

VISUAL OVERVIEWS OF ONLINE NEWS STREAMS

A PROJECT REPORT

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER SCIENCE

UNIVERSITY OF REGINA

By

Kenneth Emeka Odoh

Regina, Saskatchewan

January 2016

Copyright © 2016 — Kenneth Emeka Odoh

Abstract

The volume of news generated on the Internet is increasing faster than our ability to read the arriving news. The list-based representation found in existing news interfaces does not work well with the manner in which news is now streamed to us. Readers miss out on recent news as more updated news arrives. This motivated the design of a news dashboard following the principles of information visualization and visual analytics. The news dashboard provides a succinct presentation of the news at a glance. The topics and named entities in the news dataset are used to organize the visual display and to automatically identify related news, trending news, breaking news, and support the faceted browsing of the news stream.

Acknowledgments

I want to thank my supervisors who are Dr. Orland Hoerber and Dr. Brien Maguire for agreeing to supervise this work. My supervisors provided immense support to guide my research with generous funding for the duration of my graduate studies. I especially want to thank the Faculty of Graduate Studies and Research for providing some of the funding during my studies. I am also grateful to my mother Mrs. Felicia Odoh for her prayers, advice, and good wishes. She provided enormous emotional support during the difficult times in my graduate studies.

I would like to thank my co-research assistants which include Maha El Meseery, Khantil Patel, and Radhika Gopi. The research group were very helpful and supportive in the research work as we shared the same office space. I would like to thank the doctoral candidate, Ziyuan Gao at the University of Regina for his support in explaining some statistical concepts that helped to complete the project.

Dedication

This project is dedicated to my family for providing emotional support which proved invaluable in the completion of this work.

Contents

Abstract	i
Dedication	iii
Table of Contents	iv
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Core Problem	2
1.3 Approach	3
1.4 Organization of Report	4
Chapter 2 Theoretical Foundations and Frameworks	6
2.1 Information Visualization	6
2.2 Dashboards	9
2.3 Information-Seeking Mantra	10
2.4 Visual Analytics Mantra	12
2.5 Information Foraging Theory	15
Chapter 3 News Interfaces	19
3.1 List-Based Interfaces	19
3.2 Timeline-Based Interfaces	21
3.3 Map-Based Interfaces	22
3.4 Relationship-Based Interfaces	23
3.5 Matrix-Based Interfaces	25
Chapter 4 Design and Implementation	27

4.1	Requirements	27
4.2	News Streams	29
4.3	Dashboard Design	30
4.3.1	Analyze First	34
4.3.2	Show the Important	36
4.3.3	Zoom, Filter	41
4.3.4	Analyze Further	43
4.3.5	Details on Demand	44
4.4	Theoretical Foundations and Interactions	45
4.4.1	Dashboards	46
4.4.2	Information Foraging Theory	46
Chapter 5	Case Studies and Discussion	49
5.1	News Dashboard Scenario	50
5.2	Search as an Alternative	58
5.3	Discussion	63
Chapter 6	Conclusions	66
6.1	Primary Contributions	67
6.2	Limitations	68
6.3	Future Work	69
	References	71

List of Figures

2.1	A chart that shows the Napoleon’s march to Moscow [67].	8
2.2	An example of a dashboard [18].	10
2.3	The default view of Flamenco browsing interface [28].	12
2.4	The sense-making loop for visual analytics [35].	14
2.5	Multidisciplinary nature of visual analytics showing the complementary nature of human and machine [35].	15
2.6	The VisGets interface [16].	15
2.7	Google News [3] showing the news about the “Syrian refugee”. . . .	18
3.1	Google News website showing the list of the news [3].	20
3.2	ThemeRiver (showing the thematic changes in the news stream) [23].	21
3.3	NewsStand (a map-based interface) [61].	23
3.4	SolarMap (showing the relationship between entities in the collection of documents) [10].	24
3.5	News stream visualization based on a matrix-based interfaces [42]. . .	25
4.1	A data stream management system [55].	29
4.2	System architecture.	31
4.3	The addition of descriptive labels to the components of the news dashboard.	32
4.4	Default view of the News Dashboard.	33
4.5	The processing pipeline of Signal Media Limited data [46].	35
4.6	A description of the stacked heatmap using “Weeks vs Days” time aggregation.	37
4.7	Types of signal changes: abrupt transient shift (A), abrupt distributional shift (B), and gradual distributional shift (C) [12].	39
4.8	The view showing the breaking news in the news dashboard.	40

4.9	Faceted browsing interface that shows the list of facet terms.	42
4.10	The view showing the related news feature in the news dashboard.	45
5.1	The default view of the news dashboard.	51
5.2	The view of the news dashboard excepting the news window.	52
5.3	The news dashboard showing the topic “Tax Simplification”.	53
5.4	The news dashboard showing the topic “Legal Political”.	53
5.5	The list of terms in the facets.	54
5.6	The list of terms in the facets and stacked heatmap, when “Apple” is clicked in the facet.	55
5.7	Screenshot showing the tooltip of the news stream.	56
5.8	Screenshot showing the breaking news on display.	57
5.9	The view showing the related news feature in the news dashboard.	58
5.10	The default view of the Google News [3].	59
5.11	The related news features of the Google News.	60
5.12	List of terms in faceted navigation of Google News.	61
5.13	List of terms in faceted navigation of Google News. “Apple” is clicked in the facet.	62
5.14	The scrolled view of the Google News [3].	63

Chapter 1

Introduction

1.1 Motivation

The Internet revolution has caused significant changes in the manner of news dissemination. The proliferation of media organizations and content providers have contributed to the massive rise in the volume of online news generated on a daily basis, challenging the ability of readers to process the news in order to recognize and understand trends in the news stream. The number of users seeking to read online news keeps increasing [48], thereby motivating the need for research into the development of news interfaces that allow users to analyze and explore increasing amount of news with minimal effort.

The design of a news interface begins with an understanding of the news data. A news article is a unique type of textual data, as it is authored by journalists and therefore has a good rhetorical structure with minimal grammatical errors [21]. News articles are a rich source of information about people, organizations, and locations [49], as they provide reports on activities in the world [21]. Humans are craving information about the happenings in the world and expect news interfaces that are suited to their information needs.

Online news has become a major source of information for people in the information age. The news is curated by aggregator sites and organized using the list-based representation of the traditional newspaper format. Aggregator sites collect news articles from diverse news sources, which are merged into a single source for redistribution to the users [43]. There are a number of aggregator sites which include Google News [3] and Yahoo News [6]. There is a growing trend among Internet users to get

their daily news from aggregator sites, despite the lack of original content in these sites. These aggregator sites are disruptive to the traditional news sources such as television and radio, as the users are increasingly turning to the aggregator sites as their primary source of news thereby threatening the user base of traditional media.

Online news interfaces have practically remained unchanged for years. The interface for displaying the news has been based on the interface metaphor of the newspaper where the traditional newspaper organizes the news according to different sections in a list. Existing news interfaces have used techniques in graphic design to enhance the aesthetic feel of the interfaces. However, the problem with existing interfaces is the limited availability of support for exploratory tasks, thereby making it difficult to perform visual exploration of the news in relation to past relevant news. Due to the volume of the news stream, it is not feasible to quickly read all the news content in the stream at a glance. The users could benefit from an interface that provides a high level overview of the news stream which summarizes the news stream thereby providing support for the information seeking activities of the users.

1.2 Core Problem

Users are overwhelmed with the deluge of news [48] delivered by multiple aggregator sites at an increasing rate [43]. The problem is exacerbated by the list-based representation of existing news interfaces which are intuitive to users, but create the problem of occlusion which occurs as recent news overlaps possibly with relevant past news. The problem of occlusion occurs when an item obstructs another from the field of view. In this case, posting old news off the bottom of the list.

There have been attempts to solve the issues with existing news interfaces. This has given rise to the development of notification schemes [29, 68] that make users aware of past news in relation to recent news. This approach works on a small scale, but the readers can be overwhelmed by the frequency of the alerts and pop-up windows. Subsequently, the approach fails to solve the issue of occlusion, which results in the user losing focus on relevant news as new data arrives.

The core research question in this project is “Can we do better than the list-based representation of a news stream as a primary news interface?”. This question has guided the research in the design of a novel interface to support real-time visualization

of the news streams. The aim of news visualization is to develop a news interface that is efficient and effective in visually displaying the news. Users are not only interested in reading the latest news, but also wish to track the temporal evolution of the news as relevant old news is overlaid with current news. The need for a compact representation of the news stream can be achieved by the design of a dashboard [18].

Due to the volume of news, people tend to think of news as items that can be read once and discarded. This is not always the case, as people want to track news about specific events over time or even catch up on news after a hiatus. The understanding of long-term trends in the news stream requires focused attention, as the users need more than glances to get the full story. In media such as television, this requires a panel of experts to analyze the news trend over a time period. This panel of experts is not available in an online setting. This calls for the interactive interface to support the visual exploration of the news stream.

1.3 Approach

The aim of this research is to develop a tool that supports real-time interactive exploration of news stream. However, the ability of the users to understand the trends in the news stream is dependent on a number of factors which include how the data is presented using visualization. This tool should provide a compact visual representation of the news stream using the principles of information visualization [69] to allow the users to pre-attentively process the news stream. Compact visual representation shows the information using a small display space, and as such has a high information density.

For this research to achieve its set goal, there is a need to understand the requirements of the user based on a task analysis. This is an expectation when designing an interactive visualization tool where the conceptual model of the designers must match the mental model of the users [54]. The role of the users in visualization can be either exploratory, confirmatory, or presentation [69]. In the exploratory role, the user can check for the presence or absence of an item of interest. In the confirmatory role, the user can verify a hypothesis. In the presentation role, the user can display the information to the users to facilitate understanding of the data.

The theoretical foundations can be classified into theories that supports the design

and use of the news dashboard. This captures all aspects of the tool, considering both how the user interacts with the system, and how the designers developed the tool. This work proceeds with the review of existing news interfaces, thereby providing the basis for the design of a novel interface.

This research work consists of designing a news dashboard that uses a stacked heatmap to show the temporal evolution of the news stream using a perceptually ordered colour scale that shows the frequency of the news within each element of the stacked heatmap. The stacked heatmap provides a 2-D representation of time, where the news stream is aggregated on the horizontal axis, which represents the width of the element of the stacked heatmap within the set interval on a macro-level, and organizing the data on the vertical axis which represents the height of the element of the stacked heatmap based on a second temporal tier on a micro-level. For example, if the news is aggregated by a days versus hours, then the horizontal axis shows the days of the month, while the vertical axis can show the data aggregated by hours. This allows the user to see the news aggregated by hours on a day in the stacked heatmap. Additionally, the news dashboard has a faceted browsing interface to explore the news stream by using recognition of terms, without the need to recall these from memory. The faceted browsing interface allows for enhancing the learnability of the dashboard by new users as the users do not need training to use the news dashboard for exploration of the news stream. Furthermore, this research has culminated in the design of a breaking news detection approach that is based on anomaly detection algorithm [12].

The interface should support data exploration where the users can form a visual map of the news stream by making use of the pre-attentive processing [25, 69] of the news dashboard. In this work, a novel interface was proposed that supports the real-time visualization of news streams in a compact visual encoding.

1.4 Organization of Report

The remaining chapters are organized as follows.

In Chapter 2, a review of theoretical foundations is provided to guide the design of the news dashboard. This provides the core principle for the design of an interactive news dashboard.

In Chapter 3, a review of existing news interfaces for news visualization is provided. This provides the basis for designing a novel news interface.

In Chapter 4, the news dashboard for real-time visualization of news streams is developed in the research. This Chapter discusses the user requirements, algorithms, and implementation details required for designing the news dashboard. This design of the news dashboard was influenced by the theoretical foundations explained in Chapter 2, and the knowledge of a number of existing news interfaces in Chapter 3.

In Chapter 5, a case study to provide scenarios for using the dashboard is provided. This is followed with a discussion of the results and suitability for deployment for public use.

Finally, in Chapter 6, the conclusions and contributions of the research are presented. The limitations of the current work are provided leading to a number of suggestions on possible future development of this research project.

Chapter 2

Theoretical Foundations and Frameworks

Visual interfaces should be designed using principles that guide the users to the information. These principles are the building blocks of interactive visualization. The theories discussed in this section include information visualization in Section 2.1, dashboards in Section 2.2, the information-seeking mantra in Section 2.3, the visual analytics mantra in Section 2.4, and information foraging theory in Section 2.5.

2.1 Information Visualization

Information visualization is the use of graphical representations of data to enhance the ability of users to gain insight from the data [69]. Visual representations allow the users to understand the data using their visual perception abilities [70]. The essence of visualization is to display a representation of the data in an understandable manner, thereby enabling the users to interactively explore and visually analyze the data [67]. Visualization can be thought of as a universal language of communication, as it allows the user to make use of their visual sensory organ to perceive information [69]. The visual variables that are commonly used in information visualization consist of position, mark, size, brightness, colour, orientation, texture, and motion [69]. The visualization is based on a number of principles that describe how humans perceive patterns and relationships. These principles are based on the colour theory, Gestalt School of Psychology [64], and pre-attentive processing [25].

An important aspect in the design of an information visualization tool is the understanding of colour theory. This is a very important visual encoding in a visualization. The abuse of colours can result in a visualization that is not informative [66].

Visual encoding of a colour scale is based on the opponent process theory [70] and perceptually ordered colour scheme [69]. The opponent process theory is where six primary colours are organized as opponent pairs along three axes: black - white, red - green, and yellow - blue [30]. This colour scheme is suitable for encoding number values that have a range of positive and negative values, when used with care. This is also suitable for qualitative data as the distinction between objects is quickly recognized [70]. For displaying numerical values that grow or decrease monotonically requires the use of a perceptually ordered colour scale [66, 70].

The Gestalt School of Psychological developed a set of abstractions for how the human brain perceives patterns and relationships of interest in information visualization and those that imply relationship, including proximity, similarity, and connectedness [40]. The law of proximity allows the collection of objects that are placed together to be perceived as a group [40]. The law of similarity allows that the objects which are visually similar (colour, shape) to be perceived as a group [59]. The law of connectedness allows the user to perceive objects that are connected by visual attributes as related, and those object not connected by visual attributes are perceived as not related [64].

Image interpretation is done in parallel with the human perception and as such supports pre-attentive processing [25], which is the subconscious assimilation of information from visual artefacts. Pre-attentive processing allows the user to use visual features to identify objects at a glance. In contrast, textual interpretation requires the use of focused attention. The goal of a good visualization to support the involuntary use of pre-attentive processing to facilitate the quick interpretation of the visualization. For all visual stimuli that users can pre-attentively processed, this process takes place in parallel in our brains. By contrast, for things we need to attend to, are proceed serially thereby resulting in the use of focused attention. Visualization is deployed in cases where relative comparison is more desirable than absolute judgement, because the human visual system has evolved to be adept at relative comparison of objects [69].

Visualization is an art that has been practised from medieval times [66, 67]. The celebrated work by Charles Joseph Minard in 1861, shows Napoleon's march to Moscow [67] in Figure 2.1. Napoleon's march [67] is represented by a chart which consists of a map and subplot that shows the temperature at specific locations during

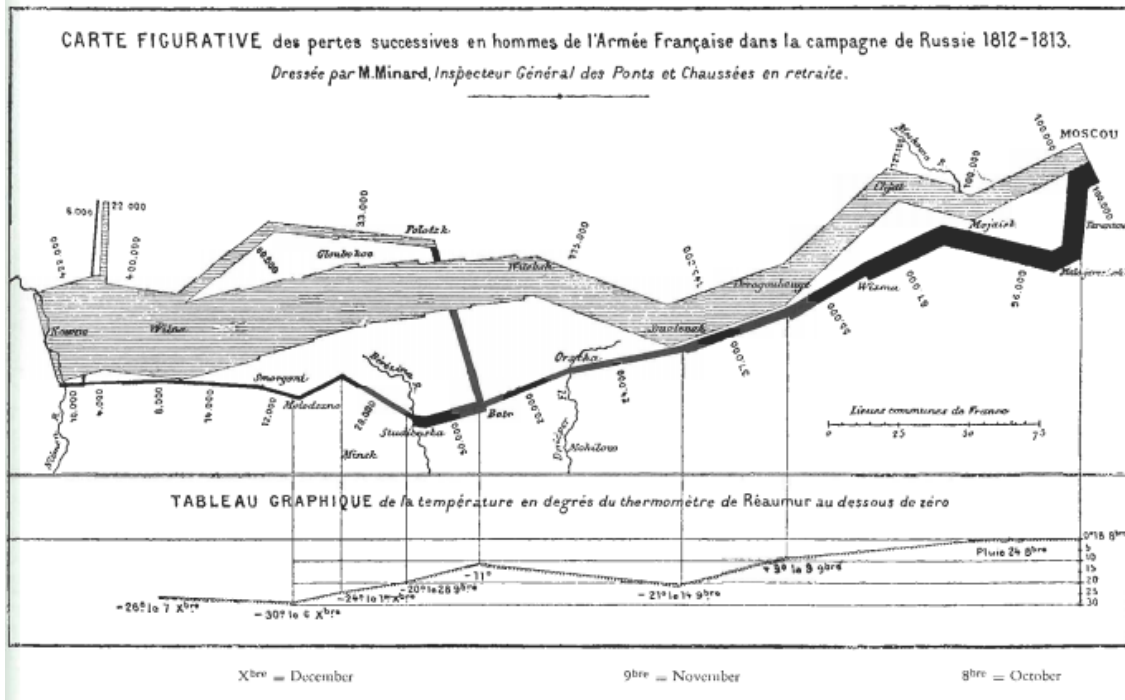


Figure 2.1: A chart that shows the Napoleon's march to Moscow [67].

the retreat. The viewers can see the number of soldiers that are alive in different geographical locations and months of the year. The physical location of the army is shown on the map and the month of the year as shown in Figure 2.1. The users viewing the chart can develop a sense of the weather conditions and distance travelled, which can explain the reasons for the deaths. Alternatively, this information can be represented in textual form such as a news report. However, it cannot be understood with a glance as it requires focused attention to read through the documents. There is also the problem of understanding the relationship between the location and weather conditions in the textual document. The information density of an image is greater than that of a textual document [67]. This information can be understood with a glance at the chart. This chart can provide situational awareness about events in the battlefield.

The problem of information overload which occurs due to the high volume of the data generated in a number of domains [63]. Information visualization provides the ability to summarize the data thereby reducing the severity of information overload [35]. Visualization can allow for visual overview of the information that give the

users a high level understanding of the data.

There are issues that arise in visualization which include distorting the data, visual clutters, and lack of understanding of colour theory. This can happen when the scale of different dimensions in the chart is altered to present misleading information [69]. The designer should avoid visual clutters which burdens the cognitive abilities of the users [67] using the principles of information visualization. Furthermore, the visualization should be done using a suitable colour coding scheme based on colour theory [25]. Colour is an important visual attribute that can be used to encode the data.

2.2 Dashboards

A dashboard is a compact visual representation that shows the information at a glance on a single screen [18], thereby allowing for situation awareness and monitoring. Due to the single-paged requirement of a dashboard, the visual representation must make efficient use of the display space. The dashboard interface has a design goal of supporting a high data-ink ratio [18] to emphasize the information on display, thereby enhancing the information content of the compact representation of the data. Coordinated views [16, 57] are a convenient way of representing the relationship among a number of different visualizations of different attributes of the same dataset, thereby allowing the user to simultaneously explore the various dimensions of the data. An application of coordinated views is for making interactive small multiples [18], which is suitable for the visualization of high dimensional data in small display area, thereby facilitating the understanding of the data by linked brushing and selection [34].

The dashboard provides a visual display that can allow the user to monitor information at a glance thereby providing situational awareness of the data. The single-paged requirement requires compact representation that provide information that meet a set criteria. The dashboard can be made to be customizable to provide support a wide diversity of users. The dashboard provides the most important information by highlighting them using visual variables to draw the attention of the users [18].

