AcademiaMap-GIV: Geo-based Information Visualization of Scholarly Conversations on Twitter

by

Jamiur Rahman

Submitted in partial fulfilment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
December 2011

© Copyright by Jamiur Rahman, 2011

DALHOUSIE UNIVERSITY

FACULTY OF COMPUTER SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "AcademiaMap-GIV: Geo-based Information Visualization of Scholarly Conversations on Twitter" by Jamiur Rahman in partial fulfilment of the requirements for the degree of Master of Computer Science.

Dated: December 7, 2011

Co-Supervisor: _____

Co-Supervisor: _____

Reader: _____

DALHOUSIE UNIVERSITY

DATE: December 7, 2011

AUTHOR: Jamiur Rahman

TITLE: AcademiaMap-GIV: Geo-based Information Visualization of Scholarly Conversations on Twitter

DEPARTMENT OR SCHOOL: Faculty of Computer Science

DEGREE: MCSc CONVOCATION: May YEAR: 2012

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions. I understand that my thesis will be electronically available to the public.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

The author attests that permission has been obtained for the use of any copyrighted material appearing in the thesis (other than the brief excerpts requiring only proper acknowledgement in scholarly writing), and that all such use is clearly acknowledged.

_____

Signature of Author

DEDICATION PAGE

-To all Visualizers who know how to visualize problems but can not figure out how to solve them.

-To all researchers of "Jigsaw" Visual Analytics project at Georgia Tech, for inspiring me to find my passion within Information Visualization.

-To all Sitcoms for making my days equally colourful.

-To all Jesters for keeping me alive.

-To all Stand-up Comedians for making me laugh.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Geo-based Information Visualizations (GIV) allow people to analyze data points based on their related geographic locations. This approach is usually adopted where a large-scale geo-referenced dataset is present and users are trying to find a way to examine hidden patterns within this data. One of the emerging trends in GIV is to visualize social media data to show how information flows between users of popular social networking sites. Due to its public nature and the large number of users, most of the visualizations in this area rely on conversational data from Twitter (Twitter.com). In this thesis, we design and implement a web-based interactive GIV system, AcademiaMap-GIV, to visualize online conversations among scholars on Twitter. A formal exploratory user study was also conducted on the target users. The study results demonstrated that most of the study respondents found the features of AcademiaMap-GIV effective in regards to visualizing information of their interests.

ACKNOWLEDGMENTS

This work was supported by the Social Sciences and Humanities Research Council (SSHRC) and NCE Graphics, Animation and New meDia (GRAND) grants. In the beginning, I would like to acknowledge and thank all of our study participants, who kindly volunteered their time and professional opinions to the usability study.

First of all, I want to thank my supervisor Dr. Anatoliy Gruzd from the bottom of my heart for introducing me to this exciting world of social media data visualization. His encouragement and continual guidance have always been an immense source of motivation for hard work towards this thesis. Dr. Gruzd's insightful commentaries, and feedbacks are responsible for my enjoyable and interesting journey as a research student. This inspiration has pushed me harder to challenge myself in every step of my self-realization. He has been the most generous even in his busiest hours as he spent his valuable time to review this thesis. I am also highly grateful to Dr. Vlado Keselj, and Dr. Kirstie Hawkey for being part of the thesis committee and providing me their invaluable inputs.

It has been a great journey together with the Social Media Lab (SML) since I joined the lab last January. My gratitude goes to all my SML team members for their continuous support and assistance, and especially to Kathleen Staves, Philip Mai, Melissa Goertzen, Sreejata Chatterjee, and Mouhcine Aitounejjar.

I am ever grateful to my parents. Today's me is a result of their sacrifice and belief. They have always been an unparallel source of love, support, and inspiration. Also, my sisters have always kept my morale high. I am honoured to be my only nephew's (Zidane) role model. But only my good doings have been disclosed to him so far! All my best wishes and hopes go towards him.

