
Electronic Thesis and Dissertation Repository

6-26-2020 11:00 AM

Triaging Twitter Users: An Exploratory Visual Analytics System

Parinaz Nasr Esfahani, *The University of Western Ontario*

Supervisor: Sedig, Kamran, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in
Computer Science

© Parinaz Nasr Esfahani 2020

Follow this and additional works at: <https://ir.lib.uwo.ca/etd>



Part of the [Artificial Intelligence and Robotics Commons](#), [Data Science Commons](#), and the [Graphics and Human Computer Interfaces Commons](#)

Recommended Citation

Nasr Esfahani, Parinaz, "Triaging Twitter Users: An Exploratory Visual Analytics System" (2020). *Electronic Thesis and Dissertation Repository*. 7112.

<https://ir.lib.uwo.ca/etd/7112>

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.

Abstract

Twitter is one of the most popular microblogging and social networking services. Many people from a wide range of backgrounds use Twitter to contribute their thoughts on different topics through postings, known as “tweets”. Analysts collect and analyze tweets to extract knowledge. To rely on tweets, it is crucial to assess Twitter users’ credibility. In recent years, researchers have proposed various techniques, especially data analytics models, for evaluating Twitter users and analyzing their behaviour; however, these techniques do not engage analysts in the process, leading to a lack of understanding and trust in results. In this thesis, an exploratory visual analytics system is designed and implemented to help with triaging Twitter users. To this end, the system can leverage analysts’ expertise and knowledge through interactive visualization to assist them in understanding the underlying information within data. Subsequently, a case study demonstrates the capabilities of the system in identifying Twitter users.

Keywords: visual analytics, triage, twitter data, twitter user group association

Summary for Lay Audience

As one of the most popular microblogging and social media services, Twitter has millions of monthly active users who publish an abundance of daily postings. This platform allows users from a wide range of backgrounds to publish their thought and opinion on various topics. Analyzing Twitter users' behaviour in terms of contributions to different topics and their group association is an integral part of investigating tweets. For instance, many automatic accounts, known as bots, are being used to propagate different content on Twitter. In addition, some associated accounts may publish fake news on Twitter. Therefore, to rely on tweets as a source of information, a journalist who aims to collect tweets for news and stay updated has to analyze and identify the users who posted about specific topics. The current approaches for identifying Twitter users do not engage the analysts in the process, leading to a lack of understanding and trust in results. This thesis proposes a system that represents information through interactive visualizations. Besides, the system takes advantage of data analysis models to support users' identification process. A case study shows that the system assists analysts in understanding Twitter users' information and detect the hidden patterns and potential similarities between the users.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Sedig, for all his guidance and continuous support. My grateful thanks are also extended to Amir Haghghati for kindly providing me with VARTTA codes and his generous help. I would also like to thank the wonderful lab-mates in the Insight Lab for their continuous support and company.

And last but not least, I want to express my gratitude to my family for their loving support, and also to my dear friends, who gave me their support and company.

Contents

Abstract	i
Summary for Lay Audience	ii
Acknowledgements	iii
List of Figures	vi
1 Introduction to the Study	1
1.1 Introduction	1
1.2 Questions and Objectives	5
1.3 Definition of Terms	6
1.4 Overview of the Structure	6
2 Literature Review	7
2.1 Twitter Data	7
2.2 Twitter Users' Identification	8
2.3 Exploratory Systems	10
2.3.1 Triage	11
2.3.2 Distributed Cognition	13
2.4 Visual Analysis	14
3 Methodology	18
3.1 Data	19
3.1.1 Tweet Topics	20
3.1.2 Twitter Users' Similarity	20

3.2	Triaging System Design	21
3.2.1	Information Space and Task Space	23
3.2.2	Topic-User Association	24
	Visualization	25
	Interaction Design	25
3.2.3	Time-User Association	29
	Visualization	29
	Interaction Design	31
3.2.4	User-User Similarity	33
	Visualization	34
	Interaction Design	34
3.2.5	Components Aggregation	34
3.3	Implementation	39
3.4	Case Study	40
	Level One: Multiple Documents	40
	Level Two: Within Documents	42
	Level Three: Further Investigation	44
4	Conclusion	45
4.1	Discussion	45
4.2	Future Work	46
	Bibliography	47
	Curriculum Vitae	53

List of Figures

2.1	Document triage sets	12
3.1	The format of Twitter data, a screen shot from some features	20
3.2	Triage process levels	22
3.3	Information space	24
3.4	Topic-User component for exploring contributions in topics/keywords	26
3.5	Topic-user component, arranging topics and keywords	28
3.6	Topic-User component, selecting and filtering actions	28
3.7	Time-User component for exploring Twitter users' activity over time	30
3.8	Designed options for interacting with information representation	31
3.9	Time-User component, inserting/removing additional stacks	32
3.10	Time-User component, collapsing/expanding	33
3.11	Twitter users' similarity component	35
3.12	Twitter users' similarity component, expanding and rearranging the Twitter users.	36
3.13	Pile of selected Twitter users	37
3.14	Drill down to see tweets	38
3.15	Triage Page on VARTTA	41
3.16	Multiple Documents	41
3.17	Multiple Documents - Filter	42
3.18	Within Document Triage - Similar Users	43
3.19	Within Document Triage - Using Similarity Component	43
3.20	Further Investigation	44

Chapter 1

Introduction to the Study

1.1 Introduction

With the advent of micro-blogging services, a massive amount of data has been generated by users of them. According to their nature, micro-blogging services allow users to express their thoughts and communicate with other users through texts, images, and video links. Among all micro-blogging platforms in today's world, Twitter is one of the most popular ones [1]. This service has attracted millions of monthly users from a variety of backgrounds who exchange up to 500 million tweets per day, in a wide range of topics [2]. Therefore, Twitter has become an invaluable source of data from which analysts can extract patterns and knowledge. Many researchers have conducted studies to make sense of tweets and their characteristics; however, only a few research has considered an exploratory approach to create a mental model of Twitter users.

As a type of social media service, micro-blogging enables users from different backgrounds to share their daily experiences, commentary, and perspectives toward various topics in the form of short postings [3]. These services facilitate their users' communication within a private network or in public, by exchanging brief messages. Compared to traditional blogging, micro-blogging limits the size of posts, which provides faster investment of generated content, and the users tend to publish updates with higher frequency [4, 5]. Several services provide micro-blogging, including Tumblr, Jaiku, Pownce, and Twitter.

Twitter is one of the most influential micro-blogging services launched in the Fall of 2006, and to date, it has attracted over 330 million monthly active users. Twitter allows its users to publish their posts, known as “tweets”, or to republish the others’ tweets. These posts would be accessible to their followers or the public, based on the users’ privacy setup. Besides, these tweets may appear as a result of a search conducted on Twitter’s searching tool for finding tweets that match certain words, keywords, dates, or places. Moreover, the results that return from search engines, such as Google, automatically include live updates of micro-blogs content; therefore, any post through Twitter can be available as public knowledge [4]. Accordingly, Twitter has become a principal medium for broadcasting information and communicating messages, which has attracted attention from the popular press and scholars [6, 7] either for publishing new content or for taking advantage of existing ones.

For analysts, Twitter is an invaluable resource that provides them with data on a wide range of topics. Users with different backgrounds such as health-care practitioners, politicians [7], humanitarian assistants [8, 9], activists [10], journalists [11–13], scholars [7, 14, 15], or employees in a workplace [16] can use Twitter for announcing events and sharing thoughts and resources with others. Simultaneously the users can read the others’ tweets to gain insight into a topic of interest or to contact or follow other contributors. In the case of journalism, Twitter is considered as a news environment that allows professional and non-professional citizen journalists to broadcast news tweets. Many professional journalists publish the content or hyperlink to a news website on Twitter, and other users can retweet that content. On the other hand, the public may be involved in news production and propagation by sharing what is going on around them or retweeting the others’ posts. Observing those tweets, professional journalists may choose to contact the authors for probing more information and covering the news [15, 17]. These are just a few examples of the indisputable role of Twitter in the dissemination of diverse information. An integral part of acquiring a high-level overview from potential knowledge among this massive data is to investigate Twitter users as the authors of tweets.

Exploring Twitter users' behaviour over time and in different topics is essential for making sense of information space and trusting tweets as a source of information. Along with all of its advantageous data, Twitter can play the role of a misinformation propagation platform [10, 18, 19] which contains unreliable content, as well. This content can be generated either unintentionally or intentionally by users, including bots, parody accounts, or real users that are possibly associated with a particular group. Therefore, an exploratory system is required to extract and assess tweets' features. Regarding the journalism, a journalist who attempts to collect information from tweets and produce credible content has to go over a large set of tweets to evaluate parameters [20] such as novelty [21], credibility, and the group association of the users. Thus, an exploratory system is required for enabling analysts to perform cognitive tasks on data for identifying Twitter users.

Accessing Twitter users and their tweets is a prerequisite for knowledge extraction from Twitter data; however, to accomplish a cognitive task based on this data, an analyst needs to go further and explore the information space for underlying patterns and evidence [22, 23]. To this end, analysts have to examine each user individually or within a group of users to detect any pattern. Generally, analysts seeking information among a large set of available documents, such as returned documents from search tools, sort through many documents to arrange them based on their relevance to a desired information. This time-constrained assessment, which usually takes place based on insufficient knowledge, is known as "triage" [24, 25]. The word "triage" originates from the French verb "trier", which means to sift or to sort. A triage process categorizes individuals from a large set of documents to facilitate grouping and cognitive activities process [26]. Accordingly, it is essential to develop a triaging system for investigating Twitter users' behaviour in terms of their contributions.

According to distributed cognition theory, rather than taking place on an individual's brain, cognitive activity ensues through a collaboration of the human and environment. This collaboration occurs over time and is the result of interactions among internal and external resources [27, 28]. For instance, an analyst aiming to make sense of Twitter users has to go through a dataset, process and interact with data, observe the results, use the background knowledge to

analyze, and understand laying patterns or features. If any new information is found, the analyst may upgrade background knowledge with this new piece of information, and presumably, start over the procedure to gain more insight from data. Accordingly, the cognitive activity of sense-making does not merely happen in the analyst's mind; instead, it takes place through a combination of factors such analytical reasoning ability and background knowledge of the analyst, computational and representational power of computers, and the interactions that the analyst has with the represented information. Therefore, to achieve a cognitive activity, a system must balance between internal and external resources.

Developing an analytical system for identifying Twitter users and their group association is challenging. Twitter users are identified with their usernames, screen names, and arbitrary profile attributes, as well as their tweets' contents and features. So the information about their behaviour, relationships, and associations should be extracted based on the patterns in this information [29]. For transferring this information to analysts, data visualization techniques are essential to thoroughly represent the massive Twitter data in different levels of granularity. However, it would not be feasible to find some patterns and features within the represented raw data, so the analysts need to take advantage of proper analytical models. For instance, to find relationships among Twitter users, their attributes should be aggregated and compared, which can barely take place without analytical models. In addition, it is crucial to design an interaction protocol that helps analysts work with visualizations.

Most studies on Twitter users exploration and identification have focused on different analytical models to detect patterns and users' attitudes, identification, and associations [12, 30–34]. Even though these approaches can extract behavioural patterns of Twitter users, as the analysts are not engaged in the process, these techniques lead to a lack of understanding, control on the process, and trust in the results. On the other hand, some research has considered information visualization techniques together with analytical models [35–37], yet these approaches do not offer a comprehensive interactive visualization for different aspects of data. This would limit analysts' ability to investigate and create a mental model of Twitter users' information.

Visual analytics (VA) is a body of knowledge that supports analytical reasoning through information visualization, human-information interaction, and data analysis techniques [38]. Deriving insight from massive multidimensional data is challenging, so information visualization offers techniques to transfer complex information into comprehensible representations and shift the burden from the cognition system to the perception system [39]. On the other hand, data analytical models can extract hidden patterns and behaviours from data; however, due to their non-transparent processing nature, these methods lead to a lower degree of comprehensibility for analysts, and a lack of confidence in the results [40]. VA systems support analysts to execute complex cognitive activities on a large set of data [29, 40, 41] by applying information visualization techniques along with analytical models on data to represent and extract patterns and features that otherwise cannot be obtained. Based on how humans interact with the information space, these systems provide interactive interfaces.

This thesis aims to develop a system for helping analysts delve into Twitter users' information to make sense of their behaviour and group association. The information is represented based on information visualization principles, and the interactions are built on the EDIFICE-AP framework epistemic action patterns [42]. This framework supports analysts to create a mental model of the information space through interactive visualizations. Besides, the thesis shows how analytical models can obtain and represent information that analysts cannot easily extract by interacting with data. The system is developed as a part of an existing VA system that supports real-time monitoring, analyzing, and making sense of tweet streams [43].

1.2 Questions and Objectives

Analysts, including scholars, journalists, and anyone intending to extract knowledge out of Twitter data, need to examine the Twitter users. These users are the agents who generate and determine the tweets' characteristics that affect the result of the analysis. To this end, analysts need to employ an exploratory system that provides them with information space visualization, the ability to interact with those visuals, and computation assistance for deriving attributes

and patterns. This system enables analysts to create a mental model of the users' behaviours during the time and their association and similarities to other users. Intending to design such an exploratory system, this research poses and addresses the following research questions:

1. How would VA be employed to implement a triaging system for exploring Twitter users?
2. How would the VA system analysts engage in a triaging procedure?
3. How would the analysts use the system to determine tweets novelty, determine a user's credibility, and select a set of users?

