Pittsburgh Pirates, which made Cardinals win the match." The framework is capable of generating different narratives as long as matching templates are ready to handle the themes classified by using complex weight matrix. If there are more than one template to handle the condition set by theme and sabermetrics, then the framework can randomize the selection process to generate different narratives every time the framework is executed. The more templates that the framework has, the richer the narrative becomes.

The major disadvantage of template-driven narrative generation is that templates are required to be created for every possible combination of events beforehand. For example, a sentence about how one team has won the match against the other team needs to distinguish situations that are similar but slightly different. Even when the inning team has won the game by scoring 8 points, there might be more than one way to describe it: all batters scored a run, one batter scored all runs for the team, three batters made consecutive home runs in one inning, two batters made gland slams, etc. In order to generate narratives that include specific details that matter to baseball fans, it requires templates to cover a wide range of possible combinations of events beforehand, but it may take too much, and probably unnecessary in most cases, efforts in preparing all possible combinations of sentences, to begin

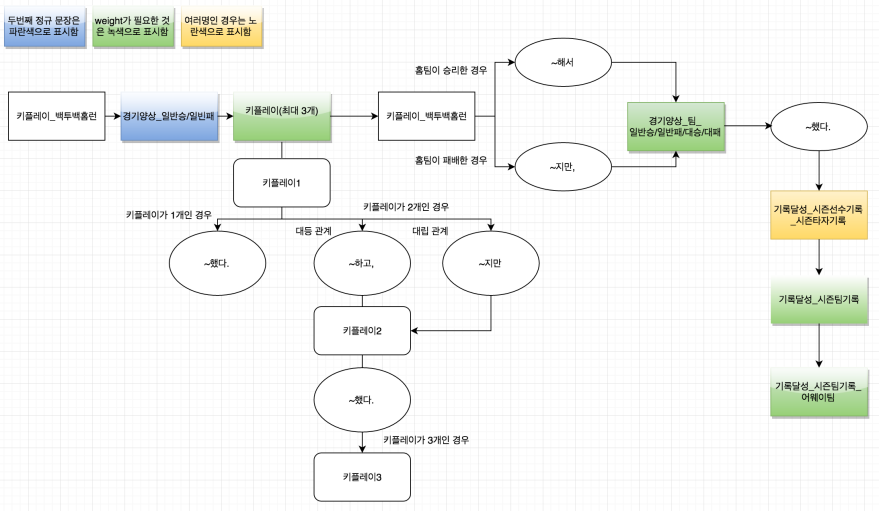


Figure 22: Varying cases of conjunctions depending on the selected theme and the weight of events after complex weight matrix

with.

The other method we implemented in the development of the framework, dynamic sentence generation, provides an alternative solution to the limitations of using templates in news generation. Dynamic sentence generation uses natural language generation algorithm to dynamically assemble lexicons such as names of players and teams, baseball terminologies, and other general subject and proper nouns for richer expression. Generating natural language using an algorithm system requires taking a series of steps for computation, which includes determining what to mention and how to make the overall structure, making lexical choices and aggregating lexicons and sentences for more natural reading, and realizing grammatically correct sentences as the deliverable of the system (Reiter et al., 2000).

In the Korean language, making lexical choices and aggregating lexicons

Table 8: Lexical and conjunction choices for sentence aggregation

Conjunction	Sentence	Weight1	Sentence
~해서/~로써/~로/~으로	경기양상-팀-결승타	100	팀플레이-팀타자-득점상황
~해서/~로써/~로/~으로	경기양상-팀-대승	100	팀플레이-팀타자-득점선수
~해서/~로써/~로/~으로	경기양상-팀-일반승	100	기록달성-연승,위닝시리즈,스윕
~해서/~로써/~로/~으로	경기양상-팀-일반승	100	기록달성-연패탈출
~해서/~로써/~로/~으로	경기양상-팀-일반승	100	기록달성-순위상승, 순위확정
~해서/~로써/~로/~으로	경기양상-팀-일반승	100	기록달성-시즌승리기록
~덕분에/~로써/~으로	기록달성-연승	100	기록달성-순위상승,순위확정,시즌기록
~덕분에/~로써/~으로	기록달성-위닝시리즈	100	기록달성-연승,순위상승,순위확정,일반
~덕분에/~로써/~으로	기록달성-스윕	100	기록달성-연승,순위상승,순위확정,일반
~덕분에/~로써/~으로	기록달성-연패탈출	100	기록달성-순위상승,순위확정,일반순위
~덕분에/~로써/~으로	기록달성-연패탈출	100	팀플레이-팀전반-득점상황
~덕분에/~로써/~으로	기록달성-연패탈출	100	선수플레이-선수타자-타격, 희생타, 번트
~덕분에/~로써/~으로	기록달성-순위상승	100	선수플레이-선수타자-타격, 희생타, 번트
~덕분에/~로써/~으로	기록달성-순위확정	100	선수플레이-선수타자-타격, 희생타, 번트

were particularly challenging due to relatively less research about Korean language, ambiguous words, and grammatical difficulties in compounding consonants and vowels. The most difficult part in generating natural language in Korean is to add postpositional particles to subject words, which changes the ending character of the previous word, and aggregating lexicons with proper conjunctions that are determined after theme selection and weighing events (Figure 22). In order to generate sentences that are different, we created 30 different type of dynamic sentence generation methods to handle various situations that are worth to distinguish. The sample sentence structure is shown in table 8, which illustrates how we designed the aggregation process for the algorithm framework, which is specifically intended to handle the winning scores made by multiple home runs.

Dynamic sentence generation works well in situations where it is almost impossible to prescribe the algorithmic activities in advance. Since it aggregates lexicons by calculating relative importance of words with complex weight matrix and the selected themes, it is more flexible and scalable to handle various situations than using pre-made templates for all incidents. However, dynamic sentence generation method also has limitations. It is less attractive that it just list words one by one and is less creative in expressing emotions that it just aggregates lexicons. We use both methods in generating narrative in the algorithm framework. The title and the introduction part of the narrative are generated using templates to emphasize the selected theme. More appealing and witty sentences are positioned early in the sentence to engage more with news readers. The middle part, or the detailed explanation of what happened in the process of scoring, we used dynamic sentence generation to cover a wide variety of situations.

The algorithm framework may publish the narrative generated using natural language generation algorithm, or it may add an interactive layer on top of the news content. The framework needs to have a pre-defined style of layout and interaction depending on the domains of news and the requirements set by the media platform where news is presented. For baseball news, interactivity layer can engage with news readers by customizing the news content that highlights certain parts of the game or remarkable play records of their favorite players. The design space of how baseball news service will be explored in the following section.

In the following section, we explored how a news system can be designed and implemented for interactive news reading, and how would news readers

evaluate the system regarding perceived news values and user experience on reading personalized and interactive news using the proposed prototype of algorithm news system. We are going to evaluate a news system that is designed to make news reading as more personalized and interactive experience by extending the capability of algorithmic computation in news generation.

6.3 Design and Implementation of PINGS

In this section, we describe the system in terms of its user interface components and the underlying algorithmic computations for automatic news story generation. We designed a system called PINGS, which adds a user-friendly interface on top of an algorithm-driven news generation process. PINGS allows users to hear more about specific moments in the game and learn about the play records of favorite players. It even changes the tone of the narrative to focus on emotionally engaging moments to provide a personalized and interactive way to consume information about sporting events (Figure 23).

6.3.1 Design Goals

We designed PINGS as a general-purpose news generation system. However, we developed an interactive prototype that deals with sports news, especially baseball news, for the study. We chose baseball as the exploratory domain for automatic news generation for multiple reasons. First, statistical methods such as ‘win expectancy’ and ‘leverage index’ from sabermetrics

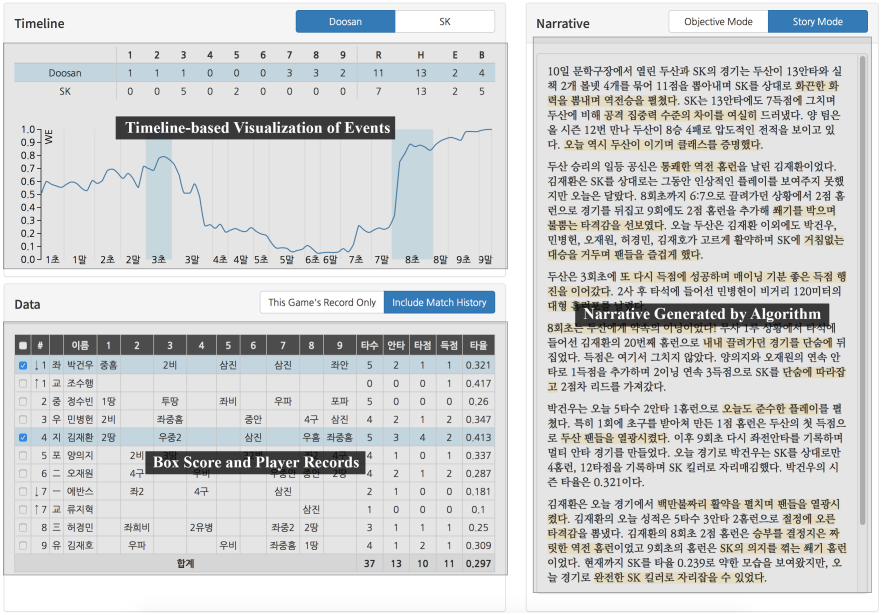


Figure 23: The user interface of PINGS, (1) Timeline Pane: Visualizes the winning expectancy of the selected team in chronological order, (2) Data Pane: Lists the name of batters for the selected team and their play for each inning, (3) Narrative Pane: Displays the new story generated by users' interaction on the timeline and data pane elements.

help to reconstruct the entire game's records into a series of events on a timeline by computing the relative importance of each record. Also, it is easier to let users specify personal interests in baseball, such as the list of favorite teams and players. Conducting an in-lab experiment is more convenient with a past baseball game.

PINGS collects play-by-play streaming data from an online sports news service in real-time, extracting information about batters from the play logs. As stated in the previous section, we built an algorithm-driven news framework that takes the raw text data, determines relatively more important events

among the entire play records, sets a mood to frame the narratives based on the key events, and generates text phrases based on the key events and mood detected in the previous stages (김동환 and 이준환, 2015).

The major design goal of this study is to explore how personalized news stories can be made by allowing users to engage in manipulating variables offered by the system actively. The system begins by automatically creating an overall summary of the game: win/loss information, information about the most valuable player in the game, and each team's updated rank in their league. The power of PINGS comes from the user interface components shown in the 'Timeline' and 'Data' panes (Figure 23), where the timeline visualization provides an easy to understand flow of the game at a glance and the list of players in Data pane shows the overall summary of stats each player recorded throughout the game.

The system takes the interactions of users into account to adjust the scores it uses internally to compute the importance of events. When a user specifies personal interest by selecting a few names of players and moments in the game, the relative importance of those data selections become larger. More personalized narratives can be automatically created as a consequence of such user interactions. Users can generate personalized news stories with PINGS in the following ways:

- Add or remove the summary of plays made at each top and bottom inning throughout the game by interacting with the timeline-based visualization of entire events

- Add or remove detailed performance description made at bat by selecting

one or more players from the roaster list

- Choose to include player-level match history details when adding specific records of selected players to the narrative
- Change the viewpoint of the news story between an objective and storytelling mode, which changes the tone of narrative dramatically

6.3.2 System Design

We have developed the system using a Ruby on Rails framework, which crawls the live streaming data from a sports news service in JSON format and generates narratives in natural language using the algorithm solution introduced in our previous research. The user interface components are built using React.js and D3.js JavaScript libraries to support real-time manipulation and reflection upon data selection and to apply animated effects for feedback. The overall design and structure of the interface are configured using Twitter's Bootstrap framework.

PINGS is designed to give richer details on various moments of the match. It also reports on the performance of favorite players depending upon selections on the user interface components. By default, the system generates a brief summary of the game in an objective narrative tone. Once the number of innings and players are selected, the narrative changes to include new details responding to the user's selections. For example, if the top of the 7th inning is selected from the 'Timeline' pane and the cleanup hitter is selected from the 'Data' pane, then the narrative is expanded to include details on how batters and pitchers performed in the selected inning

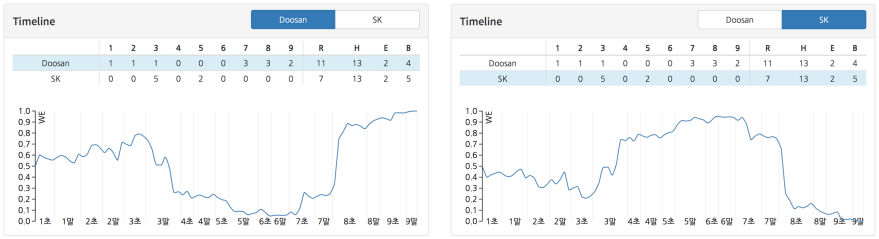


Figure 24: Selecting a team changes the chronological visualization of probability to win the game

and how the cleanup hitter played throughout the game. This user interfaces architecture provides myriad options for its users to generate narratives that meet their interests and needs.

Also, the option to change perspective to the more emotionally-engaging story mode adds an opportunity to create content that counts users' mental model of how the game went. If a user is disappointed with the result of the match, particularly when some players performed significantly below average, then the narrative can be customized to express criticism about the selected players. The system, therefore, is programmed to generate narratives in the tone of either angry or overjoyed user when in story mode.

We implemented PINGS as an interactive prototype to explore how sports news readers would explore data to create a news story when a proper tool is given. Moreover, performing a user evaluation on the quality of content in the context of a system for personalized and interactive news has not been heavily explored in academia, and we hope this paper be a stepping stone for such research.

Timeline-based Visualization of Events

The event timeline visualizes the chance of winning the match for the selected team, based on the method introduced in the book about sabermetrics (Tango et al., 2007). The rise and fall of the line in the graph depicts real-time changes in the percentage of winning the game after each at-bat play. If a batter scores and extends a team's lead, then the expectancy of winning the game increases. In contrast, if a batter misses a chance to hit a run when all the bases are loaded, the winning expectancy drops. The graph ends with either 100% or 0%, which indicates the win or loss of the game for the selected team. This part of visualization includes the detailed view of the changes in the winning expectancy for both teams (Figure 24). By default, the visualization is set to show the graph of the home team. However, selecting the other team using the button provided in the upper right corner changes the visualization in the opposite way: the winning expectancy of the selected team.