Finally, I would like to acknowledge all the cast and crew who are responsible for the wonderful sitcoms that accompanied me through some bad times. Those TV shows helped me overcome my personal struggles thousands of miles away from my family.

CHAPTER 1: INTRODUCTION

We now live in the age of online social media and networking sites such as Twitter, Flickr, Facebook, and many more. Online social media is changing the way we live as people become more focused on their virtual lives and regularly spend hours on various social media and networking sites. Consequently, people around the world produce a large amount of data every day on these sites, mainly in the form of text, images, and audio. For instance, Twitter (Twitter.com) users alone contribute millions of messages daily, which is only a portion of the total from all online communities (Twitter Blog: #numbers, 2011).

The huge amount of data produced by various online communities can be overwhelming; however, it is valuable for research communities across different fields to understand the patterns and trends of activities going on inside social media. As a result, researchers around the world are focusing their work on examining these digital footprints to gain insights into social media data. However, simply analyzing the data by looking at each and every text sequentially is a highly tedious job, and quite impossible considering the large scale of dynamic data. Moreover, this kind of task becomes more complex and cumbersome if the need arises to join each piece of information together, to derive new insights on underlying patterns and trends. When dealing with such massive and dynamic datasets, information visualization plays an important role. It allows people to analyze data visually and aids in their decision making. Information visualization presents abstract data interactively on the screen by using different graphical forms and shapes for the purpose of data analysis.

The focus of this thesis work is a special form of Information Visualization - Geospatial Visualizations or Geo-based Information Visualizations (GIV). GIV is used to allow people to analyze data points based on their related geographic locations. In particular, this work will focus on developing an effective GIV to represent online conversations happening in social media. Social media and networking sites are heavily used around the globe, and they make it easier for geographically distant people to communicate and collaborate. Hence, there is a growing need to understand how distant connections are emerging and being maintained by individuals or organizations online. Because of the distributed nature of these sites, GIV is an effective tool for studying connections and communication among their users.

The intention of this research is to visualize the underlying communicative patterns in online groups, using online conversations of scholars on Twitter as a case study. This research is part of a larger initiative at the Social Media Lab at Dalhousie University, which is developing a web system called AcademiaMap (http://AcademiaMap.com) to analyze and visualize scholarly connections and communicative patterns on Twitter. This thesis work is specifically focused on designing, developing, and evaluating a web-based interactive GIV interface for AcademiaMap, further referred to as AcademiaMap-GIV.

The main research questions of this work include

- What are the best approaches to visualizing conversational data from social media in general; and conversational data of an academic community on Twitter in particular?

- What are the best practices for creating geo-based visualizations of social media data that are commonly used?
- What are the limitations of the existing geo-based visualizations? And how can these limitations be addressed?

To address these research questions, first, an extensive literature and Internet review of web-based geographical information visualizations was conducted. The goal of this review was to learn about the range of effective approaches to visualizing information on geographical maps. Second, AcademiaMap-GIV was designed and developed using an iterative prototyping technique by involving a group of potential users in each development cycle. AcademiMap-GIV has incorporated well regarded geo-based visualization techniques, addressed some of the limitations of the existing visualizations and introduced some unique features. Finally, an exploratory user study was conducted in the final iteration to evaluate the current version of AcademiaMap-GIV and to explore the design space for future improvements to the systems.

In sum, the main outcome of this work is AcademiaMap-GIV, a geo-based interactive visualization that offers researchers a new way to follow popular topics discussed by any scholarly community on Twitter and to study how information is being disseminated among members of this community geographically and temporally. The visualization techniques applied and tested in this work are also applicable to visualizing non-scholarly communities on Twitter and on other social media and networking sites.