1.3 Definition of Terms

Some of the terms in this research may seem ambiguous as they can describe different concepts, or there may be multiple terms to define the same concept. So to avoid confusion, selected terms and meanings are clarified as follows.

- **Tweet:** refers to a post made on Twitter.
- **User:** also known as Tweeter and Twitterer, refers to one who posts tweets.
- **Analyst:** refers to a potential user of the VA system.
- **User's behaviour:** refers to contributions of a specific user during the time and within different topics.

Hereafter, these words will be used for referring to the mentioned concepts.

1.4 Overview of the Structure

This research is organized as follows: Chapter 2 gives a brief overview of the background on Twitter data analysis, cognitive activities, and VA systems. The third chapter describes a new exploratory system for triaging Twitter users and, through a case study, illustrates how the designed VA system assists analysts to fulfill the cognitive tasks. Subsequently, chapter 4 presents a discussion on the triaging Twitter users system and suggests some ways to extend the research.

Chapter 2

Literature Review

In the literature, there has been a variety of experiments on Twitter users' identification. Even though most of these experiments use data analytics models, some research debate the capability of the sheer use of these techniques to address the need for information exploration when more investigation is required. In this regard, the focus of the review is on exploratory, specifically triage systems. According to the literature, visual analysis is an indispensable part of implementing an exploratory system. Therefore, this chapter presents an overview of research on Twitter users, exploratory systems, and the principles of visual analytics.

2.1 Twitter Data

To retrieve Twitter data, analysts such as researchers, companies, and developers can access or purchase public data through Twitter's application programming interface (API) or existing repositories. The source of the dataset may affect the quality of data. For instance, mostly the datasets provided by free services are noisy and untidy [43]; however, different techniques such as data cleaning methods can be employed to identify incomplete or irrelevant parts of data and modify or remove it [44] based on the requirements of the problem that has to be solved.

An important parameter of Twitter streams is their content topic. Using this attribute, analysts can filter the massive amount of data and focus on tweets in a field of interest, which is crucial for both analyzing real-time and historical data. Therefore several techniques are sug-

gested to detect the underlying topic of tweets. In a survey, Ibrahim et al. [45] studied different systems and approaches for detecting topics of data from Twitter streams. Pointing out the use cases of this attribute in different analyses, the authors claim each technique has different outcomes and performance. They introduce a taxonomy of the existing techniques. As the survey suggests, one of the conventional approaches for topic detection is to describe each topic by a set of keywords.

2.2 Twitter Users' Identification

In recent years there has been substantial interest in studying Twitter users as the agents that generate or propagate content on Twitter. Many researchers attempted to classify Twitter users based on criteria such as users' demographic information or association with a specific group [15]. According to the literature, the most prominent characteristics of the users are their network (such as following and follower accounts), their behaviour (contributions in different topics during the time), and their association to potential groups. The first characteristic is accessible within Twitter's API dataset; however, the available Twitter data does not contain many aspects of the users' behaviours and background. Hence, for accessing these features, a more in-depth investigation is required. The existing literature indicates that identifying Twitter users is crucial for analysts; therefore, many researchers endeavour to develop approaches to extract this information.

Most studies have focused on data analytics models for attaining Twitter users' derived attributes; however, along with using such techniques, a considerable number of approaches need to explore the information space and compare the individual entities by human agents. For instance, to use scientific contents on Twitter, Hadgu and Jäschke [14] developed a classification approach to distinguish researcher users from non-researchers ones. For training their model, they collected a set of researcher users by filtering and finding relevant tweets based on their profile information and tweet contents. They mentioned that for extending the work, a group of researchers would help to distinguish researchers association to different research areas and identify expertise and connections between disciplines. So exploring Twitter users

and their tweets is inevitable.

Likewise, various approaches have been proposed algorithms to detect the bots. Since using bots for distributing information has been widespread among news and information services [46], detecting such automated accounts is critical. Exploring bots helps to understand their intention and group association and differentiate between malicious and safe bots. Referring to the possible harms of malicious bots, Chavoshi et al. [30] suggest a real-time unsupervised approach for detecting bots by correlating their activities. The authors demonstrate a video capture of two unrelated users and a group of bot accounts. The authors indicate that the agents' highly correlated activities can only be justified as if they are controlled automatically as bots. Later, for verification of their method, the researchers asked human judges to compare and find the relationships and patterns in their detected bots.

As a further example, Lokot and Diakopoulos [46] studied the role of journalism bots in information dissemination on Twitter. They manually collected a set of known bots, based on the account titles, handles, and bios, to examine these automated accounts' functionality. The samples were limited to a set of known bots that did not hide their identity, and the existing bot detector classifier they used failed to detect most of the actual bots. Indicating the non-transparency of news bots and their creator's intention, the authors emphasize the necessity of understanding the credibility of these automated accounts by journalists and other news consumers. In another work, Chen and Subramanian [32] developed a bot detector that filters Twitter users based on the keywords in their tweets. According to the duplication of contents, the algorithm classifies the users and subsequently inspect their URL to discover the source. These strategies can help analysts ascertain aspects of underlying information about the users; however, as mentioned by Kahng et al. [47], understanding these methods' functionality, interpreting the results, and trusting them remains a challenge for the analysts.

Accordingly, to rely on Twitter data analysts need to identify Twitter users and their association. To address this concern, most of the reviewed studies offered data analysis models. Some of these works required human agents to explore the users' information individually at

different stages, such as collecting a set of training data or validating the results. Whereas some other methods defined criteria and thresholds to automatically collect the users, the absence of analysts in the process may lead to an inadequate understanding of the process, misinterpreting the results for them, and lack of trust and adaptability [40]. Besides, these techniques do not take advantage of analysts' expertise in the process. Many researchers have studied the role of analysts and approaches to involve them in the process concerning these challenges. The next section discusses these experiments and systems for exploring Twitter data and making sense of them in the context of a triaging task.

2.3 Exploratory Systems

White and Roth [48] defined exploratory search as a type of information seeking and a sense-making activity concentrated on gathering and using information when analysts are unfamiliar with a domain of interest or unsure of how to achieve it. Explaining the essential information activities, the authors assert that information visualization is “an important element in hypothesis and insight generation and learning information landscape.” In another study on Twitter users' exploratory search behaviour on social media and the web, Choi et al. [21] claimed that exploring social media provides analysts with a more focused set of relevant documents and resources. They classified other scholars' work into two categories: some scholars investigated the nature of exploratory searches, including uncertainty, innovation, knowledge discovery, learning, and investigation, while others evaluated information with measures such as mean reciprocal and novelty.

Referring to the role of investigative searches in achieving a high-level mental model, Marchionini [49] declares three main search activities, including lookup search, learning search, and investigating search. Lookup searches are closed-ended for retrieving known information, and question/answering. Exploratory search is more akin to the learn and investigating search [21]. Learning search involves returning sets of objects in multiple iterations that require cognitive processing and interpretation, and investigate search takes place to accomplish a complex cognitive activity. As Dimitrova et al. mentioned [50], the relationship among these phases

is not linear. Instead, analysts go through a combination of stages to extract discoveries from presented information, and the goal is to help analysts build relationships among discovered items. For instance, in an exploratory search, an expert has to sort through a set of documents and determine which resource may be useful, and meanwhile, the expert may wish to change the exploratory style into a more effective one. Therefore, in the design of exploratory systems, there should be a certain degree of flexibility to give analysts more control over moving between stages. This idea is well examined in the distributed cognition theory and employed in visual analytics. In the following, the exploratory search is discussed in the form of triage, and then there is a brief overview of distributed cognition theory.

Twitter users create an invaluable online repository of data with their tweets. According to Chang et al. [51] to make sense of this available online data, analysts need to sift through a large set of documents, explore based on desired aspects, and find documents that meet their criteria. This procedure is known as triage. In this process, the analysts should be able to employ additional criteria. To this end, the researchers suggested employing visual explanation and a more in-depth exploration of search results to help analysts gain insight into the information space. For designing an exploratory system, distributed cognition theory is discussed after a brief overview of triage.

2.3.1 Triage

A plethora of digital documents are accessible online. Even though these documents have created a useful repository to find information on a topic of interest, one has to go through a myriad of documents to find the most relevant. To do so, analysts have to rapidly skim, scan, possibly compare, evaluate, and select different documents. This process is known as document triage. Bae et al. [24] define document triage as “the practice of quickly determining the merit and disposition of relevant documents”. The authors make a distinction between triage activity and other document findings. They point out that in a triage process, there are sets of focal and peripheral relevant documents, and the analysts seek the document sets to find the most relevant.

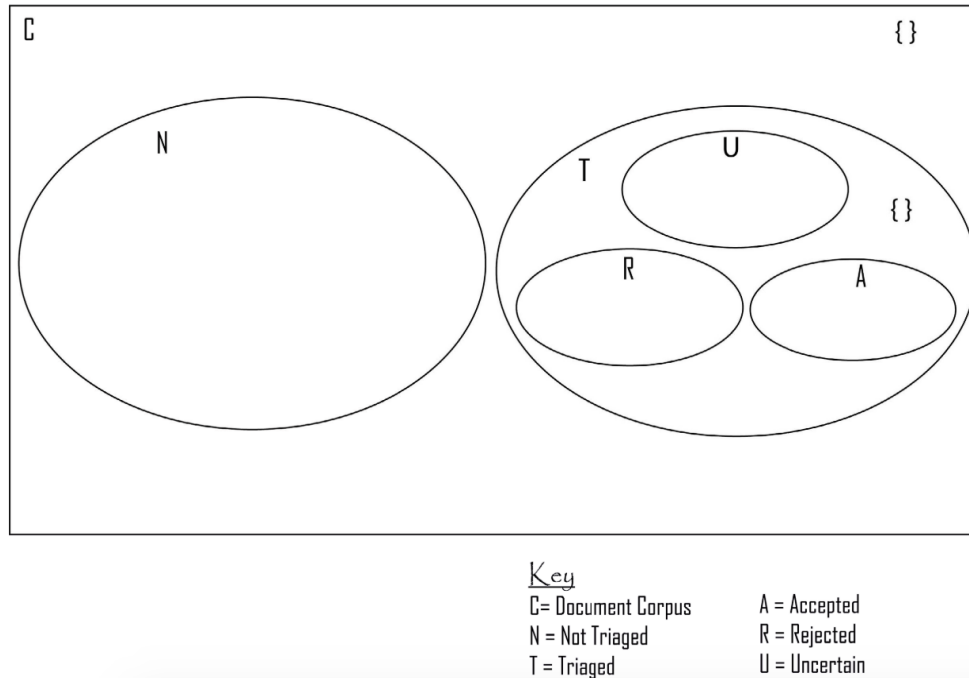


Figure 2.1: Document triage sets

The documents classified as triaged (T) or Non-triaged (N). The moment the triage process begins, the document set that the analyst works with would be classified as triaged. During the triage process each document would be categorized as accepted, rejected, or uncertain [25] ¹

The authors show that analysts mostly evaluate the documents based on their content in comparison with metadata, so when they find a set of useful documents, they spend more time on it.

With a focus on manual processes in document triage, Loizides and Buchanan [25] employ a set theory perspective to illustrate the process. As they describe, the process begins when documents are presented to the information seeker. Initially, all documents belong to the not triaged set. While triaging, documents move to the triaged set, partitioned in three possible sets: accepted for the relevant document, rejected for the irrelevant document, or uncertain documents for documents that need further examination (Figure 2.1) ¹. The process ends when all documents are examined, and the documents that are never evaluated are implicitly considered as rejected. During the triage process, if a change in knowledge status leads to a change in information needs, the process starts over.

¹The publisher's permission for reusing the figure is granted.

The first level of document evaluation in Loizides and Buchanan [25] study is “multiple document” triage [52]. First, an initial document set is selected without an in-depth examination, usually as a result of a query. In this level, the documents would be accepted as relevant or rejected. The next level is “within-document”, also known as “individual document” [52], in which analysts rapidly inspect some parts of individual accepted documents’ properties. After passing these two levels, the analysts need to read the relevant documents thoroughly to extract and decide on the documents’ relevance to the needed information. The authors introduce this stage as the “Further-Reading” level. While the analysts update their knowledge, they may circulate through these levels to explore documents. For developing a system to triage Twitter users according to their tweets, the system has to support analysts to move between these levels and gain insight into the information. To this end, distributed cognition theory should be adopted as a framework for design.

2.3.2 Distributed Cognition

In their seminal article, Hollan et al. [27] describe distributed cognition theory as an essential means of understanding human cognitive activity and designing effective interactions among humans and technologies. The challenges of supporting complex tasks, designing interaction, and managing exploitation of information have arisen with the advancement of human-computer interactions. To address the challenges, researchers require a theoretical foundation for an effective design of a human-centred interaction. The authors distinguished distributed cognition from the traditional view and emphasized that a cognitive process involves functional relationships among internal and external resources in the distributed cognition view. Moreover, a cognitive process may be distributed over time, as an external resource, in that the result of earlier events in a process affects the next events.