An example of a dashboard is shown in Figure 2.2 which is designed for web marketing analysis that displays the frequency of visits at specified times in the

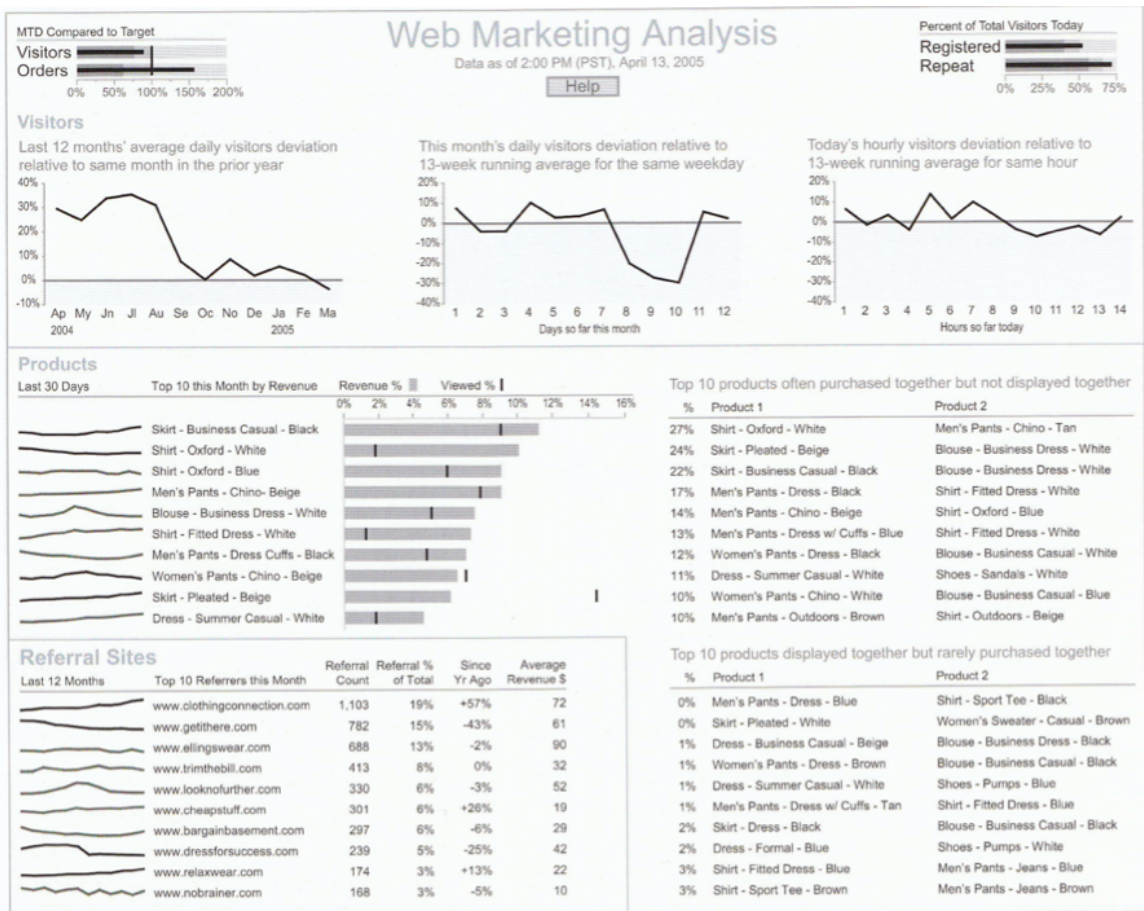


Figure 2.2: An example of a dashboard [18].

chart. The trend of different products are shown with sparklines and bullet graphs. The textual representation of the information is written in legible fonts which display the name of the products and labels for the charts. The analyst can view the health of the marketing campaign at a glance on a single-paged display.

2.3 Information-Seeking Mantra

An interface that support tasks that involve looking for information by the users can be designed by using the information-seeking mantra. This is a foundational principle that guides the design of interfaces that support information-seeking activities [63]. The mantra states “overview first, zoom and filter, then details on demand” [62]. The mantra supports the searching of documents to identify the hidden

patterns in the collection of documents [62].

The mantra is best explained by decomposing the sentence into a collection of descriptive words as modes. The interface at the “overview first” mode should provide a visual summary of the data allowing the user to form a mental model to guide his or her information-seeking activities [14]. The interface at the “zoom and filter” mode should allow for navigation into varying degrees of aggregation of the data [63]. The ability to view the visualized data at different resolutions can uncover some interesting patterns in the data. The filtering operation allows the user to focus on a smaller subset of the data to identify patterns, because it is easier to locate patterns within a subset of the data set. The interface at the “details on demand” mode should allow the users to request either the detailed view or overview depending on the information-seeking activities [14, 63] of the users.

The information-seeking mantra provides a principled approach to data exploration that works for data using an overview to show the user a high level summary of the data. The user can select a subset of the data by filtering to show a subset of the data. The user can change the resolution to view the data at different granularities. The interface can allow for changing of settings as needed to guide the exploration process to provide details on demand.

The Flamenco [16] interface is based on the information-seeking mantra shown in Figure 2.3. The interface supports faceted browsing [28] with hierarchical categories in order to fine-tune queries. The users can drill down within categories in building up queries which are user-controlled, thereby providing a step-by-step top-down approach to generating queries to guide the users towards their information needs. The faceted browsing functionality enhances the ability of the users to filter the data to obtain a subset of the data matching set criteria. The default page of Flamenco [16] provides the overview of the information, thereby allowing the recognition of the facet term, without the need to recall from memory. The user can zoom into a subset of the data for focused exploration, when required to display the details of the information on demand.

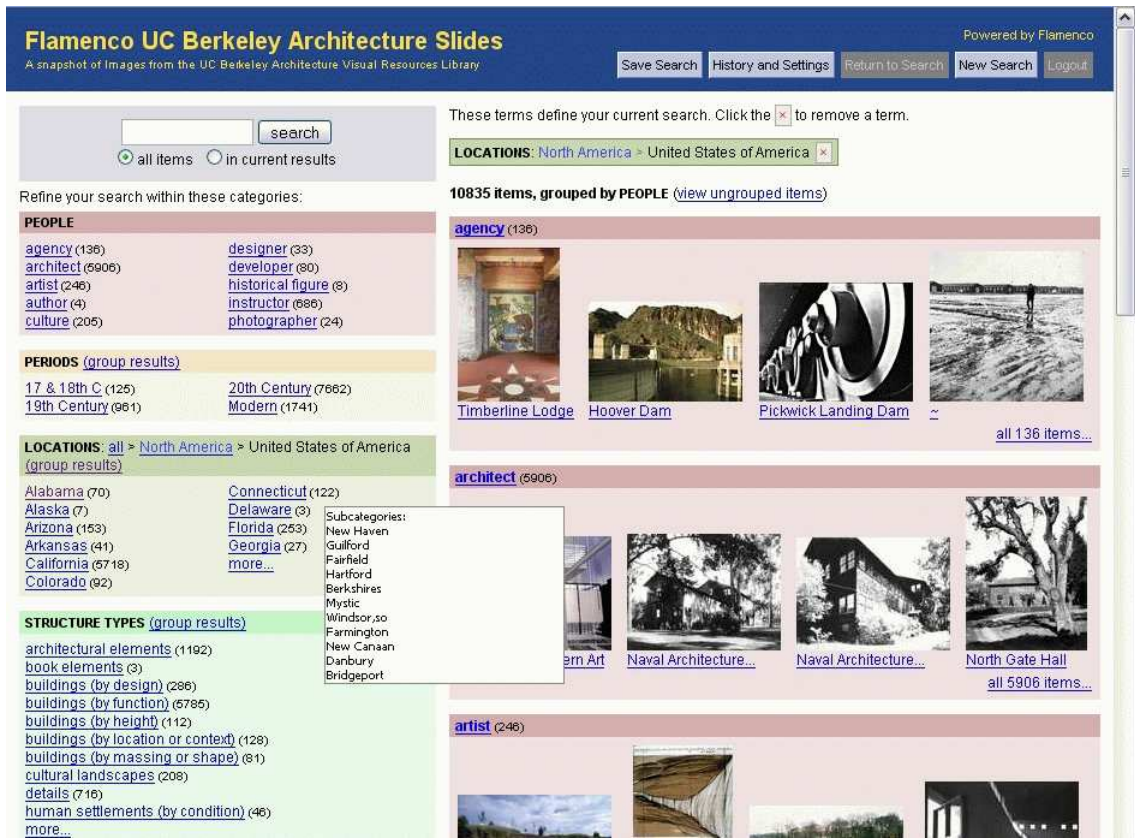


Figure 2.3: The default view of Flamenco browsing interface [28].

2.4 Visual Analytics Mantra

Visual analytics is analytical reasoning based on an interactive visual interface [65]. The core of visual analytics is to perform decision-making based on conclusions that are derived from interacting with the visual interface [36]. Visual analytics is a mechanism that combines the strengths of humans and computers for generating insight from exploration of the data. The visual analytics mantra is a modification of the information seeking mantra to support exploration of complex data sets that requires both automatic processing and user interaction to derive insights from the data [35]. The visual analytics mantra can be stated as “analyze first - show the important - zoom, filter and analyze further - details on demand” [35]. The focus of the information-seeking mantra is about resolving the information seeking goal, in contrast with visual analytics, which is focused on analyzing the data.

As with previous section, the mantra is best explained by decomposing the sentence into a collection of descriptive words as modes. The interface at the “analyze first” mode should involve some preprocessing which may include data wrangling, aggregation, machine learning, and statistical analysis which extract meaningful information from the data. The interface at the “show the important” mode is where the extracted information is visualized and interesting aspects of the data are highlighted for emphasis. At the “zoom, filter and analyze further” mode, the user can view the data at varying levels of granularity. The system can allow for further analysis based on the request of the user. This is the phase where the data analysis occurs as the processing can provide more information to the user. The interface at the details on demand mode should allow the users to request a view with varying amount of details based on information-seeking activities [63].

Visual analytics is multidisciplinary in nature with ideas from information visualization, statistics, and data mining [35]. This can be applied in several fields where professionals want to derive insights from complex data sets [7]. Visual analytics allows the users to confirm the expected and discover the unexpected from the data set [36], and allows the users to view the relationships between entities in the data set [35]. Visual analytics differs from information visualization with the inclusion of exploratory visualization, where analyses are performed at an initial phase and further analysis can be done on demand to support the decision-making of the user.

Visual analytics is a better approach than a purely automated method for data analysis. Automated methods are appropriate for managing well-defined problems, conversely, visual analytics is suitable for solving ill-structured problems [7] which are common in reality and require human intervention to drive the exploration. These problems require human input to support the task of finding a working solution, which has necessitated the need for interactivity in the interface, thereby allowing the user to support the derivation of insight during the visual exploration. Interaction can be enhanced by supporting the visual exploration of the data using techniques such as dynamic projections, filtering, zooming, linking, and brushing [34].

The visual analytics process can be described in three steps, which are preprocessing, summary, and visualization [35]. The preprocessing step includes performing data cleaning, and data wrangling. The summary step involves the use of statistical, machine learning, and data mining methods to obtain the extraction of information

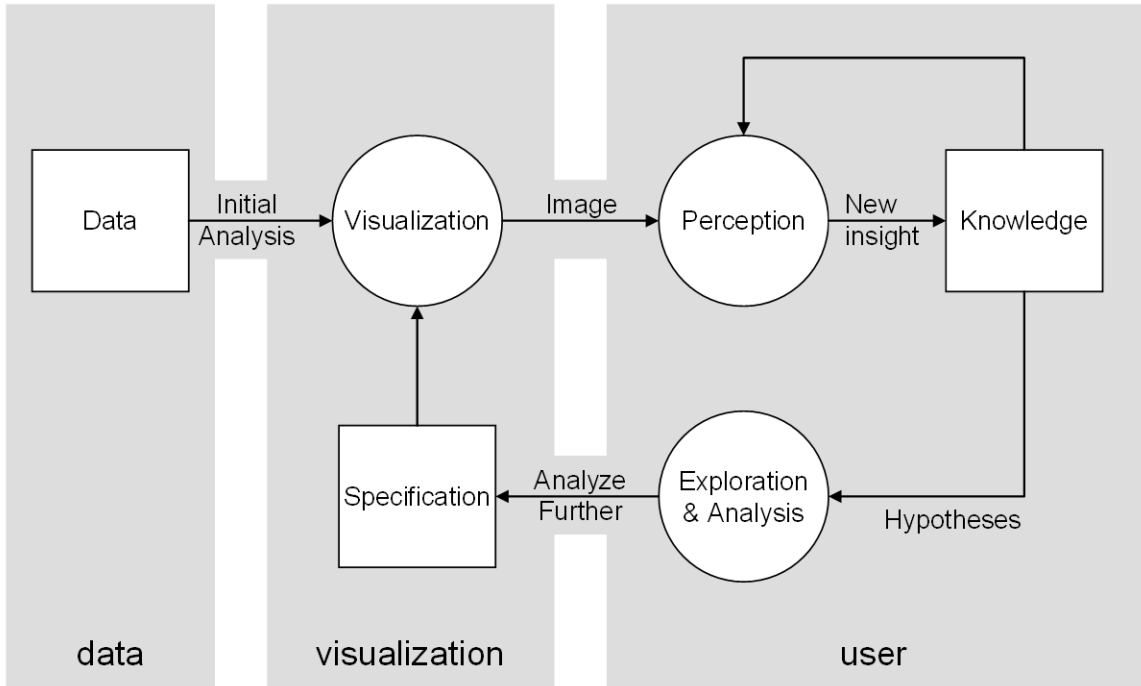


Figure 2.4: The sense-making loop for visual analytics [35].

from the data which can be rendered in the visualization stage [35, 36]. Visual analytics is a process that guides the users to exploring the data in order to gain insight from the data. This process of sense-making is shown in Figure 2.4, which shows the relationship between the data, visualization and user. The visualization is done using the data, while the user explores to understand the data using the visualization thereby resulting in knowledge for the user.

There is a proliferation of ill-structured problems in real life [7]. Visual analytics will provide the right framework for solving hard problems where domain knowledge would be incorporated in the solution. The versatile applicability of visual analytics requires knowledge from multiple disciplines to solve hard problems [35]. The scope of disciplines that can be used in visual analytics is shown in Figure 2.5

The VisGets [16] interface is an example of a visual analytics tool which consists of a bar chart, map, sliders, and word clouds as shown in Figure 2.6. The user can get the location from the map, and make a connection to the actual term in the word clouds. The slider allows the filtering of the data based on the time extent. Multiple coordinated views can allow the user to explore the data through different

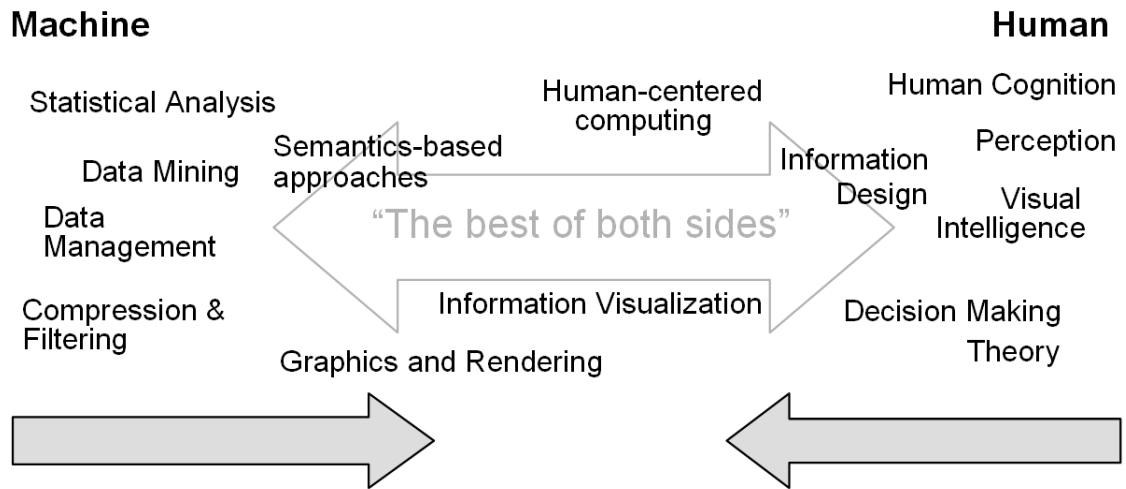


Figure 2.5: Multidisciplinary nature of visual analytics showing the complementary nature of human and machine [35].

representations, thereby providing the users with an in-depth understanding of the data to the user. This has been used to enable interaction by linking views to allow for visual exploration of the data.

2.5 Information Foraging Theory

Information foraging theory is useful for developing an information-seeking interface as it is based on the analogy of optimal foraging [52]. The optimal foraging

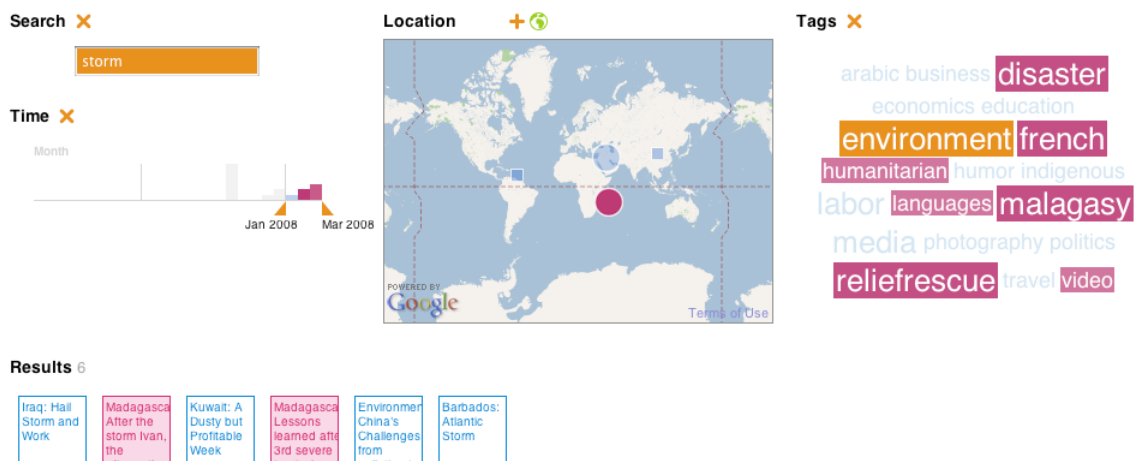


Figure 2.6: The VisGets interface [16].

analogy is a biological phenomenon where an animal scavenges for food in a field. The animal tends to continue searching for food in the field until a point is reached where it is counter-productive to continue searching for food in the same field. The animal has to make a decision to leave the field in search of greener pastures as the point of diminishing returns is reached in the current field.

Information foraging theory can make use of the enrichment versus exploitation tradeoff. The enrichment is the process of moving from one patch to another. The exploitation is the process of maximizing the amount of food that can be grazed in the given patch. The animal must optimize to obtain a sweet spot between the enrichment and exploitation of a given patch. The users can minimize the cost of going between patches, while maximizing the amount of food that can be found on a patch [52, 53].

Animals use scent to guide them to food sources. Humans make a prediction about expected benefit, before choosing an exploratory path. The animal can follow the scent to the food source. Information scent is the property of an item that attracts the attention of the user to the source. This involves moving to a productive patch with the abundance of food [53]. Information flow can follow the patch model, where the data is clustered in some area and the users can navigate to the patch to find what they need. The animal has to make the decision whether to continue grazing on a patch or to move to the next field with the hope of getting more food [52], where the scent can show the amount of information in the patch.

The foraging of animals can be likened to user performing task in an interface. People adjust their behaviour with the aim of absorbing maximal information with minimal effort [53]. The interface should be designed to support information-intensive tasks [53], as users wander in the direction of their loosely defined goal [52]. There is a tradeoff between cost and benefit which is evident in the interface [52, 53]. The organization of the information on the interface can influence how the user can search for information. The user can become frustrated if they cannot fulfil their information-seeking needs and may move to use other interfaces that better suit their needs. The users decide on the use of an interface if the benefits outweigh the cost of use.