The timeline visualization of chronological events is often introduced in research that deals with an interactive information visualization system, such as Twitter event visualization from TwitInfo system (Marcus et al., 2011). A timeline-based visualization is often used when the volume of events changes (y-axis) over time (x-axis). Diakopoulos et al., (2010) also proposed a graph that shows the overall volume changes in messages on the flow of time in their research to design a visual analytic tool called Vox Civitas. It aims to help journalists to utilize massive social media data by visualizing the change of interest and acts as an interactive user interface

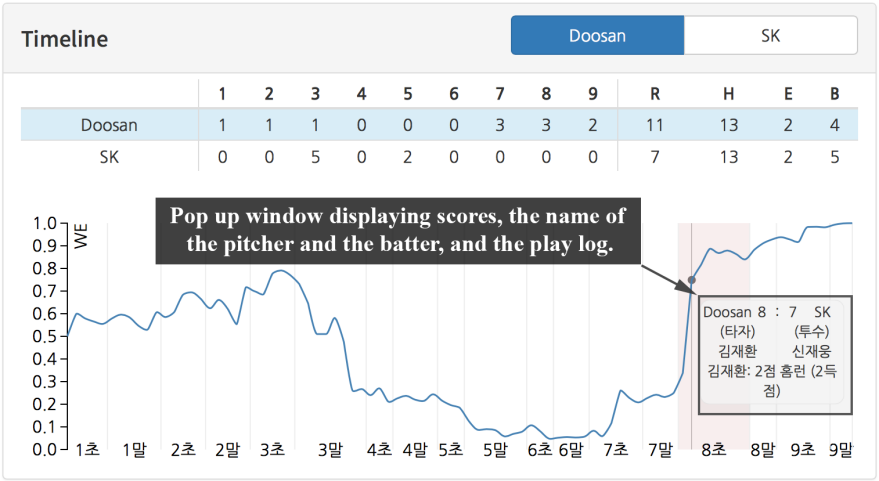
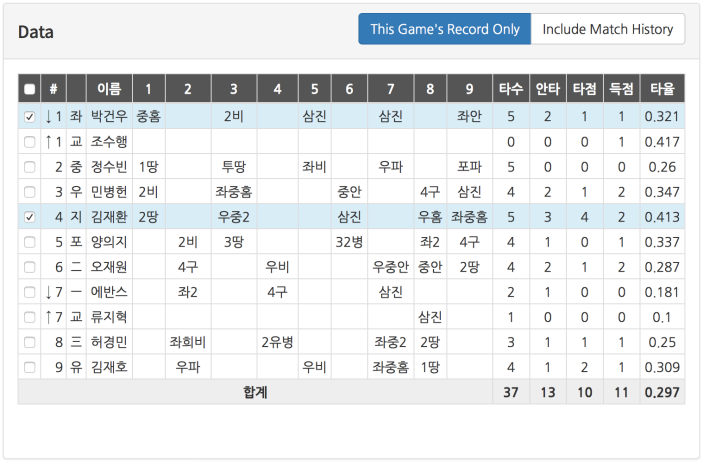


Figure 25: Hovering mouse pointer over any point on the line displays an overlaying pop up window that contains at-bat information at the selected time

element that clicking on the graph syncs the video and messages of the content component to that moment (Diakopoulos et al., 2010).

Similarly, the timeline visualization in PINGS depicts the probabilistic percentage of winning on the y-axis and the flow of time on the x-axis. We added another interactivity feature in the system, a mouse-over event, so that hovering the mouse over any point on the line shows basic at-bat information, such as the score for each team, the name of the batter and pitcher, and whether it is a hit or an out (Figure 25). When the mouse pointer is located in the timeline window, the top and bottom inning is automatically set on the focus. Hovering the mouse is an easy and useful interaction technique to gain a brief detail on a specific moment from the overall view.

Lastly, selecting an inning from the timeline adds the detailed description



	#		이름	1	2	3	4	5	6	7	8	9	타수	안타	타점	득점	타율
☑	↓1	좌	박건우	중홈		2비		삼진		삼진		좌안	5	2	1	1	0.321
☐	↑1	교	조수행										0	0	0	1	0.417
☐	2	중	정수빈	1땅		투땅		좌비		우파		포파	5	0	0	0	0.26
☐	3	우	민병헌	2비		좌중홈			중안		4구	삼진	4	2	1	2	0.347
☑	4	지	김재환	2땅		우중2			삼진		우홈	좌중홈	5	3	4	2	0.413
☐	5	포	양의지		2비	3땅			32병		좌2	4구	4	1	0	1	0.337
☐	6	二	오재원		4구		우비			우중안	중안	2땅	4	2	1	2	0.287
☐	↓7	一	에반스		좌2		4구			삼진			2	1	0	0	0.181
☐	↑7	교	류지혁								삼진		1	0	0	0	0.1
☐	8	三	허경민		좌희비		2유병			좌중2	2땅		3	1	1	1	0.25
☐	9	유	김재호		우파			우비		좌중홈	1땅		4	1	2	1	0.309
합계													37	13	10	11	0.297

Figure 26: The list of players and their play records for the match

of what happened in that inning to the narratives. This interaction is made in real-time: the selected inning becomes highlighted in blue, and a paragraph is inserted in between the existing text blocks in the narratives panel. The text, the detailed description of what happened in that inning, is inserted with an animation effect: pushing the existing paragraphs down and the new paragraph appears with a time delay of 0.5 seconds. The animation is added to inform users about the action they just made. Users can freely choose any moment in the game and the details on that specific moment of the game are added to the narrative of the game, which means they just customized a news story to meet their interest. With PINGS, adding personally interested plays into the news story is just a single click away.

Expanding Data Selection to Include Further Details

The data pane displays the list of batters for the selected team (Figure 26). The list includes the names of players, their positions, play records throughout the innings, the number of hits and runs in the game, and the hitting average. Comparing each player's at-bat plays and the number of hits and runs gives a sense of how well the player performed in this game. Users can select one or more players by clicking on the checkboxes to add a paragraph of their detailed play records into the narrative. The inserted paragraph is removed when the boxes are unchecked.

The data pane offers two options in generating the narrative: write about the events that happened in this game event only, or include match history records for each player against the opposing team. When a player is selected with 'this game's record only,' the added paragraph will only include the summary of the player's records for this specific event. For instance, when the first batter in the list is selected, then the narrative generated for this player is about two singles he made in this game.

However, when the match history button is selected, the paragraph will also include the number of hits and runs the player made against the opposing team during this entire season. For example, if the first batter has higher batting average when he was playing against the opposing team, then this player can be said to have made relatively worse performance than his average records. The background color changes to blue once players are selected, and it reverts back to the original color if deselected. As such, the narrative might expand to include a more detailed description when the

Figure 27: How PINGS generate narrative: (1) add performance details of selected players to the narrative, (2) add their records against the opposing team for this season

match history button is selected (Figure 27).

If the other team is selected, then the line-up of players also changes accordingly. The selection of the players is also reset, but the system remembers the selection once users choose the original team. The narrative, which is generated by selecting timeline and data points, would also reflect the changes on these panes in real-time.

Changing the Tone of Narrative

By default, PINGS creates a narrative story based on the home team's perspective. The story for the event is automatically generated in the narrative pane, which includes a brief summary of the result of the game, the most valuable player of the game, and the updated ranks for the teams in their league. This default story gives a brief summary of the game to provide base-level detail about the current match. Users are allowed to select various user interface components to add more information to the story.

According to the selections made on the timeline and data panes, text phrases are added to the narrative pane in real-time. Upon data selection, new paragraphs are inserted in the order of time (the flow of innings) and the order of the batter lineup. If the third batter is chosen after selecting the fifth batter in the list, then the paragraph for the third batter is inserted above the paragraph for the fifth batter. The newly added paragraph is brought into focus in the narrative pane (with an animated insertion effect) to raise attention to the changes made by their interactions.

Users can select the tone of the narrative's text by switching the buttons between objective and story mode. The story mode changes the tone of the

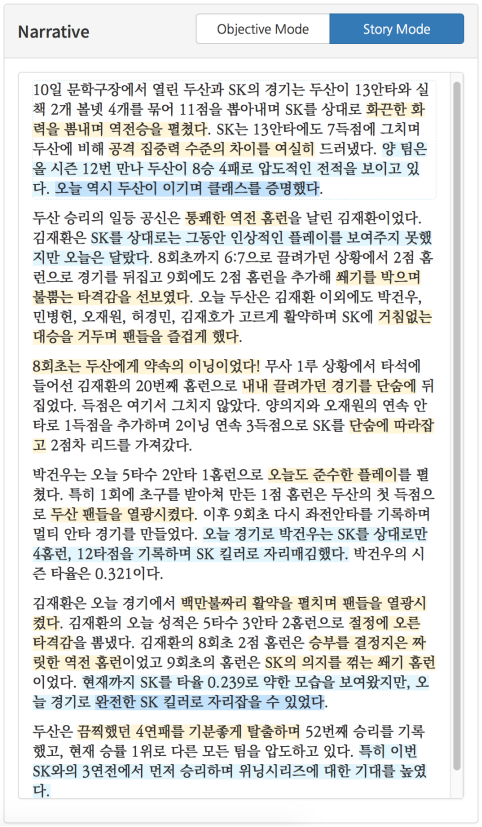


Figure 28: Narrative generated after selecting specific moments and players for personalized news story generation

text to include more emotionally engaging and detailed explanations. For example, if a player makes multiple strikeouts and keeps on missing chances to score, the paragraph in the story view describes the situation in a scolding or sarcastic manner, saying such things as 'the only thing the player knows is how to miss so many chances' or 'he must be embarrassed with his score, what a shame!' These expressions are highlighted in yellow so that users become aware of the changes in the tone (Figure 28).

6.3.3 Interactive Narrative Generation

We introduce how the narrative changes upon interactions news readers make on the user interface of PINGS. We selected the match between Doosan Bears and SK Wyverns as an exemplary case of news content generated by PINGS. The match resulted in a win for Doosan, and the key player of the game was Jaehwan Kim, who scored the winning points with a double home run on the 8th inning.

1. The most brief and objective narrative generation

PINGS is capable of generating narrative content in four different ways. First, PINGS can generate a straight and objective news article that follows the style of sports news from ordinary news media. The most objective and summative description of the match between Doosan Bears and SK Wyverns is:

> Doosan Bears took 11-7 win over Sk Wyverns at 2016 TireBank KBO League at Moonhak stadium on 10th. Doosan won the match by scoring 11 points with 13 hits, 4 base-on-balls, and 2 errors, while SK only scored 7 points. The key player was Jaehwan Kim. He made a double home run in the beginning of the 8th inning against Jaewoon Shin, SK pitcher. Doosan was able to turn the table from 6:7 to 8:7 with Jaehwan's double scores. Doosan added 3 more RBIs with timely hits, and widen the gap by 4 points. After the match, Doosan put an end to the losing streak and maintained the first place in the league with

the 52nd win (70% winning rate).

10일 문학구장에서 열린 2016 타이어뱅크 KBO리그 두산과 SK의 경기는 양팀이 최종 스코어 11:7로 두산이 승리를 가져갔다. 두산은 13안타 4볼넷 2에러를 더해 11점을 만들며, 7득점에 그친 SK에 승리를 거뒀다. 두산의 수훈 선수는 김재환이었다. 김재환은 6:7로 끌려가던 8회초 SK 신재웅을 상대로 투런 홈런을 날리며 8:7로 역전에 성공했다. 두산은 이 후에도 3점을 추가해 11:7로 스코어를 벌린 후 SK의 공세를 차단하며 오늘 경기를 극적으로 가져갔다. 오늘 경기의 결과 두산은 4연패를 탈출하며 시즌 52회째 승리를 기록했고, 현재 승률은 0.703으로 1위를 기록 중이다.

2. Narrative generated by adding multiple data points

There were some key moments in the match that most Doosan fans would love to highlight to celebrate their victory. The key and most exciting moment of the match was the 7th inning, where Doosan finally came-from-behind and took the lead by 2 points. The biggest threat was the 3rd inning that SK scored 5 points in a single inning, and the probability of winning the match for Doosan dropped 50%. When key moments are selected from the timeline-based visualization, the narrative expands to include the summative descriptions on what happened in the chosen innings (Figure 29):

At the bottom 3rd inning, SK made a big inning against Heekwan Yoo, the starting pitcher from Doosan. SK made their first score by back-to-back hits from Jung Choi and Euiyoon Jeong, and a

Figure 29: Some key moments in the match are selected from the timeline graph

double from Jungkwon Park. Jaewon Lee made a sacrifice hit, and Sunghyun Kim made a double home run. The pitcher Yoo handed 5 RBIs just in a single inning.

3회말 유희관은 SK에 빅이닝을 내주었다. 최정과 정의윤이 안타를 치고 나간 상황에서 박정권이 2루타를 치며 최정을 홈으로 불러들이며 오늘 첫 실점했다. 이어지는 무사 2, 3루 상황에서 이재원의 희생플라이와 최정민의 1타점 적시타로 동점을 허용했고, 김성현에게 2점 홈런까지 맞으며 3회말에만 5실점했다.

At the beginning of the 8th inning, Doosan finally turned the tables with Jaehwan Kim's double home run. Jaehwan walked into the batter box with the first base filled and made a home

run to the right outfield. Doosan took the lead by 8:7 for the first time in the game. Afterward, Doosan added 3 more points with timely hits from Jaewon Oh and another double home run from Jaehwan Kim on the 9th inning.

8회초 두산은 김재환의 2점 홈런에 힘입어 역전에 성공했다. 무사 1루 상황에서 타석에 들어선 김재환은 SK의 새로운 투수인 신재웅을 상대로 우익수 뒤 홈런을 기록하며 2점을 추가해 8:7로 역전했다. 이 후 두산은 양의지의 2루타와 오재원이 적시타까지 더하며 8회초에만 3득점하며 SK와의 점수차를 2점으로 벌렸다. 9회초 2사 후에 김재환이 또다시 2점 홈런을 날리며 2득점했다. 김재환의 연타석 홈런! 점수는 4점차로 벌어졌다.

3. Added depth to the narrative by adding historical match details

PINGS also support users to add descriptions on their favorite players or the ones who made impressive or disappointing plays. Moreover, the narrative can be made to include historical performances of selected players against the opposing team. We highlighted the sentence that was generated by computing the historical and personal records of the selected player:

In this season, Doosan is showing an overwhelming winning rate against SK, 67% of winning the matches. Jaehwan Kim scored 3 hits and two home runs and was selected as the key player of the match. Jaehwan made 4 out of 11 scores alone. **His performance against SK was less than impressive in this season with 0.239 batting average, but today he was a**

different man.

두산은 올 시즌 SK와의 경기에서 승률 67%로 압도적인 전적을 보이고 있다. 김재환은 오늘 5타수 3안타 2홈런을 기록하며 승리에 결정적으로 기여했다. 김재환은 3회초 우중간 2루타를 기록했으며 8회초와 9회초에서는 연타석 2점 홈런을 기록하며 두산의 11점 중 4점을 홀로 책임졌다. **현재까지 SK를 상대로 0.239의 타율로 약한 모습을 보여왔지만, 오늘은 이를 상회하는 경기력을 보여주었다.**

4. Fully personalized story generation

Lastly, We generated fully expanded and completely personalized news story. We highlighted expressions in the narrative that are generated using historical records and utilizing data from other databases. The narrative is constructed in the perspective of Doosan's fans, and it includes detailed play description happened in the 3rd and the 8th inning and two additional players who made contrasting plays (one made the winning score, and the other made disappointing plays). The full news story is illustrated in figure 30, and some text from the narrative were translated as below:

On 10th, Doosan Bears showed off an unending batting-power and crushed SK Wyverns by a thrilling come-from-behind victory with 13 hits, 2 errors, and 4 base-on-balls. SK made the same number of hits but only scored 7 points. **It looked like they all looked too boggled to stand up against Doosan's pitchers. Doosan, again, proved to be SK killer!**

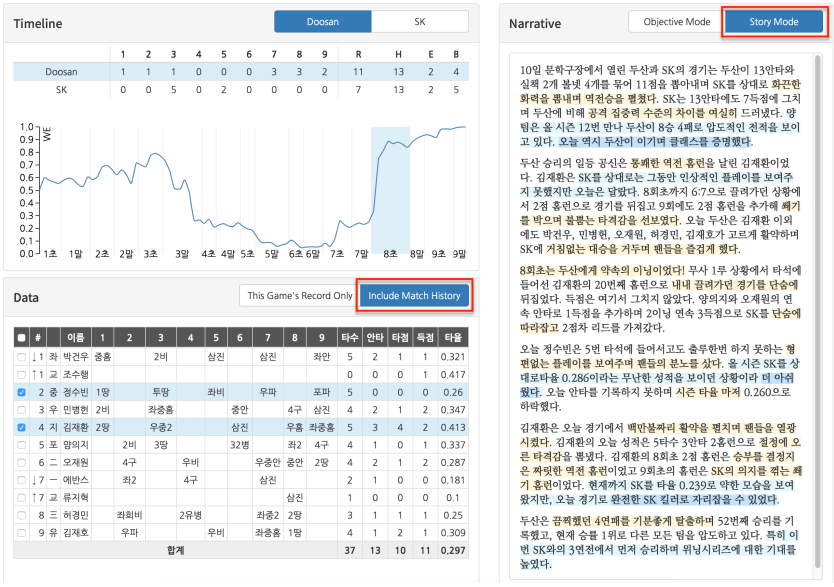


Figure 30: The narrative was fully personalized by adding various data points and historical match records, and changing the tone of narrative mode.