Distributed cognition has been proposed [28] as a robust framework for information visualization systems. Liu et al. point out that the traditional view of human-information processing encompasses three stages of perception, cognition, and action. Based on this view, cognition is

a process inside an individual's brain, and environment elements are acknowledged as a source of content. These elements are stored in memory as representations to assist the thinking process and associate with the perception stage. Therefore, in the traditional view, visual systems are considered as cognition amplifiers but do not take part in the cognitive process. However, in the distributed cognition view, cognition is an emergent property of interaction that couples external representations such as visual systems with the internal representations. Based on this view, in an information visualization system, individuals have to gain a mental model of the information space through interaction with visualizations. An exploratory visual analytics system built on distributed cognition can balance the internal and external resources of cognition.

2.4 Visual Analysis

Visual analytics is a body of knowledge that expedites analytical reasoning through interactive visual interfaces [29]. This multidisciplinary field encompasses information visualization, human-information interaction, and data analysis techniques for supporting analysts to accomplish a complex cognitive activity [40, 42, 53, 54]. A cognitive activity is considered as complex, provided it deals with complex information in complex circumstances [42], such as making sense of massive, dynamic, and uncertain data on Twitter. According to a survey by Wu et al. [54], numerous studies are exploring visual analytics methods for social media data. Examining various types of systems, the authors classify these studies in two classes of information gathering, including information retrieval and triage [43], and understanding users' behaviour.

If employed suitably, Information visualization techniques can transform unsuited complex raw information to comprehensible representations into which analysts can gain insight [55] and detect underlying facts and patterns. In recent years, information visualization has received a great deal of attention. Many studies have implemented visualization designs that are focused on specific purposes. The examples include Rotta et al. s' [56] work on visualizing Twitter users' clusters to support rumour detection and the graph visualization offered by Jus-

sila et al. [57] for representing the behaviour of scholar Twitter users attending a conference. These works show the information in a more comprehensible manner to analysts; however, they do not support researchers and developers in designing coherent visual analytics systems for various information spaces and scenarios.

On the other hand, the same data could be visualized in different ways that afford different levels of understanding. In his introductory to information visualization, Spence [55] presents some of the available representation techniques. He claims that if the techniques are not used appropriately, they can lead to misunderstanding. In addition to attempts to develop systems for specific scenarios, some researchers have endeavoured to develop frameworks for supporting analysts from different backgrounds to obtain a thorough view of a domain [58]. In their work on the design of visualizations for human-information interactions book, Sedig and Parsons [59] introduced a pattern-based framework. This framework allows designers to identify appropriate visual marks for representing information, along with describing the syntax for blending patterns, based on the objectives. These techniques can be applied to both static and interactive visualization. The hidden information should be visualized in two respects: externally, to transform information into understandable representations and visualizations, and internally as a mental model for the analysts who work with the VA system. A challenge for designing practical visualizations is to know how human works with information and reasons. These subjects are discussed in the human-information interaction field.

To accomplish a complex cognitive activity, analysts need to engage with visualization interfaces to control, modify, and transfer visualizations where needed. According to distributed cognition theory, complex cognitive activities such as decision making, problem-solving, sense-making, knowledge discovery, interpreting, planning, and investigating would be attained by the interaction of analysts with information visualizations. Even though interaction has an indisputable role for visualization approaches, literature shows that interaction did not receive enough attention in comparison with visualization techniques. Yet, as Tominski [60] reviews, there are some studies suggesting frameworks such as Sedig and Parson's [42, 59] EDIFICE-AP framework. This framework allows designers to predict interaction needs, and

creatively design and develop interactive interfaces applicable to different technologies, analysts, activities, and visualization. The authors illustrate that to accomplish a cognitive activity, analysts should engage the visual representation of information space for perception, and then, as tentacles apply actions to interact with information space and modify it as desired. So with breaking down a complex cognitive activity into more straightforward cognitive tasks and then into interactive and visual tasks, the framework introduces a comprehensive pattern-oriented list of interactions and visual marks.

In the case of Twitter users triage, analysts should be able to move among triage process levels. As analysts improve their knowledge, they may change their direction and strategy of exploration. To this end, a robust system is required to enable them to interact with information. The combination of information visualization and human-information interaction is useful in offering the opportunity to develop such potent systems that enable analysts to gain insight into the information; nonetheless, there still might be some aspects of information that analysts cannot attain without the help of data analytics methods. Therefore, VA systems leverage data analytics models to extract underlying information that analysts cannot achieve without them. In the triage process, after filtering all documents, analysts need to compare multiple documents and assess their relevance to the topic of interest, yet there is no explicit attribute in Twitter data to allow analysts to accomplish such an assessment. Accordingly, VA systems employ data analytics models such as machine learning methods together with information visualization and human-information interaction to empower analysts to create a high-level internal picture of information [40, 42].

In recent years there has been a growing interest in developing exploratory systems for analyzing Twitter data. The majority of research tended to focus on tweets sentiment and the users' networks based on their following and follower accounts, and there is not much work on using the VA system to triage Twitter users. In a survey on VA systems for social media [54], the authors introduce some works on extracting networks, social-temporal, and textual information. These works focused on extracting aspects of underlying information on Twitter data. In another study by Sana et al. [61], researchers developed a VA system that leverages linear

discriminant analysis (LDA) method for topic modelling and showing discussed topics and the trends on Twitter. This system provides analysts with an overview of the topic flow, though the approach does not consider contributor users in visualization. Likewise, Abdelsadek et al. [62] proposed a VA system for detecting users' communication. This system visualizes Twitter users' clusters based on topics they contribute in and provides interactions that allow analysts to drill down into the information visualizations to learn more; however, this work does not consider the time parameter of tweets, which is an essential parameter in detecting Twitter users' group association.

In a study on tweets analysis, Haghghati and Sedig [43] present a visual analytics system for monitoring and sense-making of real-time Twitter data (VARTTA). This system empowers analysts to gain insight into the tweet topics, content theme, and users' group. Further, since any analysis technique provides a different result [45], the system allows the analysts to apply various machine learning algorithms for sentiment analysis and topic detection, and compare the outcomes. As mentioned, VARTTA provides interactions to assist analysts in gaining insight into the most influential Twitter users, their groups, and tweets sentiments; however, the relationships of individual users with topics and the pattern of contributions over time remain ambiguous. Even though this information is available through different representations in the system, the required element for triaging tasks are scattered on different screens. Accordingly, designing a VA system to triage Twitter users for understanding their behaviour and group association is essential.

Chapter 3

Methodology

Delivering the underlying information within the Twitter data to analysts through interactive external and internal visualizations is challenging. First of all, Twitter users' behaviour toward different topics and their group associations are not explicitly available in the original data; instead, they are reflected in the contributions and actions of the users. Therefore, a Twitter user's behaviour must be defined according to the user's tweets as a whole. Another challenge in this regard is that due to a large number of available Twitter users' data and metadata, extracting and using such information requires techniques for assigning values to the users. This value would equip analysts with the ability to compare the users. Information visualization techniques can potentially engage the analysts in the process of knowledge extraction and take advantage of analytical models to address the challenges and create a deeper understanding of the information space and trust in the results.

In this thesis, an exploratory system is designed and implemented based on a framework for visualizations of human-information interaction techniques [59]. The framework provides conceptual aids for systematically designing systems, bringing the analysts in the process, and allowing them to apply their knowledge and to change representations as desired. Respectively, the system employs the comprehensive list of interactions presented in the EDIFICE-AP framework [42] for predicting the possible interaction that an analyst might need. Besides, the system is implemented as a part of VARTTA [43], since the design of VARTTA is compatible with the Twitter users' triaging system design. In addition, in a body, these systems can

serve analysts to achieve the ultimate goal of sense-making and understanding the information space. Eventually, to determine whether the developed system would assist the analysts in accomplishing cognitive activities, a case study is designed to illustrate how the system would help analysts identify Twitter users and their associations.

This chapter presents the design of triaging Twitter users system. The first section describes data as well as the steps for extracting attributes that are not available in the initial data. The next section illustrates the stages of constructing a VA system that translates the data into knowledge. Subsequently, the implementation description is provided in the following section. Finally, a showcase of the system provided to demonstrate how the designed system helps the analysts gain insight into Twitter users' behaviour.

3.1 Data

Having an overview of the literature, to explore Twitter users' behaviour, analysts need to consider their activities on different topics over time and possibly compare them with other relevant users. For this purpose, analysts require pertinent features with which they can assess and identify Twitter users. This research has used free samples of Twitter API data, available in the form of JavaScript Object Notation (JSON) files. These files encompass different types of features and metadata that show aspects of data such as location and description; however, these are arbitrary attributes determined by the Twitter users and may remain void or contain unreal values. So in this study, the focus is on the given attributes that are relevant to the Twitter users' behaviour and affiliation.

In general, tweets are identified with features such as id, release time, content, used symbols and hashtags, author, and other metadata (Figure 3.1). Likewise, the available features for identifying authors are id, screen name, network information (e.g. number of followers and followings accounts), profile information, joining date, and other metadata regarding the account. Other features such as Twitter users similarity and the topic of contributions are implicitly available in the content and features of the dataset and can be extracted with analytical

```

"created_at" : "Thu Apr 06 15:24:15 +0000 2017" ,
"id_str" : "850006245121695744" ,
"text" : "1\ / Today we\u2019re sharing our vision
"user" : {
  "id" : 2244994945 ,
  "name" : "Twitter Dev" ,
  "screen_name" : "TwitterDev" ,
  "location" : "Internet" ,
  "url" : "https://dev.twitter.com/" ,
  "description" : "Your official source for Twitte
} ,
"place" : {
} ,
"entities" : {
  "hashtags" : [
] ,

```

Figure 3.1: The format of Twitter data, a screen shot from some features

techniques [43].

3.1.1 Tweet Topics

As mentioned in the previous chapter, different approaches have been suggested to extract tweet topics. One of these approaches is to describe each topic by a set of keywords [45]. The focus of this research is not on the topic detection technique; instead, it focuses on how to represent each topic's connection to the Twitter users. So with this general design, regardless of the employed topic extraction method, topics are assigned to the tweets and the corresponded Twitter users. In this thesis, topics have been assigned to the tweets based on the used keywords in the content. In other words, each topic group contains some keywords so that a topic is assigned to a tweet, provided any corresponded keyword has appeared in the tweet content.

3.1.2 Twitter Users' Similarity

To analyze Twitter users, the other essential derived attribute is a criterion for comparing them. In this thesis, each Twitter user is treated as a document that contains all of the user's tweets.

Hence, the analysts can compare the users by comparing their corresponded documents. In the thesis, the TF-IDF method has been used to calculate the users' similarity measure; however, other applicable methods can be used, alternatively.

TF-IDF is one of the most popular methods for reflecting each term's importance in a document corpus. To more precisely evaluate each term in a document, TF-IDF weights the term based on the number of its appearance in a document divided by the number of documents in the corpus that contain the term. Accordingly, this approach declines the importance of the terms that appear in most of the documents, such as propositions. TF-IDF assigns a vector to each document, and the vectors' entries are the measured weights of the terms in a document. The similarity of the two documents can be determined by calculating the similarity of their vectors. For this purpose, the cosine similarity metric is used [63].

$$W_{m,i} = freq_{m,i} \times \log \left(\frac{N}{n_m} \right) \quad (3.1)$$

TF-IDF calculation, $freq(m, i)$ is the frequency of term t in document i . N is the number of documents in the collection, and n is the number of documents containing the term t . Also, the equation for cosine similarity of two document p and q with n words is as follows:

$$cosinesimilarity(p, q) = \cos\theta = \frac{p \cdot q}{\|p\| \|q\|} = \frac{\sum_{i=1}^n p_i q_i}{\sqrt{\sum_{i=1}^n p_i^2} \sqrt{\sum_{i=1}^n q_i^2}} \quad (3.2)$$

It should be noted that in comparison with other document similarity metrics, TF-IDF is relatively simple and easy to measure. However, Since the design is modular, other methods are applicable as well. For this, the documents' vectors and their similarity should be calculated based on the algorithm and assigned to the data.

3.2 Triaging System Design

As mentioned before, a triage process takes place in three stages. The first stage is filtering a large number of available documents to collect a smaller set of relevant ones. This set can

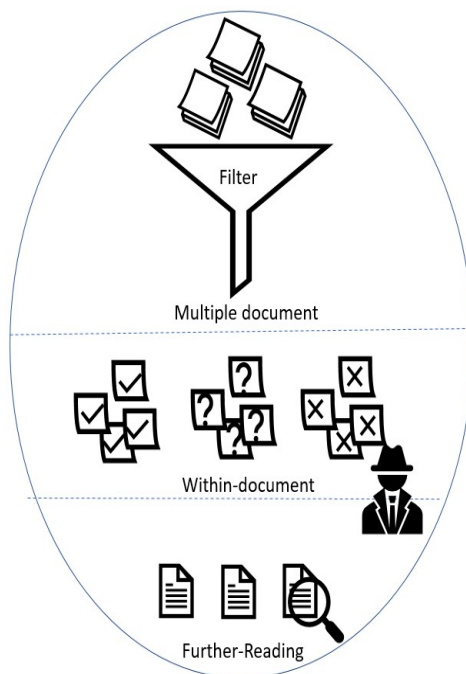


Figure 3.2: Triage process levels

- (1) filtering multiple Twitter users with some criteria (2) inspecting selected Twitter users based on their attributes (3) scanning individual Twitter users' profile and tweet contents

be obtained after querying Twitter users by different attributes, such as Twitter users' profile attributes or a specific topic they have written about. At this stage, some documents would be rejected, and another part would be passed to the next level for a more in-depth inspection. Since this information is digital, the analysts need an interface to interact with the information so that by reviewing and changing the information, they can decide to reject some documents and select the others for further review. Seeking for more detailed information, the analysts scan each document to understand actual contents (Figure 3.2). This step takes place by displaying and interacting with document text.