The remainder of this thesis will proceed as follows. In Chapter 2, various GIV applications from the literature as well as from online applications are discussed in detail to show how researchers and businesses around the globe represent their data geographically. Since the focus of the research is on the social media data, the literature review section primarily focuses on different kinds of visualizations of social media data. In Chapter 3, based on the previous work surveyed, our approach to visualizing online conversational data geographically is presented. This chapter details the methodology of the system implementation and design. In Chapter 4, the methodology of the formal usability evaluation of AcademiaMap-GIV is provided, followed by a detailed discussion of the results of the study. In Chapter 5, concluding remarks are provided, which summarize the key contributions of this research, along with some recommendations for future improvements.

CHAPTER 2: PREVIOUS WORK

Geospatial or Geo-based Information Visualizations (GIV) are widely used in different fields such as: investigative analysis, social network analysis, health situation awareness, business analysis, traffic management, tourism, environmental science, weather forecast, news updates, and geo-information retrieval. For instance, visualizing Twitter conversations about an election can provide important insights about supporters of different parties and where they are coming from (CND Election 2011 Social Media Lab.Ca, 2011). In general, GIV are especially useful when users are trying to uncover non-visual relationships and hidden patterns within the data to gain better insights to support their decision making. This approach is often adopted when large-scale geo-referenced data is present. In part, it is because of the growing popularity of mobile devices with GPS capabilities and social media data that creates the abundance of such geo-referenced data online.

In this chapter, various GIV applications from the literature as well as from online applications are discussed in detail to show how researchers and industries represent their data geographically. Since the focus of the research is on the social media data, specifically conversational data from Twitter, this review will primarily focus on different visualizations of Twitter data. Some visualizations based on non-social media data are briefly discussed in Appendix F to acknowledge their existence and provide additional information to the readers.

One of the emerging trends in GIV of social media data is to visualize how information flows between users of popular social networking sites (e.g. Facebook, Twitter, and Flickr) across the globe (Visualizing social media | VizWorld.com, 2009). Due to its public nature and the large number of users (who collectively generate the enormous amount of data), most of the visualizations in this area rely on conversational data from Twitter, a popular microblogging service for exchanging short messages. Twitter provides its users with a text box to type up to 140 characters so that they can post their current thoughts as a status on their accounts. Their messages (also known as *tweets*) are also delivered to all of their account subscribers (also known as *followers*). In Twitter, users can mention each other's name in their conversations (also known as *mention*), and can retransmit interesting messages posted by others (also known as *Retweet* or *RT*) to their own followers. To track and indicate the primary topic of their tweets, Twitter users often include topic specific tags (also known as *hash tags*) into their tweets. Each hash tag is a single word that started with the "#" character. For more Twitter specific terms and definitions, see Appendix A.

Below is a review of different GIV systems that rely on data from Twitter, beginning with applications specifically designed to support certain tasks in two sample domains: situation awareness (Section 2.1) and political analysis (Section 2.2), and followed by a review of general purpose applications (Section 2.3).

2.1 SITUATION AWARENESS

A natural disaster like an earthquake, tsunami, flood, storm, or hurricane often leads to a spike in social media usage in the affected areas. For instance, during the recent earthquake, which happened in the US on August 23, 2011, many Twitter users in the east coast got the news regarding the earthquake first on Twitter before the earthquake hit their region (Earthquake: Twitter Users Learned of Tremors Seconds Before Feeling Them, 2011). Social media usage has also been very high among users at the time rebellions happened in Egypt and Tunisia (Twitter, Facebook and YouTube's role in Arab Spring (Middle East uprisings, 2011). During the recent rebellion in Egypt, for example, social media played a huge role in allowing the people within Egypt to disseminate their outlooks and also to make the whole world aware of the turmoil in their country. Through social media, people find a common platform to share their own perspectives and ideas about a particular incident with the whole world. In this section, different applications that visualize the data for the purpose of situation awareness are discussed in detail.

Situation awareness application like **AudioBoo** (Audioboo / The BooMap, n.d.) visualizes online audio from around the world on Google Maps. For instance, during the rebellion in Egypt in 2011, people used Audioboo to record their voices to disseminate updates during the insurgency. Anyone can record their voice describing an interesting and moving issue and upload and post it as a message through their AudioBoo account. The users of this visualization can see the audio icon on the map and can click on the icon

to play the audio file. This visualization uses heat maps (see Figure 2.1) to cluster icons according to their proximity to each other.