Information foraging theory is aimed at directing the users to the source of the information by understanding the relationship between people and the information environment and designing better ways to search for the information. It studies

how people interact with the information system as they proceed in achieving their information-seeking needs [53]. This allows for the creation of a model to predict the behaviour of people seeking information. The best interfaces are those that provide visual cues to the user about what they can expect by following a particular path when analyzing the data [52].

The best interface should allow the user to maximize the amount of information consumed with limited effort. The user considers the cost of finding and analyzing the information on the screen [53]. The information scent analogy can be used to describe how an interface supports the users in their tasks. The scent is imperfect information which can be likened to the process of reading a textbook. The table of contents can be the scent that guides the user to reading the needed sections for a given task.

Let us consider the news interface shown in Figure 2.7; the list of news item are the “information scent”. The interface provides visual cues in the form of information scent to guide the users by directing them to click on the news links that are relevant to their information-seeking goals. The designer of the interfaces ensure that the news can be quickly read with minimal effort.

Google

News

Canada English edition Modern Personalise

+ Refugees

Top Stories

- Russia
- Refugees
- Kabul
- Jason Heyward
- Star Wars
- Muammar Gaddafi
- Burundi
- Iraq
- In vitro fertilization
- Chicago Blackhawks

News near you

Suggested for you

- World
- Canada
- Business
- Technology
- Entertainment
- Sports
- Science
- Health
- More Top Stories

Syrian refugees now in Toronto look forward to 'beautiful future'
 CBC.ca - 12 hours ago
 For newly arrived Syrian refugee Samer Barkel, hearing his kids laugh and play after their arrival in Canada made the 13-hour journey from Beirut, Lebanon worthwhile.
 Alexander Panetta, The Canadian Press CTV News
 5 challenges faced by Syrian refugees arriving in Canada Globalnews.ca

See realtime coverage

Opinion: Canada's Warm Embrace of Refugees New York Times

Conservative critics seek details about Canada's refugee airlift
 The Globe and Mail - 9 hours ago
 Immigration Minister John McCallum knows the days are ticking by, and that key details remain up in the air, but he insists he is still hoping to get 10,000 Syrian refugees into the country by the end of the year. One day after welcoming the first ...

'My country is now your country': Canadians tell Syrian refugees #WelcometoCanada
 CTV News - 12 hours ago
 As the first plane carrying Syrian refugees arrived in Toronto, Canadians were on hand to welcome them to their new country.

Change text size for the story
 London Free Press - 6 hours ago
 "We know the community groups are going to need to be referring the refugees once they arrive to expert social services," Lockie said after the announcement at London's Cross Cultural Learner Centre, which will be a major recipient of the new fund.

Syrian refugees get warm welcome in Calgary despite Alberta's economic woes
 CBC.ca - 7 hours ago
 Naheed Gilani's investments have been pummelled by the crude price collapse, but the Calgarian says he hasn't hesitated for a moment to contribute thousands of dollars to sponsor a Syrian refugee family. "My reasoning was, 'we can't afford to wait for ..."

Welcoming refugees to province 'nothing new,' says Sask. government
 Saskatoon StarPhoenix - 7 hours ago
 Although many questions still exist about what the federal government's plan to bring 10,000 refugees to Canada by the end of the year will mean for the province, Cleary said she isn't worried.

Figure 2.7: Google News [3] showing the news about the “Syrian refugee”.

Chapter 3

News Interfaces

News interfaces are medium for displaying the news items on a screen, which contain visual representations of the news stream. There are different types of interfaces for displaying news with a number of benefits and limitations. The ideal choice of news interface should suit the single-paged constraints of the news dashboard, which requires an interface with high information density to represent information in a compact form. The knowledge of existing news interfaces have motivated the design of a novel interface that provides a compact representation of the news stream.

The news interfaces discussed in this Chapter include list-based interfaces in Section 3.1, timeline-based interfaces in Section 3.2, map-based interfaces in Section 3.3, relationship-based interfaces in Section 3.4, and matrix-based interfaces in Section 3.5.

3.1 List-Based Interfaces

An example of list-based representation is shown in Figure 3.1, which shows the news content that is filtered according to top stories organized in a list. This interface allows for personalization as the user can filter the news by locations and languages.

List-based interfaces are common in popular news media such as BBC [1], CBC [2], and Google News [3]. The news items are organized according to different themes in a list. Despite their popularity of the list-based representation in mainstream media, the problem of occlusion remains in these news interfaces. It occurs as recent news is overlaid with old news in a frequently updated news stream, which occurs when the list-based representation is used as the primary means of news visualization.

The list-based display is static, providing limited ability for the users to interact

The screenshot shows the Google News interface. At the top, there are navigation options for 'Canada English edition', 'Modern', and 'Français', along with a 'Personalise' button. The main content area is divided into 'Top Stories' and 'Suggested for you'. The 'Suggested for you' section features four news items, each with a thumbnail image, a headline, a source, and a timestamp. The first item is about Manchester United's Louis van Gaal, the second is about Mario Balotelli, the third is about Google's new CEO, and the fourth is about Dr. Dre. Below this is a 'World' section with a news item about Bangkok. On the right side, there is a 'Recent' section with a news item about China stocks, a 'Weather for Regina, Saskatchewan' section with a table of weather forecasts for Today, Wed, Thu, and Fri, and an 'Editors' Picks' section featuring 'THE VANCOUVER SUN' and several news items. At the bottom right, there is a 'Personalize this news source' section for 'Vancouver Sun' and a 'Most popular' section with a news item about Conservatives.

Figure 3.1: Google News website showing the list of the news [3].

with the news item. Due to the display constraints of the news dashboard, this may require hardcoding the number of news items to be displayed per page to keep the user focused on the subset of the explored news.

The list-based representation provides an intuitive way of sequentially organizing documents in a list. This interface can become problematic for the limited display space of a dashboard due to the large amount of screen space required for showing the news items. As more news becomes available, a scrollable list is required, which result in the user having difficulties in finding correlations between the current news in view and the updated news in the scrollable list.

The list-based visualization is not suitable for showing the relationships between

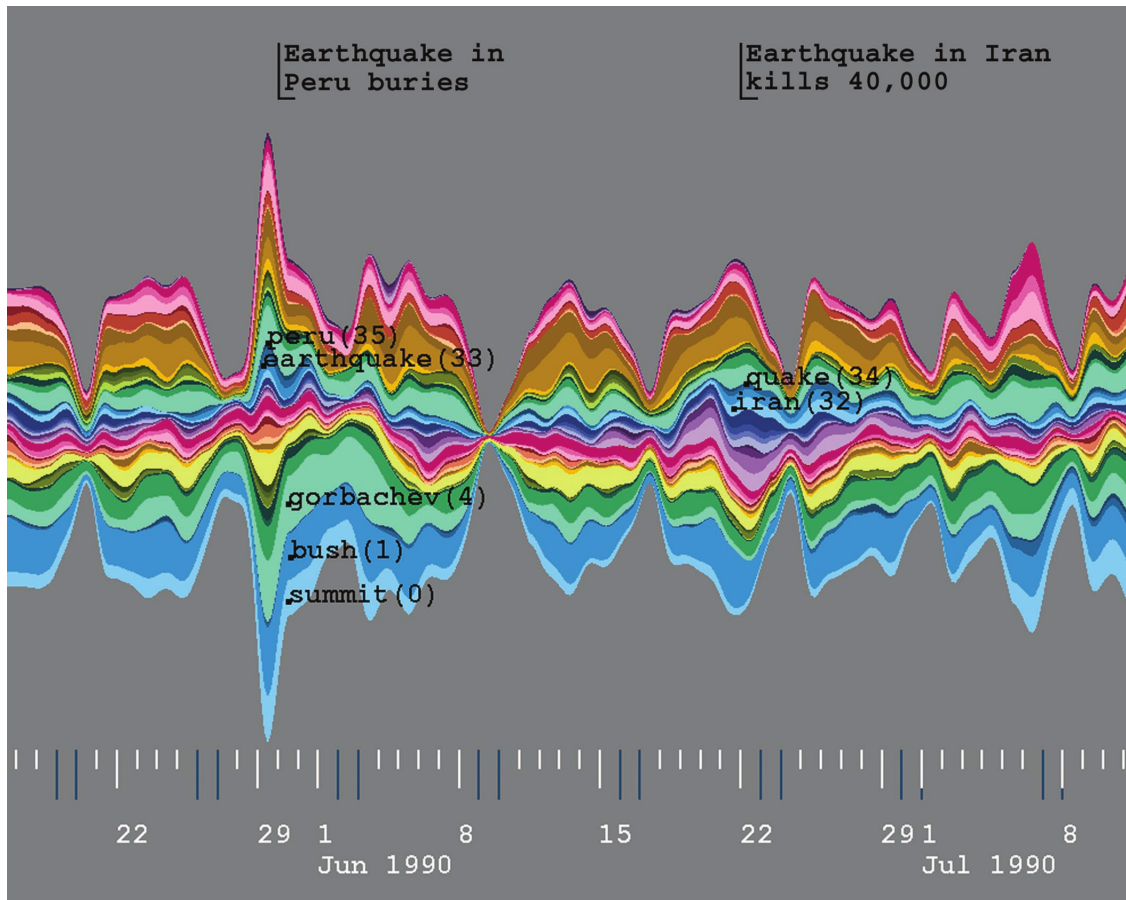


Figure 3.2: ThemeRiver (showing the thematic changes in the news stream) [23].

the news items. This is because news items can be grouped into a number of subjects, thereby making it hard to see the relationship between news in different groupings. Therefore, the list-based visualization is not suitable for uncovering relationships between the news placed in different groups in the news stream.

3.2 Timeline-Based Interfaces

An example of timeline-based visualization is shown in Figure 3.2, which shows the frequency of topics in different days of the month. This shows the collection of news about earthquakes between the months of May 1990 and July 1990. The timeline shows the popularity of topics related to earthquakes in a collection of news articles.

The timeline-based representation provides a simple way for displaying the news

content of the news stream based on timestamps. There are interesting trends in the data that will show up in the visualization due to the temporal aspect of the data. This can show how different items in the data are related by visual comparison of the trends.

There are examples of the timeline-based visualization systems that show the thematic changes in the collection of the document such as ThemeRiver [23] and TIARA [71]. There are a number of methods for event-based visualization and detection using the temporal aspect of the data. These include CloudLines [41], LeadLines [17], and the Stanford dissertation browser [13].

This interface type is ideal for showing the temporal nature of the news stream [23]. There is a limitation in timeline-based interfaces where the less frequently occurring topics are not clearly visible when overlaid with other more frequently occurring topics due to their small width in the chart. This limitation results in overemphasizing the more frequently occurring news at the expense of less frequently occurring news.

3.3 Map-Based Interfaces

An example of map-based visualization is shown in Figure 3.3, which shows the news with the context of the location using a map. This give the user awareness about the location of the news item.

The location of the news can be used to enhance the news visualization by giving users the ability to understand the news in the context of the locality. This can give the users the impression of the events happening in specific locations. A news application that uses this approach is NewsStand [61], which allows people to search for news using a map-based interface.

The users can perform both location-based and feature-based queries on the data rendered on the map [61]. A location-based query is a method of querying where the coordinates are provided and all the news matching the coordinates that are returned to the users [61]. The feature-based querying is a transformation of the location-based querying where the news are supplied and the system returns the locations matching the query [61]. The collection of news items emanating from different locations can show interesting behaviour that can reveal hidden patterns in the news stream [47].

There are a number of known issues with map-based interfaces. The reliability



Figure 3.3: NewsStand (a map-based interface) [61].

of the location attribute of the news comes into question. The location may refer to any of the following: the location of the journalist at the point of writing the news, the location of the news publication, and the location derived from named entity recognition of the news textual content. There is also the problem of occlusion when there are overlapping news items that make the news underneath the top layer inaccessible to the user.

3.4 Relationship-Based Interfaces

An example of relationship-based visualization is shown in Figure 3.4, which shows the relationship between multiple facets in a collection of documents. This tool shows the relationships between entities in different documents.

News streams are known to be a rich source of entities [49], where the relationships can be more important than the individual news items for visual analysis. These relationships between entities can provide a valuable source of information. The visualization of the relationships between entities in the news stream can reveal hidden

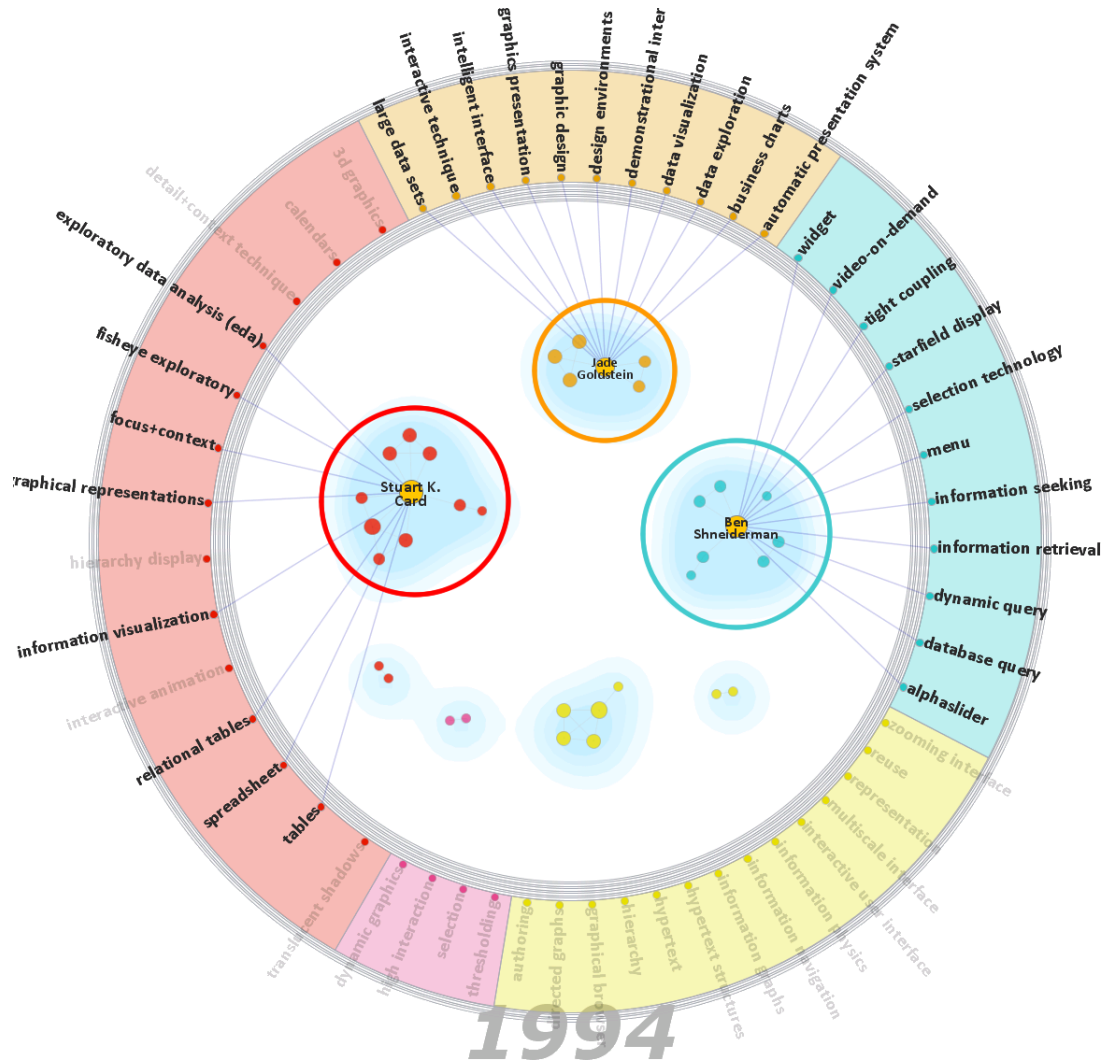


Figure 3.4: SolarMap (showing the relationship between entities in the collection of documents) [10].

patterns in the news stream, which can facilitate the realization of insight from the data.

There has been work done in the area of relationship-based interfaces which include FacetAtlas [11], SolarMap [10], and Hierarchical Chords [38] which provides a method for visualizing the relationship in the data set with hierarchical structures. This can show the relationship between metadata which can be very informative.

The relationship between each news in the news stream can create a complex graph-like structure, which can become messy and result in visual clutter which could result in a bad visualization. This requires interactivity for the user to understand the



Figure 3.5: News stream visualization based on a matrix-based interfaces [42].

information presented in the relationship-based visualization. The relationship-based interfaces can show the news that contains the same information, thereby allowing the users to make the connection between the news item in the news stream.

3.5 Matrix-Based Interfaces

An example of matrix-based visualization is shown in Figure 3.5, which shows the compact representation of the news stream. This provides an effective visualization of a collection of documents.

The matrix-based interface is suitable for visualizing large collections of documents due to its compact representation. An example of a matrix-based news interface is the Galaxy of News [56], which allows for succinct representation of the data using small display space. Matrix-based visualizations have high information density, which makes them ideal for the single-paged constraint of a dashboard.

The matrix-based visualization has inspired the design of modern news visualization systems for users that show the evolution of concepts within the data stream by

aggregating the group of news into article threads [42]. Research by Hearst shows that despite the intuitive nature of the matrix-based interface, it has not been proven to enhance the ability of the users to understand the data [27]. Therefore, the usability has to be enhanced by the inclusion of interactivity to support data exploration.

The matrix-based visualization is compact and commonly used for visualization of a collection of documents. However, the elements on display may not be obvious to the end users especially when the news stream is sparse. The visualized object on the display can be enhanced by means of visual boosting [51] to draw the attention of the user to certain news in the news stream. The matrix-based representation is used to make a compact visualization of the news stream in a small display space of the news dashboard. This was used for visualizing the news showing the temporal evolution of the news. This interface can allow the user to see and understand the trends in the news stream.

Chapter 4

Design and Implementation

This Chapter discusses the steps in building a news dashboard that meets the needs of the users utilizing the theoretical foundations and frameworks discussed in Chapter 2. A discussion of the properties of a news stream that influenced the choice of appropriate preprocessing methods suitable for the work is provided.

The Chapter is organized into different sections, where the user’s requirements are discussed in Section 4.1, the discussion of the news stream is provided in Section 4.2, the preprocessing steps are discussed in Section 4.3.1, the description of the features of the news dashboard are discussed in Section 4.3, and use of the theoretical foundations are discussed in Section 4.4.

4.1 Requirements

This section describes the tasks to be supported by the news dashboard based on an informal task analysis performed among three friends and colleagues in their mid-twenties, and as such the results may not be generalizable to the entire population. However, as the people who use the Internet as a source for news are young adults, the task analysis would generalize to the intended audience of users for the dashboard, and as such supported the development of functionalities that can allow the user to see the value in the use of a news dashboard to explore news streams. The user in this requirements is named “Jane”, who works as a top-level executive in a software company.

Requirement 1: Users want a visual summary of the news stream

Jane wants to obtain a visual overview of the news stream without the need to read through the textual data of the news. She may also want to see at what times of the day are news about a person, topic, and organization occurring, which may show hidden patterns in the data at a glance.

Requirement 2: Users want to read the news item alongside the related news, and breaking news

Jane is interested in reading news, breaking news, and related news. Jane sees a news item about the Paris attack. She would want to be notified that the news is a breaking news given that it is a rare event. Jane would also like to follow up on the development surrounding the Paris attack in the form of related news.

Requirement 3: Users want to filter by persons, topics, and organizations

Jane wants to filter for news about certain entities in the data stream. However, if Jane wants to see all the news about “Larry Page”, then she will filter by person. The retrieved result would contain all the news matching the query criteria. Jane can filter the news stream by organization. For instance, if Jane wants to see all the information about an organization named “Google”, then Jane can see all the news about the entity in the news stream.

These requirements can be satisfied by using the principles and theories previously discussed, include information seeking and the visual analytics mantra to guide the design of a news dashboard [18] that supports data exploration. The design of the news dashboard is based on theories discussed in Chapter 2. The users can form a visual map of the news stream by making use of the pre-attentive processing [25] of the news dashboard.