10일 문학구장에서 열린 두산과 SK의 경기는 두산이 13안타와 실책 2개 볼넷 4개를 묶어 11점을 뽑아내며 SK를 상대로 화끈한 화력을 뽐내며 역전승을 펼쳤다. SK는 13안타에도 7득점에 그치며 두산에 비해 공격 집중력 수준의 차이를 여실히 드러냈다. **양 팀은 올 시즌 12번 만나 두산이 8승 4패로 압도적인 전적을 보이고 있다. 오늘 역시 두산이 이기며 클래스를 증명했다.**

The MVP from 2015 Korean-Series, Soobin Jung, **is making a disappointing play in this season.** He entered the batter box for five times but failed to advance to the first base. Soobin's batting average against SK was 0.286 but is dropped to 0.260 after today's match.

오늘 정수빈은 5번 타석에 들어서고도 출루한번 하지 못하는 **형편없는 플레이를 보여주며 팬들의 분노를 샀다.** 올 시즌 SK를 상대로타율 0.286이라는 무난한 성적을 보이던 상황이라 더 아쉬웠다. 오늘 안타를 기록하지 못하며 시즌 타율 마저 0.260으로 하락했다.

Jaehwan Kim's performance against SK was less than impressive in this season, but today he was a different man. As Doosan took the first of the three consecutive match series, the fans of Doosan now expect to have a winning series against SK Wyverns.

김재환은 오늘 경기에서 백만불짜리 활약을 펼치며 팬들을 열광시켰다. 김재환의 오늘 성적은 5타수 3안타 2홈런으로 절정에 오른 타격감을 뽐냈다. 김재환의 8회초 2점 홈런은 승부를 결정지은 짜릿한 역전 홈런이었고 9회초의 홈런은 SK의 의지를 꺾는 쐐기 홈런이었다. 현재까지 SK를 타율 0.239로 약한 모습을 보여왔지만, **오늘 경기로 완전한 SK 킬러로 자리잡을 수 있었다.**

6.4 Evaluation of PINGS

In the previous sections, the design and implementation of personalized and interactive news generation system have been introduced. In this section, we investigated how the algorithmic news generation system and its output content were perceived by news readers and domain experts to address the last research question (RQ3), "How is the algorithmic news generation system

perceived by news readers and domain experts?" There was two important type of domain experts in our research: one was a group of experts who are familiar with the process of news production (e.g. journalists, editor), and the other group was the system experts who were involved in designing the user interface and interactions or developing the web or mobile service.

The major objective of this research was to measure the perceived quality of news content generated by the system as well as to evaluate the system itself. The experiment in this thesis consisted of three sessions:

1. The first study was to compare the perceived news values between algorithm-generated and journalist-generated news content. We also examined if knowing the source of news creator affects the evaluation score by adding another condition. The objective of the first study was to examine the effect of news source and source perception on the perceived news values.

2. The second study was to examine how news readers would evaluate personalized news content that was generated using the algorithm system we propose. We adopted the evaluation criteria from the first study to identify the types of news values that personalized news effect the most.

3. The third study was conducted by taking interviews with various user groups including ordinary news readers, news experts, and system experts. The objective of this last study was to explore how algorithmic news system is perceived regarding the system quality and experience of using the system for personalized and interactive news generation.

6.4.1 Method

We conducted mixed design studies for the evaluation of the system and content generated by the system. For the first and second study, which were to measure the perceived news values on the human and algorithm-generated news, and to compare it with the measurements of the news content generated by the algorithmic news generation system.

We designed a 2 (news source told: algorithm vs. journalist) x 2 (news given: algorithm vs. journalist) between-subject conditions to investigate if participants' perception of the source of news affects the evaluation score.

To provide background knowledge on how to read a baseball news article, we opened a match result and summary page from a sports portal web service (Figure 31). We asked participants to look at the summarized result of the game (such as box scores, player records, and play-by-play records) to become familiar with the information shown on the website for five minutes.

For the first study, we selected two baseball matches for selecting news article and the system development: matches between Doosan Bears vs. SK Wyverns and Hanhwa Eagles vs. Samsung Lions. For human-generated news, we selected a news article for each match events that was the most popular news from the online sports portal website. For algorithm-generated news, we selected news that was published on 'Baseball News Robot' page on Facebook[2].

We randomly assigned participants to conditions and handed in a printed

[2]https://www.facebook.com/kbaseballbot

Figure 31: The match summary page from an online baseball broadcasting service

news article to match with the experimental conditions: a human-written article given to one group and an algorithm-written article given to the other group. For each condition, contrary to what we told them, half of the participants were given a human-written article while the other half were given news written by an algorithm.

After taking the first study, we asked the study participants to perform guided tasks specifically chosen by the researchers in order to make them explore and learn about different parts of the system:

- Find the moments and players in the game that are most worthy of mention. What impressed you the most?
- Switch between ‘This game only’ and ‘Include match history’ modes. How well did players perform against the opposing team?

Table 9: Demographics (N = 116)

	Mean or % (N)	S.D.
Gender:		
male	70% (81)	
female	30% (35)	
Age:	23.9	2.52
Level of Interest:		
high	47% (55)	
low	53% (61)	

- Change the narrative mode between 'objective' and 'story' modes. How does the narrative content change?

The goal of these tasks was to let participants explore interactive elements in the interface to be familiar with the interface. After completing these tasks, we asked participants to create a news story by interacting with various user interface components. They have been invited to evaluate the perceived news values of the content generated by the system.

The second study was to conduct a within-subjects survey to assess the quality of personalized news generated by PINGS. Again, participants were given with the same survey questions, and we compared the result with the scores from the first evaluation study. The objective of the second study was to compare the difference in how participants perceive news values of personalized content over neutral content algorithm-generated.

6.4.2 Result

We recruited 116 participants for the in-lab experiment by posting a message on the university's web-board. We posted pre-survey questionnaire

and collected information about demographic variables including gender, age, and the level of interest in baseball news (Table 9). To calculate the degree of interest, we asked how often they watch baseball games and if they actively search for the match result. Participants who only watch and read the news a few times per year (less than two or three times per month) were regarded as less interested. The experiment took about 45 minutes for each participant, and we compensated them with $10 gift vouchers.

Effect of News Source and Source Perception

There have been a couple of prior studies, which compared how people measure the news values of algorithm-generated news against that of human journalists (Clerwall, 2014; 김영주 et al., 2015). The results from these research reported that the perceived news values were not significantly different, which meant that people treat algorithm-generated news as a legitimate source of news information. In Clerwall's study, the scores for the algorithm-generated news were higher in criteria related to credibility and trustworthy, while human-generated news was significantly more pleasant and less boring to read (Clerwall, 2014).

The overall structure of the first study was similar to the design of the previous studies, but we updated study conditions and evaluation criteria in our design. The former studies mostly focused on comparing the algorithm news as another type of text-based news article. However, we wanted to examine the perceived news values in a broader set of evaluation criteria, since the major objective of our study was to explore how to generate more valuable news stories that are personalized to the news readers and are read

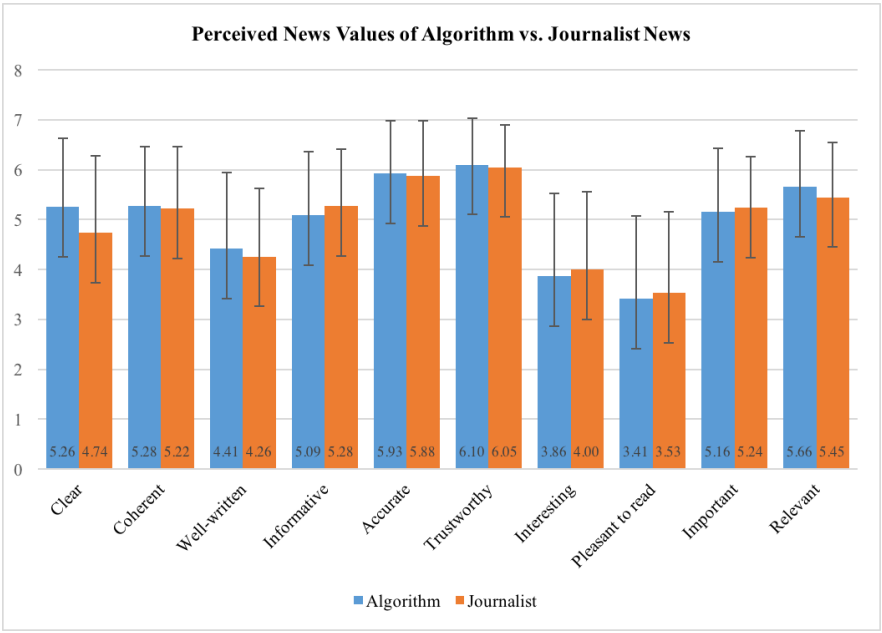


Figure 32: Perceived news values of algorithm vs. journalist news

with added interactivity.

To produce appropriate evaluation criteria for scoring, we reviewed the metrics found in previous studies on the perceived news value (Sundar, 1999; Clerwall, 2014). Since the goal of this thesis was to design and develop a personalized news system, we filtered out some of the evaluation criteria that do not match with the current system design, which includes objectivity, lively, timely, etc. We derived ten evaluation criteria with each measured between 1 and 7 Likert scale, and the criteria were clear, coherent, well-written, informative, accurate, trustworthy, interesting, pleasant to read, important, and relevant. This evaluation criteria used in the first study was also used in evaluating the personalized news content in the second study.

We conducted our first evaluation study, which was to evaluate a news

content that is generated by algorithm or human, with 116 participants in four different conditions. When we simply compared news values of an algorithm-generated news against the news from a journalist by the mean, we were able to find that the news values in these two conditions do not differ significantly (Figure 32).

To study the effects of all factorial conditions, we conducted 2 (news source told: algorithm vs. journalist) x 2 (news given: algorithm vs. journalist) analysis of covariance (ANCOVA) for each news values. The type of news became the dependent variable, and we set sex, age, and the level of engagement as covariate values. The result showed that there is no significant effect of news types on news values after controlling for the covariate variables (Table 10). Therefore, we were able to derive the conclusion that the news values do not differ in both algorithm-generated and journalist-generated conditions.

In sum, the result revealed that there were no significant main effects for the types of news sources. Unlike Clerwall's research result, what were known to be the strength of human in news writing, such as well-written, pleasant to read and more enjoyable (less boring) content creator, were scored no differently, or sometimes rated slightly less. Also, the scores for accurate and trustworthy contrasted with the scores for interesting and pleasant to read, which indicated that the participants generally thought both news are trustworthy but not as interesting to read.

Effect of Personalization

For the second study, we conducted another survey on the perceived news values of the news content generated by PINGS. The major objective

Table 10: Combined Table of Analysis of Co-Variance for News Values

	Algorithm-Algorithm Type		Algorithm-Journalist Type		Journalist-Algorithm Type		Journalist-Journalist Type		
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*F*
Clear	5.01	0.27	5.42	0.28	4.67	0.27	4.94	0.27	1.36
Coherent	5.35	0.24	5.32	0.24	5.41	0.23	5.06	0.24	0.41
Well-written	4.31	0.28	4.63	0.28	4.42	0.27	4.26	0.27	0.38
Informative	4.96	0.24	5.27	0.24	5.31	0.23	5.38	0.23	0.58
Accurate	6.10	0.21	5.81	0.21	5.85	0.21	5.81	0.21	0.41
Trustworthy	6.13	0.17	6.06	0.17	6.11	0.17	5.92	0.17	0.31
Interesting	3.86	0.31	3.95	0.31	4.10	0.30	4.10	0.31	0.15
Pleasant to read	3.41	0.32	3.47	0.33	3.70	0.32	3.50	0.32	0.16
Important	5.19	0.22	5.14	0.23	5.42	0.22	5.05	0.22	0.52
Relevant	5.89	0.22	5.46	0.22	5.41	0.21	5.59	0.22	0.97

* $p < .05$.

of this study was to examine what are the changes caused by personalizing news content especially when participants were actively engaged in generating the news with interactive user interface elements. We conducted paired t-test with the participants who evaluated algorithm-generated news in the first study. As displayed in Table 11, there were statistically significant differences, at .05 significant level, in the evaluation of personalized algorithm news for informative, trustworthy, interesting, pleasant to read, and important. We found no significant differences between the two news types for clear, coherent, accurate, and relevant.

Most notable differences were found between 'interesting' and 'pleasant to read' factors. In our study, interesting factor was scored much higher for personalized news ($M = 6.07, SD = 1.01$) than for objective algorithm-generated news ($M = 3.86, SD = 1.66$), $t(57) = 8.53, p < .001, d = 1.66$. The scores on pleasant to read were again much higher for personalized

Table 11: Descriptive statistics and t-test results for objective and personalized algorithm news

	Objective-Algorithm		Personalized-Algorithm			95% CI for Mean				
Outcome	*M*	*SD*	*M*	*SD*	*n*	Difference		*d*	*t*	*df*
Clear	5.26	1.37	5.07	1.18	58	-0.21	0.59	0.15	0.95	57
Coherent	5.28	1.18	5.05	1.44	58	-0.30	0.75	0.17	0.86	57
Well-written	4.41	1.53	5.03	1.08	58	-1.05	-0.19	-0.48	-2.9	57
Informative	5.09	1.27	5.93	0.75	58	-1.19	-0.50	-0.84	-4.97*	57
Accurate	5.93	1.06	5.91	0.88	58	-0.35	0.39	0.02	0.09	57
Trustworthy	6.10	0.93	5.64	1.09	58	0.15	0.78	0.46	2.98*	57
Interesting	3.86	1.66	6.07	1.01	58	-2.73	-1.69	-1.66	-8.53*	57
Pleasant to read	3.41	1.65	5.98	1.03	58	-3.05	-2.08	-1.91	-10.6*	57
Important	5.16	1.27	5.81	0.69	58	-1.03	-0.28	-0.67	-3.51*	57
Relevant	5.66	1.13	5.81	0.83	58	-0.52	0.21	-0.16	-0.8	57

* p < .05.

news ($M = 5.98, SD = 1.03$) than for objective algorithm-generated news ($M3.41, SD = 1.65$), $t(57) = 10.6, p < .001, d = 1.91$. In the first study, where we compared the scores for journalist-generated news against algorithm-generated news, there were no significant differences between the two groups. Rather, the mean value of the scores for journalist-generated news was slightly higher. The overall enjoyment of reading a news content was found to be greater in the personalized algorithm-generated news.

When the algorithm followed the writing style of a human journalist, news readers did not show any preference on its output content. However, when PINGS generated news, participants found the personalized content and interactive storytelling news much more interesting and pleasant to read. In other words, when algorithm learned to create more emotionally rich and

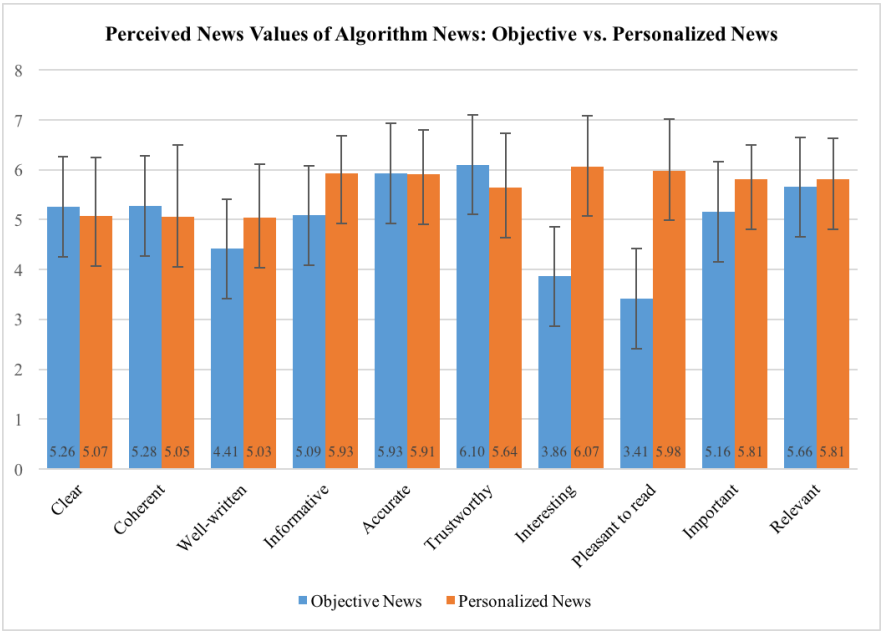


Figure 33: Comparison between personalized algorithm news and plain-objective algorithm news

tailored stories, it became an engaging storyteller that is capable of creating much more appealing stories to its readers.