Similarly, **Disaster Response and Assistance** (Disaster Response and Assistance, n.d.) visualizes Youtube videos (as links), Twitter tweets, and Flickr images on a map to show the 2011 unrest in Egypt.

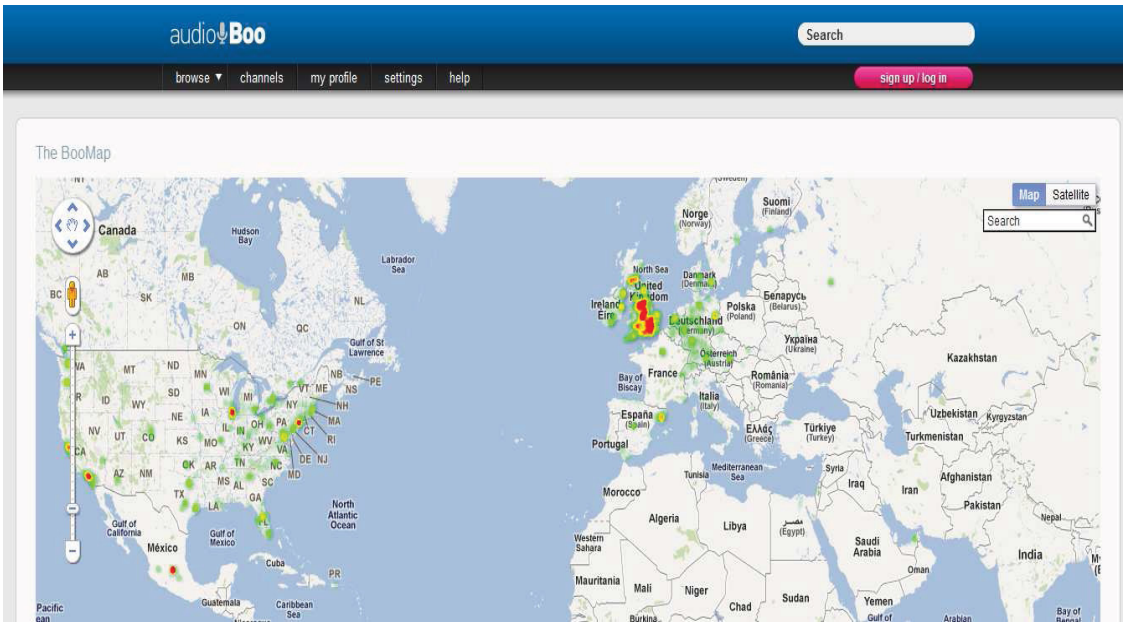


Figure 2.1: AudioBoo uses heap maps to show the intensity of data by frequency on the map, where those data originated from (Disaster Response and Assistance, n.d.)

SensePlace2 (also known as **Geo-Twitter Analytics**), developed by GeoVista Centre, at the Pennsylvania State University, is a web-based search interface to visualize data spatially on a map. It is intended for use as a geo-visual analytics application for crisis management and situation awareness during earthquakes, storms, floods, tsunamis, etc., for persistent monitoring, and regulating assistance and relief for people of the affected regions. SensePlace2 uses a crawler for collecting tweets about events of interest, limited to previously decided terms and topics, and continuously updates over time whenever