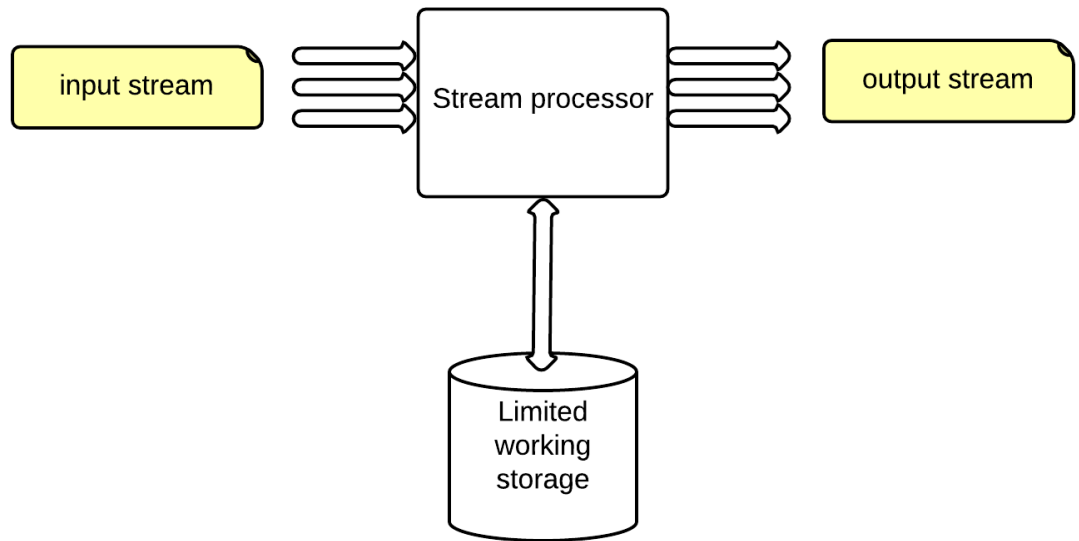


Figure 4.1: A data stream management system [55].

4.2 News Streams

News streams are ordered sequences of data with streaming properties. The stream data may arrive in an unpredictable manner, as the rate of data arrival is beyond the control of the local system because the data is coming from an external source. The nature of the data distribution can change with time in a process known as concept drift [37]. For example, a news stream containing the frequency of topics with time, may experience change in the data distribution as the popularity of topics changes in relation to events in the real world.

News streams require stream processing, which is different from the traditional database architecture where the data are saved in a data store and queried in order to retrieve relevant results. The traditional database has guarantees on the arrival of data because they are accessed locally in a database, where the rate of data arrival is limited by the speed of memory access to the data store. Stream processing is a convenient manner for handling large data sets that are too big to be saved at once in memory. A description of a data stream management system is shown in Figure 4.1. As the data arrives and a summary of the data is saved in a traditional data store, and the output stream is processed to solve the problem.

There are a number of challenges that must be addressed in order to effectively manage data streams, which include the following.

1. Time and space constraints: The data stream model requires algorithms that are efficient and process data that are not fully stored in memory.
2. Online algorithms: The algorithms for managing the data stream should have a linear time complexity to guarantee fast processing. This can allow the pattern in the data to be observed in real-time.

The understanding of the news data is the first step towards designing a news visualization system. The properties of the news data can provide a justification for the preprocessing step discussed in Section 4.3.1. This can influence the choice of the visual encoding that would convey useful information about the preprocessed news stream.

4.3 Dashboard Design

This dashboard was designed to provide a compact visual representation of the important information in the news stream on a single screen [18]. The news dashboard provides real-time visualization of a news stream using the architecture shown in Figure 4.2. This data was preprocessed by Signal Media Limited in order to extract attributes (topics, persons, and organizations) which are added as metadata to the news data. The metadata is used for organizing the display, faceted browsing interface, breaking news detection, and for finding related news.

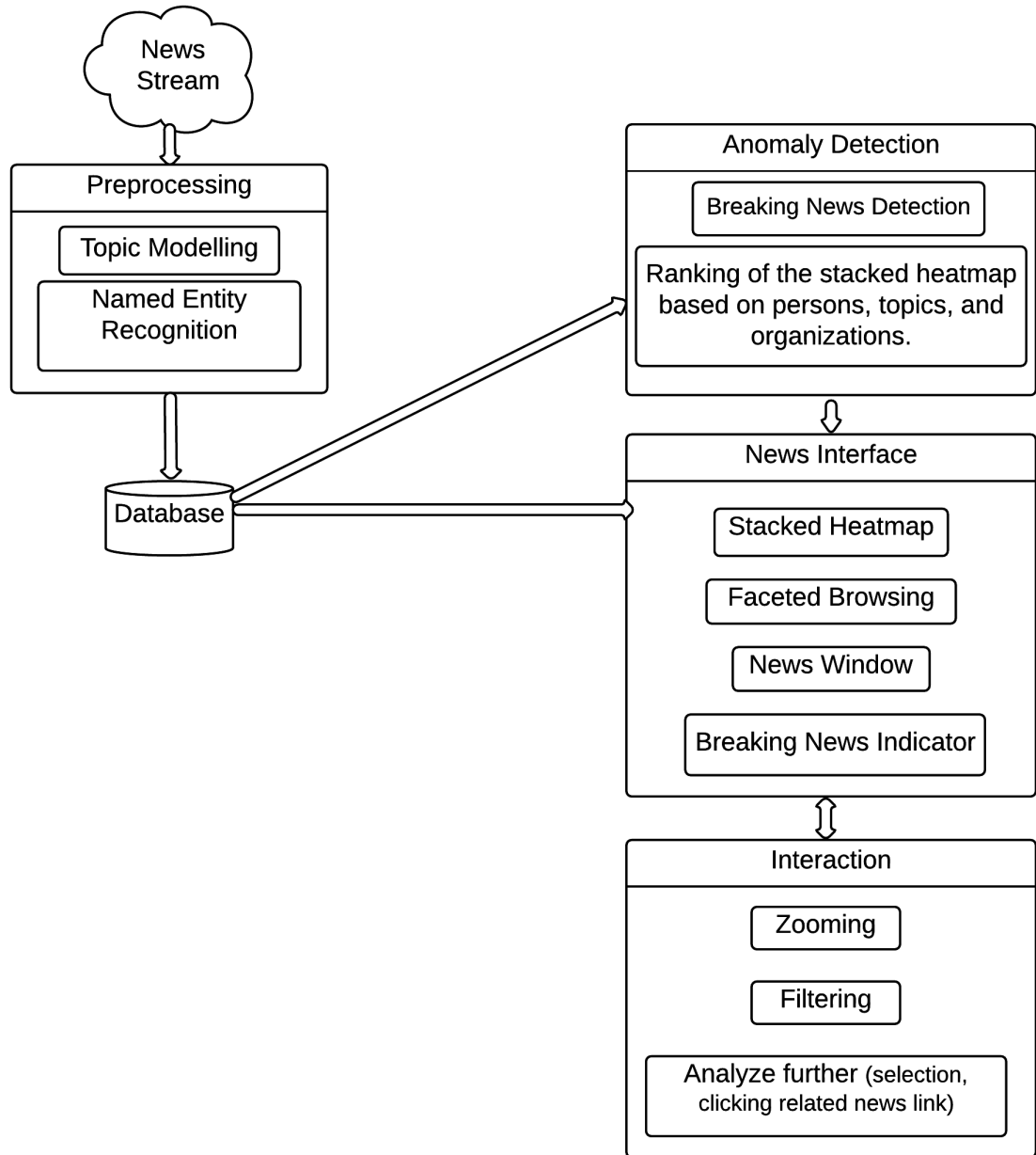


Figure 4.2: System architecture.

The news dashboard was designed using a client-server architecture where the client-side module is implemented using D3.js [9] and the server-side module was implemented using the PHP programming language. The client-side module consists of the visualization display, while the server generates the news data that is sent to

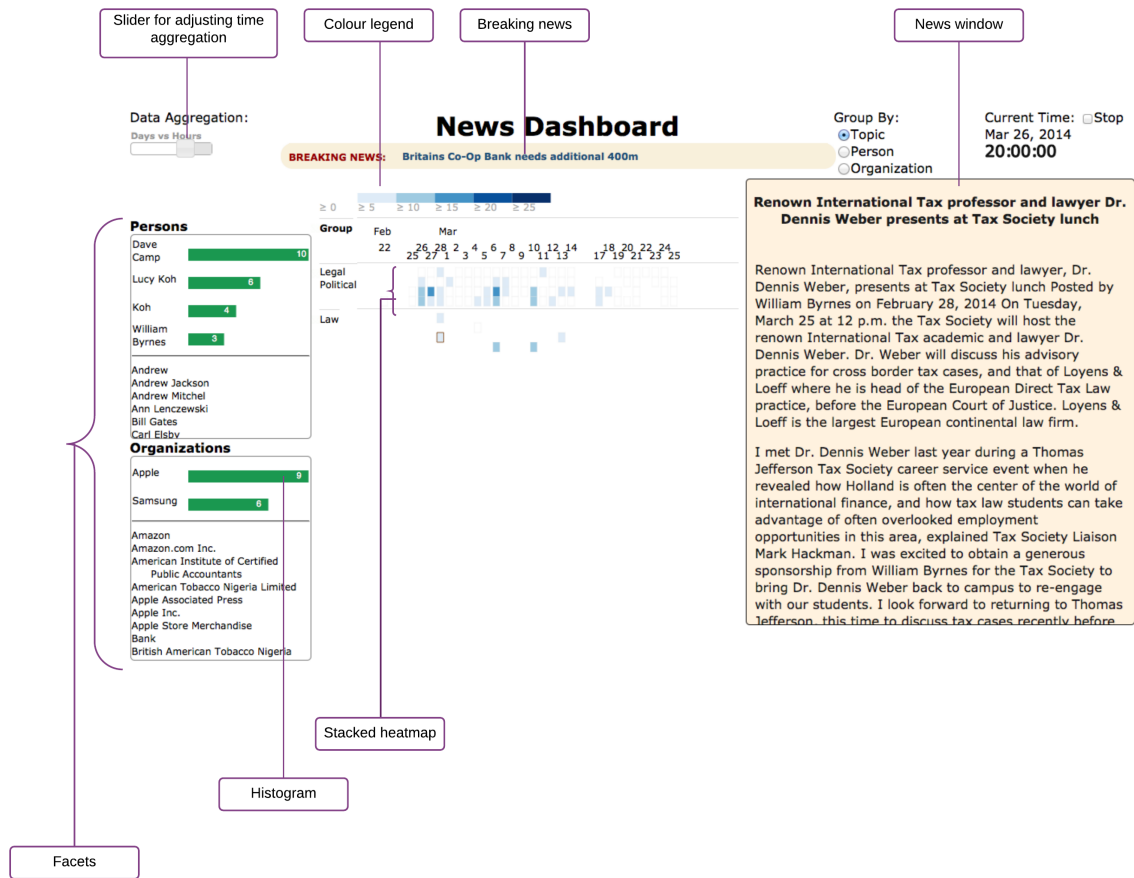


Figure 4.3: The addition of descriptive labels to the components of the news dashboard.

the client for visualization.

A screenshot of the news dashboard has been enhanced with the inclusion of descriptive labels to improve the readability of this document as shown in Figure 4.3. The labels are drawn in purple colour which is different from the colours in the background of the news dashboard in Figure 4.3.

The dashboard has a colour legend for displaying the colour scale that encodes the frequency of the news in different blocks. Underneath the colour scale is the label of the month and week to track the blocks, and the news window shows the news content in Figure 4.4. The set of controls for adjusting the visualization settings to support customization of the dashboard which consist of a slider for adjusting time aggregation, a radio button for grouping to support organizing the interface by a

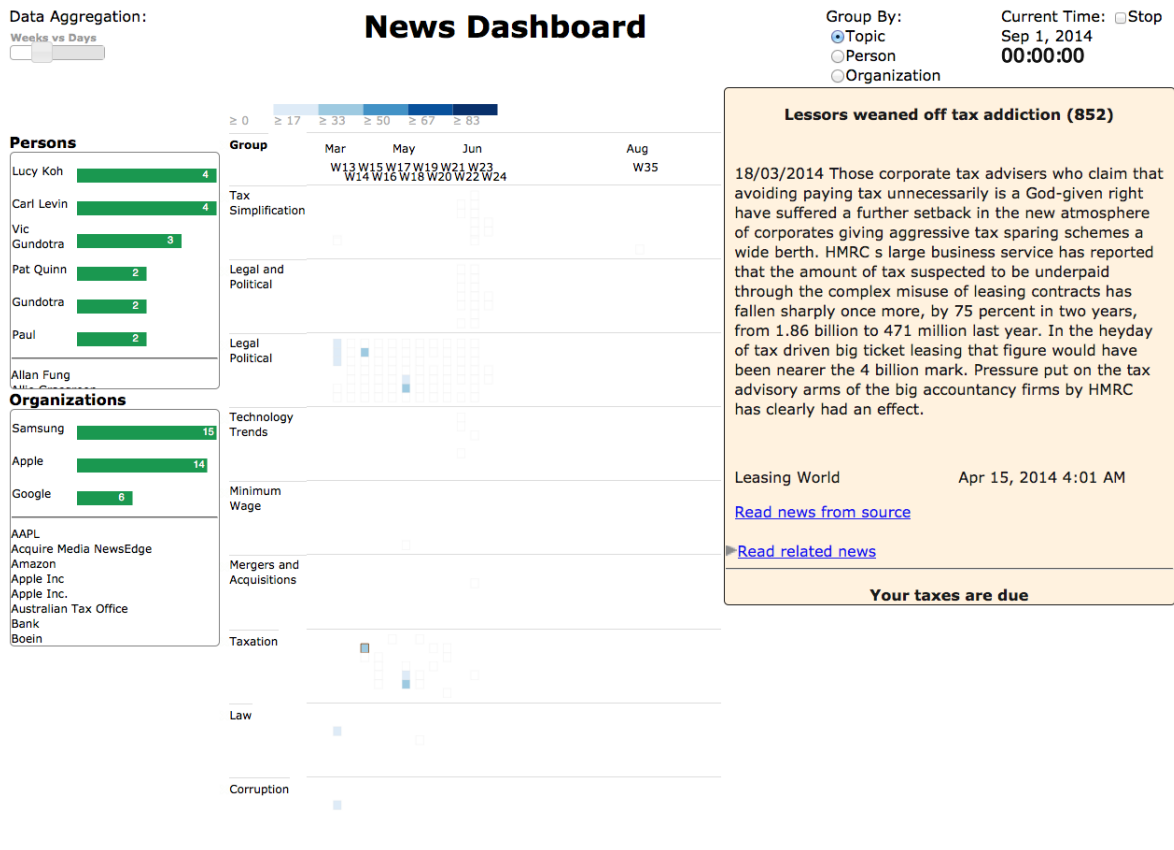


Figure 4.4: Default view of the News Dashboard.

person, topic, and organization, and a display of current time of the news stream.

The components of the news dashboard are the following:

- Stacked heatmap (Section 4.3.2).
- Faceted browsing (Section 4.3.3).
- Related news (Section 4.3.5).
- Breaking news (Section 4.3.2).

The design of the news dashboard will be discussed using the visual analytics mantra, which is thus stated as “analyze first - show the important - zoom, filter and analyze further - details on demand” [35].

4.3.1 Analyze First

The news stream in the raw form is not suitable for visualization, therefore, it is converted into a suitable data form by preprocessing.

Preprocessing

The news data has been preprocessed by Signal Media Limited [5] using text analytics methods. The company has made their preprocessing pipeline known to the scientific community in form of publications [8, 44, 45, 46] as shown in Figure 4.5. The company used text analytics methods for identifying patterns from the textual news data, based on research that draws inspiration from disciplines such as information retrieval, data mining, and computational statistics.

Signal Media Limited makes use of a cloud-based architecture to perform real-time processing of news streams from diverse news sources [46]. The company adopted a flexible pipeline with a “plug and play” philosophy which results in a system that can be customized with minimal engineering effort [44]. This loosely coupled system was achieved by the use of a queuing scheme between the processes in the pipeline that allows for scalability and also modification of a module independently of the total system [46].

The preprocessing of the data by Signal Media Limited begins by converting the text into a mathematical representation for further computation. There are a number of representations such as bag-of-words [60] and probabilistic model [58]. The exact representation used by Signal Media Limited is not available in any scientific publication. This is a very important step in preprocessing the data, as it transforms the data in a form that is suitable for computation.

The company prefers the use of semi-automatic text classification because it provides the best results based on their experience. Automatic classification is faster but inferior to human classification [45]. Nevertheless, human classification is expensive and does not scale with the data. This calls for a tradeoff between human and automatic classification where humans classify documents that could be misclassified by computers [45]. Signal Media Limited’s pipeline utilizes semi-automatic text classification to combine both human effort with automated methods [45]. The users have the ability to correct a mislabelled topic or named entity in the form of a

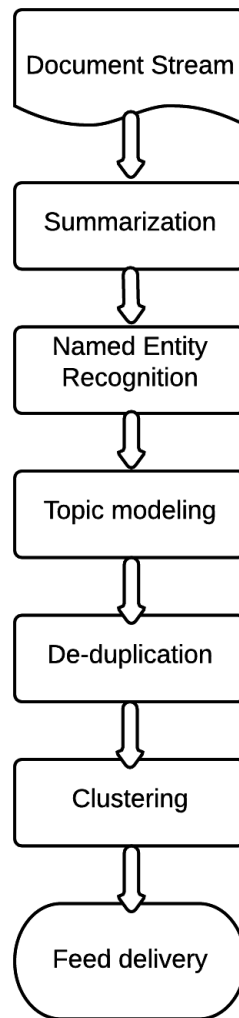


Figure 4.5: The processing pipeline of Signal Media Limited data [46].

semi-automatic classification of the text.

Signal Media Limited has set forward the vision of building a pipeline that is optimized for searching for documents with the ability to perform different queries [44]. The current pipeline allows the users to track news, matching some chosen keywords and named entities (person, location, and organization) [8, 46]. The pipeline consist of summarization, named entity recognition, topic modelling, de-duplication, and clustering [46]. There are different news about the same information from different news sources. The topic modeling module identifies the topics in the news,

where the topic of a document can serve as a summary of the textual data. The topics can be used for organizing and searching for a subset of documents from the total collection. Signal Media Limited has used named entity recognition methods to identify the entities mentioned in the news data [8, 39]. The entities in the text can be the names of persons, organizations, and locations [8]. The preprocessing of the news stream is a very important process because it will not be possible to design a useful news dashboard without the metadata generated in the procedure. The news content is labelled with the metadata that include person, topic, and organization is used to organize the stacked heatmap in a list. The persons and organizations extracted from the news stream is used to develop the breaking news detection and related news functionality.

4.3.2 Show the Important

This provides a description of the functionality of the news dashboard that highlights the most important aspect of the data. This is because not all data are relevant to the user's current information needs, so the relevant data has to be emphasized.

Stacked Heatmap

Given the volume of data in a news stream, it is not feasible to show every news item in the stream. Therefore, there is a need to use aggregation for summarizing the news data within intervals into a set of data points. The compact representation of a matrix-based visualization makes the stacked heatmap an ideal choice for visualizing the news stream as shown in Figure 4.6. A heatmap is a visual encoding of frequency or value within a dense structure using colour. The structure could be dictated by time, but could also be dictated by location in a document such as in TileBars [26] or list of search results as in HotMap [33].