Displaying a myriad of available tweets and Twitter user's data on limited screens is not feasible on one shot. Besides, interacting with this amount of information is challenging and can barely lead to knowledge extraction. So on the first stage of Twitter user triage, the analyst has to filter data before displaying it. Therefore, based on criteria such as time and subject, a reasonable amount of data would be displayed, and the analyst can limit selected documents

by applying more filters. Creating such an environment requires visualization techniques along with appropriate interactions. In the final stage, the analyst needs to drill down and actually read the text of a user's tweets to decide on the concept and its relevance to other documents.

Literature shows that to determine the Twitter users' group affiliation, patterns of activity in different topics, contributions during a time frame, and their similarity with each other are essential factors to be considered. Therefore, the relevant components are designed and illustrated in this section. As mentioned, the design of these components is based on the pattern-based framework suggested by Sedig and Parsons [59]. First, the information space, activities, and tasks that the analysts are likely to perform are discussed. According to this stage, three major components can represent the Twitter user's activities. Next, the process of adopting visualization techniques for representing information as well as designing interactions for moving within visualizations and changing them are explained respectively for each component. It should be noted that these stages are interrelated, and the decisions made on each stage affect the others. So even though it is presented sequentially, the designing process took place by moving back-and-forth between the stages.

3.2.1 Information Space and Task Space

In the dataset, each tweet is sent by a Twitter user, so there may exist multiple tweets of the same user on different topics and time. If a tweet has particular keywords, it can be assigned to the corresponded topic category. Therefore, according to the information space, each topic may contain several keywords, and each keyword has been used in several tweets (Figure 3.3). For instance, the Twitter users who used the keyword "Dems" in their tweets are implicitly associated with the "democratic" topic. Accordingly, information entities are topics, keywords, and tweets, and the relationship between them. The attributes regarding creation time and similarity measures are attributed to the tweets. The initial data is heterogeneous and contains features with various types such as images, texts, and numerical values; however, non-textual information is not the concern of this thesis. The information space is open to new tweets, and upcoming data could be visualized and analyzed.

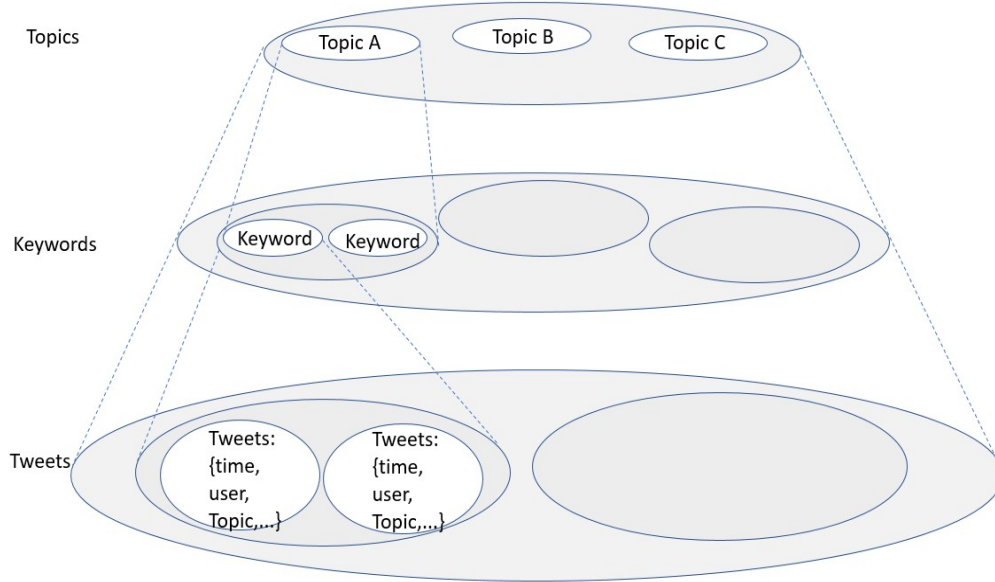


Figure 3.3: Information space

Each topic encompasses some keywords, each keyword is associated to a number of tweets, and each tweet is published by a user.

Three major components are designed to separately represent the underlying information about Twitter users' relationship with the mentioned attributes, including topic-user association, time-user association, and user-user similarity. These components would lead to a more concentrated observation of an attribute to find potential patterns. However, the components are not fully detached, and while the analysts interact with each component, the other components would update, respectively. To explore the relationships, analysts require an interactive visualization that enables them to perform associating, comparing, detecting, discriminating, finding, focusing, identifying, observing, recognizing, and scanning the entities. The next sections describe these components.

3.2.2 Topic-User Association

To explore Twitter users based on the tweets' content, analysts need to extract patterns within the relationship of each user to a topic or keyword. So this information should be accessi-

ble through the represented information. In this regard, individual information items such as entities, attributes, relationships, and systems are assigned to appropriate visual marks. The patterns of components are selected based on the offered visual patterns list [59]. Accordingly, each of the topics, keywords, and Twitter users must be displayed individually onto a single unitized visual representation, known as a token. Also, the relationship between Twitter users with contributed topics and keywords and the connection between topics and corresponded keywords have to be reflected.

Visualization

As Meirelles [58] suggests, there are different ways to demonstrate the hierarchy in information space. One of these visualization techniques is using a sunburst, also known as a radial icicle. Figure 3.4 shows the designed component for representing information of Twitter users and topic association, with a sample data of political tweets. Information items (i.e. Twitter users, topics, and keywords) are visualized with tokens. The relationship between a user and contributed topics is depicted with links. Also, colours discriminate each topic group and their keywords, and the size of the corresponded arc has a direct relation with the number of tweets in that subject. Therefore, the keywords and topics popularity is demonstrated by their arc portion, so by observing the diagram, an analyst can identify the most popular topics and keywords. As discussed, the topic-user association is represented as well as the keywords that a topic encompasses. To go further and see the related keywords to a Twitter user, the analyst requires to interact with information space and filter it. The next part discussed the list of designed interactions.

Interaction Design

For working with represented information, the analyst has to interact with visualizations to change or modify them. In this thesis, the list of possible interactions is designed according to the proposed catalog of epistemic action patterns by Sedig and Parsons [42]. For predicting the required interactions to accomplish the triage task, different scenarios have been considered. For instance, an analyst might want to see the relationship between the Twitter users and the used keywords. Also, the analyst might need to create a pile of Twitter users for more inves-



Figure 3.4: Topic-User component for exploring contributions in topics/keywords. Topics and keywords are depicted as arc with different colours, and the ribbons connects them with the Twitter user tokens.

tigation in other components. Besides, an analyst should be able to select all Twitter users associated with a topic or a keyword.

Likewise, the other required epistemic actions to accomplish the triaging task are determined. Before explaining the predicted interactions for the topic-user component, it should be noted that Twitter users are the communal parameter in all the components. Therefore an additional component is designed to aggregate the interactions regarding Twitter users' behaviour and profiles that are mutual among all three components. Interactions that are merely related to a single component are explained for this component. Subsequently, the design of communal interactions is discussed separately. Another important statement is that the interactions are not independent and may overlap. So an interaction may serve several epistemic actions.

The predicted interactions for the topic-user component involve the epistemic actions, including comparing, drilling, storing/ retrieve, selecting, arranging, and filtering. The first four actions are communal and are discussed in the next sections. As mentioned, the analysts may want to see the direct relationship between the keywords and the users. That can be accomplished by rearranging the place of topics and keywords. As shown in Figure 3.5, the Twitter users are directly linked to the keywords.

In this component, selecting and filtering actions overlap. The analyst can move the mouse cursor on the topics, to highlight all associated keywords and Twitter users, and dim the rest visuals (Figure 3.6). The same interaction is possible for keywords and Twitter users. By selecting any of these visuals, the analyst selects all related entities, and by repeating the action, the analyst can deselect them. Filtering is defined as “displaying a subset of visual representation elements according to certain criteria”. Accordingly, filtering Twitter users based on the topics/keywords they used and filtering topics/keywords based on the Twitter users who used them can happen through selecting entities.

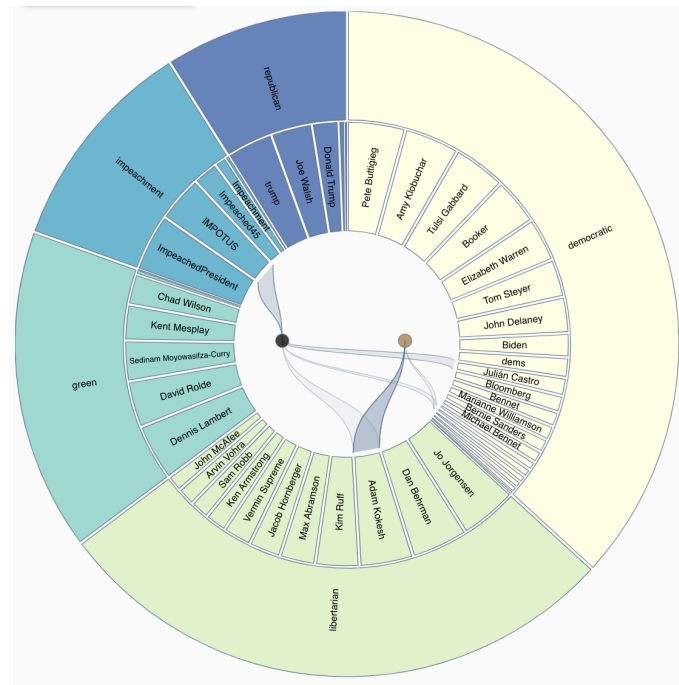


Figure 3.5: Topic-user component, arranging topics and keywords
 By arranging the place of keywords and topics, analysts can change the level of granularity.

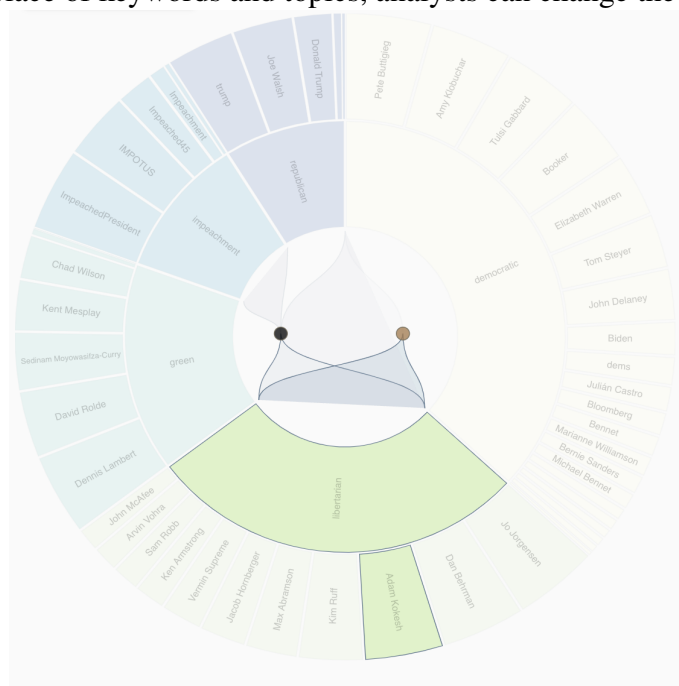


Figure 3.6: Topic-User component, selecting and filtering actions
 The analysts can hover the mouse pointer on the entities to select them or filter data based on selected topic/keyword.

3.2.3 Time-User Association

The other critical attribute of tweets for the triaging process is the time factor, according to which important conclusions such as novelty or association with other tweets and Twitter users can be inferred. As mentioned in the previous chapter, examining analysts' behaviour patterns over time can help analysts conclude group association detection, especially for bots. As a result, the analysts need to examine the time of tweets during different time slots and on different scales. This section explains the design of such a component.

In this component, Twitter users are information entities, that scatter on different time based on the tweet's creation time. The users are atomic entities; however, the time factor has different scales. For instance, working with the component, an analyst might notice a correlation between two accounts over the past year and may wish to inspect the tweet's time more precisely and focus on the pattern of tweets on a day. So the system should enable the analyst to change the time scale and see how Twitter users acted on a day. The goal of this component is to display a potential relationship between users in terms of tweeting time. Therefore, the track of a Twitter user's activity during a time slot should be detectable. In the design, for a specific time slot, each tweet is assigned to a track, and the Twitter user's activities are coordinated based on tweets' creation time. The next part explains the visualization of this structure.