Moreover, PINGS-generated news was scored significantly higher on two other criteria, 'informative' and 'important' source of news information. For informative, the differences of scores were higher in personalized news ($M = 5.93, SD = 0.75$), $t(57) = 4.97, p < .001, d = 0.84$ than algorithm-generated news ($M = 5.09, SD = 1.27$). For importance of news content, personalized news ($M = 5.81, SD = 0.69$) was scored higher than for algorithm-generated news ($M = 5.16, SD = 1.13$), $t(57) = 3.51, p < .001, d = 0.67$. From these results, we were able to address the added value of showing historical match result when introducing a summary of players' performances

in the narratives. When we asked if consolidating a wide variety of data to support the narrative is necessary for raising the news values, most survey participants left comments that they enjoyed reading more relevant data when a new piece of the narrative was added. *"Knowing more about the players I added to the narrative helped me to understand more"* (P102), *"It will especially help the ones who need to catch up with previous games"* (P109)

Another notable result from the evaluation was that participants perceived PINGS news as a less trustable source of news information than when algorithm followed the writing style of a human journalist. The scores on trustworthy factor were lower for personalized news ($M = 5.64, SD = 1.09$) than for algorithm-generated news($M = 6.10, SD = 0.93$), $t(57) = 2.98, p < .005, d = 0.46$. The result reinforced what participants mentioned during the session. Some mentioned that the personalized narrative was written more like a casual blog or social media posts shared among the fans of their favorite team rather than a typical news article. *"It is still news, but not for everyone."* (P6)

In sum, evaluating news values of personalized news from the algorithmic news generation system revealed the strength of having personalization in news writing. By making the content more personalized and interactively made, we were able to generate more interesting, pleasant to read news, informative, and important news content than typical sports news found in today's media. The point that it is a less trustworthy source of news information may have been caused by the fact that the content was heavily highlighted stories from one side and was different from a typical news

content. However, the personalized news generation system showed it is capable of generating news that news readers would enjoy and engage.

Evaluation on Interactivity

The third part of the experiment was to conduct a general evaluation of the user experience of PINGS and the potential value and limitations of the current system. In order to collect diverse views, we recruited three different user groups for the interviews: news readers, news experts, and system experts. For news readers, we recruited 26 university students who read sports news regularly. For news experts, we recruited five journalists for their professional journalistic opinions on the quality of the system and its output. For system experts, we conducted interviews with three user experience and user interface designers and a platform programmer for their experience in using the system for news generation.

The focus of interviews with news experts was to take their perspectives on the significance and news values of the content generated by the system. In general, journalists believed that PINGS would be potentially competitive news provider in the domains where the stories are made based on data. In terms of news writing, the most difficult part of their job was to compute and check numbers and historical records in writing data-centric stories in a short time. This system utilized a wide variety of data with speed and accuracy, which made it especially useful tool for journalists. However, they raised issues that this kind of system might not be appropriate for the news domains where real-time or open-access data are scarce. Also, the use of right and non-provocative vocabularies and expressions are critical in

politics and community news, but the way the system writes a story seemed to need an extra editing stage since these kinds of decisions are difficult for a machine.

From the system experts, we were able to gather their professional opinions on the usability of system in terms of manipulating user interface controls and their experience of using the system for news generation. First of all, all experts seemed to find the system intuitive to use. They found visualized user interface elements such as the timeline graph was very useful since the visualization itself was an exceptionally good tool for storytelling. The downside of the system design was that the UI controls, for the most part, were not flexible enough to support various use cases. Instead of listing players in the batting order in data selection panel, the system might have been more usable if it allowed to filter and rearrange data upon other indexes and to edit the narrative content automatically generated by the system. With such inflexibility, it might not satisfy heavy news readers. Also, it would not always be useful for having more data, but it would be more important to have qualified data for enhanced user experience with the system.

6.4.3 Implications

We were able to derive a number of implications that we learnt from news readers and experts, and listed the following implications from the interviews:

What Becomes News is Personalized

We compared the types of events included in a news story generated by participants using the system to list the data points they mentioned must be included in a news article on the game event (before the system was given). When simply reading game summaries and play-by-play records, participants mostly took notice of the moment their team started to take the lead (19 out of 26 participants) and points scored throughout the game (20 participants). The types of events they selected mostly corresponded to the list of events an algorithm would also select.

However, we observed more diverse generation patterns when PINGS was given to them. Participants appreciated various user interface components and often selected data points that would make the article two to three times longer than a typical news report. Many of them changed the tone of the narrative to a more excited or scolding tone to emphasize specific events of interests: *"I want to add text that criticizes those who severely lowered our win rate, and text to cheer those who made great plays" (P6)*. We made an open-ended system for narrative generation, and participants responded with various versions of news stories, each describing the game from a different perspective.

Individual's Perceived Importance Weighs Heavier

A typical algorithm-driven system is optimized to identify statistically important events. However, we found PINGS helped to derive personal context from users, who were able to adjust the scoring matrix based on their explicitly

expressed interests and preferences. We observed some participants intentionally avoided selecting events that were objectively important but included all of the build-up events that took place prior to the major event: *"Although they took the lead in the 3rd inning, we wanted to add the scores made in the first and second innings to describe how they built up the lead" (P9).* The result of the survey on the perceived news values of PINGS news also sheds light on how the system allows users to generate more interesting and readable stories. One of the main conclusions from the evaluation of PINGS is that an algorithm-driven news generation system must prompt users to explicitly express their priority of events about how they would use the user interface.

How Data is Presented Shapes the Whole Story

Regardless of participants' level of interest in and knowledge of baseball games, we found they were able to understand the overall flow of the game much easier and faster with the help of various data points we provided in the system. Some participants noted that the timeline visualization was especially helpful:

> *"This timeline helps me understand the whole game at a glance." (P26)*
>
> *"I can see when things were looking positive or negative by looking at these spikes. I found the moment I started to win the game. I am definitely adding this part!" (P5)*

Our original intention of using a timeline graph was to emphasize data-driven analysis of the game events, such as using stiff curves to represent critical moments. However, we found that timeline visualization was also helpful in revealing the small changes in the game that contributed toward victory. We found the method in which data was provided influenced users not only in understanding what was happening in the match but also in developing their perspectives on how to interpret the whole event in their stories.

Guide Novice and Maximize Controls for Experienced

We observed differences in how much the user interface components were fully utilized between avid baseball news readers and those with less knowledge in this subject domain. Novice participants took more time and were less aggressive in pouring data points into the news story. One of the fewer experience participants complained that the system is difficult to start with at first: *"It's difficult to begin: this kind of user interface is unfamiliar to me" (P4)*. Many recommended that we provide more preset options, which would allow them to generate stories more easily: *"I wish it summarizes all the 'important plays', which would make story generation easier" (P12)*.

However, more interested baseball news readers demanded interface controls that could describe the game event in more detail:

> *"It doesn't show the teams' seasonal performance information." (P11)*
>
> *"I wish I could add the challenge call made by our manager."*

(P13)

Perceptions of the Algorithm Alters Usage Patterns

A participant with less trust in algorithm was found to spend more time reviewing the interpretation of data and nitpicking on the details of descriptions made by the system: *"Let me check if the data is good ... See? I found an error" (P16)*. Unfortunately, the system happened to provide wrong details about one of the batter events, and this participant heavily criticized the algorithm as an active news creator. This pattern aligns with the findings from the research on the perception of algorithms, where people lost trust when they found errors in an algorithm (Dietvorst et al., 2015a). In contrast, some participants had strong trust in the computational capacity of machines: *"It (the stats) must be correct, it is a computer" (P20)*. We found those with higher trust were more experimental in story generation without questioning much on the underlying algorithmic computation.

On the other hand, we found participants were more generous about awkward interpretations in the narratives. Most of them gave positive feedback about the algorithm's ability to create emotionally engaging stories. In our analysis, participants were more forgiving of small errors in the writing, which can be interpreted as differences in personal preference. However, their tolerances were affected by their internal attitude toward technology. Participants with higher tolerance tended to spend less time in checking the numbers but jumped straight into the investigation on algorithm's capacity to create more creative and human-like expressions.

Narrative Text is Rather an Interactive Element Than a Final Output

PINGS is an automated news generation system that can adjust the level of details in a story by taking users' active engagement into account. Although PINGS is still a prototype, participants were excited to be given a system that can generate personalized news stories using various data points. One particular insight we had was that users do not recognize the text generated from the system as the 'final output.' Rather, they perceived the text as an element of the interface they could further interact with. They were excited to see the narrative change in real-time when a data point was selected and tried to find a better combination of text phrases by actively exploring data points that matched their preferences.

Many participants tried to edit or rearrange the text phrases generated in the narrative pane: *"I don't like how it starts the story. I want to cut to the chase, and begin the 9th inning." (P17), "I want to insert a 'commentator's comment' in between these two paragraphs. It would help the fans of other teams." (P24)* We found that those who follow baseball news closely do not take the narrative generated by the system as the final output, but as an interim product that can also be further manipulated.

Moreover, most of the dissatisfaction with the system was based on how it did not support direct manipulation of the text itself. It is not just about the matter of functional deficiency, but their desire to use the narrative as a medium through which to express their opinion of the game when communicating with others. For this reason, we believe PINGS showed

great potential to become a content-management system (CMS) for sports news readers, a tool that gives all the required details for gaining knowledge of the news, and as a vehicle for sharing narratives with others who would be interested in their personal interpretation of the event.

6.5 Discussion

In Chapter 6, we designed and developed an algorithmic news generation system with the underlying algorithm framework, and examined the quality and experience of the system in the perspectives of news readers, news and system experts. We discussed the issues we found throughout the process in terms of the key concept proposed in developing the algorithm framework, which was (1) to expand the source and variety of input data for enhanced algorithmic computation, (2) to frame the news content with personalized context and viewpoints, and (3) to present news as a service for interactive and self-driven storytelling.

We chose baseball as the exemplary case of algorithmic news generation system in this research for multiple reasons: (1) data in a baseball match is recorded and broadcasted that are open for crawling, (2) baseball data are transcribed in a structured way and terminologies follow baseball conventions, (3) a wide variety of datasets are available such as decades of performance records, players' and teams' historical information, user-generated data on social media, and other news from media, and (4) sophisticated baseball statistics such as sabermetrics.

However, there were also limitations in constructing a consolidated

database. Some data were not captured: interviews with players and coaches, unrecorded mistakes made by fielders, or players' personal affairs that might have affected their performance. Some diachronous data were not available: historical records of other leagues (MLB, NPB). Some synchronous data such as weather were not analyzed statistically due to the shortage of research resources. The algorithmic capacity, therefore, was limited to the report made during or after the game as the summary of performance records. Even for baseball news, more research on various types of news and content exploration need to be made for future studies.

For the study on the effect of personalization, we presented news readers with a news system that had various user interface elements for exploring various data points. Most news readers found the system was capable of generating much more interesting and enjoyable news stories. What we found from the study of the effect of personalization were that the system was appreciated for generating more engaging and pleasing news reading experience than just reading news for the masses. Reading news stories that narratives are added upon user interaction and changed the tone of voice to meet their personal preferences seemed to enhance the overall news reading experience.

The downside of the result was that we only observed how news readers behave and answer the questions we asked for the purpose of research. Since the system we built worked as an interactive prototype rather than a fully working news service for everyone, we were not able to collect the log data upon their real system usage nor update the content based on their personal reading patterns. In a research perspective, we were able to observe

how sports news readers would respond and play with an interactive news system, but the ability for our system to reflect personalized preferences and interests was limited in making machine-driven personalization into the news generation process.

For the issues in the interactivity, we found the participants particularly enjoyed the presence of timeline visualization for many reasons throughout the evaluation study. Some enjoyed the visualization that it conveys the overview of what has happened throughout the game and visually highlight critical moments in the match even without looking into the actual data. The participants, in general, perceived steep curves made in the graph are the key moments of the match event that would worth to look into in the narrative. Also, they appreciated visualization that it did not only provide the overview of the entire events in time but also provided detailed information as an overlay window when hovering the mouse over a specific moment in the game.

However, we were able to observe sports fans wanted more flexible and customizable user interface controls. For instance, the timeline visualization of the winning expectancy was designed to serve only the selected team's perspective of the game. We predicted that the fans of Doosan Bears would like to look at the graph regarding their team's ups and downs, and read narratives that deal with the performance summary of their team's batters and pitchers. We did found many participants liked the controls, but we also heard the voices from participants that they wanted to expand the data selection to include the play records of the opposing team to get into more details on how their team won (or lost) the game. It was also evident in the

data exploration, where they wanted to zoom in and out of various moments to meet their perception of personalized news stories.

The participants with higher baseball knowledge especially liked the interactivity function that brings up the additional information. They used this feature as a way to forecast how the narrative will be framed and to confirm the system's credibility whether the algorithm had correctly interpreted the scene in the narrative. Participants constantly explored how narrative contents are made and can be altered to reflect better on their mental model of the news story. Some of the functions they requested were missing from the prototype: they wanted to change the order of paragraphs, add a comment to or change words in a sentence to adjust the level of emotional expression, etc. In other words, the narrative generated from the system can be presented as a tool or user interface element that is further manipulated for deeper personalization.

The user interface we prepared was limited in providing the personalized customizing options. However, it was also important to provide user interface controls that are usable for everyone. The user experience would not go along with more controls. There was a conflict from the system design point of view on whether to match with the needs of heavy information seekers by providing more controllable user interface elements or to provide data for quick and easy generation of news stories. What are the kinds of user interface controls and the depth of details that make a good personalized news generation system need to be further investigated.

In this research, we only employed interactivity functions that helped news readers to customize the narrative content generated by the system. In

other words, we only explored person to machine interactions. However, we could expand the scope of research to include person to person interactivity. Adopting more communicative or conversational interactivity features for exchanging news information would be a good topic for a follow-up study.

Chapter 7

Discussion for Algorithmic News Generation

We explored how to design an algorithm framework for personalized news generation and evaluated the news system built upon the underlying framework. In this chapter, we listed the following points of discussion for an algorithmic approach to news generation. Also, we listed contributions we made to the field of journalism and HCI research and limitations of the current research.

7.1 Discussion

One of the most frequent questions we receive from journalists is whether an algorithmic system would replace human beings in the journalism industry. From the review of the current practices and technological advances made by algorithm companies, we found that algorithms are increasingly involved in news generation process. In the following sections, we are going to review and discuss the expanded role of the algorithm in journalistic processes mediated by algorithmic activities.

Algorithm for Journalistic Processes

In the case of Korean' news media industry, there is a rising concern about the expanded role of the algorithm in news generation process. As

more algorithms run on the servers of news media, we are facing an exponentially growing number of news articles. The quality, on the other hand, of algorithm- generated news are still at summative and straightforward data report, and the overly generated news articles are regarded as spams for the most cases. In Korea, most of the news articles are distributed to readers via big web portals, Naver and Kakao with more than 80% of entire traffic in the news business, and therefore the problem is worse for Korean news readers in terms of finding news with fine quality.