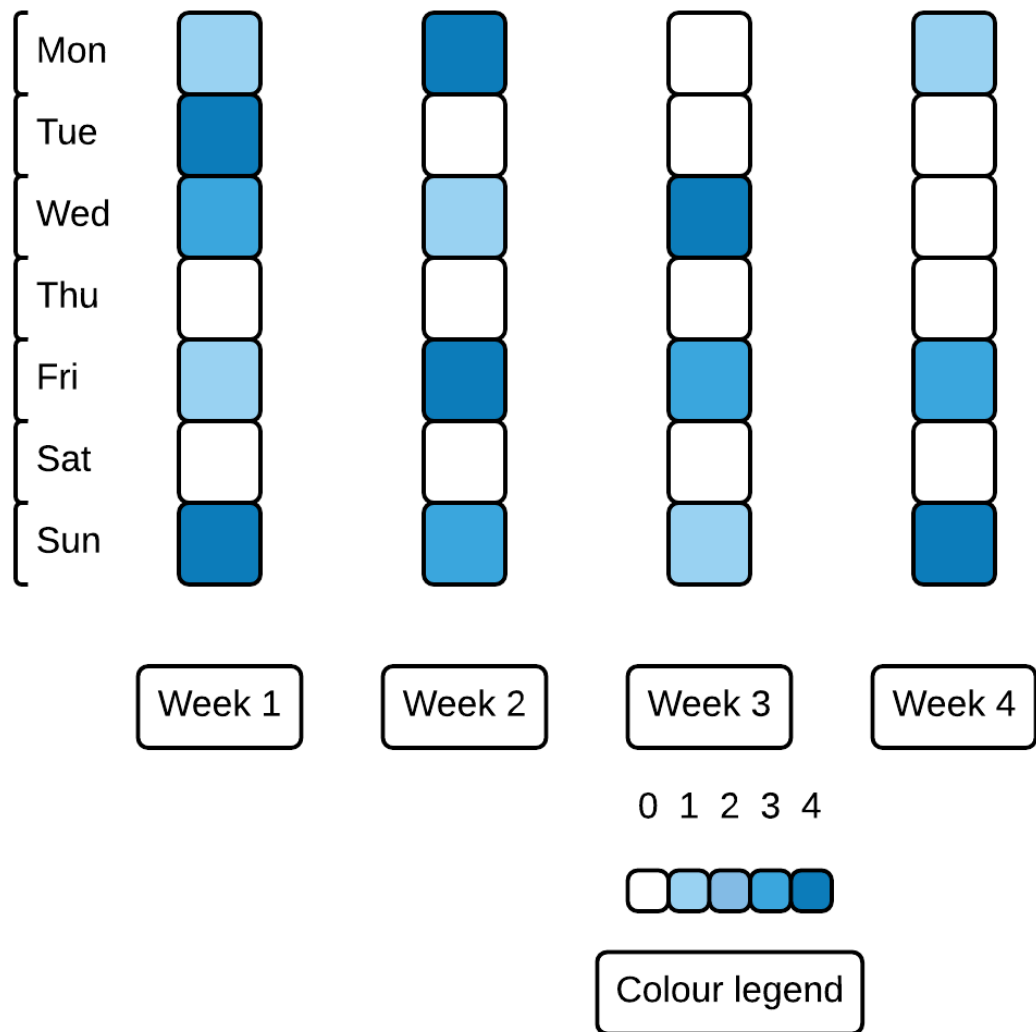


Figure 4.6: A description of the stacked heatmap using “Weeks vs Days” time aggregation.

The stacked heatmap provides a 2-D representation of time, where the news stream is aggregated on the horizontal axis, which represents the width of the element of the stacked heatmap within the set interval on a macro-level, and organizing the data on the vertical axis which represents the height of the element of the stacked heatmap based on a second temporal tier on a micro-level, drawing inspiration from a browsing history interface called BrowseLine [32]. The stacked heatmap allows aggregated news

articles to be visually encoded within an element. For example, in the diagram of the news dashboard showing the time aggregation set to “Weeks vs Days” as shown in Figure 4.6, the horizontal axis shows the weeks of the month, while the vertical axis can show the data aggregated by days of the week. This allows the user to see the news aggregated by days on a week in the stacked heatmap.

The colour of each element of the stacked heatmap shows the frequency of the aggregated news in the set interval. The number of colours was set in the range of 4 to 6 colours for representing the frequency of the news within each block in the stacked heatmap. The colours are perceptual ordered [25] where the lighter colours depict lower frequency of news and darker colours show higher frequency of news in each elements of the stacked heatmap.

The news stream is filtered by persons, topics and organizations. There are a stacked heatmap for each terms in the news dashboard. The Figure 4.4 show that there a one-to-one mapping between each topic extracted from the news stream and a stacked heatmap. This list of topics consist of “Tax Simplification”, “Legal and Political”, and “Legal Political” among others.

Breaking News

The detection of breaking news can be solved as an anomaly detection problem. It is known that every breaking news is important; however, not all important news are breaking news. There are a few conditions for breaking news.

1. This must be the first occurrence of the news in the collection of news of similar information content (same persons and organizations). The first report of the news tends to have higher information content.
2. The breaking news should be relatively new which must be determined within a set time offset from the current time.

There are different types of anomalous data in time series modeling as shown in Figure 4.7. The anomaly detection algorithm can be aimed at identifying any of the signal changes which include abrupt transient shift, abrupt distributional shift, and gradual distributional shift [12].

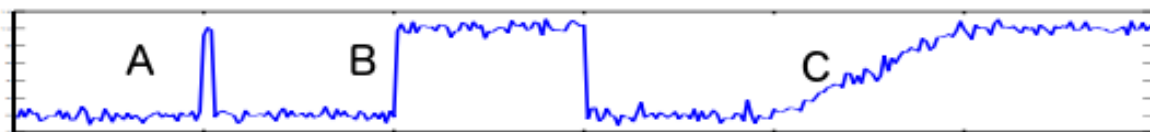


Figure 4.7: Types of signal changes: abrupt transient shift (A), abrupt distributional shift (B), and gradual distributional shift (C) [12].

Online algorithms are useful for real-time application, as they operate incrementally and do not store historical data, which is ideal for analyzing the news streams. The online algorithm receives an input in an incremental manner and makes a decision based on an updated parameter which conveys the current state of the news stream. This philosophy contrasts with offline algorithms that assume the entire data is available in memory. The issue with an offline algorithm is that the data may not fit in memory. The online algorithm should be both time and space efficient.

Anomaly detection algorithms may work in diagnosis or accommodation mode [31]. The diagnosis method identifies the outlier in the data for further processing of the outlier. The outlier is removed from the data sample so it does not skew the distribution. This is useful when the exact parameters of the distribution are known, so the outlier is excluded from further estimation of the parameters of the distribution [31]. The accommodation method identifies the outliers and uses them for estimating the parameters of the statistical model. This is suitable with data streams that account for the effect of concept drift [15], as the news varies in relation to world events.

The Probabilistic Exponentially Weighted Moving Average (PEWMA) [12] algorithm works in accommodation mode. The algorithm allows for concept drift [15], which occurs in news streams by updating the set of parameters that convey the state of the news stream. Breaking news detection can be solved as an anomaly detection problem because there are a small number of negative samples and large number of positive samples. The breaking news data follows the pattern of abrupt transient shift, which makes the PEWMA [12] suitable as an anomaly detection algorithm.

The breaking news can also be estimated by using the source of the news. The implementation of the breaking news detection is to generate a time series data of the number of news sources and the average of the timestamp of every news in the collection of news of similar information content (same persons and organizations).

The time series can be assumed to be normally distributed based on the central limit theorem [24], given that the number of news sources is large. An example of the breaking news display in action is shown in Figure 4.8.

BREAKING NEWS: US government requests for Google user data jump 120% since 2009

Figure 4.8: The view showing the breaking news in the news dashboard.

The parameters of the anomaly detection algorithm consist of X_t the current data, μ_t the mean of the data, \hat{X}_t is the mean of the data, $\hat{\alpha}_t$ the current standard deviation, P_t the probability density function, \hat{X}_{t+1} the mean of the next data (incremental aggregate), $\hat{\alpha}_{t+1}$ the next standard deviation (incremental aggregates), T the data size, and t a point in T . Initialize the process by setting the initial data for training the model $s_1 = X_1$ and $s_2 = X_1^2$, the process can be updated as described in Algorithm 1.

Algorithm 1 Probabilistic Exponential Weighted Moving Average [12]

Require: $X_t, \hat{X}_t, \hat{\alpha}_t, T, t$
Ensure: $P_t, \hat{X}_{t+1}, \hat{\alpha}_{t+1}$
 incremental Z score
 $Z_t \leftarrow \frac{X_t - \hat{X}_t}{\hat{\alpha}_t}$
 probability density function
 $P_t \leftarrow \frac{Z_t}{\sqrt{2\pi}} e^{-\frac{Z_t^2}{2}}$
if $t < T$ **then**
 increment standard deviation (training phase)
 $\alpha_t \leftarrow 1 - 1/t$
else
 increment standard deviation
 $\alpha_t \leftarrow (1 - \beta P_t)\alpha$
end if
 moving average
 $s_1 \leftarrow \alpha_t s_1 + (1 - \alpha_t) X_t$
 $s_2 \leftarrow \alpha_t s_2 + (1 - \alpha_t) X_t^2$
 incremental mean
 $\hat{X}_{t+1} \leftarrow s_1$
 incremental standard deviation
 $\hat{\alpha}_{t+1} \leftarrow \sqrt{s_2 - s_1^2}$

The news stream is sampled using a sliding window based on the set time aggregation. The processed data is fed to the anomaly detection algorithm with the

parameter $\alpha = 0.98$, $\beta = 0.98$, and $\tau = 0.0044$. The thresholds are chosen for determining outliers that are greater than 3 times the standard deviation in a normally distributed data.

4.3.3 Zoom, Filter

This provides a description of the functionality of the news dashboard that supports the user in focusing on a subset of the data, as the user explores for the information that suits his / her needs. The zooming is supported by the use of a slider that allows the users to zoom in and out of the data to allow for data exploration at different levels of temporal resolution.

Faceted Browsing

The faceted browsing interface [28] provides a mechanism for multiple filtering of the data according to persons, topics, and organizations in a hierarchical ordering. This is a top-down approach where the users can incrementally build the queries that guide the users in pursuit of their information-seeking goals. The faceted browsing interface shows the available facets as shown in Figure 4.9.

The design for the news dashboard favours the use of browsing rather than searching for entities because the browsing does not require the use of recall, but rather recognition from a list of browsing terms will exert less cognitive burden on the users of the news dashboard.

The user can browse for topics, persons, and organizations in the news stream using the faceted browsing interface [28]. The user clicks on a facet term to highlight the matching terms in the stacked heatmap as discussed in Section 4.3.2. Another click on the same term unhighlights the facet term and unhighlights the elements of the stacked heatmap. This allows for recognition, rather than recall [50]. The user does not have to memorize large chunks of text; the interface provides the text for the user to select based on recognition from past experience.

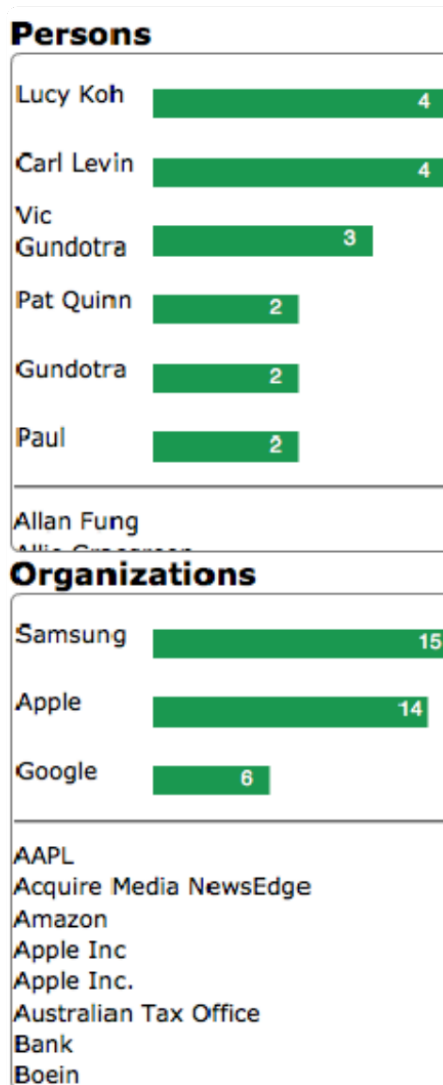


Figure 4.9: Faceted browsing interface that shows the list of facet terms.

The available facets can be topics, persons, and organizations depending on the visualization settings. One of the facets is used to organize the stacked heatmap. The two remaining facets are used to provide the browsing terms. The browsing terms that can be used for changing the path of exploration. For example, if the user clicks on the “Topics” radio button, then the available facets are “Persons” and “Organizations”. This works similarly for groupings by “Organizations” or “Persons”.

The facets are shown using a combination of histograms and terms based on the count of terms in the persons, topics, and organizations. The user can compare a number of terms in the facets using length with the use of a histogram. The term count increases in proportion to the size of the list of browsing terms, thereby resulting in single count terms that are not descriptive of the information in the news dashboard. This problem has influenced the use of a threshold to determine a suitable cut-off point. After experimentation, the threshold was set at 30% of the maximum count of terms in the persons, topics, and organizations in the news stream. The term count below the set threshold are lexically sorted, whereas those above the threshold are represented using histograms.

4.3.4 Analyze Further

This provides a description of the functionality of the news dashboard that supports the further analysis of the data during the visual exploration of the news stream.

Interactive Exploration of News Stream

A change in the temporal aggregation during the process of visual exploration would result in the recomputation of the stacked heatmap. This results in showing the temporal trends of the news streams that may not be visible in the previous time aggregation settings.

The radio buttons of “Group By” is a component of the dashboard controls which consist of “Topics”, “Persons”, and “Organizations” buttons. A click on the any radio button would organize the display of the stacked heatmap in a list, after the news stream has been filtered by the label of the clicked radio button. The remaining two unchecked radio buttons are used for designing the faceted browsing interface.

A click operation on the “Read Related News” link will retrieve all the news items that share similar information, thereby providing a context of the current news on display. This can provide a further analysis of the current news on display. This is fully discussed in Section 4.3.5.

A scrolling operation on the news text window highlights the matching elements of the stacked heatmap. A click operation on the stacked heatmap loads the news text window and adds a bounding box on the cell of the clicked stacked heatmap.

4.3.5 Details on Demand

Not all relevant information are useful at every step in the exploration of the news stream. This provides a description of the functionality of the news dashboard that allows the user to access the data only when it is required, thereby making the visualization space available for satisfying the current information needs.

Related News

The news dashboard allows the users to explore the news in relation to relevant past news. The news stream is aggregated using a sliding window based on the set time aggregation. The users can see the relationships between the current news and past news by reading the news that are related to the news on display. This can give the user a overview of the news about a collection of news with the same information. The related news functionality shows the current news in relation to past news about similar information as shown in Figure 4.10. The news window labelled “A” show the default view, the click on the “Read Related News” expand to show the related news in news window labelled “B” in Figure 4.10.

Galtung et. al. developed the earliest model for estimating the newsworthiness of a news item based on the named entities [19]. The newsworthiness model was improved by the work of Harcup and O’Neill [22]. This motivated the creation of a gazette list for estimating newsworthiness. The gazette list consists of the names of the leaders of countries, wealthy individuals, popular actors, musicians, athletes, and popular organizations. This is an intuitive way of measuring the importance of the news data in real-time because people are interested in news about popular entities.

News about the same person and organization is called a collection of news about the same information. Each datum is grouped into a list based on similar information content using locality sensitive hashing [20]. This is used for grouping items that are similar into the same bucket; dissimilar items are placed in different buckets. The news in the window is aggregated into a list of news containing similar information.

The collection of news about the same information (same person and organization) are the related news which can be displayed in news window which shows the title, content, source, URL, and date. This could be beneficial information to the user. The title gives the context for the new content. The URL is provided to allow the

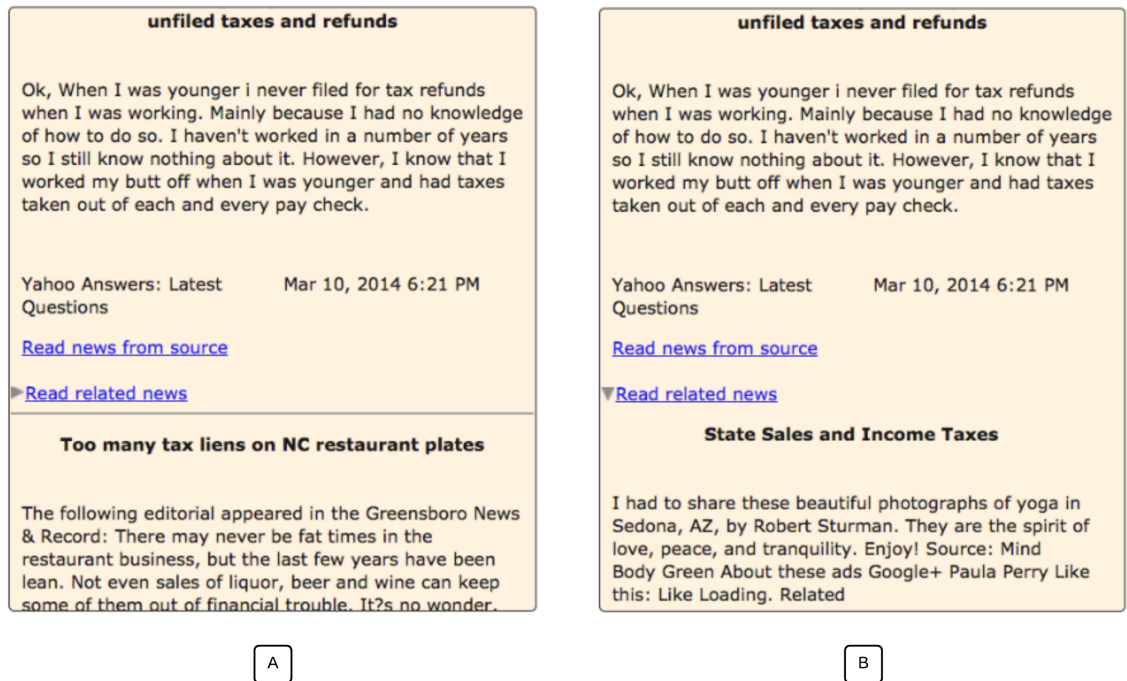


Figure 4.10: The view showing the related news feature in the news dashboard.

user to check the action news content if the user decides to verify the news before viewing it from the source.

4.4 Theoretical Foundations and Interactions

The section discusses how the theoretical foundations in Chapter 2 were employed for designing and using the news dashboard. The principles of dashboard design and visual analytics are the guiding principles for developing a news dashboard that supports the visual exploration of the news stream. Information foraging theory describes how the users utilize the news dashboard to meet their information needs.

The fundamental principles are discussed as dashboard in Section 4.4.1 and information foraging theory in Section 4.4.2.

4.4.1 Dashboards

The dashboard makes use of a matrix-based visualization to show the frequency of the news within a set time interval. The dashboard ensures that the most relevant information is visible when needed. The user can see at a glance the popularity of the persons, topics, and organizations in the news stream based on the pre-attentive processing of the stacked heatmap.

The breaking news is placed in the center area of the topmost control, directly below the title of the dashboard. The breaking news is animated to avoid change blindness [25], which could occur if the breaking news were stationary. The animation directs the attention of the users to the breaking news on display.

The scrolling operation on the news window highlights the stacked heatmap until the next update of the news stream. The click operation on the element of the stacked heatmap loads the news window and adds a bounding box on the element of the clicked stacked heatmap. Hovering over each element of the stacked heatmap displays a tooltip that shows the list of titles of the news aggregated in the given time range. A click on the element of the stacked heatmap loads the news window. The stacked heatmap is highlighted using a bounding box over an element of the stacked heatmap. The choice of the highlighting colour is also based on the distinguishability of colours [25]. The highlighted bounding box should be clearly visible from the background of the news dashboard.

The dashboard provides real-time visualization of online news stream; in addition, the user can pause the incremental nature of the visualization to allow for detailed exploration of historical news data. The news interface allows the grouping of the news by either topics, persons, or organizations by clicking on a labeled radio button. The user can resume the visualization using the current time by unchecking the radio button.

4.4.2 Information Foraging Theory

The use of information foraging theory was aimed at enhancing the usability of the news dashboard. The design of the interface has supported the user to maximize of the utility of the news dashboard, allowing the users to pursue their goal of managing information-intensive tasks in the news interface.