Visualization

The topic-User component is designed as a circular diagram that encompasses several stacks of tracks. Each stack is associated with a particular time slot (e.g. this year, last year, etc.); however, all the stacks have the same number of tracks associated with each Twitter user. In other words, for each Twitter user, there is an imaginary track in each of the time slots. Twitter users are diagnosable with their colour and their relative position towards other Twitter users. Each token on a Twitter user's track is associated with a tweet sent by that Twitter user.

Figure 3.7 demonstrates the sample of the topic-user component for two users. For each year, tweets of each user are located on an imaginary circular track. As shown, the red user

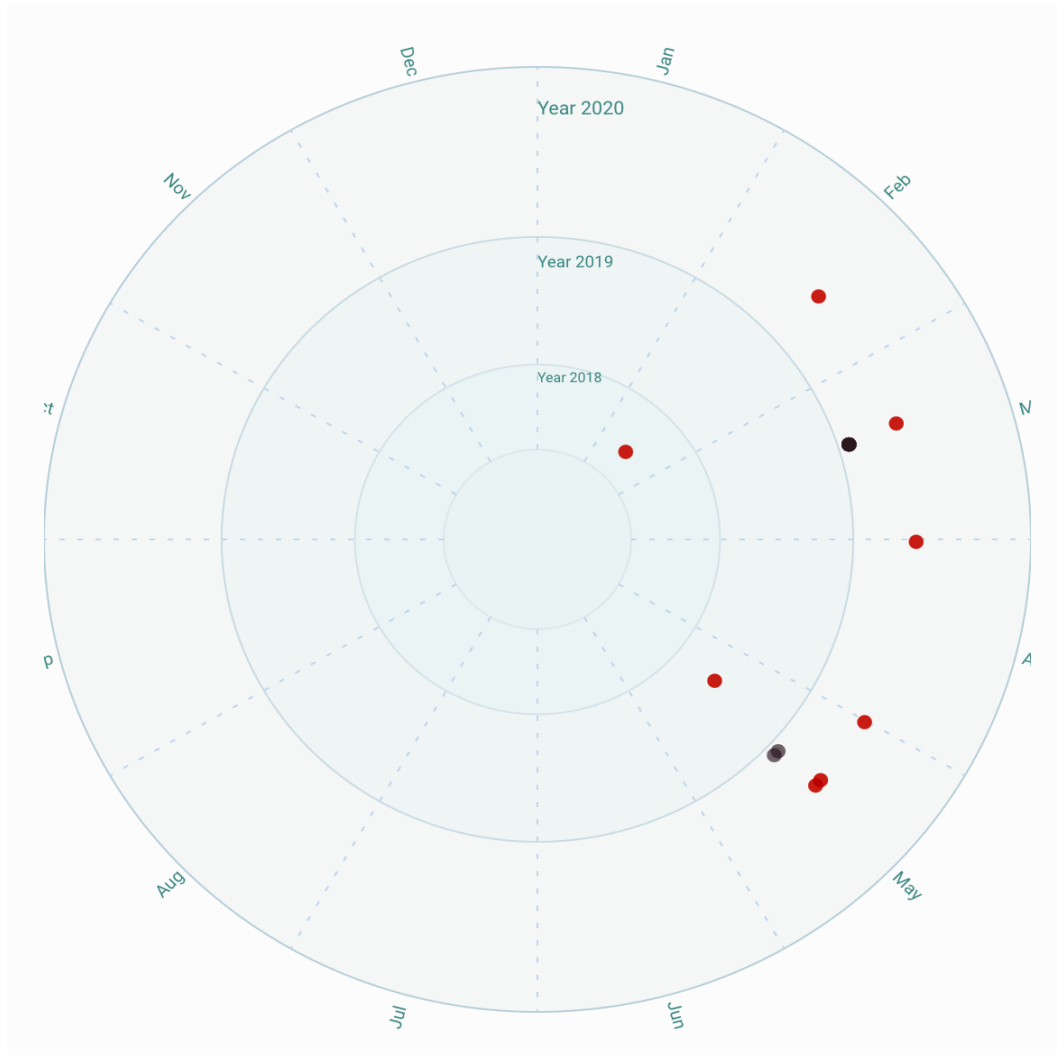


Figure 3.7: Time-User component for exploring Twitter users' activity over time. The component consists of several time tracks, and on each one, Twitter users' tracks are stacked. Each Twitter user is assigned to a specific colour.

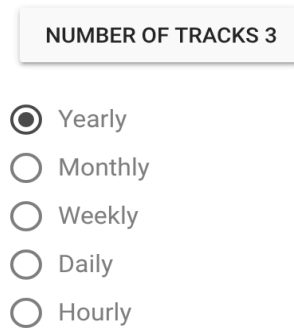


Figure 3.8: Designed options for interacting with information representation
Analysts can increase the number of tracks to see farther years. However, due to the screen display limitation to show all available information, the maximum number of stacks is limited. Also, analysts can select different time unit scales.

contributed one tweet in February 2018 and one in May 2019, and six tweets in 2020. It is noticeable that all three black Twitter user's tweets are concurrently posted with some of the red user's tweets.

Interaction Design

Similar to the topic-user component, different scenarios have been considered to predict required interactions with this component for a triage activity. For instance, in the previous example, the analyst may notice the correlation between the two users and want to inspect their behaviour on the scale of hours or minutes. Or the analyst might want to check the users' activity in more than three years. So the component should empower the analyst to interact and change the representation to see the information as needed. As mentioned, possible communal interactions are discussed separately in the next parts.

The designed component enables analysts to control the represented information through interactions (Figure 3.8). For instance, analysts can insert new time slots to see the information about Twitter users' activity at a farther time. On the other hand, analysts may want to focus on new time slots. For that purpose, they can remove additional stacks (Figure 3.9). Besides, after observing Twitter users behaviour in the past years, the analyst may want to see the

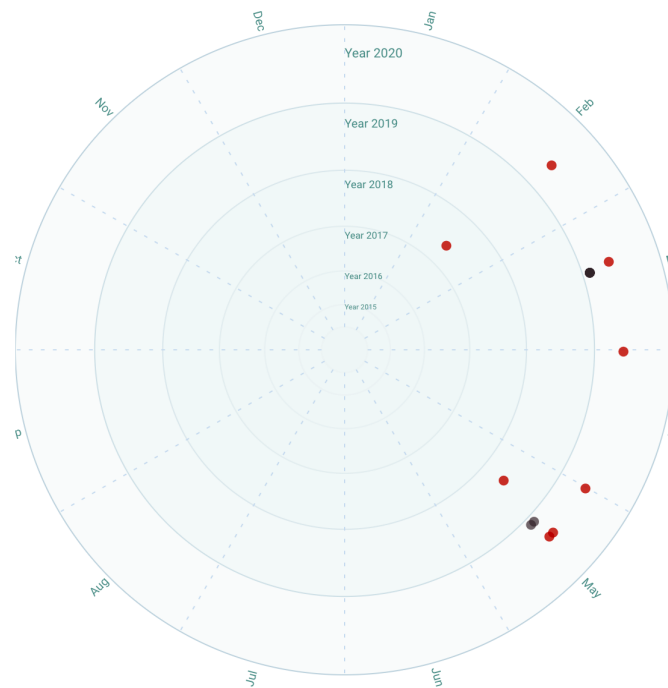


Figure 3.9: Time-User component, inserting/removing additional stacks
analysts can see additional information about the Twitter users' background by inserting
information of more years.

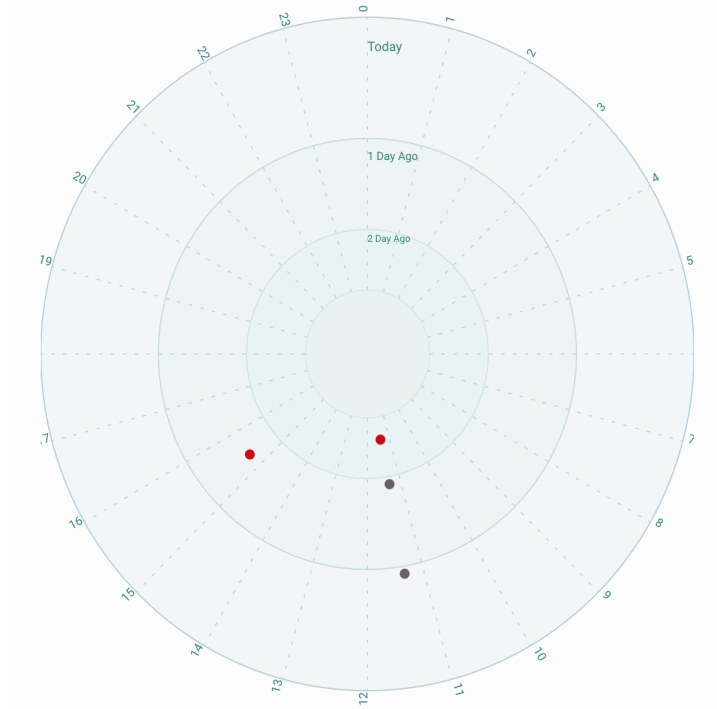


Figure 3.10: Time-User component, collapsing/expanding
By interacting, analyst can change the time scale from year to day.

behaviour over the past month, weeks, days or hours. Therefore, the analyst can interact with the component to expand the scale and see the tweets' distribution on a more precise scale. Since the data is coming in real-time, so the focus is on recent times. In turn, if the analyst did not find any pattern, they can collapse the time scale into a more broad scale unit (Figure 3.10).

3.2.4 User-User Similarity

Due to a massive amount of available data, analysts have to filter the dataset based on the mentioned criteria; however, this may lead to losing potential Twitter users that are similar in terms of other attributes. For instance, a bot may propagate fake news in a topic with the same temporal pattern, but in different keywords; if on the early stages analysts filter the data based on one of those keywords, they may overlook the underlying correlation. Therefore, employing data analysis techniques to find and select similar Twitter users is essential. The similarity component is designed to demonstrate analogous Twitter users resulted from analytical techniques.

In the users' similarity component, information entities are the Twitter users with different distances to a particular user. These entities are atomic. The component represents the result of an analytical model, to complement other interactive information visualizations and to support the analysts in finding similar users. Therefore, the similarity between users and a selected user should be distinguishable.

Visualization

The similarity component is designed as a list of Twitter users. Each row is associated with a user, distinguishable by their screen name and profile pictures. Besides, for each user, the degree of similarity to a particular user is mentioned. Figure 3.11 demonstrates a sample of the designed component for a selected Twitter user. In this component, Twitter users are sorted based on their similarity to the selected user.










Interaction Design

Likewise, various scenarios have been considered to predict the required interactions for this component. These interactions involve selection, arranging, drilling, filtering, composing, storing, and collapsing/ expanding epistemic actions. For instance, the analyst may want to select a user and the most similar users to it and store the users to drill down for more exploration. To this end, interactions are designed to support analysts in accomplishing the tasks. Again, the collective epistemic actions regarding Twitter users are described in the next section.

3.2.5 Components Aggregation

In a Twitter users triaging process, the analysts work with the three described components to examine the information space. According to the triage process stages, the analyst has to select a set of related Twitter users by examining different attributes. Therefore, these three components must work together and enable the analyst to see how changing one parameter affects the others. To this end, this section explains how to sync the components. Furthermore, the communal interactions between the three components are discussed.

User Similarity (TF-IDF - Compared To: @darkstar_logan)

Username	Cosine Similarity ↑
 Eric Edwards @UCFONEBIGOHANA	0.0225727242890508
 Osman Naqs @jinkinggon	0.03755206962060437
 ron @ron84750909	0.04393365606748257
 Tommy @birdman8272	0.028814997263305903
 Bobby From The Bronx @newkingofmedia	0.052589671160308285
 Lucalbio, the North-East Recluse @Lucalbio	0.02326644482454577
 VoteForChristsakes @Melanie78436358	0.04338230059387121
 CL Moller @cieloyla	0.04210263339902969
 Ray Robbins @tradewind35	0.10763026158558492
 @mistbeer	0.019638094939021383

Rows per page: 10 ▾ 1-10 of 61 < >

Figure 3.11: Twitter users' similarity component
 For a selected Twitter user, the value of cosine similarity with other users is determined based on their TF-IDF vectors.