interesting events occur. This application is integrated with multiple sources, for instance, news, RSS feeds, and blog posts for the purpose of data collection. This system is used as a search user interface, where users enter a specific query in the search box to refine the real time data relevant to that query. A query is defined as a combination of terms that could be: places, incidents, people, or organizations (MacEachren et al., 2011). SensePlace2 differentiates between tweets about and from any location visually using two different icons, which represent those categories separately on the map. In situation awareness, for instance, during an earthquake, it is sometime necessary to find out whether someone from that affected region is tweeting about the current situation or not. Users, most importantly, analysts, benefit from the gathering of more accurate and relevant information regarding the situation in the affected regions. The square icon encodes all the tweets about a particular location that are relevant to the queried terms. In this case, the density of pink shades (see Figure 2.2) on the map represents the intensity of tweets about a particular region. The icon (circular) representing tweets from a particular region gets bigger or smaller depending on the frequency of tweets generated from that particular region over time. A time slider is provided to filter out the data with respect to the user-specified timeline. Also, the same shading of pink used for representing tweets about location feature is reflected on the time slider to help users relate the frequency of tweets about a location with the current timeline. A tag cloud of frequently used terms (see Figure 2.2, left bottom) in the tweets is presented at the bottom of screen to give an overview of topics over time to the users. Real tweets are provided (see Figure 2.2, left) by arranging them sequentially in a panel. Users can also sort tweets

by their relevance, time of origin, or place, to get more relevant information associated with their query.

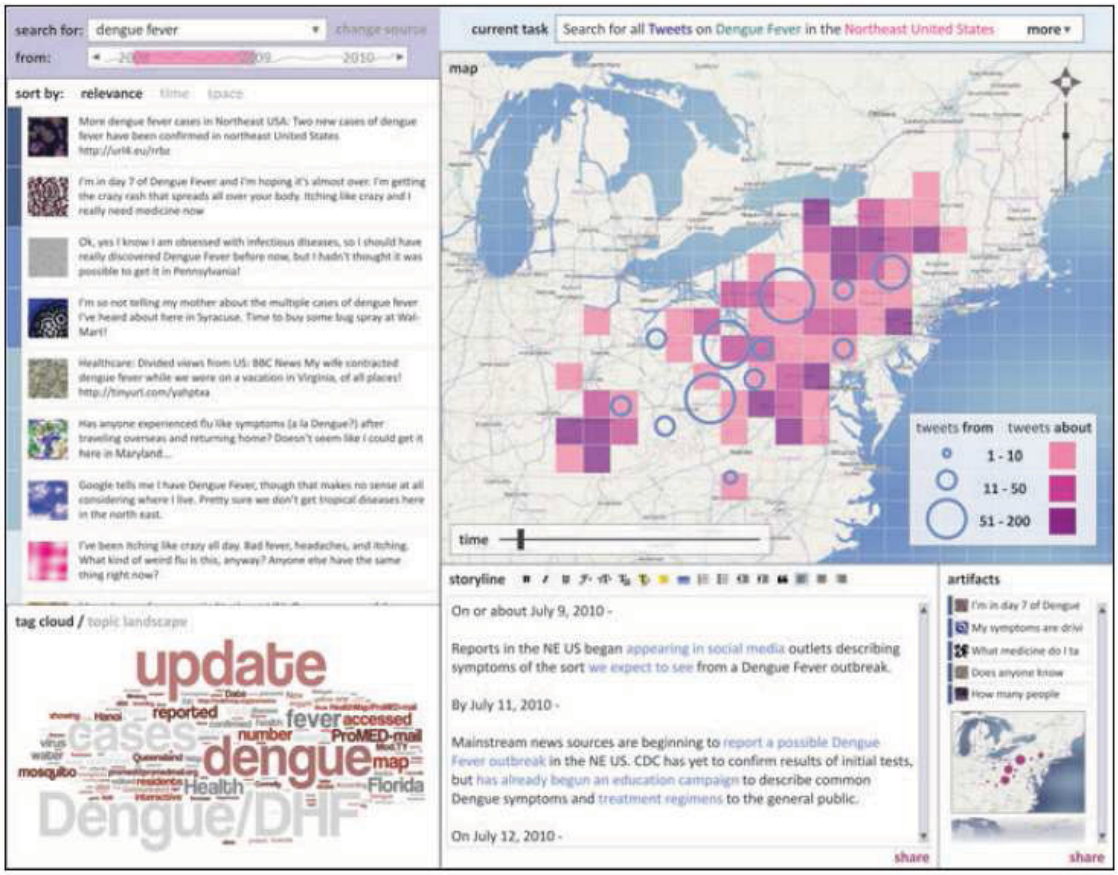


Figure 2.2: Different views of SensePlace2 visual interface during user-interaction (MacEachren et al., 2011)