The user can perform quick scanning of the interface to get an overview of the data. However, If the user decides to read the news about a term, then the user can click on the facet term on the interface. The faceted browsing allows the user to find news using recognition rather than recall [50] to form a mental model of the news stream. This automatically extract terms that can be used for browsing for more information. The interface provides the user with multiple paths of exploration. The user can decide to continue exploring through the interface in search of his/her goal or decide when to stop if it does not make sense to continue exploring, in connection with the enrichment versus exploitation tradeoff as discussed in Section 2.5.

The dashboard is designed with the navigation within the temporal aggregation by the use of a slider that allows the users to zoom in and out of the data to allow for data exploration at different levels of granularity which include “Months vs Weeks”, “Weeks vs Days”, “Days vs Hours”, and “Hours vs Minutes”. The interface allows the grouping of the news by topics, persons, and organizations by clicking on a labeled radio button. For example, if the user clicks on the “Topics” radio button, then the available facets are “Persons” and “Organizations”. This works similarly for groupings by “Organizations” or “Persons”.

The highlighted linking between the stacked heatmap, faceted browsing interface, and news window directs the attention of the users to the source of the news. This occurs when the facets are clicked, as there is highlighting of the matching elements of the stacked heatmap to draw the attention of the user to the news. The news dashboard has been designed to allow the users to see the connectedness between the news window, stacked heatmap, and faceted browsing interface. This was made possible by the use of multiple coordinated view, where the users can understand the visualization as a whole, instead of a collection of individual visualizations using the Gestalt Principles of proximity and similarity.

There are a number of information scents emanating from the news dashboard. These consist of the breaking news indicator, related news link, and histograms in the faceted browsing. The breaking news indicator can serve as an information scent to guide the user to reading the news by clicking on the indicator. The related news feature can provide more information about the news on display. The histogram on the faceted browsing interface tells the user how much they can find by following an exploratory path.

The highlighted portion of the news interface also serves as information scent. They provide visual cues to the users as they explore for news in the dashboard. The users will follow the scent provided to the news in the stacked heatmap on a click. These interactions that are supported in the news dashboard are summarized in Table 4.1.

Table 4.1: Set of Interactions

Object	Operation	Result
Each element of the stacked heatmap	click	Open up the news window.
Top corner of stacked heatmap	click	Deletes the row of stacked heatmap.
Facet terms	click	Highlights or unhighlights the facet term. Highlights or unhighlights the stacked heatmap thereby filtering the news stream.
News window	click and scroll	load windows and highlights the stacked heatmap

The theoretical foundations have inspired the design of a news dashboard that support the ability of users to interactively explore news stream. A number of scenarios that describe how users can benefit from using the news dashboard is shown in Chapter 5.

Chapter 5

Case Studies and Discussion

In order to show the usefulness of the news dashboard in comparison to Google News [3], a number of scenarios are provided to highlight the potential benefits of using the news dashboard. The news dashboard makes use of a simulated news stream derived from data provided by Signal Media Limited [5]. However, the comparison of the two news interfaces can be done using data sets with similar information, as the news articles are discussing the same world events. The scenarios of the news dashboard in Section 5.1 are discussed, leading to the description of scenario using Google News in Section 5.2, and concludes with the discussion of the merits of using the news dashboard as opposed to Google News in Section 5.3.

The aggregated news from a number of news organizations, as curated by Signal Media Limited [5], is filtered to extract the news that provides information about business-related events in the world. The recurring themes in the news include: the state of the economy, taxation, legal implications of business decisions, stocks, firm acquisitions, and business opportunities among others. The intended consumers for the aggregated news are business executives, who want to know about the news on financial market conditions.

The collection of news aggregated by Google News comes from a number of news organizations. This accumulated news data contains information about diverse categories of news. The Google News website provides information about the state of events in the world to the readers. The intended consumers of the aggregated news are users of all ages, who have familiarity with basic use of computers.

The user in this scenario is named “Jane”. She is a top-level executive in a software company that specializes in producing software applications for mobile phones. Jane

reads the news in order to understand the market and incorporates this as part of her daily routine. The information from the news can provide latest trends, which can help her decide on which platform is the most profitable for creating platform-specific software. Understanding trends in the news stream is crucial to a complete comprehension of the market forces. The news dashboard is ideal for market analysis as it provides instant information about trends in the news stream. It also provides a visual summary of news stream about business-related events. Jane is trying out both the news dashboard and Google News in order to decide which is best suited for her information needs.

5.1 News Dashboard Scenario

The dataset for the news dashboard is a collection of news between February and August, 2014. The total number of news items in the collection is 1008. The visualization settings include the time aggregation set to “Weeks vs Days”, grouping the news by “Topic”, with the current time being August 17, 2014 at 00:00 hours GMT as shown in Figure 5.1.

Jane is trying to see the current market trends after a day’s work. She opens up the news dashboard to get situational awareness of the market by exploring the news stream. She proceeds to view the compact representation of the news stream in the form of a visual summary as shown in Figure 5.1.

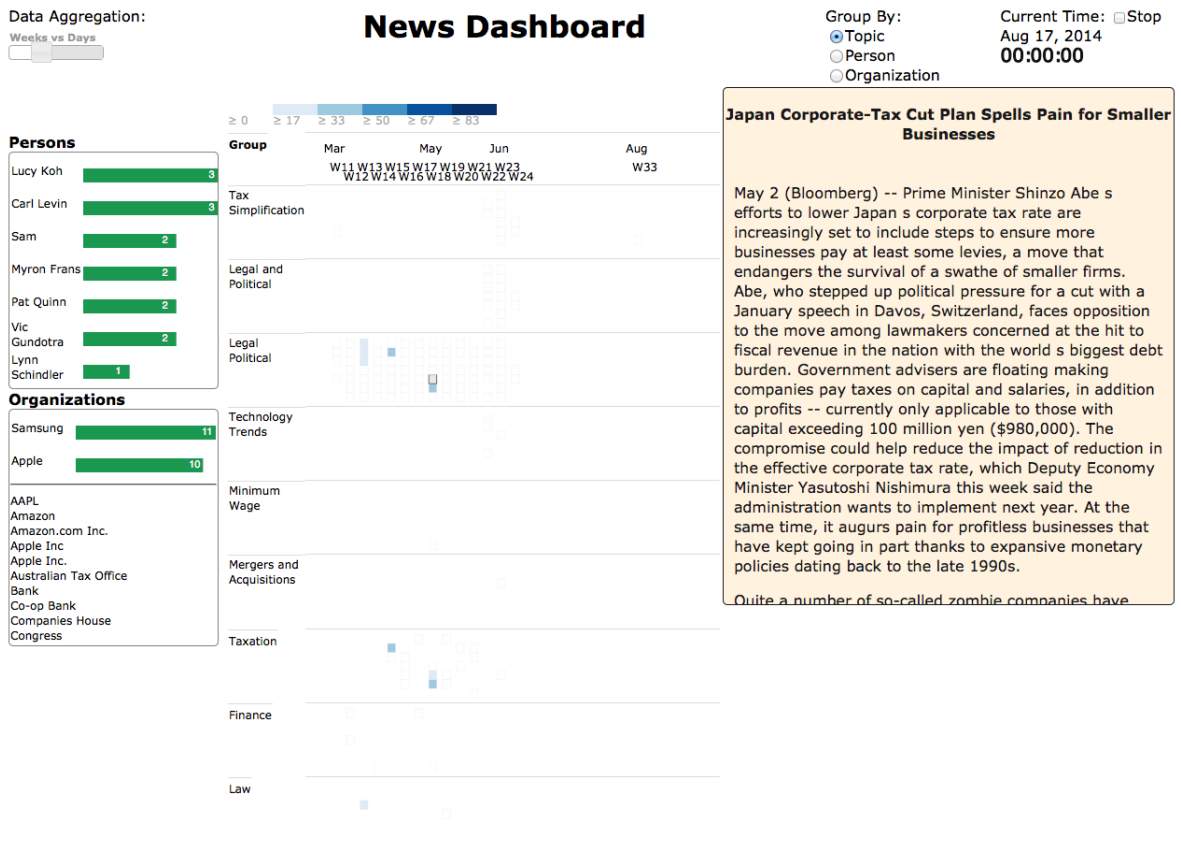


Figure 5.1: The default view of the news dashboard.

The initial assessment of the news dashboard by Jane shows the stacked heatmap has visually encoded the popularity and distribution of news about various topics which include “Tax Simplification”, “Legal and Political”, “Legal Political”, “Technology Trends”, “Minimum Wage”, “Mergers and Acquisitions”, “Taxation” on a single-paged display, as shown in Figure 5.2. As Jane is seeking to understand the trends in the news stream, she experiences the frustration of the current tax declaration process that is overly complex. She picks up interest in the story about attempts by the government to simplify the tax declaration process and proceeds to explore the news.

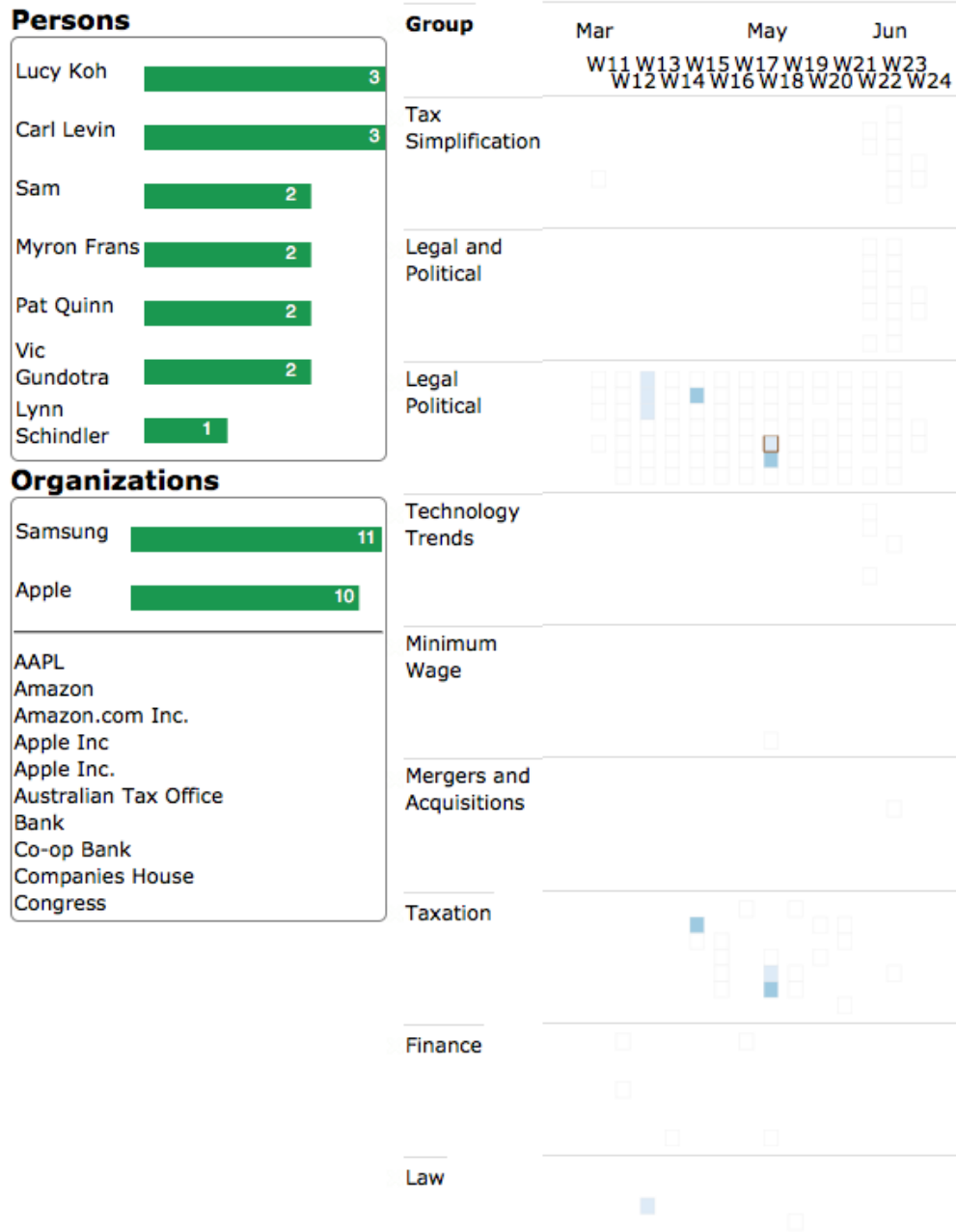


Figure 5.2: The view of the news dashboard excepting the news window.

Jane observes the topic “Tax Simplification” in the stacked heatmap, where she can tell at a glance that there is no discussion about this in the media until the month of June 2014. This can allow her to make correlations between external events in the real world, and understand how the topic became of interest during the month

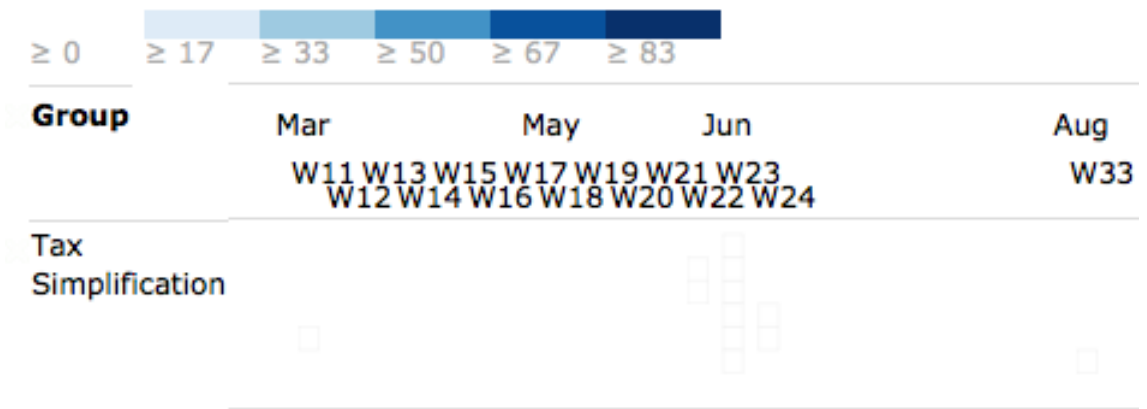


Figure 5.3: The news dashboard showing the topic “Tax Simplification”.

as shown in Figure 5.3.

Jane became attracted to the stacked heatmap, which is the most popular topic named “Legal Political” in the news stream. The visual encoding can show the topic, which occurs more frequently due to the number of elements of the stacked heatmap that is coloured different from the white colour in the background as shown in Figure 5.4.

Jane is also attracted to the list of facet terms as shown in Figure 5.5 and uses the faceted browsing interface in the news dashboard.

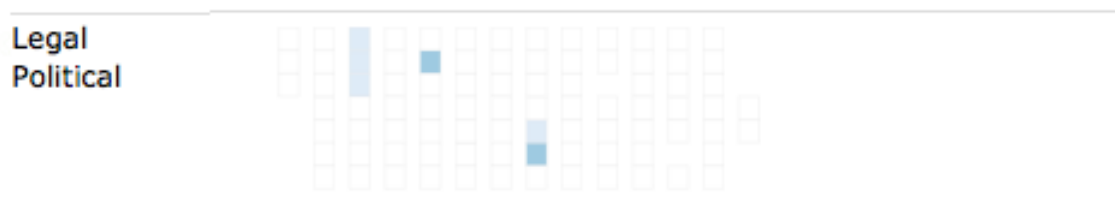


Figure 5.4: The news dashboard showing the topic “Legal Political”.

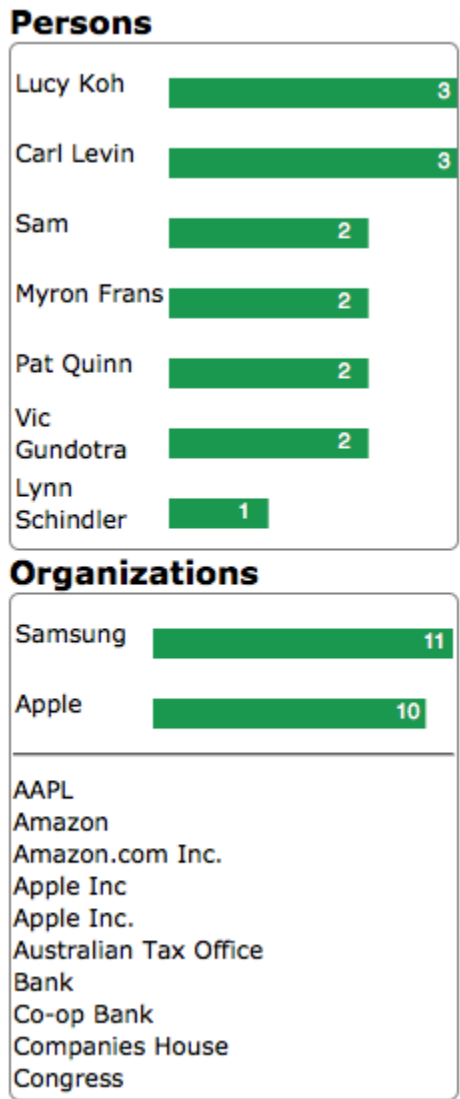


Figure 5.5: The list of terms in the facets.

She is interested in software companies that produce mobile phones because she is an executive in a software firm. She decides to browse for the term “Apple”, where she then clicks the term in the facets named “Organizations” as shown in Figure 5.6 .

The faceted browsing provides a mechanism for visual browsing of the stacked heatmap to see the news, in relation to other news, in the stream as the matching elements of the stacked heatmap are highlighted as shown in Figure 5.6. Jane can see that “Apple” is a very popular entity that occurs very frequently in the news stream.

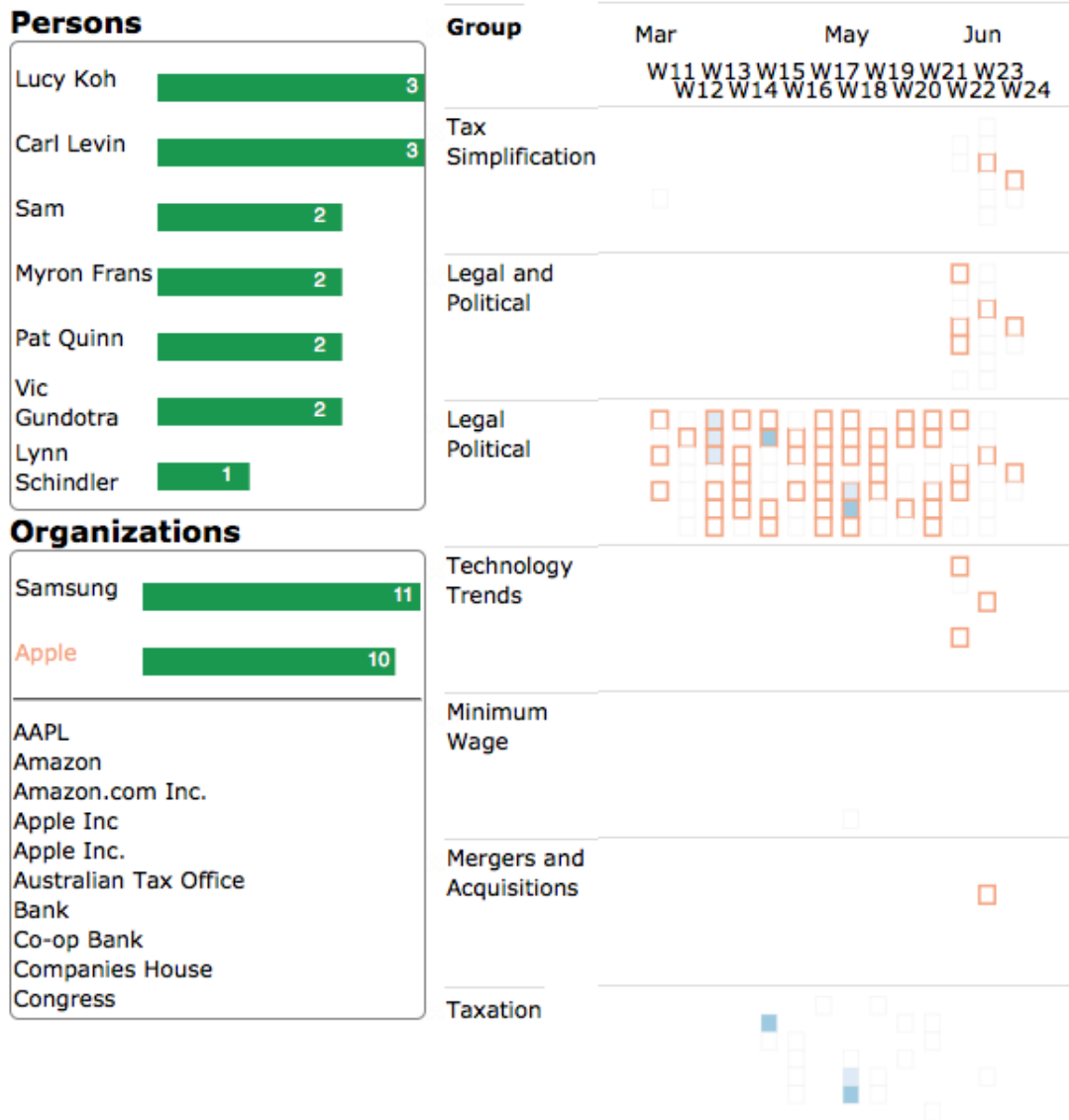


Figure 5.6: The list of terms in the facets and stacked heatmap, when “Apple” is clicked in the facet.