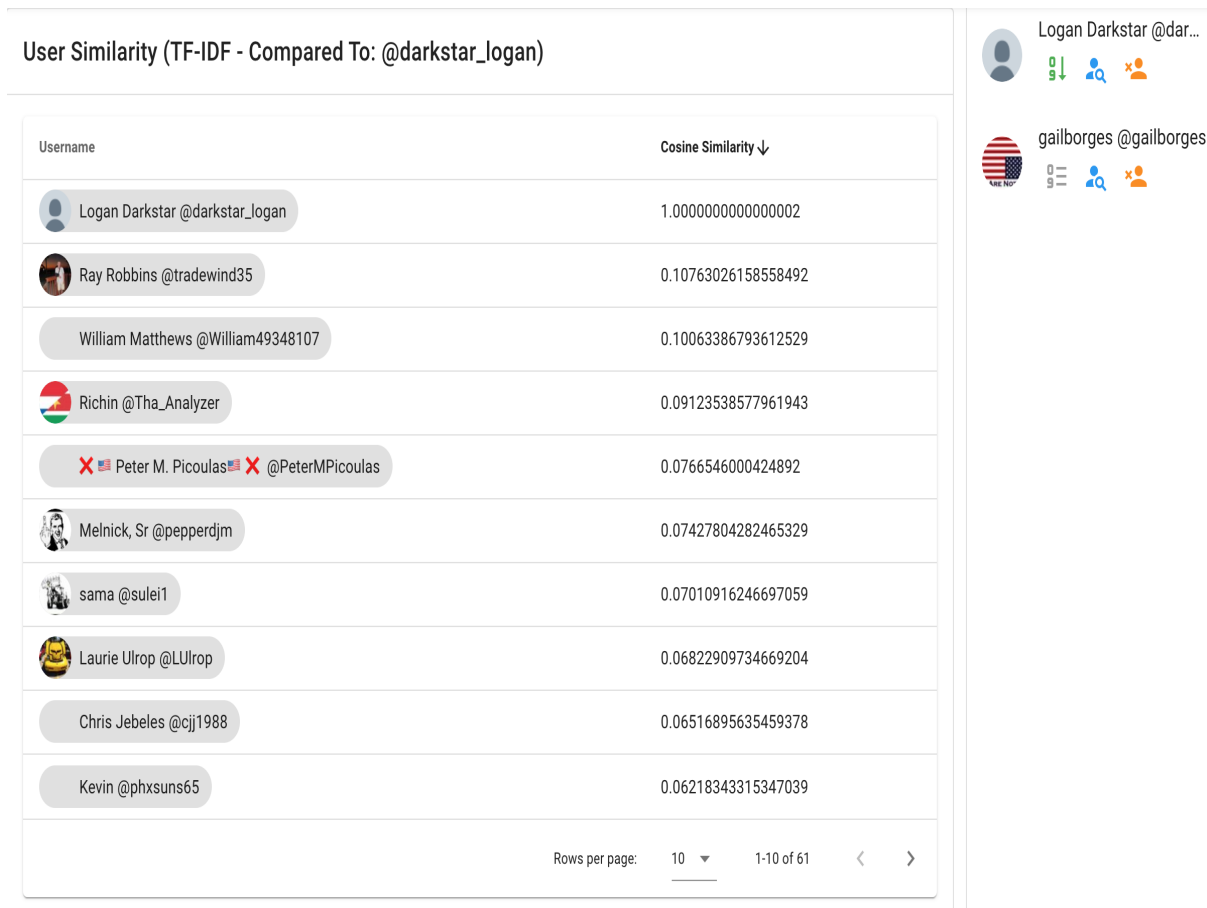


Figure 3.12: Twitter users' similarity component, expanding and rearranging the Twitter users. By interacting with the visualization, the user can rearrange the order of users, extend the number of users in the list, and see more users' similarity value.

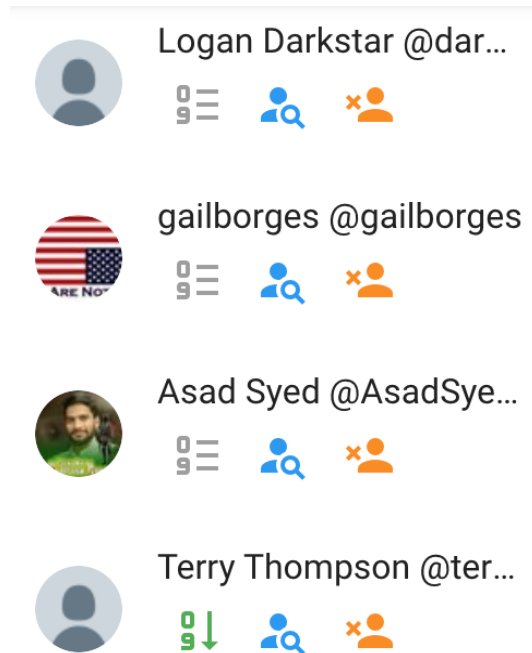


Figure 3.13: Pile of selected Twitter users

By selecting any Twitter user, their names would be shown in a pile. The analysts can interact with the pile to remove, select for inspecting similar users, and drill down for a more in-depth inspection.

As explained, the tasks that an analyst might need to perform triaging were predicted. That was done by examining the interaction list and considering different scenarios. Specific actions for each component were described in the previous parts, and in this section, the collective epistemic actions involved in all three components are discussed. These interactions include selecting, comparing, drilling, and storing/retrieving. For more integration between the three components, these interactions are considered and designed together. To perform each action, the analyst has to select one or multiple Twitter users. So by selecting each user, the corresponded tokens would simultaneously be highlighted in all three components.

Performing a triage process, the analysts can compare the Twitter users on each component based on the corresponded attributes. However, to examine and compare those Twitter users situation in terms of other attributes, the components should be sync with each other. Therefore, a component is designed that allows the analyst to store Twitter users or remove them from the

The screenshot displays a user selection interface. At the top, there is a table with two columns: user names and numerical identifiers. The first row shows 'Ray Robbins @tradewind35' with the identifier '0.0035775425008885827'. The second row shows '@mistbeer' with the identifier '0.06318182236406092'. Below the table, there is a 'Rows per page' dropdown set to '10' and a pagination indicator '1-10 of 61'.

Below the table is a 'User Details' section. It contains three columns, each showing a user's profile and a tweet. The first column shows 'LOGAN DARKSTAR @DARKSTAR_LOGAN' with a tweet from Logan Darkstar (@darkstar_logan) 23 weeks ago. The second column shows 'GAILBORGES @GAILBORGES' with a tweet from gailborges (@gailborges) 26 weeks ago. The third column shows 'ASAD SYED @ASADSYED05' with a tweet from Asad Syed (@AsadSyed05) 17 weeks ago. The tweet content includes a retweet from @MaleehaHashmey and mentions of #DonaldTrump and #Pakistan.

At the bottom of the tweet view, there is a section for user similarity. It shows two users: 'ibm' with a similarity score of 0.0 and 'naturaljs-afinn165' with a similarity score of 0.067.

Figure 3.14: Drill down to see tweets

The analyst can use the second icon on the pile to drill down and read a Twitter user's tweets. This is considered as the third level of the triage process.

list. As shown in Figure 3.13, the designed component contains the list of selected Twitter users along with three buttons. The first button allows the analysts to select a user and inspect its similar users in the user-user similarity component; the similarity measures will be shown based on the last selected user. By clicking the second button for a user, analysts can drill down and see the tweets content of the corresponded user (Figure 3.14). Using the third button, analysts can remove the user from the pile and consequently deselect the user from all other components.

3.3 Implementation

The triage process is sequential; however, on each level, the components of the designed abstract system work concurrently. The system is implemented as a part of a visual analytics system for making sense of real-time Twitter data, VARTTA. This system aims to enable analysts to understand the underlying information within the tweets, such as the tweet topics, content theme, and Twitter users group. The system allows analysts to apply various ML techniques for topic detection and compare the outcomes. Also, analysts can drill down through interactions with the system and see the tweets of a Twitter user. VARTTA is adaptable with real-time analysis of tweets in different domains such as public health, smart city, or U.S. elections.

Even though VARTTA provides the analysts with some aspects of Twitter users information, as mentioned in the previous chapter, the system does not offer a triaging component with a focus on Twitter users and their attributes. Therefore, a triage system is required to collect different aspects of users' information and allow analysts to explore this information. Consequently, VARTTA, together with the designed Twitter user triaging system, can reach the analysts to the ultimate goal of gaining insight into the Twitter information and accomplish their cognitive activity.

VARTTA analyzes real-time Twitter data and is robust to the incoming data streams and gets updated dynamically. Also, it provides high-quality interactions, so the analysts can easily navigate between the pages to analyze data. For this, VARTTA has been implemented as a progressive web application (PWA) and uses the Vue.js framework. Vue.js is a programming framework developed in JavaScript that supports building high-performance web applications. This framework manages the reactivity of the components. Also, VARTTA takes advantage of Nuxt.js package structure, which provides code efficiently, and makes it possible to add new pages and components. As a page on VARTTA, the designed Twitter user triaging system is implemented using vue.js and nuxt.js, and the diagrams are drawn using the d3.js library.

3.4 Case Study

In this section, a case study is presented to illustrate the effectiveness of the designed and implemented VA system. In this case study, the focus is on demonstrating the triage process levels for collecting, analyzing, and identifying Twitter users and their group association. To this end, tweets regarding the U.S. 2020 presidential election are collected and their topics and similarity measures have been assigned to them. The data of these tweets is visualized, and the analysts are able to accomplish each level of the triaging process through interactions.

Referring to the journalism example in the first chapter, an analyst (a journalist in this case) should be able to analyze Twitter users' behaviour and create a collection of related users for further investigation. As mentioned, the analyst has to go through three stages of multiple documents, within-documents, and further reading to accomplish a triaging process. Therefore, the VA system has to enable analysts to perform tailored tasks and actions. This section examines the required activities that an analyst has to accomplish at each level.

Level One: Multiple Documents

In the first level of triage, all the documents are available to the analyst (Figure 3.16). So the analyst has to filter data to create a subset of relevant users based on different criteria. For instance, the analyst looks for users whose tweets are associated with democrats or the ones who tweeted last month. As shown in Figure 3.17, the analyst can select the users based on their topic or the creation time of the tweets.

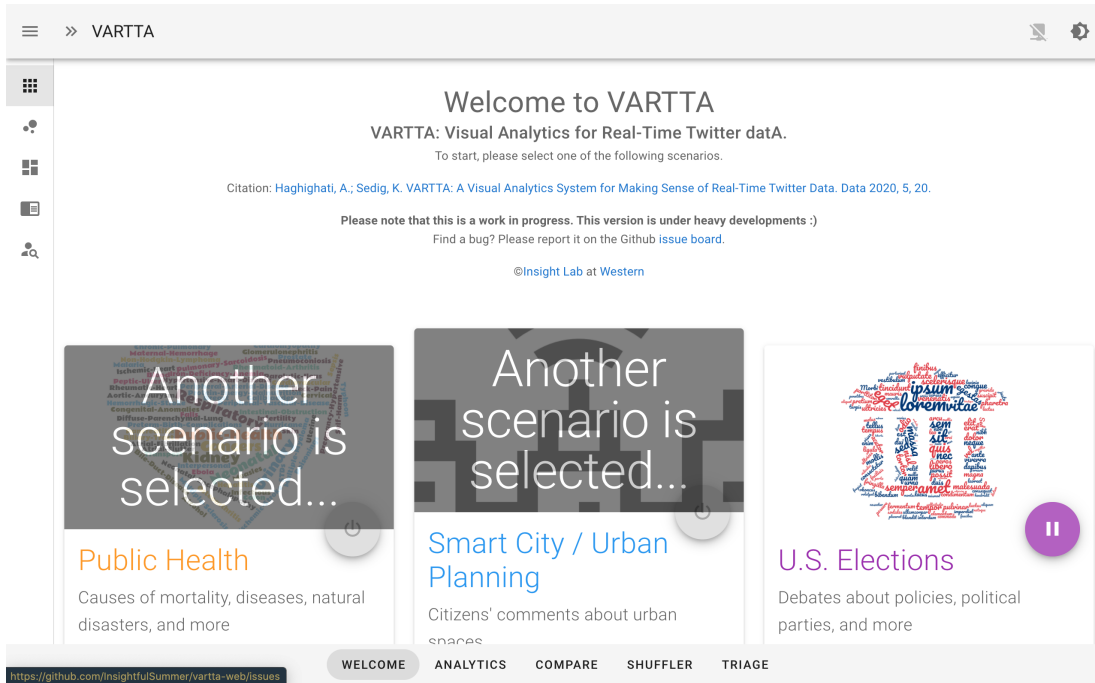


Figure 3.15: Triage Page on VARTTA
The analyst has to select one of the scenarios and navigate to the Triage page.