Health situation awareness is another popular area in which GIV applications are used. Visualization of health related data from different social media plays an important role in the case of situation awareness. **Monitoring Swine Flu using Twitter** (Monitoring Swine Flu using Twitter, n.d.) is an online application that provides several key topics of swine flu in categories found in related tweets. All of the categories are arranged at the bottom of the page with different colours, and the frequency of tweets for each category is placed right after it (see Figure 2.3). Tweets are represented as dots of the same color

as the topics that appear in categories. It is a real time application, so it updates results quite frequently. Sample Twitter topics with frequencies are given in parenthesis, “swine influenza OR flu"(31), “swine vaccine OR shot” (30), “Tamiflu OR oseltamivir” (22), “Zanamivir OR Relenza” (2), and “Amantadine OR Rimantadine” (2).

Figure 2.3: Sample Twitter topics with frequencies given in parenthesis are found at the bottom of the map (Monitoring Swine Flu using Twitter, n.d.)

2.2 POLITICAL ANALYSIS

GIV applications are becoming popular in visualizing different societal issues and events such as: elections, employment and unemployment statistics, GDP (Gross Domestic Product) growth, and income and expense ratio. National events such as upcoming elections are also becoming a popular reason for the rise of social media usage among people. Several GIV applications designed to visualize election-related social media activities are discussed below.

The New York Times (Tracking Twitter Traffic About the 2010 Midterm Elections, n.d.) developed a visualization called **2010-Twitter-candidates** that visualizes Twitter activities around the 2010 election for the Senate and governors in the United States (US). This visualization (see Figure 2.4(a)) provides an overview of the Twitter account activities of each candidate with the timeline. This application counts online activities of candidates by taking into account the candidate's tweets, the tweets that have been re-tweeted by others, and even the tweets mentioning the candidate's name. It encodes the relative frequency of the Twitter user's activities to the size of each user's icon placed on the screen.