The content generated by PINGS are different from that of the traditional news business, and we believe that PINGS will be a complementary tool for journalistic activities performed by human journalists. PINGS can potentially provide a direction for algorithmic news generation process in many ways. First, PINGS generated news content not by automatically assembling premade texts and conditions but using personalized data points and customized data selections.

PINGS can generate a tailored message using the context determined by news readers, and such messages were found to be more effective at engaging and persuading the audiences (Rimer and Kreuter, 2006; Hawkins et al., 2008; Roberto et al., 2009). The content generated by PINGS is unique regarding how the content are generated and targeted to meet the needs of each specific news reader. Also, providing interactive user interface elements to news readers invited news readers to participate in modifying the form and content. Since this machine-human interactivity functions offer more controls and functions on the reception side of information, PINGS is capable of providing news stories that evolve as it automatically collects and

updates its database with an ongoing stream of live data.

Second, traditional news media services could make use of PINGS as a system that further customizes the news article generated through their processes. Collecting the explicit and implicit user engagements would be a big plus for journalistic investigation on their output products. According to the journalists we interviewed, we found that the system was perceived as a potentially useful tool for the professionals in the news industry in the following ways.

The machine generated news is best suited as the first story for breaking news. In the digital platform, the most important news values for the first story are fast and accurate reporting. The digital content drives more traffic and user engagements, such as liking, commenting, or sharing, if published faster than any other competitors. One of the journalists we interviewed commented that some journalists might sacrifice a bit of accuracy of reporting to become the first author of the related news event. The algorithmic news system, on the other hand, almost takes no time to publish articles even when the data are not ready, and the accuracy is guaranteed for the most cases. With the algorithmic system, journalists would yield the first story to the machine and work on more detailed and analytical stories.

The system can help journalist in some ways. Computers are much faster in finding relevant data than any human can do. For example, when a journalist writes a baseball news, calculating an updated batting average or the changes in the winning percentage caused by the batter take reasonable time even for the baseball experts. However, a machine can find the answer instantly if it is programmed to calculate such metrics beforehand. In other

words, the algorithmic news system extends the ability of a human to seek out more relevant data if customized to meet the wants and needs of the journalist and the news domain.

Moreover, the system can be an ultimate content management system (CMS) tool for a journalist. News media companies usually construct their CMS system, which allows journalists to fill in the templates given by the system such as title, content, and metadata, and also provides access to photos and videos from the media server to insert into the articles. The algorithmic system can be designed to support journalist by automatically filling in the empty templates and finding relevant multimedia content. With the interactive user interface components, journalists can explore different parts of data and start writing up on a news article by editing the system made narratives rather than starting from the scratch. The system will be much more useful if it supports customizable user interface where journalists can add their preferred data selection options to the interface, or the system learns the usage pattern of its users and automatically rearrange the controls for each user.

Algorithmic as Content Creator

The goal of traditional news business was to reach as many readers as possible with the fixed news content. Too targeted or too narrowly focused on a small group of interested readers did not meet the business model. After going digital, the way news readers encounter news content diversified. News media companies such as ProPublica and OPB offered news services that collected news readers' location to present relevant information. The

New York Times and Quartz went a bit further and offered news content as if they are sending chat messages. These services invited users' interactions and massively customized to the targeted user groups.

The algorithmic news generation system went a step further by personalizing news content for each reader. The narratives by PINGS were designed to reflect the user interaction and change upon the data selections. The result of the effect of personalization proved that the direction was right: news articles generated in a personalized and interactive manner are interesting and pleasant to read (Figure 33). However, many participants asked for more controls especially in the narrative generated by the system, such as to change the order of paragraphs, add or hide details on specific points, and edit the narratives. The limitation of PINGS was that it allowed readers to play with data interactively but it could not reflect their preferred style of writing or news reading patterns before they came into the lab for the experiment.

With the ability of an algorithmic system to generate news content instantly, an ideal algorithmic news generation system is expected to generate different layers of stories by mixing relevant data from scratch. The real opportunity for an algorithmic approach to news generation come from its power to make automated decisions on what would news readers like to read and tailor the content and structure of the narrative to match with their preferred reading style, rather than aggregating or curating ready-made information to the readers. PINGS is not at the level of fully personalized story generation system, but once it runs as a news service with real users, then we expect PINGS to learn from the patterns of news readers and generate

more tailored stories based on the activity data.

We found that participants particularly enjoyed the option to change the tone of narrative between an objective mode and story mode. The system generated the narratives that praised players who made hustle plays and wrote the details on the bonehead plays in a sarcastic manner seemed to make very different news reading experience. Some participants raised an interesting issue that is worth to discuss: “Can the narrative written in the story mode be treated as news?” There also were similar opinions from participants such as “A news should be less biased”, “It is still news, but not for everyone.”

To answer the questions from the research perspective, we believe the algorithm-generated content that is tailored to the wants and needs of news readers are not within the boundary of traditional journalism. However, the distinction between what is news and what is not will be fading with the extensibility of algorithms that allow the birth of more hybrid news content (Manovich, 2013). In the evaluation chapter, we compared the perceived news values of content generated by a human journalist and algorithm-generated news content. There were not any notable differences, and these results were predictable since the algorithm followed the writing style of a human journalist.

In digital media, news on a digital platform was remediated from text-based news articles into interactive news services. Although the interaction was limited to simple navigation or exploration, news presentation changed to adapt its surrounding software environment. With PINGS, news readers actively manipulate the graph, table, and the narrative itself and remixing

static text into playable media content. In this regard, the software architecture of PINGS should expand to include more open and user-driven customizable user interface components and make a wide variety of data sets available.

A Universal News Generation System

The algorithm framework behind PINGS was designed to serve the construction of news information not just in the field of sports but to include other domains such as finance, weather and disaster reporting, and any other fields that make heavy use of data and deliver data-driven storytelling. The financial news such as stock market and earnings report are one of the first news domains that algorithm played the role of content creator. The Forbes and the Associated Press partnered with leading narrative-generation algorithm developers in such cases.

Just like baseball news, PINGS could generate financial news using the current user interface structure. The timeline visualization could be designed to show the ups and downs of the stock index throughout the day. The market opens at 9:00 am and closes at 4:00. We could keep track of changes in the index in every minute, and hovering the mouse over at a specific moment in the day could give an overlay pop up of the market index for both KOSPI and KOSDAQ, for example. The data selection pane could provide options to select the type of industries and the name of companies to add a more detailed description of preferred areas in the narratives. Also, there could be a button to expand the data selection to include the figures from earnings reports and historical performance records of the selected companies for the more synchronous rendering of data.

In general, PINGS could be a universal news generation system for events that happen in chronological order and have identifiable and distinct entities for data selection. The most relevant news domains for PINGS, other than sports and financial news, would be the weather and disaster reports since they both deal with timely events that involve obvious data points such as the name of places for gathering and presentation of news information. These would be the news domains that personalizing the narratives with interaction would make the most sense.

In contrast, news experts believed that PINGS-like news generation system would not be the best candidate for the creator of news information for the domains where heavy personalization should be avoided. They mostly believed that the least relevant news domains for PINGS were political and community news. The tone of voice in the story mode would not be the best way to deliver political views since the narratives were generated in a biased manner. If the goal of generating news content was set to provide an objective and representative interpretation of political activities, then the way PINGS generate the narratives would be too personal. Furthermore, the added interactivity in data selection for customizing narratives would only limit the scope of issues confined to the prior knowledge of news readers, if not properly guided.

However, the objectivity and credibility of news content that an algorithm generates could be rated higher when the system lists data points and user interface controls appropriately. We found that personalized and interactive news from PINGS are as much accurate and informative compared to news content created by the algorithm for the straight and summative report.

Also, we received a couple of feedback on many interviewees that the data visualization component was itself self-explanatory, and therefore it is reliable and trustworthy.

7.2 Contributions

The main contribution of this thesis is the demonstration of a new algorithmic news generation system for more engaging and compelling news reading experience. Throughout the research we investigated how a new algorithm framework can be designed and implemented in order to maximize algorithmic capacity, and how the insights and implications learned from the process can be applied in developing a fully-functioning news generation service. Taken together, we elaborated the contributions this thesis made in the fields of journalism and HCI.

Journalism and News Industry

This thesis explored a variety of exemplary news products that are generated by algorithmic computation. The seven most common cases of news products were analyzed through the classification of news types, and common algorithmic attributes were also derived from the analysis. The list of news types and attributes not only helped journalists to get an overview on how algorithm mediate news reading experience in today's media environment, but also help to classify further upcoming news products that are generated by algorithms. Also, we made discussions on the limitations of the current algorithm-generated news products which would become opportunity areas

for other news algorithm developers.

We derived three key opportunity areas from the review of current practices. This thesis offers a conceptual framework that aligns algorithmic activities to journalistic processes, which utilizes research findings from various other fields to define and specify the key terminologies such as data consolidation, personalization in context, and machine-man interactions. This thesis attempts to bridge the gap by incorporating journalism and technology adoption research. We present the following list of contributions from the implementation of algorithm framework:

- We designed the algorithm framework to handle not only the event-related data but also historical match records and events from other baseball parks. We demonstrated how the depth and breadth of the algorithmic computation become deeper and wider by making heavy use of related databases.

- We implemented a working prototype of algorithm framework that generates machine-driven personalized and user-initiated customization of news content via interacting with the news system. We depicted how narrative content is framed to meet the personalized context using complex weight matrix. This thesis had conceptually demonstrated and specified technical processes on how the algorithm can automatically generate personalized content.

- Furthermore, this thesis demonstrated how news readers could be actively involved in customizing the news content by interacting with the user interface elements offered by the system. The system is designed

to take a real-time request from news readers on specifying and adding narrative content based on interactivity functions provided, and the system showed a potential algorithmic news system that generates tailored messages for each news reader with the interactive customization.

Another significant contribution is that we presented how a news generation system can be evaluated. We reviewed and analyzed various research papers to establish comprehensive evaluation criteria for both content and system-wise analysis on an algorithmic news generation system. We derived ten evaluation criteria for algorithm news, which would be more appropriate to evaluate news values of personalized and interactive news content. In discussion part, we proposed implications and made discussion on how a personalized news generation system should function, especially when algorithms play the central roles, as a guide for further research on algorithmic news generation.

HCI and Interface Design

This thesis also made a contribution to the field of HCI by designing and implementing an algorithmic news generation system that news readers can interact upon and generate narratives that meet their personal interests. PINGS is one of the first active news systems that is capable of generating personalized news content based on the activities of algorithmic computation, and providing various interactivity functions for user-driven customization. The design space and interaction methods implemented in this thesis will be beneficial to other researchers when building an active news generation system.

We introduced a news system that offers an interactive user interface components for personalized narrative generation. We explored various information visualization methods, such as timeline visualization and a list of data points in a table form. Our exploration of the design space of algorithmic news generation system helped to discover insights and implications not only in the generation of narrative content but also in visually representing the structured data into an interactive storytelling tool.

This thesis demonstrated how an algorithm system could be an active creator of news content by introducing various narrative generation techniques, such as template-based or rule-based text processing, and natural language generation algorithm. These narrative generation techniques require more attention regarding HCI research agenda as the leading narrative-generation algorithm developers such as Narrative Science, and Automated Insights adopt their solutions to the generation of information content and user-ended services. The expanded role of the algorithm in information content generation was studied in this thesis.

Furthermore, we evaluated the system on behalf of usability and usefulness by recruiting different groups of experts: news experts and system experts. We expanded the spectrum of experts' review by incorporating professionals with varying journalistic and technology skills, which was particularly important when designing an interactive system for news readers. Both expert groups evaluated the system as an easy to use and intuitive tool for information gathering and content generation. The process of experts' evaluation and the result are expected to demonstrate a way to conduct an evaluation study that can be applied to design and development of a news

generation system.

Through quantitative and qualitative evaluation of the system, we derived insights on the implications for designing and developing an interactive news system. The design implications we found throughout the research include the importance of the presentation of data and interactivity functions for personalized and interactive news reading experience. These implications are expected to apply to other research projects not only for news services but broader information system in general.

In sum, this thesis made contributions by demonstrating how to generalize the process of framework development and system implementation in the design of an automated information generation system.

7.3 Limitations

Although PINGS has shown the potential to be a useful system for news generation, there also are limitations in the design of algorithmic news generation system and an algorithmic approach to news generation in general.

Criticism on News Personalization

Along with the emergence of various news services that filter and tailor news for targeted users, the acquisition of biased information and the problems associated with it are also constantly being raised. Since the key strategy of personalization is to frame information in a context that is meaningful to the recipient to increase attention, interest, and motivation (Hawkins et al., 2008), such practice of blocking out certain information might mislead news

readers on perceiving the importance of issues that have opposing perspectives. It is especially critical when an algorithm is a key actor in filtering news content, which is known as filter bubble (Pariser, 2011).

The problems raised by filter bubble are the process of information screening is not transparent, and it takes away the freedom of choices in selective consumption of information. It limits the interaction between people with diverse perspectives and opinions and the opportunity to solve problems via deliberative processes (Bozdag and van den Hoven, 2015).

Since PINGS operates on behalf of algorithmic activities, it cannot be free from the problems raised by filter bubble. PINGS is designed to collect and interpret data from various sources. When it only generates news content for baseball, an automatic information filtering would not be a problem. If it is expanded to generate news for other domains, then it would require more transparent selection criteria to become more trustworthy and credible source of news information.

Also, the ways to express emotions and the vocabulary in the expression should change depending on the news domain. It will be much more critical when PINGS is designed to generate politics and community news. The narrative in the story mode must be checked concerning the facts the system tries to convey and the opinions it adds on top of the facts. The current expressions were tailored to meet the interests of baseball fans, but news readers for political news might be offended. Therefore, further research is required to examine the breadth and depth of personalization in other news domains.

Limitations in the Current Narrative Generation Methods

The more sophisticated narrative-generation algorithm may be applied in future work. Currently, we used both template-based text processing and natural language generation algorithm in compositing the narratives. The software codes that we developed for natural language generation is capable of handling the influx of text broadcasting data and pairing words and phrases with the given rule at scale. However, there are rapid technological advances made in various machine learning and neural network algorithms that would make the narrative generation process much richer.

With the recent advancement in machine learning and neural network algorithms, the technical sophistication behind the leading software vendors may go well beyond artificial intelligent system for natural language generation. Recently, Google and Microsoft both announced language translating services operated by their neural network algorithms. According to Google, their Neural Machine Translation (NMT) algorithm is capable of encoding input word into a list of vectors and decodes into the output word using their internal weight computation system (Wu et al., 2016). The quality of the output content, as of the year 2016, do not yet meet the level of professional translators, but the quality may go up as more people use their service. It is only of the major advantage of machine learning algorithms that the quality of algorithm goes up as there are more training data to learn. Eventually, Google's NMT might even generate more complicated articles.

However, we believe using rule-based language processing techniques backed up by our complex weight matrix system have a competitive advantage

over neural network machines. The complex weight matrix utilizes both static and dynamic language generation upon the weights the system assigns. It computationally determines the key events and themes that match with the appropriate mood of the event. The process of the mood determination is equal to how human journalists set an angle in writing news articles. It is more important that the narrative generated by a machine meet the syntactic need of such news content in delivering the information with the right tone and manner, and we believe the complex weight matrix be as competitive shortly.

We also have plans to complement the system by applying neural network algorithms that meet our needs. We believe neural network algorithms could help us to amplify the variety of themes for more sophisticated mood selection and find the most trending style of writing from the large mass audiences. Our future research will include developing neural network algorithm that works for/with the complex weight matrix.

The Level of Interactivity

PINGS is designed to take direct user engagements such as click, mouse hovering, data selection, and perspective changing by manipulating various user interface elements. The system would be much appreciated for heavy users, or enthusiastic baseball fans in this case, for its adaptive and dynamic news generation method. However, it cannot be the news system for everyone. Light users, or even some heavy users with a limited timeframe for news reading, would still prefer an autonomous news generation system. They would still like to see personalized content, but the level of interaction they

are willing to make would be much lower than the degree of interaction required for manipulating UI components in PINGS.