She can proceed to read detailed information by clicking the element of the stacked heatmap to view the news window.

Jane makes a close examination of the element of the stacked heatmap on June 12, 2014, with a topic named “Tax Simplification”. Showing the tooltip of titles of news in the elements of the stacked heatmap as shown in Figure 5.7, she can then proceed to open the news window for further exploration. The news item about “Apple” is loaded with the title “EU takes on Starbucks Apple tax break”, which

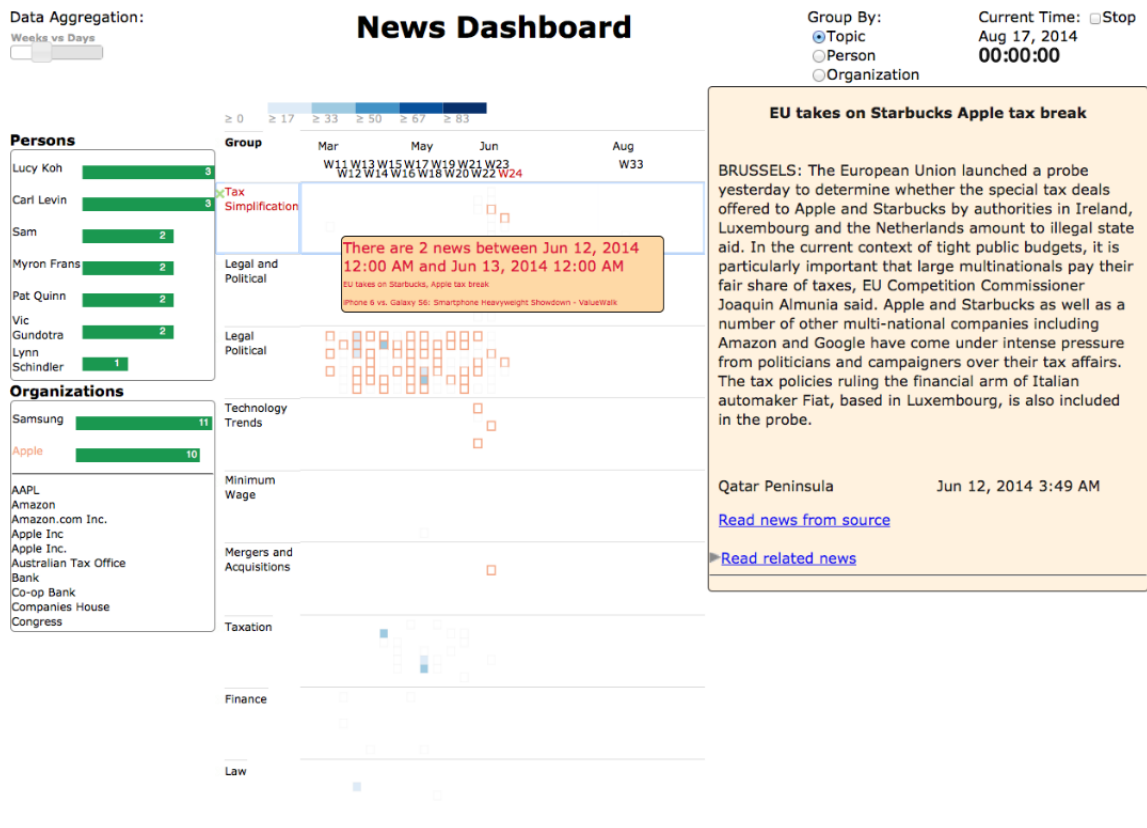


Figure 5.7: Screenshot showing the tooltip of the news stream.

discusses government scrutiny over the tax practices of the Apple company. The next news article on the text window is titled “Iphone 6 vs. Galaxy S6: Smartphone Heavyweight Showdown - ValueWalk”, which describes the marketplace competition between Samsung and Apple.

Jane explores the news on the dashboard, and decides to change the time aggregation from “Weeks vs Days” to “Days vs hours”, thereby allowing her to obtain a daily report of the news. Shortly after making this changes in settings. She is attracted to the breaking news that pops up about “US government requests for Google user data jump 120% since 2009”. Jane cares about the security of the users using her firm’s software, so she clicks on the breaking news indicator to view the news to get a full understanding as shown in Figure 5.8.



Figure 5.8: Screenshot showing the breaking news on display.

Jane identifies on an element of the stacked heatmap that is of interest, and clicks on it. This default view of the news in display is shown in the Figure 5.9 labelled “A”, which discusses “unfiled taxes and refunds”. She can proceed to read the news that are related to the current news on display as clicking on “Read Related News” to read the part of the screenshot labelled “B”. She can tell that the news with the title “State Sales and Income Taxes” is related to the news with the title “unfiled taxes and refunds”. The related news feature allows the user to have access to information that is not immediately shown on the screen, but has a relationship with the item displayed on the screen.

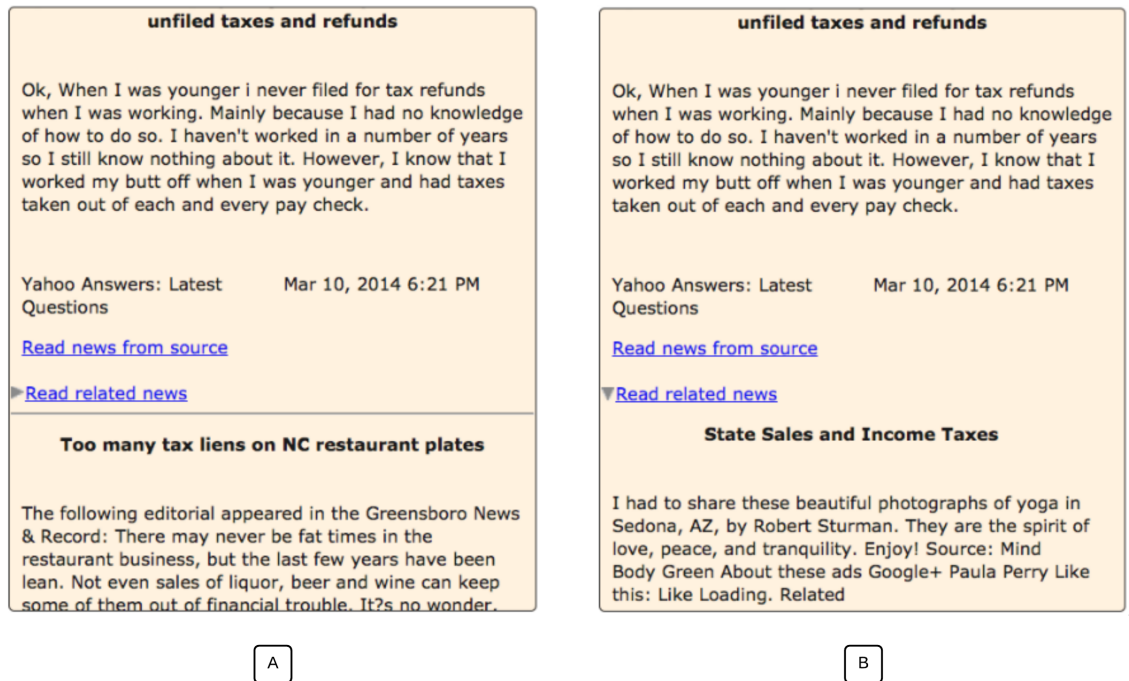


Figure 5.9: The view showing the related news feature in the news dashboard.

5.2 Search as an Alternative

Google News [3] is an aggregator news service that has a list-based representation as the primary means of presenting the news. It allows for personalization with options for modifying the location and a slider for weighing the proportions of news about different topics in the news stream. This interface has a faceted navigation, thereby allowing Jane to filter the news stream by the selected term. At the point of writing this report, Google News is unable to support the querying of past news archives. This limitation was resolved in our research by using the Internet Archive to obtain past news between two specified dates. The dataset for Google News is a collection of news from Internet Archive [4] with the date of March 25, 2014.

Jane is trying to see the trends in the current market after a day's work. She opens up Google News to read the news stream. She decides to look for quick information show the distribution of the news within the time interval. The interface does not provide a compact summary of the news stream because of the limited amount

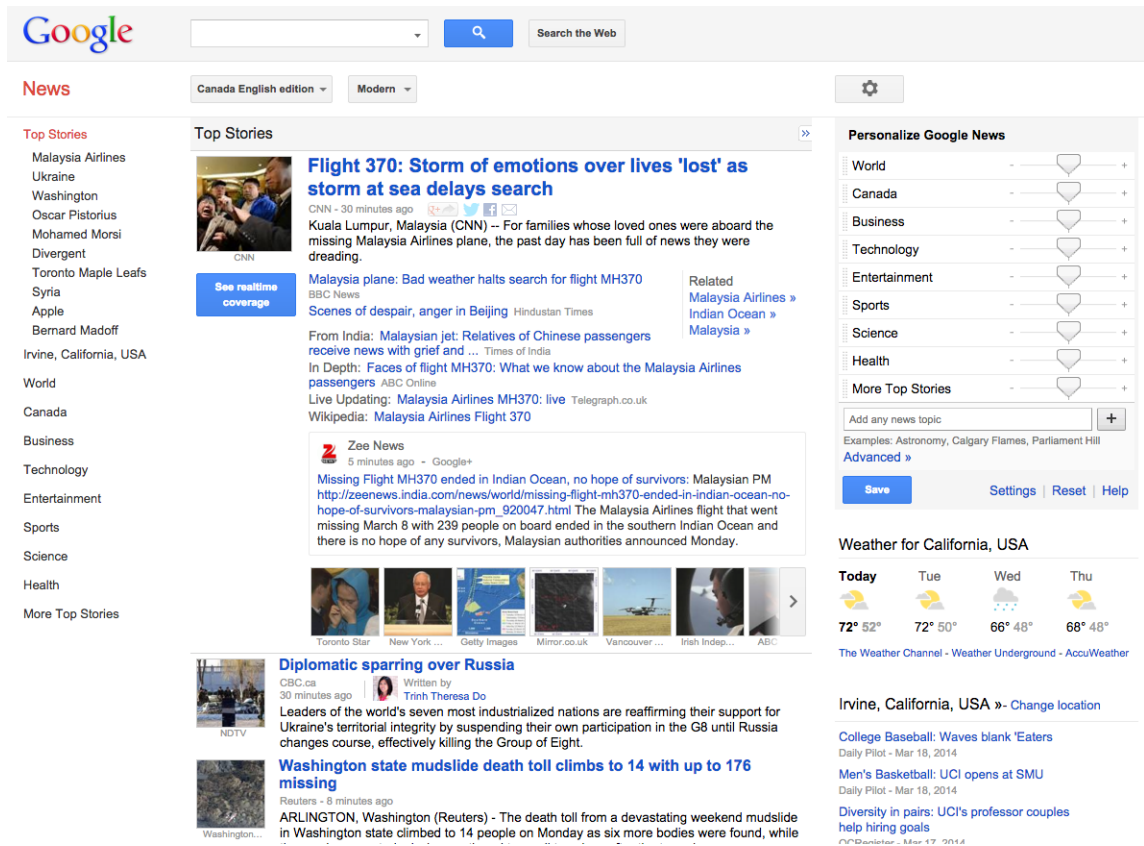



Figure 5.10: The default view of the Google News [3].

of display space required to show the news articles. Google News requires lengthy vertical scrolling to view the entire data in the news stream as there is a limit on the number of news articles that can be displayed at one time on the screen. The overview of the news stream is shown in Figure 5.10 and is only a tiny fraction of the news stream, which cannot be said to be a summary of the news stream.

Google News provides a related news functionality, which allows Jane to have access to information that is not immediately shown on the screen, but has a relationship with the item displayed on the screen as shown in Figure 5.11. The list of related news to the news titled “Flight 370: Storm of emotions over lives ‘lost’ as storm at sea delays search” consist of “Malaysia Airlines”, “Indian Ocean”, and “Malaysia”.

Top Stories >>



CNN

Flight 370: Storm of emotions over lives 'lost' as storm at sea delays search

CNN - 30 minutes ago + t f ✉

Kuala Lumpur, Malaysia (CNN) -- For families whose loved ones were aboard the missing Malaysia Airlines plane, the past day has been full of news they were dreading.

See realtime coverage

Malaysia plane: Bad weather halts search for flight MH370 BBC News

Scenes of despair, anger in Beijing Hindustan Times

From India: Malaysian jet: Relatives of Chinese passengers receive news with grief and ... Times of India

In Depth: Faces of flight MH370: What we know about the Malaysia Airlines passengers ABC Online

Live Updating: Malaysia Airlines MH370: live Telegraph.co.uk


Wikipedia: Malaysia Airlines Flight 370

Related

[Malaysia Airlines »](#)

[Indian Ocean »](#)


[Malaysia »](#)





Zee News

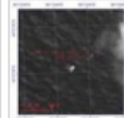
5 minutes ago - Google+


[Missing Flight MH370 ended in Indian Ocean, no hope of survivors: Malaysian PM](http://zeenews.india.com/news/world/missing-flight-mh370-ended-in-indian-ocean-no-hope-of-survivors-malaysian-pm_920047.html)
http://zeenews.india.com/news/world/missing-flight-mh370-ended-in-indian-ocean-no-hope-of-survivors-malaysian-pm_920047.html The Malaysia Airlines flight that went missing March 8 with 239 people on board ended in the southern Indian Ocean and there is no hope of any survivors, Malaysian authorities announced Monday.

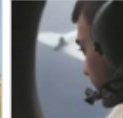

Toronto Star


New York ...


Getty Images


Mirror.co.uk


Vancouver ...


Irish Indep...



ABC

Figure 5.11: The related news features of the Google News.

Jane is attracted to the list of facet terms as shown in Figure 5.12 which provides all the exploratory path of exploration in the Google News interface.

Top Stories
Malaysia Airlines
Ukraine
Washington
Oscar Pistorius
Mohamed Morsi
Divergent
Toronto Maple Leafs
Syria
Apple
Bernard Madoff
Irvine, California, USA

World

Canada

Business

Technology

Entertainment

Sports

Science

Health

More Top Stories

Figure 5.12: List of terms in faceted navigation of Google News.

The faceted navigation interface can provide Jane with a new exploratory path to further explore the news stream. Jane works in the software industry and as such, she takes interest in the firm named “Apple”, she clicks on the term in the facets named “Top Stories”. The faceted navigation interface allows her to filter the news stream based on the selected facet term as shown in Figure 5.13.

-
- Top Stories**
- Malaysia Airlines
 - Ukraine
 - Washington
 - Oscar Pistorius
 - Mohamed Morsi
 - Divergent
 - Toronto Maple Leafs
 - Syria
 - Apple**
 - Bernard Madoff
 - Irvine, California, USA
- World
- Canada
 - Business
 - Technology
 - Entertainment
 - Sports
 - Science
 - Health
- More Top Stories**

Figure 5.13: List of terms in faceted navigation of Google News. “Apple” is clicked in the facet.

The resulting news articles that match “Apple” as shown in Figure 5.14.

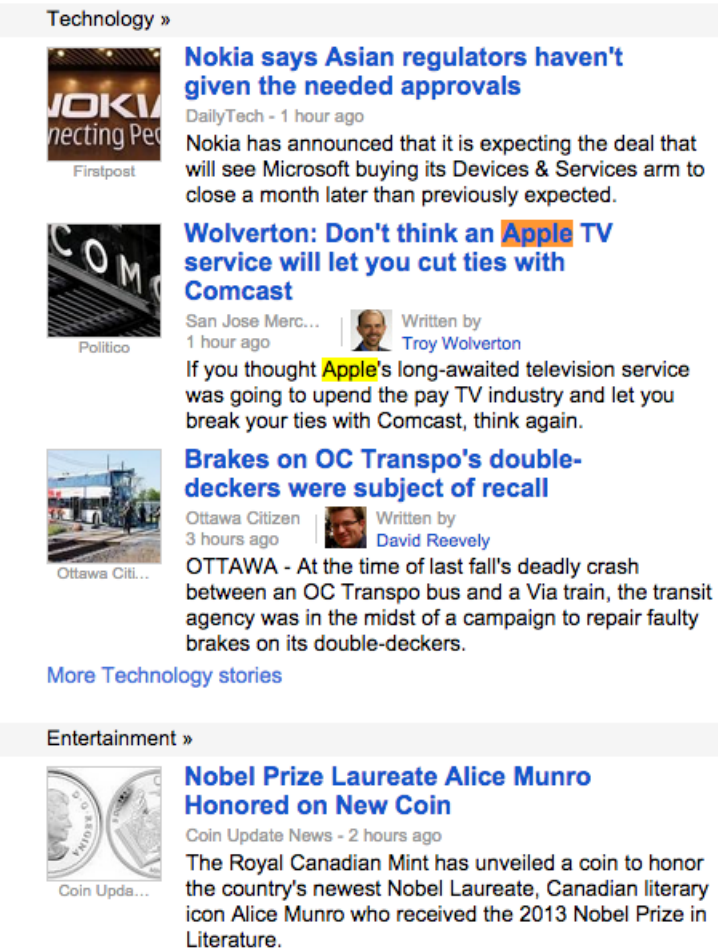


Figure 5.14: The scrolled view of the Google News [3].

5.3 Discussion

Google News [3] is similar to our news dashboard with the support for trending news. The proposed news interface and Google News, both support the related news functionality. However, there are differences between the two systems. For example, while Google News supports the real-time coverage of the news, it does not support breaking news notifications to the users. While our proposed interface provides room for the limited adjustment of visualization settings, Google News allows for more robust customization leading to personalization of the news interface. The news dashboard provides a faceted browsing interface, while the Google News provides a faceted navigation interface.

Within this scenario, the news dashboard has allowed the user to interactively explore the news stream using a single screen. The single screen view allows the user to see more previous news about the news stream. This is not the case in Google News as the news stream cannot be shown on a single screen, and requires lengthy vertical scrolling of the page. The list representations are static and have low information density as there is a fixed number of documents per page. The matrix-based visualization was adopted in this research as the primary interface because its compact representation requires only a small display space. These requirements are met with the design of a news dashboard.