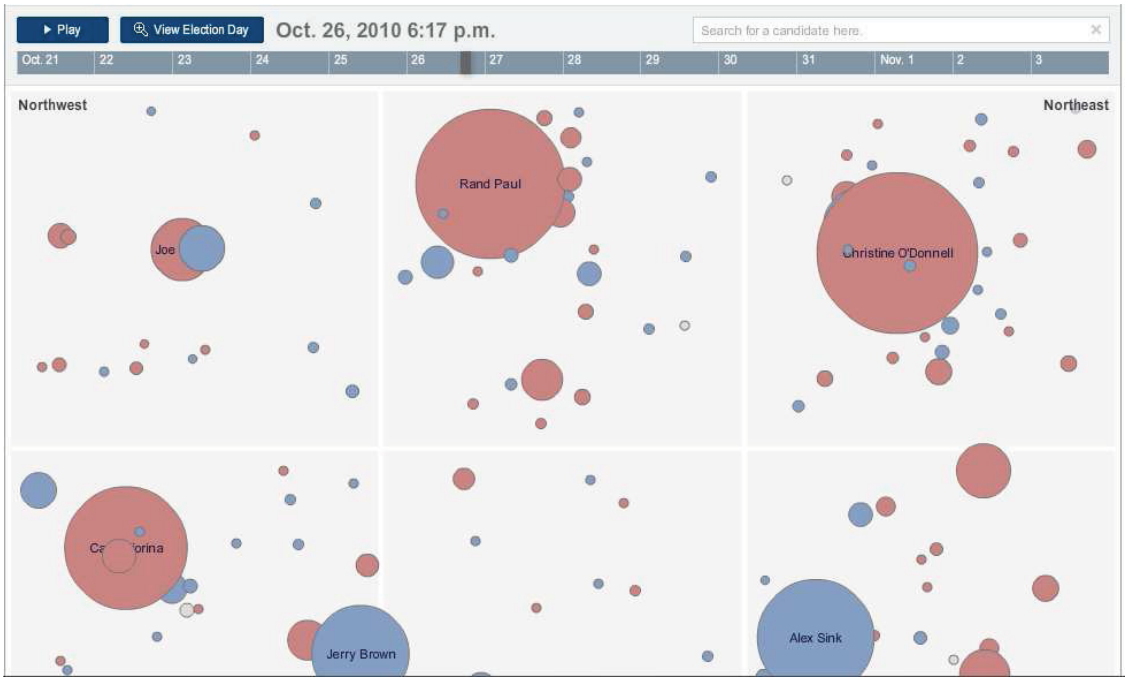


Figure 2.4(a): **2010-Twitter-candidates** provides an overview of the Twitter account activity of each candidate over time (Tracking Twitter Traffic About the 2010 Midterm Elections, n.d.).

This visualization is interesting because of its activity animation. Users can animate the timeline to see the activity as time progresses. The circular icon for each candidate shows the rate at which it grows and shrinks with respect to the activities happening around it.

The animation also shows smaller circles in lighter gray moving toward a candidate's icon to represent the tweets that have mentioned this candidate. Re-tweets of a candidate's original tweets are represented by a small circle in darker grey moving away from the candidate's icon. Circles coloured the same as the candidate's main icon color (orange means republican, blue means democratic) that are moving away from the candidate's icon represent original tweets. Users can also select a candidate (see Figure 2.4(b)), doing so will hide all other candidates and allow the user to examine one particular politician with all available features. Moreover, any two candidates' activities can be compared simultaneously by adding another candidate after selecting the first.