To lower the barrier for light news readers, other types of interactivity methods could be adopted. As in the case of face-to-face communication, a less demanding news system would be designed to deliver information that meets the context of news readers just like having a conversation with friends who know their interests and preferences. The level of interactivity would be much lower if the system automatically generates messages that reflect their baseball news reading patterns. PINGS is only at an interactive prototype and therefore collecting log data, and usage patterns were limited at this time. Our future research will explore how to diversify the strategies in designing and implementing interactivity method for varying types of news readers.

Chapter 8

Conclusion

In this last chapter, we summarized this thesis on how we designed and implemented algorithmic news generation framework and an interactive news generation system. We conclude this research with the list of opportunities for future research.

8.1 Summary of Work

We started this research by exploring the changes made in the practices of journalism caused by technological advances. While technology has always influenced journalism at some degree (Pavlik, 2000), there are increasing cases of the application of algorithm in the practices of journalism, especially in the creation of news and news related content (Napoli, 2014; Dörr, 2016).

The type of algorithm-generate news expanded from a short and straight report to include more sophisticated news services such as messaging, news game, and personalized news reports. Also, the leading narrative generation algorithm companies published patent documents about the ways to generate more personally meaningful narratives by automatically processing data. They heavily collaborate with various media companies and other information-service vendors to provide automated solutions for converting data into information with context.

In this thesis, we made the following research activities:

- We reviewed many news products that are generated by algorithms to classify algorithmic attributes commonly found, and derived implications on how to build more comprehensive algorithm framework to overcome the limitations of the current practices.

- We proposed a new conceptual design of algorithm framework that extends the capacity of algorithmic news generation in three process stages: (1) data synthesizing, (2) narrative framing, and (3) presentation of news information.

- We developed fully-functioning algorithm framework that generates baseball news as an exemplary case. We crawled play-by-play text broadcasting data and extracted key events by algorithmic computation, and depicted how narrative content is framed and generated using complex weight matrix and dynamic sentence generation algorithm.

- We designed and implemented a personalized and interactive prototype of news generation system. The user interface elements and style of interaction were designed upon the implications found from the review of existing practices and framework development. The system is capable of generating personalized news stories that reflect manipulations made by news readers and proved to be a good and competent news creator in the domain of baseball news.

- We presented how a news generation system can be evaluated. We recruited ordinary news readers as participants for a survey for news values and also interviewed news and system experts to evaluate the

system. We reviewed and analyzed various research papers to establish comprehensive evaluation criteria for both content and system-wise analysis on an algorithmic news generation system.

- We proposed implications on how a personalized news generation system should function, especially when algorithms play the central roles. We also identified opportunities and limitations in designing a personalized news system as general implications for other researchers.

8.2 Opportunities for Future Work

The evaluation of PINGS revealed its potential to be a useful and usable news generation system for sports news readers. While it might still be questionable whether the system could become a legitimate news content provider in news domains other than sports, we believe the system is capable of automating news generation process in domains where the narrative structure is similar, and events can be arranged in chronological order. In other words, if a news article deals with a series of events and any event can be interpreted as a part of a larger context, then PINGS is capable of automating news generation process and be a useful tool for personalized news generation in that domain. The exemplary news domains for PINGS include finance and stock market reports, streams of media coverage on the local and national news, alerts about natural disasters and follow-up reports, etc.

Our first and foremost future research agenda would be to prove if PINGS could be capable of generating news content for other news domains, and is appreciated as a usable and useful news provider. Depending on

the news domain, how the algorithm framework constructs a consolidated database, how the system derives appropriate and contextual narrative framing based on algorithmic computation, and how news information could be disseminated and distributed to news readers would change.

As much as its domain can be diversified, the form of output content might also go beyond plain text but to include graphics, photos, audios, videos, and any other forms of media elements. Google published a research paper about the automatic creation of text descriptions of images found on the web (Vinyals et al., 2014), which not only helps them to search images using search terms but also widens the capacity of algorithms in changing the modality of information for an enhanced news reading experience. Also, some of the recent news algorithm development trends include summarizing text-based news articles into image-based contents (Ha et al., 2015), and a web service for the automatic creation of video contents from text documents[1]. These technological advances enable the creation of online news packages that access databases of elements in multiple forms and provides new opportunities for users to engage with news (Royal, 2010).

In conclusion, we believe this research on personalized and interactive news generation using algorithm system has just opened up an academical research, both journalism and human-computer interaction point of view, on the growing influence of algorithm in the practices of information content generation and distribution. Our future research will continue to explore how to apply algorithmic activities in journalistic processes and examine theories and methodologies in automatic news content generation.

[1] Wibbitz, http://www.wibbitz.com/

References

Allen, N. D., Templon, J. R., McNally, P. S., Birnbaum, L., and Hammond, K. J. (2010). Statsmonkey: A data-driven sports narrative writer. In *Proceedings of the AAAI Fall Symposium: Computational Models of Narrative*.

Allen, R. C. (2013). *U.S. Patent No. 8,515,737*. Washington, DC: U.S. Patent and Trademark Office.

Anderson, C. (2006). *The long tail: Why the future of business is selling more for less*. Hyperion, New York.

Beam, M. A. (2014). Automating the news: How personalized news recommender system design choices impact news reception. *Communication Research*, 41(8):1019–1041.

Birnbaum, L. A., Hammond, K. J., Allen, N. D., and Templon, J. R. (2013). *U.S. Patent No. 8,355,903*. Washington, DC: U.S. Patent and Trademark Office.

Blom, J. (2000). Personalization: A taxonomy. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '00, pages 313–314, New York, NY, USA. ACM.

Blom, J. O. and Monk, A. F. (2003). Theory of personalization of appearance: why users personalize their pcs and mobile phones. *Human-Computer Interaction*, 18(3):193–228.

Bogost, I. (2007). *Persuasive games: The expressive power of videogames*. MA: The MIT Press.

Bogost, I., Ferrari, S., and Schweizer, B. (2012). *Newsgames: Journalism at play*. MA: The MIT Press.

Bolter, J. D., Grusin, R., and Grusin, R. A. (2000). *Remediation: Understanding New Media*. MA: The MIT Press.

Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679.

Bozdag, E. and van den Hoven, J. (2015). Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, 17(4):249–265.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117.

Brusilovsky, P. and Maybury, M. T. (2002). From adaptive hypermedia to the adaptive web. *Communications of the ACM*, 45(5):30–33.

Carlson, M. (2015). The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digital Journalism*, 3(3):416–431.

Chen, C.-H., Lin, J., Yücesan, E., and Chick, S. E. (2000). Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270.

Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. (2010). Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30. ACM.

Chua, R. (2010). Structured journalism. Retrieved from http://structureofnews.wordpress.com/ structured-journalism/.

Clerwall, C. (2014). Enter the robot journalist: Users' perceptions of automated content. *Journalism Practice*, 8(5):519–531.

Cohen, S., Hamilton, J. T., and Turner, F. (2011). Computational journalism. *Communications of the ACM*, 54(10):66–71.

Cole, J. S., Kennedy, M., and Ben-Avie, M. (2009). The role of precollege data in assessing and understanding student engagement in college. *New Directions for Institutional Research*, 2009(141):55–69.

Cormen, T. H. (2009). *Introduction to algorithms*. The MIT press.

Cox, M. (2000). The development of computer-assisted reporting. *Informe presentado en Association for Education in Jornalism end Mass Comunication). Chapel Hill, EEUU: Universidad de Carolina del Norte*.

Craig, T. and Ludloff, M. E. (2011). *Privacy and big data*. O'Reilly, Sebastopol, CA.

Daniel, A. and Flew, T. (2010). The guardian reportage of the uk mp expenses scandal: A case study of computational journalism. In *Record of the Communications Policy and Research Forum 2010*, pages 186–194. Network Insight Pty. Ltd.

Das, A. S., Datar, M., Garg, A., and Rajaram, S. (2007). Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, pages 271–280. ACM.

Davison, J. and Guiro, B. (2014). *U.S. Patent No. 8,821,271*. Washington, DC: U.S. Patent and Trademark Office.

Diakopoulos, N. (2011). A functional roadmap for innovation in computational journalism. Retrieved from https://www.nickdiakopoulos.com/2011/04/22/a-functional-roadmap-for-innovation-in-computational-journalism.

Diakopoulos, N. (2014). Algorithmic accountability reporting: On the investigation of black boxes. *Tow Center for Digital Journalism, Columbia University*.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3):398–415.

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

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015a). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015b). Overcoming algorithm aversion: People will use algorithms if they can (even slightly) modify them. *Available at SSRN 2616787*.

Dörr, K. N. (2016). Mapping the field of algorithmic journalism. *Digital Journalism*, 4(6):700–722.

Flew, T., Spurgeon, C., Daniel, A., and Swift, A. (2012). The promise of computational journalism. *Journalism Practice*, 6(2):157–171.

Fox, S. and Duggan, M. (2013). Pew internet and american life project. *Health online*, 2013.

Garrison, B. (1998). *Computer-assisted reporting*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Gillespie, T. (2014). The relevance of algorithms. In Gillespie, T., Boczkowski, P. J., and Foot, K. A., editors, *Media Technologies: Essays on Communication, Materiality, and Society*, pages 167–194. The MIT Press.

Graefe, A. (2016). Guide to automated journalism. *Tow Center for Digital Journalism. Janeiro*.

Gray, J., Chambers, L., and Bounegru, L. (2012). *The data journalism handbook*. O'Reilly Media, Sebastopol, CA.

Gynnild, A. (2014). Journalism innovation leads to innovation journalism: The impact of computational exploration on changing mindsets. *Journalism*, 15(6):713–730.

Ha, J.-W., Kang, D., Pyo, H., and Kim, J. (2015). News2images: Automatically summarizing news articles into image-based contents via deep learning. *RecSys 2015,Vienna, Austria*.

Hamilton, J. T. and Turner, F. (2009). Accountability through algorithm: Developing the field of computational journalism. In *Report from the Center for Advanced Study in the Behavioral Sciences, Summer Workshop*, pages 27–41.

Hawkins, R. P., Kreuter, M., Resnicow, K., Fishbein, M., and Dijkstra, A. (2008). Understanding tailoring in communicating about health. *Health Education Research*, 23(3):454–466.

Hermida, A. (2010). Twittering the news: The emergence of ambient journalism. *Journalism Practice*, 4(3):297–308.

Hoffman, D. L. and Novak, T. P. (1996). Marketing in hypermedia computer-mediated environments: Conceptual foundations. *Journal of Marketing*, 60(3):50–68.

Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York: NYU Press.

Karat, J., Karat, C.-M., and Ukelson, J. (2000). Affordances, motivation, and the design of user interfaces. *Communications of the ACM*, 43(8):49–51.

Karlsson, M. (2011). The immediacy of online news, the visibility of journalistic processes and a restructuring of journalistic authority. *Journalism*, 12(3):279–295.

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, 331(6018):719–721.

Kitchin, R. and Dodge, M. (2011). *Code/space: Software and everyday life*. MA: The Mit Press.

Kizilcec, R. F. (2016). How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2390–2395, New York, NY, USA. ACM.

Kowalski, R. (1979). Algorithm= logic+ control. *Communications of the ACM*, 22(7):424–436.

Kramer, J., Noronha, S., and Vergo, J. (2000). A user-centered design approach to personalization. *Commun. ACM*, 43(8):44–48.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Lantz, B. (2013). *Machine learning with R*. Packt Publishing Ltd, Birmingham.

Lavie, T., Sela, M., Oppenheim, I., Inbar, O., and Meyer, J. (2010). User attitudes towards news content personalization. *International Journal of Human-Computer Studies*, 68(8):483 – 495.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8595–8598. IEEE.

Lewis, S. C. and Usher, N. (2013). Open source and journalism: Toward new frameworks for imagining news innovation. *Media, Culture & Society*, 35(5):602–619.

Liu, J., Dolan, P., and Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 31–40. ACM.

Lokot, T. and Diakopoulos, N. (2016). News bots. *Digital Journalism*, 4(6):682–699.

Mager, A. (2012). Algorithmic ideology: How capitalist society shapes search engines. *Information, Communication & Society*, 15(5):769–787.

Manovich, L. (2013). *Software takes command*, volume 5. New York: A&C Black.

Manurung, H. (2004). *An evolutionary algorithm approach to poetry generation*. PhD thesis, University of Edinburgh.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*.

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., and Miller, R. C. (2011). Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 227–236, New York, NY, USA. ACM.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., and Barton, D. (2012). Big data. *The management revolution. Harvard Bus Rev*, 90(10):61–67.

Meyer, P. (1973). *Precision journalism: A Reporter's Introduction to Social Science Methods*. Indiana University Press.

Murray, J. H. (1997). *Hamlet on the holodeck: The future of narrative in cyberspace*. (2001 ed.). MA: The MIT Press.

Napoli, P. M. (2014). Automated media: An institutional theory perspective on algorithmic media production and consumption. *Communication Theory*, 24(3):340–360.

O'Banion, S., Birnbaum, L., and Hammond, K. (2012). Social media-driven news personalization. In *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web*, pages 45–52. ACM.

Pariser, E. (2011). *The Filter Bubble: What The Internet Is Hiding From You*. Penguin UK.

Pavlik, J. (2000). The impact of technology on journalism. *Journalism Studies*, 1(2):229–237.

Reiter, E., Dale, R., and Feng, Z. (2000). *Building natural language generation systems*, volume 33. Cambridge University Press, Cambridge.

Rimer, B. K. and Kreuter, M. W. (2006). Advancing tailored health communication: A persuasion and message effects perspective. *Journal of Communication*, 56:S184–S201.

Roberto, A. J., Krieger, J. L., and Beam, M. A. (2009). Enhancing web-based kidney disease prevention messages for hispanics using targeting and tailoring. *Journal of Health Communication*, 14(6):525–540. PMID: 19731125.

Rogers, S. (2013). *Facts are sacred*. The Guardian Books, London.

Royal, C. (2010). The journalist as programmer: A case study of the new york times interactive news technology department. In *International Symposium on Online Journalism*.

Segel, E. and Heer, J. (2010). Narrative visualization: Telling stories with data. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1139–1148.

Seib, P. M. (2002). *Going live: Getting the news right in a real-time, online world*. Oxford: Rowman & Littlefield.

Sohn, D. (2011). Anatomy of interaction experience: Distinguishing sensory, semantic, and behavioral dimensions of interactivity. *new media & society*, 13(8):1320–1335.

Steiner, C. and Dixon, W. (2012). *Automate this: How algorithms came to rule our world*. New York, NY: Portfolio.

Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, 42(4):73–93.

Sundar, S. S. (1999). Exploring receivers' criteria for perception of print and online news. *Journalism & Mass Communication Quarterly*, 76(2):373–386.

Sundar, S. S. and Marathe, S. S. (2010). Personalization versus customization: The importance of agency, privacy, and power usage. *Human Communication Research*, 36(3):298–322.

Tango, T. M., Lichtman, M. G., and Dolphin, A. E. (2007). *The Book: Playing The Percentages In Baseball*. Potomac Books, Inc.

Thurman, N. and Schifferes, S. (2012). The future of personalization at news websites: lessons from a longitudinal study. *Journalism Studies*, 13(5-6):775–790.

Tuzhilin, A. (2009). Personalization: The state of the art and future directions. In Adomavicius, G. and Gupta, A., editors, *Business computing*, volume 3, chapter 1, pages 3–43.

van Dalen, A. (2012). The algorithms behind the headlines: How machine-written news redefines the core skills of human journalists. *Journalism Practice*, 6(5-6):648–658.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wojtkowski, W. and Wojtkowski, W. G. (2002). Storytelling: its role in information visualization. In *Proceedings of the Fifth European Systems Science Congress*, pages 16–19.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Young, M. L. and Hermida, A. (2015). From mr. and mrs. outlier to central tendencies. *Digital Journalism*, 3(3):381–397.