The news dashboard was designed using the knowledge of how people consume news in the real world. The users have an incomplete idea of what they want, then they use their mental model to derive a set of actions that can lead to their goal. The user can obtain an overview of the data in the news dashboard over a wide time interval due to compact display space; Google News is not suitable for exploring past news due to the limits of the list-based representation that results in a limited number of news articles that can be shown in a screen. The user can modify the time aggregation in the news dashboard to allow the exploration of the news at different levels of granularity, thereby enabling exploration of the current news in relationship to existing news in the news stream. This feature is missing from Google News at the point of writing this report, thereby limiting the ability to explore past data. The faceted interface allows for filtering the news in relation to certain persons, topics, and organizations. Finding news articles using the facet browsing interface is supported in two stages of interaction. In the first stage, the user views the facet terms and chooses a subset of the terms on display. In the next stage, the user clicks on the term to see the highlighted element of the stacked heatmap. The click event on the stacked heatmap shows the news matching the query. This is a feature in both interfaces under comparison. This allows the user to explore the subset of the news stream and allow for detailed analysis of the news stream.

The user's needs can change during the exploration process as the user can modify their plan until the required news articles are obtained from the news streams. In some cases, the user can partially satisfy the goals during the exploration process, or even abandon the current plan in favour of a new goal that can be set during the exploration process. Users are looking for a collection of documents that may meet

their goal obtained during the exploratory process. The strength of the information scent encourages the user to continue towards their initial goal. Conversely, the diminishing of the information scent results in the user abandoning the original plan, in favour of a new goal. This information scent can take the form of identifying a relevant facet term that modifies the query of the interface to filter the result in Google News, or the highlighting of the elements of the stacked heatmap in the news dashboard. The information scent is also present in the form of the “Read Related News” link that directs the users to the news articles surrounding the events of the current news on display.

The visual representation of the news stream is easier to understand than reading a textual document as can be seen in Figure ??, as the matching elements of the stacked heatmap are highlighted to further guide the exploration of the news stream. The users are attracted to the news article in the new stream that is relevant to their goals and diverted away from news that is not relevant to the exploratory goals. The aim of the users is to uncover the news buried deep within the news stream. The related news feature in the news dashboard and Google News provides more context of the event that surrounds the current news on display.

The Google News [3] has a major limitation when exploring past news using the real-time coverage, because it is not a live event. The news dashboard is better for showing trends in the news stream, when compared to the Google News due to the compact graphical representation of the news stream. The users are more capable of comparing relevant old news as recent news arrives. Despite the small size of data, it is difficult to understand the trends in the news in its textual form at a glance without the use of the news dashboard. A data of comparable size in Google News will result in a lengthy scroll of the page to locate and read the news content. Therefore, resulting in the loss of the ability to understand aspects of the news stream at a glance by using the textual representation of the news stream.

Chapter 6

Conclusions

This work has utilized the theoretical foundations and frameworks discussed in Chapter 2 to guide the design of a news dashboard. The dashboard was chosen as a convenient means of representing the information due to the compact use of display space, thereby providing a visual summary of the news stream. It makes use of information visualization techniques to present information about the news stream in a manner that can be easily understood by the users. The information-seeking mantra guided the design of a news interface in order to support the users in their information-seeking activities. The visual analytics mantra guided the design of the news dashboard by supporting the ability of the users to analyze the data at multiple resolutions on demand. The goal is to guide the users in the process of finding information that meets their needs. In addition, information foraging theory has supported the users in their information-seeking activities in the news dashboard in the interactive exploration of the news stream as the users fulfil their information needs. A review of existing news interfaces was provided in Chapter 3 leading to the design of a novel hybrid interface that addresses the limitations of existing news interfaces. The set of requirements for the news dashboard are described in Section 4.1, which guided the implementation of the news dashboard shown in Section 4.3. A number of scenarios are provided to show examples of using the dashboard for news exploration in Chapter 5.

6.1 Primary Contributions

This research has resulted in the design of a novel real-time dashboard that supports the interactive exploration of news streams. The news interface can support information-seeking activities of the users by creating a visual map of the news stream, which meets the single-paged requirement of a dashboard [18]. The main contributions of the work are the following:

- The design of a news interface that supports a compact representation of the news stream, which results in a glanceable and pre-attentively processed news interface. The interface supports the visual encoding of the frequency of the news within elements of a stacked heatmap using colour. The colours of the elements of the stacked heatmap are chosen to support the perceptual ordering of the frequency of news items contained therein. The users can benefit from the use of visualization to form a visual map of the news stream, thereby reducing the cognitive burden when passively or actively observing trends in the news.
- The faceted browsing interface [28] provides a framework for automatically filtering the news stream. This can provide visual cues to guide the exploratory process of uncovering hidden trends in the news dashboard. This provides a hierarchical classification of the news streams according to persons, topics, and organizations. The faceted browsing interface allows the user to modify their exploratory plan in order to satisfy their information needs. The faceted browsing interface includes histograms which shows the frequency of the facet terms in the news stream. The visual encoding of the histograms allows the user to know at a glance the frequency of occurrence of an entity in the news stream.
- The addition of a breaking news detection feature based on the anomaly detection [12] is a core part of the news dashboard. The breaking news detection system for the news dashboard uses the Probabilistic Exponentially Weighted Moving Average (PEWMA) [12] algorithm. The news interface provides the ability to identify breaking news on specific persons, topics, and organizations independently of the other, which is an important feature in a news dashboard. This information can keep the user abreast of developments in the news stream.

6.2 Limitations

The benefits of using the news dashboard for the interactive visualization of news streams can be shown by using a number of scenarios in Chapter 5. Despite the benefits of the dashboard, there are still some limitations that are discussed in this section.

The lack of support for panning interaction to alter the time extent of the current stacked heatmap in the visualization space is a limitation. The current implementation has a fixed time extent, which is set in relation to the time aggregation. For example, if the time aggregation of the news dashboard is set to “Days vs Hours”, then the time extent is set to a month of news data. Similar, if the time aggregation of the news dashboard is set to “Hours vs Minutes”, then the time extent is set to a day of news data. The system lacks the ability to adjust the time extent beyond the extremes of the extents. The interface also lacks a focus+context interface for altering the time extent of the visualization.

One drawback of the news dashboard is the lack of an option to save settings [29] of the visualization. The current implementation makes use of settings that are session-based; as a result the settings are only valid for the run of the visualization. The ability to save settings could lead to news personalization. This feature allows the user to utilize past visualization settings in future exploration tasks.

The faceted browsing interface has limited interactivity which is a limitation. The current implementation provides a list of terms that highlight the elements of the stack heatmap containing the news about the term. It does not support dynamic deletion or addition of terms, which can affect the ability to fine-tune the navigation to smaller subsets of the data. The faceted browsing interface can provide a way of customizing the interface to contain only the facet terms that are relevant to their information needs.

The current implementation of the news dashboard makes use of textual data with temporal attributes that helps the organization of news in the stream showing the temporal evolution of the news stream. The system lacks support for complex data, such as geospatial data which results in the inability of the users to understand the news in the context of the geographic location. This feature can allow the user to understand the trends about news from certain locations.

The anomaly detection algorithm for breaking news can identify breaking news that are not a global phenomenon. The breaking news tends to show regional bias, as an collection of news of similar information content (same persons and organizations) of interest to a geographical area may be reported by more news outlets in the area, thereby resulting in the detection of the breaking news may not be a global phenomenon.

The faceted browsing interface on the news stream makes use of the persons, topics and organizations as facets. The use of the source of the news for faceted browsing interface is not currently supported in the news dashboard. The source of the news can serve as a measure of credibility of the news item. This functionality would support the filtering by source of the news.

The project made use of inspection heuristics [50], which are suitable for identifying usability issues, but does not consider users in real life conditions. Therefore, the absence of a user study which describes the usefulness of the news dashboard is a known limitation of this work.

6.3 Future Work

This section provides pointers for further development of the ideas developed in this research. These are possible advancement of the knowledge obtained from the current work, and the addressing the limitations in the news dashboard.

The modification of the news dashboard to provide for panning interaction, would allow the user to navigate beyond the currently displayed extents of the news dashboard. The user can navigate to historical news at a distant time in the past, thereby allowing the users to view the current news in relation to old news. This can enhance the ability of the news dashboard to keep sufficient data about the news stream and make the users see the long term trends in the news stream. This gives the users the ability to go back in time to view historical data, thereby providing functionality that support the historical exploration of news streams.

Modification of the news dashboard may be made to provide the ability to save settings of the visualization. The users can recreate similar exploratory conditions at a later time. The ability to saving settings [29] can lead to news personalization, as the user can modify the default settings of the visualizations, which include slider

for time aggregation and group by radio button. Further, this can allow the user to continue exploring the news stream using previously saved settings. The addition of the saved settings functionality can prevent the cognitive burden of having to recollect from memory the past setting to observe some trends in the news stream

Further modifications of faceted browsing interface may support the adding and removing of facet terms to modify the query. This can allow the users to focus on a smaller subset of the news stream for focused attention. The news dashboard can be enhanced by the use of a focus+context interface that can allow for altering the time extent of the sliding window of the visualization. This can provide the ability to make fine-tuned navigation to analyze varying subsets of the visualized data.

The news dashboard can be modified to support geospatial data which can allow the users to understand the news in the context of the geographical location. The dashboard can be modified to accept different formats of location data such as those commonly used in the Geographic Information Systems (GIS). This can take the form of integration with a map interface that shows where the current news has occurred.

The de-duplication module in the preprocessing step can be designed with higher accuracy and precision. Alternatively, the news can be weighted by the population of the location. This will provide a measure that can be incorporated in the breaking news detection algorithm thereby minimizing the chance of obtaining breaking news with localized effects.

The system may be modified to support the use of the source of the news in faceted browsing interface. This upgrade involves modifying the news dashboard to support filtering of the news stream by the source of the news in the news dashboard.

Finally, the design and execution of experiments will allow for the empirical evaluation of the system using real users. This will provide a measure of the level of user acceptance and usefulness of the news dashboard in real world situations. The participants will perform realistic tasks under laboratory conditions to mimic the normal user, and as such the demographics of the users must match to the intended audience of the news dashboard. This is provide a validation of the design assumptions used in the design of the news dashboard.

References

- [1] British Broadcasting Corporation. <http://www.bbc.com/>. Accessed: October, 2015.
- [2] Canadian Broadcasting Corporation. <http://www.cbc.ca/news>. Accessed: October, 2015.
- [3] Google News. <http://news.google.ca/>. Accessed: October, 2015.
- [4] Internet Archive. <https://archive.org/index.php>. Accessed: October, 2015.
- [5] Signal Media Limited. <http://signal.uk.com/>. Accessed: October, 2015.
- [6] Yahoo News. <https://news.yahoo.com>. Accessed: October, 2015.
- [7] Samar Al-Hajj, Ian Pike, and Brian Fisher. Interactive dashboards: Using visual analytics for knowledge transfer and decision support. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 1–8, 2013.
- [8] Maha Althobaiti, Udo Kruschwitz, and Massimo Poesio. Identifying named entities on a university intranet. In *Proceedings of the 4th Computer Science and Electronic Engineering Conference*, pages 94–99, 2012.
- [9] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.
- [10] Nan Cao, David Gotz, Jimeng Sun, Yu-Ru Lin, and Huamin Qu. SolarMap: Multifaceted visual analytics for topic exploration. In *Proceedings of the 11th International Conference on Data Mining*, pages 101–110, 2011.

- [11] Nan Cao, Jimeng Sun, Yu-Ru Lin, David Gotz, Shixia Liu, and Huamin Qu. FacetAtlas: Multifaceted visualization for rich text corpora. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1172–1181, 2010.
- [12] Kevin M. Carter and William W. Streilein. Probabilistic reasoning for streaming anomaly detection. In *Proceedings of the Statistical Signal Processing Workshop*, pages 377–380, 2012.
- [13] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452, 2012.
- [14] Brock Craft and Paul Cairns. Beyond guidelines: What can we learn from the visual information seeking mantra? In *Proceedings of the 9th International Conference on Information Visualization*, pages 110–118, 2005.
- [15] Gregory Ditzler and Robi Polikar. Incremental learning of concept drift from streaming imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 25(10):2283–2301, 2013.
- [16] Marian Dörk, Sheelagh Carpendale, Christopher Collins, and Carey Williamson. VisGets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1205–1212, 2008.
- [17] Wenwen Dou, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle Zhou. LeadLine: Interactive visual analysis of text data through event identification and exploration. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 93–102, 2012.
- [18] Stephen Few. *Information Dashboard Design: The Effective Visual Communication of Data*. O’Reilly Media, Sebastopol, CA, USA, 2006.
- [19] Johan Galtung and Mari Ruge. The structure of foreign news: The presentation of the Congo, Cuba, and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research*, 2(1):64–91, 1965.

- [20] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases*, pages 518–529, 1999.
- [21] Marko Grobelnik and Dunja Mladenic. Visualization of news articles. *Informatika (Slovenia)*, 28(4):375–380, 2004.
- [22] Tony Harcup and Deirdre O’Neill. What is news? Galtung and Ruge revisited. *Journalism Studies*, 2(2):261–280, 2001.
- [23] Susan Havre, Beth Hetzler, and Lucy Nowell. ThemeRiver: Visualizing theme changes over time. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 115–123, 2000.
- [24] William Hays. *Statistics*. Cengage Learning, Boston, MA, USA, 5th edition, 1994.
- [25] Christopher G. Healey and James T. Enns. Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188, 2012.
- [26] Marti A. Hearst. TileBars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 59–66, 1995.
- [27] Marti A. Hearst. User interfaces and visualization. In Ricardo Baeza-Yates and Berthier Ribeiro-Neto, editors, *Modern Information Retrieval*, pages 257–323. Addison-Wesley Longman Publishing, Boston, MA, USA, 1999.
- [28] Marti A. Hearst. Design recommendations for hierarchical faceted search interfaces. In *Proceedings of the ACM SIGIR Workshop on Faceted Search*, pages 26–30, 2006.
- [29] Marti A. Hearst. *Search User Interfaces*. Cambridge University Press, New York, NY, USA, 2009.
- [30] Ewald Hering. *Outlines of a Theory of Light Sense (Grundzge der Lehr von Lichtsinn, 1920)*. Harvard University Press, Cambridge, MA, USA, 1964.

- [31] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.
- [32] Orland Hoeber and Joshua Gorner. BrowseLine: 2-D timeline visualization of web browsing histories. In *Proceedings of the 13th International Conference on Information Visualization*, pages 156–161, 2009.
- [33] Orland Hoeber and Xue D. Yang. HotMap: Supporting visual exploration of web search results. *Journal of the American Society for Information Science and Technology*, 60(1):90–110, 2009.
- [34] Daniel A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.
- [35] Daniel A. Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. In Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North, editors, *Information Visualization: Human-Centered Issues and Perspectives*, pages 154–175. Springer-Verlag, Berlin, Heidelberg, 2008.
- [36] Daniel A. Keim, Florian Mansmann, and Jim Thomas. Visual analytics: How much visualization and how much analytics. *SIGKDD Explorations Journal*, 11(2):5–8, 2010.
- [37] Mark G. Kelly, David J. Hand, and Niall M. Adams. The impact of changing populations on classifier performance. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 367–371, 1999.
- [38] Taimur Khan, Henning Barthel, Achim Ebert, and Peter Liggesmeyer. Visualization and evolution of software architectures. In *Proceedings of the Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling, and Engineering Workshop*, pages 25–42, 2012.
- [39] Artjom Kochtchi, Tatiana Landesberger, and Chris Biemann. Networks of names: Visual exploration and semi-automatic tagging of social networks from newspaper articles. *Computer Graphics Forum*, 33(3):211–220, 2014.

- [40] Kurt Koffka. *Principles of Gestalt Psychology*. Harcourt, Brace and Company, New York, NY, USA, 1935.
- [41] Milos Krstajic, Enrico Bertini, and Daniel A. Keim. CloudLines: Compact display of event episodes in multiple time-series. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2432–2439, 2011.
- [42] Miloš Krstajić, Enrico Bertini, Florian Mansmann, and Daniel A. Keim. Visual analysis of news streams with article threads. In *Proceedings of the 1st International Workshop on Novel Data Stream Pattern Mining Techniques*, pages 39–46, 2010.
- [43] Angela M. Lee and Hsiang I. Chyib. The rise of online news aggregators: Consumption and Competition. *International Journal on Media Management*, 17(12):3–24, 2015.
- [44] Miguel Martinez-Alvarez. Descriptive modelling of text classification and its integration with other IR tasks. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1317–1318, 2011.
- [45] Miguel Martinez-Alvarez, Alejandro Bellogin, and Thomas Roelleke. Document difficulty framework for semi-automatic text classification. In *Proceedings of the 15th International Conference On Data Warehousing and Knowledge Discovery*, pages 110–121, 2013.
- [46] Miguel Martinez-Alvarez, Udo Kruschwitz, Wesley Hall, and Massimo Poesio. Signal: Advanced real-time information filtering. In *Proceedings of the 37th European Conference on Information Retrieval Research*, pages 793–796, 2015.
- [47] Andrew Mehler, Yunfan Bao, Xin Li, Yue Wang, and Steven Skiena. Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):765–772, 2006.
- [48] Amy Mitchell. State of the news media. <http://www.journalism.org/2015/04/29/state-of-the-news-media-2015/>. Accessed: April, 2015.

- [49] David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Analyzing entities and topics in news articles using statistical topic models. In *Proceedings of the International Conference on Intelligence and Security Informatics*, pages 93–104, 2006.
- [50] Jakob Nielsen. Enhancing the explanatory power of usability heuristics. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 152–158, 1994.
- [51] Daniela Oelke, Halldor Janetzko, Svenja Simon, Klaus Neuhaus, and Daniel A. Keim. Visual boosting in pixel-based visualizations. *Computer Graphics Forum*, 30(3):871–880, 2011.
- [52] Peter Pirolli. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press, New York, NY, USA, 2007.
- [53] Peter Pirolli and Stuart K. Card. Information foraging. *Psychological Review*, 106(4):643–675, 1999.
- [54] Jenny Preece, Yvonne Rogers, and Helen Sharp. *Interaction Design*. John Wiley & Sons, New York, NY, USA, 2002.
- [55] Anand Rajaraman and Jeffrey D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2nd edition, 2014.
- [56] Earl Rennison. Galaxy of News: An approach to visualizing and understanding expansive news landscapes. In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology*, pages 3–12, 1994.
- [57] Jonathan C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Proceedings of the 5th International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pages 61–71, 2007.
- [58] Thomas Roelleke, Hany Azzam, Marco Bonzanini, Miguel Martinez-Alvarez, and Mounia Lalmas. The D2Q2 framework: On the relationship and combination of language modelling and TF-IDF. In *Proceedings of the Lernen, Wissen & Adaptivität Workshop*, pages 33–40, 2013.

- [59] Mary B. Rosson and John M. Carroll. *Usability Engineering: Scenario-based development of human-computer interaction*. Morgan Kaufmann Publishers, San Francisco, CA, USA, 2002.
- [60] Gerard Salton, Andrew Wong, and Chungshu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
- [61] Hanan Samet, Jagan Sankaranarayanan, Michael D. Lieberman, Marco D. Adelfio, Brendan C. Fruin, Jack M. Lotkowski, Daniele Panozzo, Jon Sperling, and Benjamin E. Teitler. Reading news with maps by exploiting spatial synonyms. *Communications of the ACM*, 57(10):64–77, 2014.
- [62] Ben Shneiderman. Dynamic queries for visual information seeking. *IEEE Software*, 11(6):70–77, 1994.
- [63] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336–343, 1996.
- [64] Barry Smith. *Foundations of Gestalt Theory*. Philosophia Verlag GmbH, Munich and Vienna, 1988.
- [65] James J. Thomas and Kristin A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.
- [66] Edward Tufte. *Envisioning Information*. Graphics Press, Cheshire, CT, USA, 1990.
- [67] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1986.
- [68] Sergio Vavassori, Javier Soriano, David Lizcano, and Miguel Jimenez. Explicit context matching in content-based publish/subscribe systems. *Sensors*, 13(3):2945–2966, 2013.
- [69] Matthew Ward, Georges Grinstein, and Daniel A. Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. A. K. Peters, Ltd., Natick, MA, USA, 2010.

- [70] Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers, San Francisco, CA, USA, 2004.
- [71] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. Tiara: A visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 153–162, 2010.