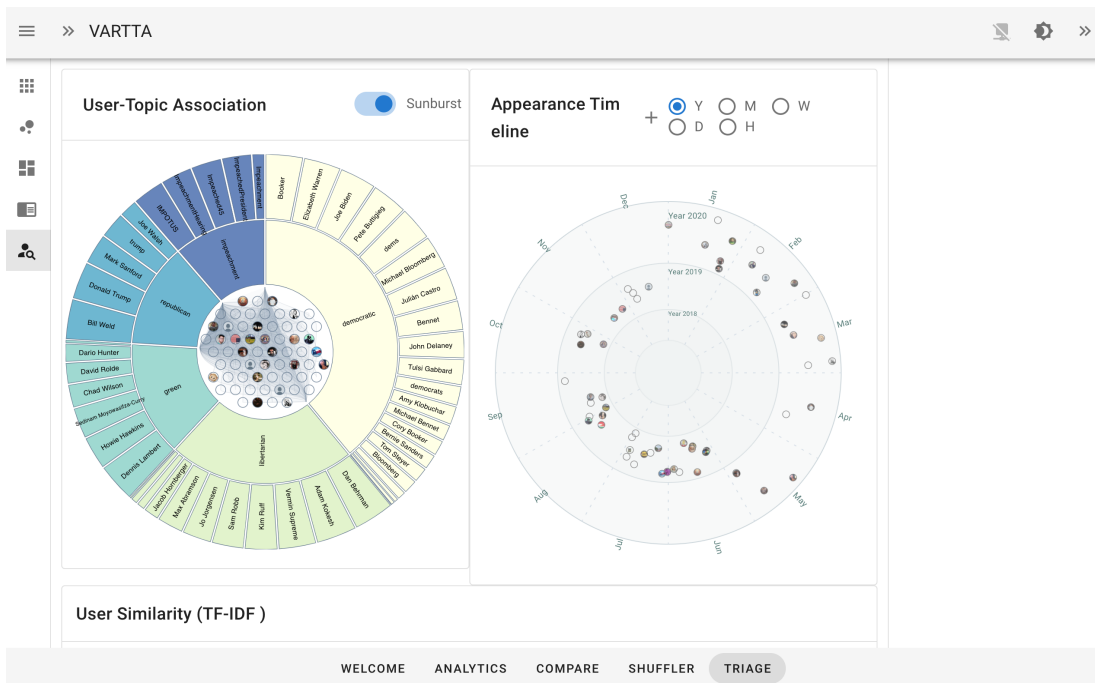


Figure 3.16: Multiple Documents
All of the Twitter users on the sample are shown on the components.

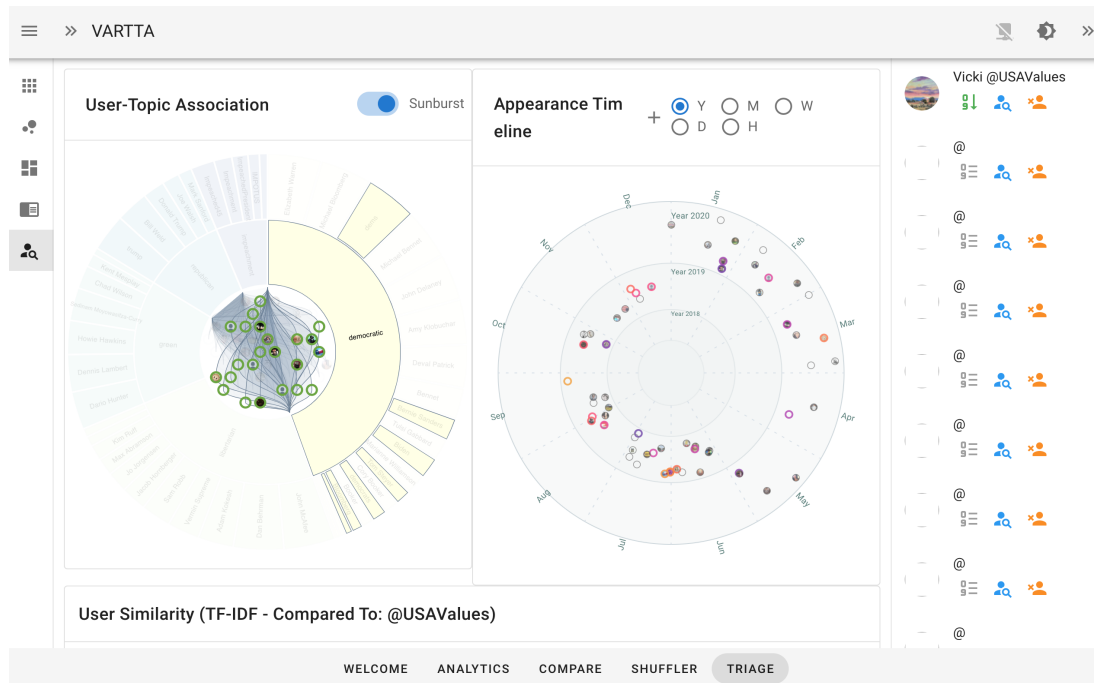


Figure 3.17: Multiple Documents - Filter

The analysts can filter the users with the topics/keywords they have contributed.

Level Two: Within Documents

In the second level of triage, the analysts interact with the Twitter users' information to select a set of related users. For this, the analysts go through different components and set the criteria to limit the users set. Shown of Figure 3.18, the analyst has chosen the three users who have tweeted about the current president of the U.S. in the last month. By investigating the similarity component (Figure 3.19), the analyst finds a high degree of similarity between two users. So the analyst decides to remove the irrelevant user from the pile and drill down for more investigation.

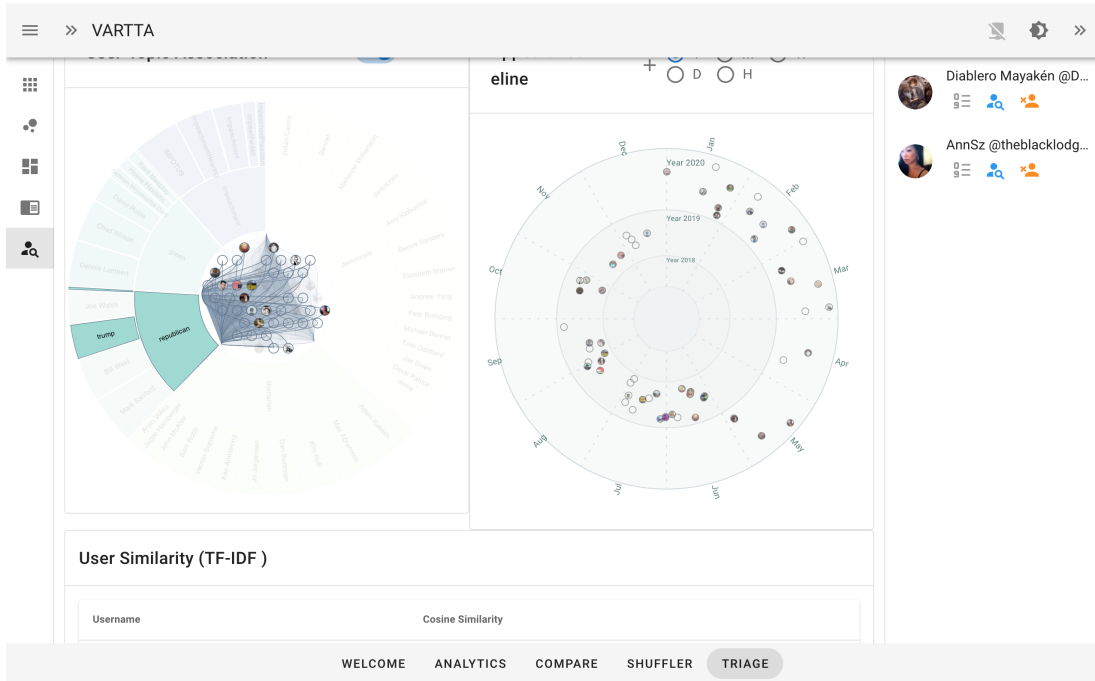


Figure 3.18: Within Document Triage - Similar Users

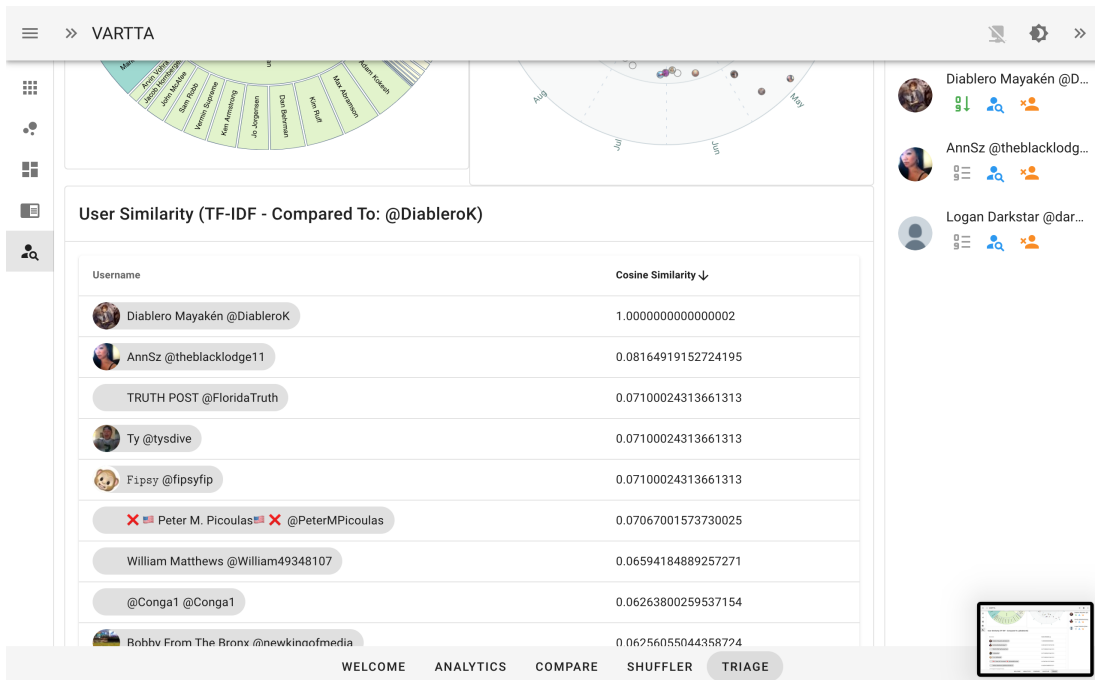


Figure 3.19: Within Document Triage - Using Similarity Component
The analyst uses different aspects of the information to determine related users.

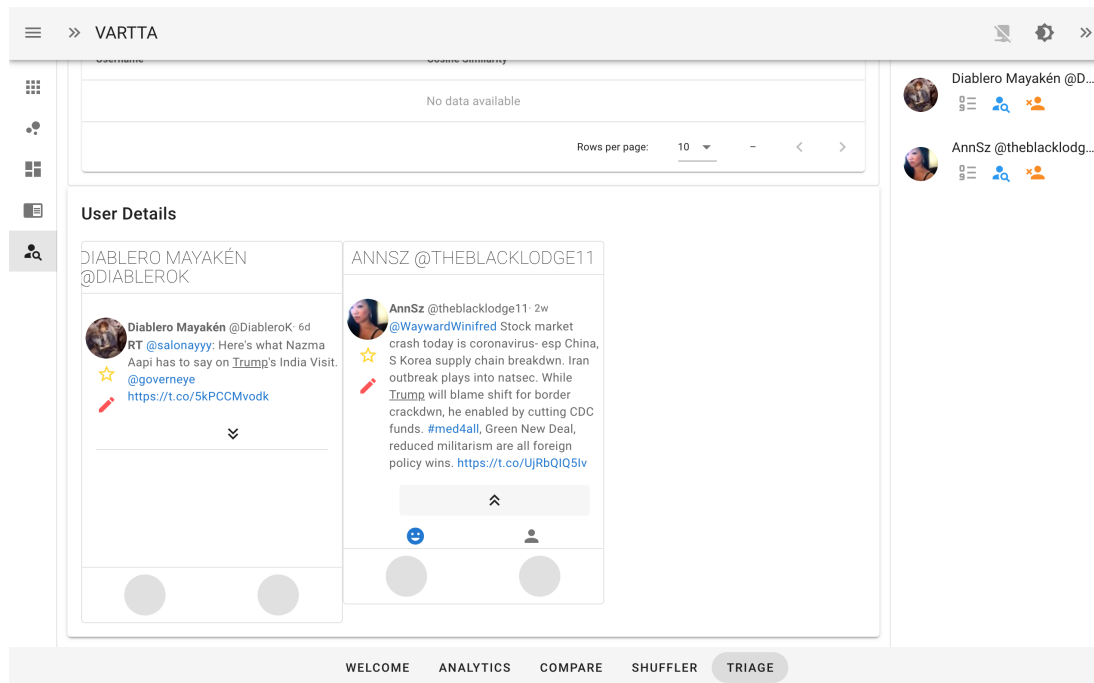


Figure 3.20: Further Investigation

The analysts can drill down to scan the content of the selected users' tweets on the previous level.

Level Three: Further Investigation

In the last level of triage, the analysts need to scan the documents, which are the tweets of each user. To this end, the analyst can create a list of users to inspect their tweets. At this level, the analyst can determine whether the tweets and their authors are related to the topic of interest or not. According to this, the analyst may keep the user in the list of selected users or remove the user as a rejected document.

Chapter 4

Conclusion

According to the literature, identifying twitter users is an integral part of analyzing tweets. Therefore, this thesis proposed and illustrated an exploratory VA system for triaging Twitter users. This system integrates data visualization, human-data interaction, and data analytics, and empowers analysts to triage real-time Twitter users' data through interaction and detect potential correlations among them. Also, the thesis showed that an abstract level design for analyzing Twitter users' information is a pragmatic strategy to form an exploratory VA system. Subsequently, a case study demonstrated that the implemented VA system could address the challenges of Twitter users' data analysis and understanding the relationship among the users.