김동환 and 이준환 (2015). 로봇 저널리즘: 알고리즘을 통한 스포츠 기사 자동 생성에 관한 연구. *한국언론학보*, 59(5):64–95.

김영주, 오세욱, and 정재민 (2015). 로봇 저널리즘: 가능성과 한계. Technical report 2015-02, 한국언론진흥재단.

오세욱 and 이재현 (2012). 소프트웨어 '페이스북'의 알고리즘 분석. *언론과 사회*, 21(1):136–183.

Appendix A: Algorithm News Products

The purpose of this appendix is to introduce the current news products we reviewed in Chapter 3 in more detail. Additional screen captures and more detailed analysis of algorithmic attributes are presented.

More Data-centric Reports

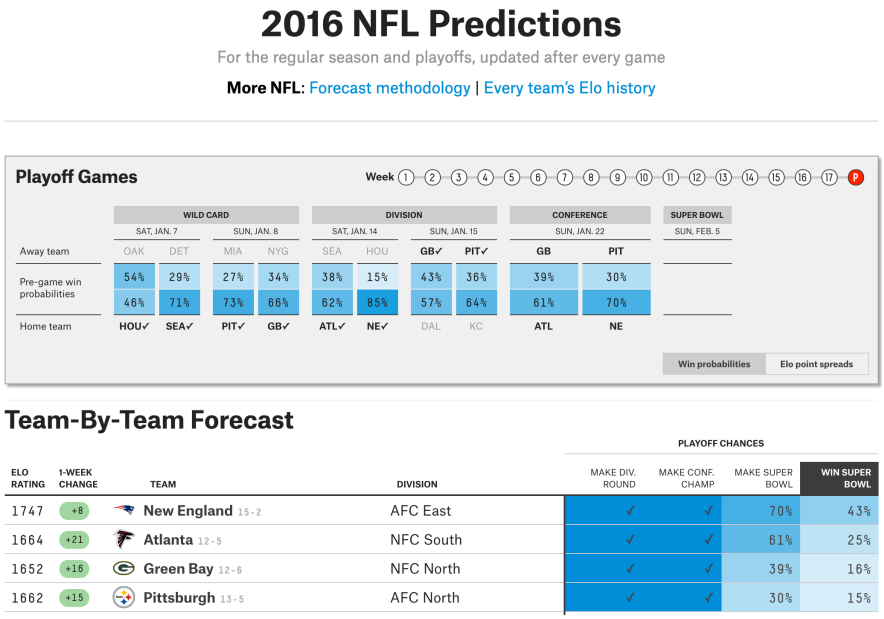


Figure 34: Data-centric report from NFL Predictions

In addition to 4th Down Bot (NYT), FanGraphs and NFL Predictions (FiveThirtyEight) are also exemplary cases of data-centric reports (Figure 34). FanGraphs is a web service that provides statistics for every player and team in Major League Baseball as well as the scoreboard and plays log

of matches in real-time. The use of the algorithm in FanGraphs seems to be trivial. However, this website presents various machine-generated performance report based on statistical analysis of current and historical baseball records.

NFL Predictions is a site that forecasts the winning probability of NFL teams for the regular season and playoff games. The service makes predictions using Elo ratings, which is a method for calculating the skill level of each NFL team using historical play records of teams and players. NFL predictions present the ranking list of teams for the super bowl and a visualized graph on each team's Elo ratings over time.

Table 12: Algorithmic attributes for data-centric reports

Stage	Attribute	Note
Gathering		Various sports-related records and broadcasting messages are the basis for generating data-centric reports.
Input Source:	Web/API	
News Domain:	Sports News	
Processing		Sabermetrics and Elo ratings are widely used for analyzing sporting events. The major objective is to deliver credible reports to news readers.
Comp. Method:	Statistical Analysis	
Narrative Angle:	Data Only	
Presentation		Data-centric reports make heavy use of numbers and graphs to deliver news information, and also provides intearctivity functions for exploring row data.
Output Type:	Report	
Lv of Interactivity:	Exploratory	

News Products from Narrative Science and Automated Insights

Narrative Science (NS) and Automated Insights (AI) are the leading algorithm developers for a narrative generation. They collaborate with various news media for generating simple and straight reports. Financial news on

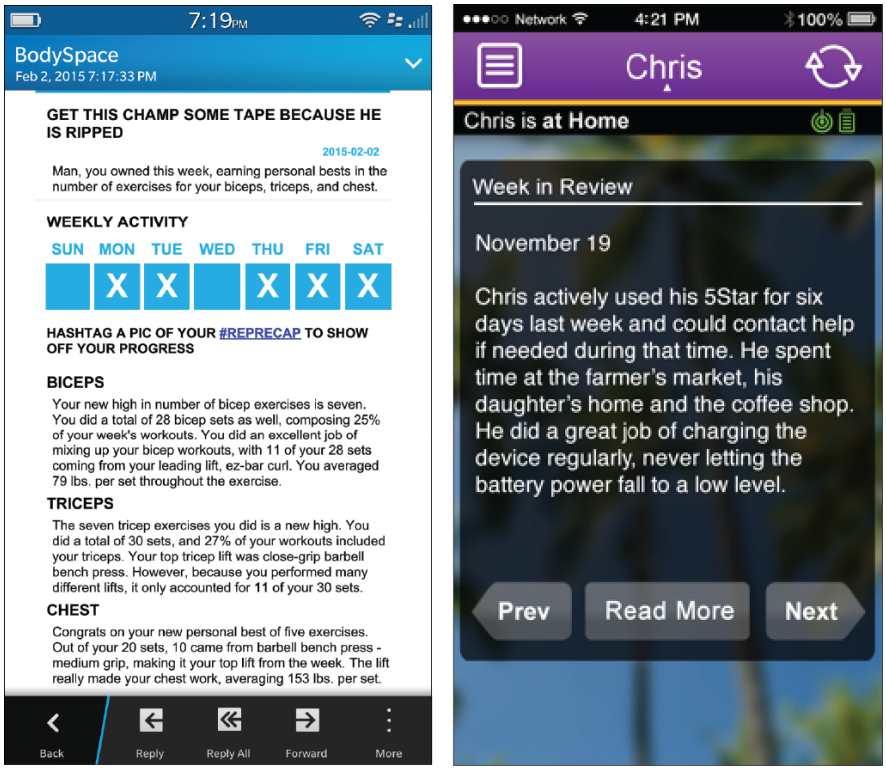


Figure 35: Products from NS and AI: BodySpace(L) GreatCall(R)

Forbes and the AP, sports news on BigTenNetwork and GameChanger are the output news products from such collaborations. However, the breadth of their news services goes beyond news media. They also work with other information vendors to generate personalized content for targeted audiences. Personalized workout report for BodySpace and sensor-based family report for GreatCall are exemplary cases of a personalized report from these algorithm developers (Figure 35).

Table 13: Algorithmic attributes for news from NS and AI

Stage	Attribute	Note
Gathering		NS and AI make heavy use of data from various sources to expand the capacity of algorithmic computation. They are capable of generating contents at various level for their business partners.
Input Source:	Web/API, Sensors, and Other data	
News Domain:	Sports, Finance, Personal News	
Processing		Their focus of technological development is to invent human-sounding narrative generation algorithms. They are also capable of generating tailored messages for news readers in many domains.
Comp. Method:	Statistical Analysis, NLG	
Narrative Angle:	Objective and Personalized report	
Presentation		Most of their products mediate human-processing of information content, and therefore, their products are mostly in a form of news article.
Output Type:	Long story (article)	
Lv of Interactivity:	Static	

News Curation Services

Google Trends and Flipboard are news curation services that collect and present news content from various sources (Figure 36). The use of algorithms in these services is not geared towards content generation, but the reason we included them into the review process is to look into how algorithms function in filtering content or generating visualizations and layout. Google Trends makes use of a vast amount of search queries that Google users make, and automatically create an interactive visualization on trending stories. Flipboard is a widely used news curation service that collects content from a variety of sources including news reader's social network services such as Facebook, Instagram, and Twitter. Flipboard introduced a dynamic layout engine called Duplo, which uses an algorithm to fit content with varying size into the appropriate layout and generates summarized text for

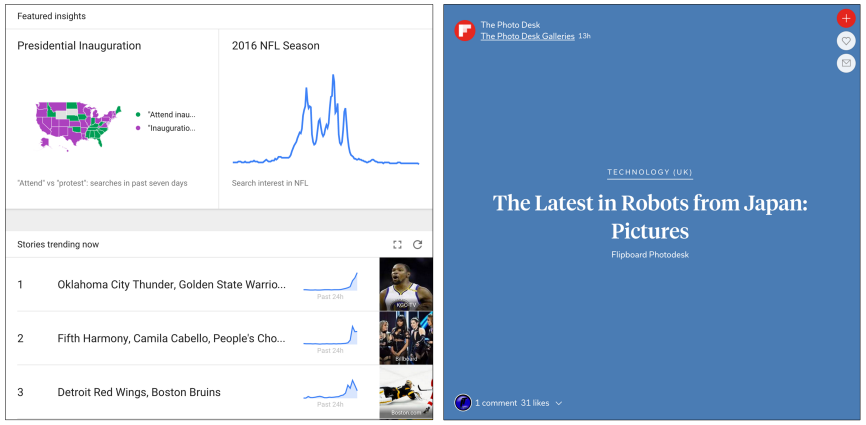


Figure 36: News curation services: Google Trends(L) Flipboard(R)

snippet view. Both services give insight on how algorithms can be adopted at various levels in the process of delivering news information.

Table 14: Algorithmic attributes for news curation services

Stage	Attribute	Note
Gathering		Curation services collects content from various news media and social network services.
Input Source:	News database	
News Domain:	No specific fields	
Processing		The objective of these services is to arrange and filter news information to the preferences of news readers, and algorithms are heavily used in filtering and presenting news in many formats.
Comp. Method:	Visualization, rule-based processing	
Narrative Angle:	Mass customized	
Presentation		They often present news with interactive visualization of data, which invites various interactivity functions and allows exploratory news reading experience for users.
Output Type:	Interactive content	
Lv of Interactivity:	Exploratory	

Appendix B: Study Materials

In this appendix, the study materials we used for the evaluation are presented. These materials include pre-survey questionnaire, survey questions for scoring news values of given news article, news articles presented to study participants, and interview questions and guidelines for experts interview.

Pre-survey Questionnaire

1. 모집 사이트 가입 시 사용한 이름을 적어주세요.
2. 나이를 알려주세요.
3. 가장 좋아하는 구단을 선택해주세요.
    - 두산
    - NC
    - 넥센
    - SK
    - LG
    - 한화
    - KIA
    - 롯데
    - 삼성
    - kt

4. 가장 좋아하는 선수 세 명을 적어주세요.

5. 프로야구 중계를 시청하는 횟수는?

    - 보지않는다
    - 일년에 서너번
    - 한달에 서너번
    - 일주일에 두세번
    - 매일보는 편이다

6. 프로야구는 언제부터 시청했나요?

    - 보지않는다
    - 이번해부터
    - 2 – 3년 전부터
    - 2008년 베이징올림픽 이후
    - 2008년 이전부터

7. 다른 경기 결과는 어떻게 접하나요?

    - 보지않는다
    - 인터넷 포털 중계창에서 직접 확인
    - 기사를 통해 (인터넷 기사 포함)
    - TV 뉴스를 통해 (스포츠 전문 채널 포함)
    - 소셜 미디어 혹은 커뮤니티를 통해
    - 친구로부터 듣는다

Survey Questions for Scoring News Values

명료하다	매우아니다	1	2	3	4	5	6	7 매우그렇다
일관적이다	매우아니다	1	2	3	4	5	6	7 매우그렇다
잘쓰였다	매우아니다	1	2	3	4	5	6	7 매우그렇다
유용한 정보를 준다	매우아니다	1	2	3	4	5	6	7 매우그렇다
정확하다	매우아니다	1	2	3	4	5	6	7 매우그렇다
믿을만하다	매우아니다	1	2	3	4	5	6	7 매우그렇다
흥미롭다	매우아니다	1	2	3	4	5	6	7 매우그렇다
읽기에 즐겁다	매우아니다	1	2	3	4	5	6	7 매우그렇다
중요한 내용을 담고있다	매우아니다	1	2	3	4	5	6	7 매우그렇다
경기 내용을 잘 반영한다	매우아니다	1	2	3	4	5	6	7 매우그렇다

Journalist-generated News (Doosan v.s. SK)

두산 베어스는 10일 인천SK행복드림구장에서 열린 2016 타이어뱅크 KBO리그 SK 와이번스와의 팀 간 4차전에서 11-7로 재역전승을 거두며 4연패 탈출에 성공했다.

홈 팀 SK는 문승원이 선발 마운드에 오른 가운데 이명기(좌익수)- 조동화(중견수)- 최정(3루수)- 정의윤(우익수)- 박정권(1루수)- 이재원(포수)- 헥터 고메즈(유격수)- 최정민(지명타자)- 김성현(2루수)로 라인업을 구성했다. 원정 팀 두산은 박건우(좌익수)- 정수빈(중견수)- 민병헌(우익수)- 김재환(지명타자)- 양의지(포수)- 오재원(2루수)- 닉 에반스(1루수)- 허경민(3루수)- 김재호

(유격수)로 라인업을 짰고, 선발 투수에는 유희관이 등판했다.

두산은 1회초 리드오프로 타석에 들어선 박건우가 상대 선발 문승원의 초구 145km/h 속구를 통타해 기선을 제압하는 중월 솔로 홈런을 터뜨렸다. 이 홈런은 두산 팀 통산 첫 번째 초구 · 선두타자 홈런이었다. 스코어는 1-0. 선취점을 올린 두산은 2회초에도 1사 이후 오재원의 볼넷과 에반스의 2루타로 기회를 잡았고, 후속 타자 허경민이 좌익수 쪽 희생플라이를 기록해 2-0으로 앞서갔다.

기세를 탄 두산은 3회초 2사 이후 민병헌이 문승원의 8구를 타격해 좌중간 담장을 넘기는 비거리 120m짜리 솔로포를 쳐내 석 점째를 올렸다. 끌려가던 SK는 3회말 경기를 단숨에 원점으로 돌려놨다. SK는 최정의 좌전 안타와 정의윤의 우중간 안타로 무사 1,3루 기회를 잡았고, 후속 타자 박정권이 좌측 담장을 맞히는 1타점 2루타를 터뜨려 한 점 추격했다. 이후 SK는 이재원의 희생플라이와 최정민의 우전 1타점 적시타로 3-3 동점을 만들었다. 여기에 그치지 않은 SK는 김성현의 역전 투런포까지 터지며 5-3으로 역전했다.

SK는 5회말 고메즈가 2루타와 상대 실책을 엮어 3루까지 진루했고, 후속 타자 최정민이 우익수 옆 1타점 2루타를 쳐내며 추가 점수를 기록했다. 계속해서 SK는 김성현의 희생번트와 이명기의 몸맞는공으로 1사 1,3루를 만들었다. 여기서 SK는 조동화가 2루수 땅볼로 실책을 이끌어내 3루 주자 최정민의 홈인을 도왔다. 스코어는 7-3.

두산은 7회초 1사 1루에서 허경민이 좌중간 2루타를 터뜨리며

한 점을 올렸고, 후속 타자 김재호가 추격의 투런포를 작렬시켜 6-7로 따라붙었다. 결국 두산은 8회초 무사 1루에서 김재환이 역전 투런 홈런을 터뜨리며 경기를 뒤집었고, 오재원의 1타점 중전 안타까지 나오며 9-7로 격차를 벌렸다. 두산은 9회초에도 김재환의 쐐기 투런포가 터지며 11-7로 스코어를 만들었다. 리드를 되찾은 두산은 8회말부터 필승조를 연이어 투입해 SK 타선을 봉쇄했고, 결국 4연패 탈출이라는 값진 선물을 얻었다.