4.1 Discussion

There are some considerations to be taken into account, which can potentially affect the system's performance and effectiveness. First of all, while designing, the triage process levels should be considered so that the analysts can appropriately work with the information at each level and move between them. Also, it is crucial to represent the information simple enough to minimize perception load; yet, the representations need to enable analysts to accomplish complex cognitive activities through appropriate interactions. For this purpose, it is essential to anticipate and create a comprehensive list of required interactions. The other salient point to consider is that due to a large amount of data, leveraging data analysis models for extract-

ing patterns out of data is inevitable. Therefore, the system should provide analysts with the ability to use analytical methods along with interactions with data. Creating a balance between these resources would equip the analysts with a wide range of analytical resources to perform cognitive activities.

4.2 Future Work

There are several limitations to this research. First of all, the free Twitter API data contains only a brief part of actual tweets, so the presented thesis has only investigated a small sample of tweets. Nevertheless, this could be the basis for future work on the information visualization design of a massive amount of Twitter user data along with high-performance techniques for processing it. The other limitation is that for analyzing Twitter users' behaviour, the study has only investigated their similarity based on their individual parameters. Future experiments have to examine the impact of considering different parameters, such as users' behaviour in a network in terms of followers and following. Finally, in designing the interaction list, diverse scenarios have been considered in order to create a comprehensive list. However, the list is open to new scenarios, and the future works may consider other interactions.

Bibliography

- [1] Mohd Najib Mohd Salleh. *Recent Advances on Soft Computing and Data Mining*, volume 287. r, 2014.
- [2] Jesús Silva, Alexa Senior Naveda, Ramiro Gamboa Suarez, Hugo Hernández Palma, and William Niebles Núz. Method for Collecting Relevant Topics from Twitter supported by Big Data. *Journal of Physics: Conference Series*, 1432(1), 2020.
- [3] Luke Sloan, Anabel Quan-Haase, Lori McCay-Peet, and Anabel Quan-Haase. What is Social Media and What Questions Can Social Media Research Help Us Answer? *The SAGE Handbook of Social Media Research Methods*, pages 13–26, 2017.
- [4] Andreas M. Kaplan and Michael Haenlein. The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*, 54(2):105–113, 2011.
- [5] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: Understanding microblogging usage and communities. *Joint Ninth WebKDD and First SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pages 56–65, 2007.
- [6] Courtenay Honeycutt and Susan C. Herring. Beyond microblogging: Conversation and collaboration via twitter. *Proceedings of the 42nd Annual Hawaii International Conference on System Sciences, HICSS*, pages 1–10, 2009.
- [7] Ehsan Mohammadi, Mike Thelwall, Mary Kwasny, and Kristi L. Holmes. Academic information on Twitter: A user survey. *PLoS ONE*, 13(5):1–18, 2018.
- [8] Lei Zou, Nina S.N. Lam, Heng Cai, and Yi Qiang. Mining Twitter Data for Improved Un-

- derstanding of Disaster Resilience. *Annals of the American Association of Geographers*, 108(5):1422–1441, 2018.
- [9] Avinika Agarwal and Edmond Fernandes. Tweeting up for humanitarian emergencies. 8164(3):160–161, 2017.
- [10] Camilla Vásquez. *Language, Creativity and Humour Online*. Routledge, 2019.
- [11] Martin J. Chorley and Glyn Mottershead. Are You Talking to Me?: An analysis of journalism conversation on social media. *Journalism Practice*, 10(7):856–867, 2016.
- [12] Alfred Hermida. #Journalism: Reconfiguring journalism research about twitter, one tweet at a time. *Digital Journalism*, 1(3):295–313, 2013.
- [13] Ulrika Hedman and Monika Djerf-Pierre. The social journalist: Embracing the social media life or creating a new digital divide? *Digital Journalism*, 1(3):368–385, 2013.
- [14] Asmelash Teka Hadgu and Robert Jäschke. Identifying and analyzing researchers on twitter. *WebSci 2014 - Proceedings of the 2014 ACM Web Science Conference*, pages 23–32, 2014.
- [15] Dhiraj Murthy. *Twitter : social communication in the Twitter age*. Polity Press, Cambridge, UK, second edition, 2018.
- [16] Salvatore Parise, Eoin Whelan, and Steve Todd. How Twitter Users Can Generate Better Ideas — MIT Sloan Management Review. 56(4), 2015.
- [17] Shamanth Kumar, Fred Morstatter, and Huan Liu. *Twitter Data Analytics*. Springer, New York, NY, 2014.
- [18] Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann, editors. *Twitter and Society*. Peter Lang, 2014.
- [19] Fang Jin, Wei Wang, Liang Zhao, Edward Dougherty, Yang Cao, Chang Tien Lu, and Naren Ramakrishnan. Misinformation propagation in the age of Twitter. *Computer*, 47(12):90–94, 2014.

- [20] Julia Metag and Adrian Rauchfleisch. Journalists' Use of Political Tweets: Functions for journalistic work and the role of perceived influences. *Digital Journalism*, 5(9):1155–1172, 2017.
- [21] Dongho Choi, Ziad Matni, and Chirag Shah. Switching sources: A study of people's exploratory search behavior on social media and the web. *Proceedings of the Association for Information Science and Technology*, 52(1):1–10, 2015.
- [22] Maria Taramigkou, Dimitris Apostolou, and Gregoris Mentzas. Supporting Creativity through the Interactive Exploratory Search Paradigm. *International Journal of Human-Computer Interaction*, 33(2):94–114, 2017.
- [23] Armin Pournaki, Felix Gaisbauer, Sven Banisch, and Eckehard Olbrich. The twitter explorer: a framework for observing Twitter through interactive networks. (732942), 2020.
- [24] Soonil Bae, Rajiv Badi, Konstantinos Meintanis, J. Michael Moore, Anna Zacchi, Haowei Hsieh, Catherine C. Marshall, and Frank M. Shipman. Effects of display configurations on document triage. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3585 LNCS:130–143, 2005.
- [25] Fernando Loizides and George Buchanan. Towards a framework for human (manual) information retrieval. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8201 LNCS:87–98, 2013.
- [26] Glenn W. Mitchell. A brief history of triage. *Disaster Medicine and Public Health Preparedness*, 2(SUPPL.1):13–16, 2008.
- [27] James Hollan, Edwin Hutchins, and David Kirsh. Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research. *ACM Transactions on Computer-Human Interaction*, 7(2):174–196, 2000.

- [28] Zhicheng Liu, Nancy J. Nersessian, and John T. Stasko. Distributed cognition as a theoretical framework for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1173–1180, 2008.
- [29] James J. Thomas and Kristine A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.
- [30] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. Identifying Correlated Bots in Twitter. 1, 2016.
- [31] Gerasimos Razis, Ioannis Anagnostopoulos, and Sherali Zeadally. Modeling influence with semantics in social networks: A survey. *ACM Computing Surveys*, 53(1), 2020.
- [32] Zhouhan Chen and Devika Subramanian. An Unsupervised Approach to Detect Spam Campaigns that Use Botnets on Twitter. pages 1–7, 2018.
- [33] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Emergent properties, models, and laws of behavioral similarities within groups of twitter users. *Computer Communications*, 150(July 2019):47–61, 2020.
- [34] David R. Bild, Yue Liu, Robert P. Dick, Z. Morley Mao, and Dan S. Wallach. Aggregate characterization of user behavior in Twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology*, 15(1):1–24, 2015.
- [35] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Processing and visualizing the data in tweets. *SIGMOD Record*, 40(4):21–27, 2011.
- [36] Carmen Zarco, Elena Santos, and Oscar Cordón. Advanced visualization of Twitter data for its analysis as a communication channel in traditional companies. *Progress in Artificial Intelligence*, 8(3):307–323, 2019.
- [37] Guilherme Coletto Rotta, Vinícius Silva De Lemos, and Ana Luiza. Exploring Twitter Interactions through Visualization Techniques : Users Impressions and New Possibilities. pages 700–707, 2013.

- [38] Chitvan Mehrotra, Nayan Chitransh, and Ajayshanker Singh. Scope and challenges of visual analytics: A survey. *Proceeding - IEEE International Conference on Computing, Communication and Automation, ICCCA 2017*, 2017-Janua(4404):1229–1234, 2017.
- [39] Ali Baigelenov, Michael Saenz, Ya-Hsin Hung, and Paul Parsons. Toward an Understanding of Observational Advantages in Information Visualization. *IEEE VIS '17: Proceedings of the 2017 IEEE Conference on Information Visualization, Poster Abstracts*, (October), 2017.
- [40] Arman Didandeh and Kamran Sedig. Externalization of Data Analytics Models: Toward Human-Centered Visual Analytics. 2016.
- [41] Daniel A. Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Gorg, Jörn Kohlhammer, and Guy Melançon. Visual Analytics: Definition, Process, and Challenges. *Information Visualization / Kerren, A. et al. (ed.). - Berlin : Springer, 2008. - pp. 154-175, 2008.*
- [42] Kamran Sedig and Paul Parsons. Interaction Design for Complex Cognitive Activities with Visual Representations: A Pattern-Based Approach. *AIS Transactions on Human-Computer Interaction*, 5(2):84–133, 2013.
- [43] Amir Haghghati and Kamran Sedig. VARTTA: A visual analytics system for making sense of real-time twitter data. *Data*, 5(1), 2020.
- [44] Sonia Saini, Ritu Punhani, Ruchika Bathla, and Vinod Kumar Shukla. Sentiment Analysis on Twitter Data using R. *2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019*, pages 68–72, 2019.
- [45] Rania Ibrahim, Ahmed Elbagoury, Mohamed S. Kamel, and Fakhri Karray. Tools and approaches for topic detection from Twitter streams: survey. *Knowledge and Information Systems*, 54(3):511–539, 2018.
- [46] Tetyana Lokot and Nicholas Diakopoulos. News Bots: Automating news and information dissemination on Twitter. *Digital Journalism*, 4(6):682–699, 2016.

- [47] Minsuk Kahng, Dezhi Fang, and Duen Horng Chau. Visual exploration of machine learning results using data cube analysis. *HILDA 2016 - Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, 2016.
- [48] Ryen W. White and Resa A. Roth. *Exploratory Search: Beyond the Query-Response Paradigm*, volume 1. 2009.
- [49] Gary Marchionini. From finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.
- [50] Vania Dimitrova, Lydia Lau, Dhavalkumar Thakker, Fan Yang-Turner, and Dimoklis Despotakis. Exploring exploratory search. pages 1–8, 2013.
- [51] Joseph Chee Chang, Adam Perer, Nathan Hahn, and Aniket Kittur. SearchLens: Composing and capturing complex user interests for exploratory search. *International Conference on Intelligent User Interfaces, Proceedings IUI*, Part F1476:498–509, 2019.
- [52] Jonathan Demelo, Paul Parsons, and Kamran Sedig. Ontology-Driven Search and Triage: Design of a Web-Based Visual Interface for MEDLINE. *JMIR Medical Informatics*, 5(1):e4, 2017.
- [53] Kamran Sedig, Paul Parsons, Hai-Ning Liang, and Jim Morey. Supporting Sensemaking of Complex Objects with Visualizations: Visibility and Complementarity of Interactions. *Informatics*, 3(4):20, 2016.
- [54] Yingcai Wu, Nan Cao, David Gotz, Yap Peng Tan, and Daniel A. Keim. A Survey on Visual Analytics of Social Media Data. *IEEE Transactions on Multimedia*, 18(11):2135–2148, 2016.
- [55] Robert Spence. *Information Visualization: An Introduction*, volume 37. Springer, third edition, 2017.
- [56] Guilherme Rotta, Vinicius Lemos, Felipe Lammel, Isabel Manssour, Milene Silveira, and André Pase. Visualization techniques for the analysis of twitter users’ behavior.

- Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, pages 748–749, 2013.
- [57] Jari Jussila, Jukka Huhtamäki, Hannu Kärkkäinen, and Kaisa Still. Information visualization of Twitter data for co-organizing conferences. *Proceedings of the 17th International Academic MindTrek Conference: "Making Sense of Converging Media"*, *MindTrek 2013*, pages 139–145, 2013.
- [58] Isabel Meirelles. *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*. Rockport Publishers, illustrate edition, 2013.
- [59] Kamran Sedig and Paul Parsons. *Design of Visualizations for Human-Information Interaction: A Pattern-Based Framework*. Morgan & Claypool, 2016.
- [60] Christian Tominski. *Interaction for Visualization*, volume 3. Morgan & Claypool, 2015.
- [61] Sana Malik, Alison Smith, Timothy Hawes, Panagis Papadatos, Jianyu Li, Cody Dunne, and Ben Shneiderman. TopicFlow: Visualizing topic alignment of Twitter data over time. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013*, pages 720–726, 2013.
- [62] Youcef Abdelsadek, Kamel Chelghoum, Francine Herrmann, Imed Kacem, and Benoît Otjacques. Community extraction and visualization in social networks applied to Twitter. *Information Sciences*, 424:204–223, 2018.
- [63] Per Ahlgren and Cristian Colliander. Document-document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, 3(1):49–63, 2009.

Curriculum Vitae

Name: Parinaz Nasr Esfahani

Post-Secondary Education and Degrees: Isfahan University of Technology
Isfahan, Iran
2012 - 2017 B.Sc.

Related Work Experience: Graduate Teaching Assistant
The University of Western Ontario
2018 - 2020