Journalist-generated News (Hanhwa v.s. Samsung)

한화는 4일 대구 삼성라이온즈파크에서 열린 2016 타이어뱅크 KBO리그 삼성 라이온즈와의 팀간 8차전 맞대결에서 8-7로 승리했다. 이날 승리로 3연승을 달린 한화는 19승(1무 32패) 째를 거뒀다. 반면 삼성은 2연패에 빠졌다.

이날 한화는 1회초 정근우의 2루타 뒤 이용규와 김태균의 진루 땅볼이 나와 선취점을 냈다. 삼성은 1회말 곧바로 화력을 집중시키며 역전에 성공했다. 선두타자 배영섭이 볼넷을 얻어낸 뒤 박해민이 2루타를 쳐 무사 1,3루 찬스를 잡았다. 이어 이승엽의 땅볼 때 배영섭이 홈을 밟았다. 이후 최형우의 적시타와 박한이의 2루타, 조동찬의 희생 플라이가 이어지면서 4-1로 점수를 벌렸다.

분위기가 삼성으로 넘어가는 듯 했지만 한화가 곧바로 동점을 만들었다. 하주석이 몸에 맞는 공으로 출루한 뒤 이성열의 투런 홈런을 날렸다. 이후 신성현의 볼넷과 이용규의 안타, 김태균의 적시타로 4-4 동점을 만들었다.

한화는 4회초 정근우와 이용규의 연속 2루타로 한 점을 치고

나갔지만, 4회말 곧바로 이지영이 솔로 홈런을 날려 경기를 원점으로 돌렸다. 삼성은 5회 무사 만루, 6회 1사 만루 찬스에서 한 점도 뽑아내지 못하면서 좀처럼 치고 나가지 못했다. 결국 균형은 7회초 한화가 깼다. 7회초 선두타자 이용규가 볼넷을 골라낸 뒤 김태균과 로사리오가 안타와 2루타를 잇따라 때려냈다. 이어 양성우도 2타점 적시타를 치면서 점수는 8-5로 벌어졌다.

삼성은 7회말 선두타자로 나온 김정혁이 데뷔 첫 홈런포를 날려 한 점을 만회했다. 그리고 9회말 삼성은 김정혁의 안타와 이지영의 사구 뒤 김재현의 희생번트가 상대 실책으로 이어지면서 무사 만루 찬스를 잡았다. 그러나 배영섭이 병살타를 치면서 한 점과 아웃카운트 두 개를 바꿨고, 결국 박해민이 삼진으로 물러나면서 경기를 뒤집지 못했다.

Algorithm-generated News (Doosan v.s. SK)

10일 문학구장에서 열린 2016 타이어뱅크 KBO리그 두산과 SK의 경기에서 두산이 26안타 6홈런 18득점을 합작하는 타격쇼 끝에 승리하였다 SK도 무려 7명의 주자를 홈으로 불러들였으나 승리를 가져가기엔 역부족이었다.

1회 초 두산은 박건우의 1점 홈런으로 1점 앞서나가기 시작했다. 1회 말 SK는 2사 2, 3루 상황에서 박정권의 볼넷으로 2사 만루 상황을 만들었으나 이후 이재원의 중견수 플라이로 공수교대가 이루어지며 차이를 좁히지 못했다. 2회 초 두산은 1사 2, 3루 상황에서 허경민의 희생플라이로 점수 차를 벌렸다. 2회 말 SK는 1사 1, 2루 상황에서 이명기의 볼넷으로 1사 만루

상황을 만들었으나 이후 조동화의 병살타로 공격기회를 소진해 차이를 좁히지 못했다.

3회 초 두산은 민병헌의 1점 홈런으로 1점을 달아났다.3회 말 SK는 무사 1, 3루 상황에서 박정권의 1타점 2루타, 이재원의 희생플라이와 최정민의 1타점 적시타로 동점을 만들었고, 김성현의 2점 홈런으로 선취 5득점했다. 그 후 4회 말에는 2사 2루 상황에서 박정권의 볼넷으로 2사 1, 2루 상황을 만들었으나 이후 이재원의 투수 땅볼로 아웃카운트를 헌납해 달아나지 못했다. 또 5회 말에는 무사 3루 상황에서 최정민의 1타점 2루타로 점수 차를 벌렸으며 이명기의 데드볼로 1사 1, 3루 상황을 만들고 조동화의 2루수 땅볼로 점수 차를 벌렸다.

7회 초 두산은 상대 좌익수 이명기의 실책으로 1점을 따라잡았으며 김재호의 2점 홈런으로 1점을 따라잡았다. 그 후 8회 초에는 김재환의 2점 홈런으로 경기를 뒤집었으며 양의지의 2루타로 무사 2루 상황을 만들고 오재원의 1타점 적시타로 1점을 달아났다. 1사 2루 득점찬스를 맞이하였으나 허경민의 2루수 앞 땅볼과 김재호의 1루수 땅볼로 차이를 벌리지 못했다. 또 9회 초에는 김재환의 2점 홈런으로 2득점하며 경기결과를 확정지었다.

경기 결과 두산은 극적으로 SK를 이길 수 있었다. 오늘 경기의 결과 두산은 4연패를 탈출했고 시즌 20회째 승리했으며 현재 1위(승률 0.667)이다. 한편 SK는 현재 3위 상위권으로 승률 0.576을 기록 중이고 13안타 1홈런 4볼넷 7타점 7득점으로 괜찮은 경기를 보여줘 결과에 아쉬움을 남겼다.

Algorithm-generated News (Hanhwa v.s. Samsung)

4일 대구구장에서 열린 한화와 삼성과의 2016 타이어뱅크 KBO 리그 결과 한화가 로사리오의 결승타에 힘입어 8:7로 신승을 거두었다. 5:5로 경기 중이던 7회 초 무사 1, 3루에서 로사리오의 적시 2루타가 1득점을 얻으며 오늘 경기의 승리를 결정지었다. 오늘 승리에 결정적 기여를 한 로사리오는 시즌 194타수 61 안타 10홈런 13볼넷 43타점 26득점을 기록 중이다.

1회 초 한화는 무사 2루 상황에서 김태균의 2루수 땅볼로 선취 1득점했다. 1회 말 삼성은 무사 2, 3루 상황에서 이승엽의 2 루수 땅볼, 최형우의 1타점 적시타와 상대 유격수 하주석의 실책으로 점수 차를 벌렸고, 백상원의 볼넷으로 1사 1, 3루 상황을 만들고 조동찬의 희생플라이로 점수 차를 벌렸다. 2 회 초 한화는 이성열의 2점 홈런으로 2점을 따라잡았다. 2사 1, 2루 상황에서 김태균의 1타점 적시타로 승부를 제자리로 돌려세웠다. 3회 말 삼성은 1사 1, 2루 상황에서 백상원의 안타로 1사 만루 상황을 만들었으나 이후 조동찬의 병살타로 공격권을 넘겨주며 주자를 불러들이는 데에는 실패했다.

4회 초 한화는 2사 2루 상황에서 이용규의 1타점 2루타로 선취 1득점했다. 4회 말 삼성은 이지영의 1점 홈런으로 경기를 원점으로 돌려세웠다. 2사 3루 득점찬스를 맞이하였으나 이승엽의 2 루수 땅볼로 공수교대가 이루어지며 점수를 내지 못했다. 그 후 5회 말에는 무사 2루 상황에서 상대 투수 박정진 야수선택으로 무사 1, 3루 , 백상원의 볼넷으로 무사 만루 상황을 만들었으나 이후 조동찬의 삼진, 이지영의 삼진과 김재현의 중견수 플라이로

주자를 불러들이는 데에는 실패했다. 또 6회 말에는 1사 2루 상황에서 이승엽의 볼넷으로 1사 1, 2루 , 최형우의 볼넷으로 1사 만루 상황을 만들었으나 이후 박한이가 땅볼로 출루 후 진루하던 배영섭이 홈승부 실패로 아웃되고, 백상원의 2루수 플라이로 점수를 내지 못했다.

7회 초 한화는 무사 1, 3루 상황에서 로사리오의 1타점 2루타로 1점 앞서나가기 시작했고, 양성우의 2타점 적시타로 3점을 달아났다. 7회 말 삼성은 김정혁의 1점 홈런으로 1점을 따라잡았다. 8회 초 한화는 1사 2루 득점찬스를 맞이하였으나 정근우의 3루수 플라이와 이용규의 2루수 라인드라이브 아웃으로 달아나지 못했다. 9회 말 삼성은 무사 1, 2루 상황에서 상대 실책으로 무사 만루 상황을 만들고 배영섭이 병살타를 기록했지만 3루주자의 홈인으로 1점을 따라잡았으나 역부족이었다.

경기 결과 한화는 아슬아슬하게 삼성을 이겼다. 오늘 경기 결과에 따라 한화는 이번 시리즈 삼성과의 경기를 스윕으로 장식했으나 현재 10위(승률 0.373) 리그 최하위이다. 한편 삼성은 현재 6위(승률 0.472)이고 12안타 2홈런 7볼넷 5타점 7득점으로 좋은 플레이를 보여줘 결과를 아쉽게했다.

Interview Questions for News Experts (Journalists)

알고리즘이 자동으로 데이터를 분석-조합해 뉴스 기사를 생성하는 로봇 저널리즘 시스템에 대한 연구를 진행하며 뉴스 전문가 여러분의 가감없는 의견과 평가를 듣고자 합니다. 다시 한번 시간과 노력에 감사드립니다.

1. 본 시스템의 기사는 뉴스 기사가 갖춰야 할 요소를 (빠짐없이) 갖추고 있나요?

2. 본 시스템은 개인화된 기사를 인터랙티브하게 만든다는 기획 의도에 부합해 동작하나요? 개선이 필요하다면?

3. 기자의 기사와 비교해 시스템의 기사가 보다 나은 점과 잘 못하고 있는 점은?

4. 본 시스템이 생성한 기사는 뉴스 기사로서의 가치가 있나요?

5. 이런 시스템에서 놓치고 있는 뉴스의 가치가 있다면 무엇입니까? 앞으로 반영해야 할 저널리즘의 기능적-산업적 가치가 있다면?

6. 데이터를 보다 폭넓게 사용하거나(통시적), 풍부한 인터랙션 방식을 제공하거나, 개인적인 관점에서 기사를 작성하는 본 시스템의 방식은 기사의 뉴스 가치를 올리는 일인가요?

7. 기자가 기사를 쓰는 방식과 비교해 개선이 필요한 점은 무엇입니까?

8. 편집자(데스크)와 같이 뉴스 조직의 다양한 역할 중 본 시스템에도 필요해 보이는 산업적-기능적 요소가 있나요?

9. 언론사 조직에 적용될 수 있는 시스템이라고 생각하나요 (e.g. CMS 툴)? 이런 시스템이 도입된다면 뉴스 기사를 작성하는 방식이 어떻게 달라질 것으로 기대하나요?

10. 본 시스템은 스포츠 이외에도 데이터를 기반으로 기사를 작성하는 다양한 분야에 적용할 수 있는 범용적인 툴을 목표로 만들어졌습니다. 이에 예상되는 본 시스템의 한계와 제한점은?

Interview Questions for System Experts (UI Designers)

1. 본 시스템의 인터페이스는 조작하기에 편한가요? 인터랙션 방식은 직관적인가요?
2. 인터페이스는 기대대로 동작하나요? 학습이 필요하거나 불친절 하지는 않은가요?
3. 인터페이스를 사용해 뉴스 기사를 만드는 데 필요한 기능이 충분히 제공되고 있나요? 기능적 완성도는 어떤가요?
4. 인터페이스 디자인은 일관적으로 구현되었나요? 디자인적 완성도에 대한 의견을 남겨주세요.
5. 추가로 필요한 인터페이스 콘트롤, 인터랙션, 기능 등이 있나요?
6. 사용자가 수행하려는 작업을 적절히 지원하나요? (원하는 데이터의 탐색과 선택, 기사 생성 등)
7. 초보자와 전문가의 다른 니즈를 적절히 지원하나요?
8. 본 시스템은 스포츠 이외에도 데이터를 기반으로 기사를 작성하는 다양한 분야에 적용할 수 있는 범용적인 툴을 목표로 만들어졌습니다. 이에 예상되는 본 시스템의 한계와 제한점은 무엇인가요?
9. 본 시스템은 사용자에게 즐거운 사용 경험을 제공할 것이라 생각하나요?
10. 본 시스템을 사용해 본 전체적인 경험에 대한 평가를 남겨주세요.

국 문 초 록

알고리즘에 기반한 개인화되고 상호작용적인 뉴스 생성에 관한 연구

김동환
언론정보학과
서울대학교 대학원

알고리즘은 온라인 상에서 생성되는 데이터가 폭발적으로 증가하고 컴퓨테이션 기술이 발전하는 추세에 맞춰 많아진 데이터를 사람의 손을 거치지 않고 자동으로 모으고 분석하는 역할을 한다. 특히 기사의 작성에 필요한 데이터의 수집에서, 분석, 처리, 작성, 배포에 이르는 저널리즘의 모든 과정에 걸쳐 알고리즘의 사용이 늘어나며 콘텐츠의 작성에 있어 그 비중은 점차 늘어나고 있다. 미국 일간지 로스엔젤레스 타임스는 지난 2014년 L.A. 근교 웨스트우드(Westwood)에서 발생한 강도 4.7의 지진을 퀘이크봇(Quakebot)을 사용해 자동으로 발행하며 알고리즘에 기반한 기사 작성의 서막을 열었다. 미국 AP통신(The Associated Press)과 경제전문지 포브스(Forbes) 역시 2014년 부터 알고리즘에 기반해 콘텐츠를 생성하는 기술 업체와의 협력을 통해 알고리즘 기사를 쏟아내고 있다. 워싱톤포스트(The Washington Post)는 2016년 리우 올림픽의 소식을 빠르게 보도하기 위해 헬리오그래프(Heliograf)라는 로봇을 활용했다고 보도한 바 있다.

본 연구에서는 알고리즘이 뉴스 기사의 작성에 활용된 여러 뉴스 기사를 검토해 알고리즘이 가진 공통적인 속성과 한계점을 도출했다. 이를

통해 알고리즘의 콘텐츠 생성 능력을 극대화시키는 세 가지 핵심 개념을 도출해 알고리즘 프레임워크의 개발에 사용했다. 이 세 가지 개념은 1) 알고리즘의 연산에 사용되는 데이터의 종류와 폭을 넓히는 통합 데이터베이스(consolidated database)의 구축, 2) 독자의 컨텍스트를 반영해 콘텐츠를 생성하는 개인화(personalization), 3) 독자가 뉴스 시스템의 유저 인터페이스(user interface)를 직접 조작해 기사를 만들고 소비하는 상호작용적(interactive) 뉴스 시스템을 만드는 것이다.

본 연구에서는 이와 같은 핵심 개념을 바탕으로 보다 개인화된 관점과 상호작용적인 방법으로 뉴스 기사를 생성하는 새로운 알고리즘 프레임워크를 제안하며, 이를 실제 동작하는 뉴스 시스템으로 개발한 PINGS (Personalized and Interactive News Generation System)를 소개했다. PINGS는 데이터 통합(data synthesizing), 내러티브 관점설정(narrative framing), 뉴스 프레젠테이션(news presentation)의 세 가지 단계를 거쳐 독자 개인에게 맞춰진 기사를 생성하는 뉴스 시스템이다. 본 연구에서는 PINGS의 유저 인터페이스 디자인 및 시스템 개발 과정과 동작 방식을 소개한다. 또한, 116명의 일반 사용자를 통해 PINGS가 생성한 개인화된 콘텐츠의 뉴스 가치를 평가하고, 뉴스 및 시스템 전문가와의 인터뷰를 통해 시스템에 대해 평가하는 사용자 실험을 통해 본 연구가 가진 디자인적 함의와 의의를 도출하였다.

주요어 : 알고리즘 저널리즘, 로봇 저널리즘, 뉴스, 프레임워크, 통합데이터, 개인화, 상호작용, HCI, 인터페이스 디자인, 시스템 개발

학 번 : 2012-30846