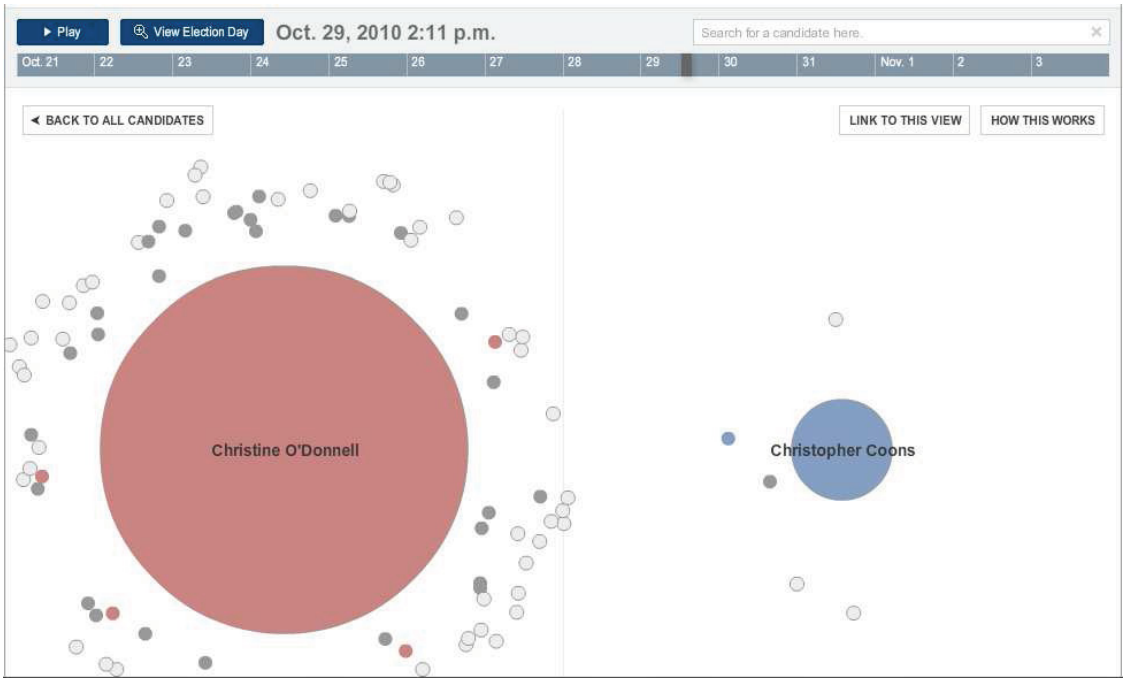


Figure 2.4(b): Comparing Twitter activities of two candidates by placing them next to each other provides more insight on their publicity in the social media realm (Tracking Twitter Traffic About the 2010 Midterm Elections, n.d.)

2.3 GENERAL PURPOSE ACTIVITIES

This section describes some of the applications that are not intended for use in any specific domain. Generally speaking, these visualizations do not have any apparent target users and in most of the cases, they are not focused on any specific dataset. Most of the applications discussed here visualize data from Twitter, Flickr, Wikipedia, and Youtube in real time; and they constantly update their visualizations by checking for any incoming recent data available from those social media sites.

In one of the most recent papers, Marcus et al. (2011) looked for prominent terms employed by users of social media networks and conducted sentiment analysis on the data over time. They developed a system called **TwitInfo** to visualize Twitter events in real time (see Figure 2.5). In its current implementation, TwitInfo works as an information retrieval system where users enter their own specific query to initiate the visualization. TwitInfo uses the Twitter API to retrieve recent events that are queried by the user and to update event specific information in real time as the event progresses. It gets updated information from the Twitter stream in the case of live events like a live televised soccer match. TwitInfo uses a multi-view visual interface of different visualizations including the visualization of topic streams along a timeline, the geo-visualization of the locations from where those tweets are generated, the display of real tweets and popular URLs (Universal Resource Locator) retrieved from the tweets, and a sentiment analysis visualization in the form of a pie chart.

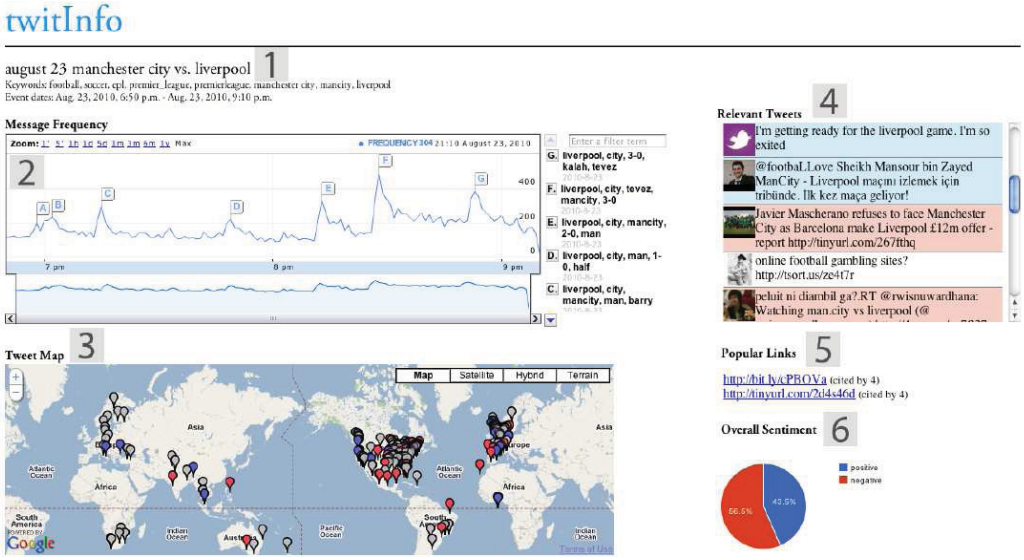
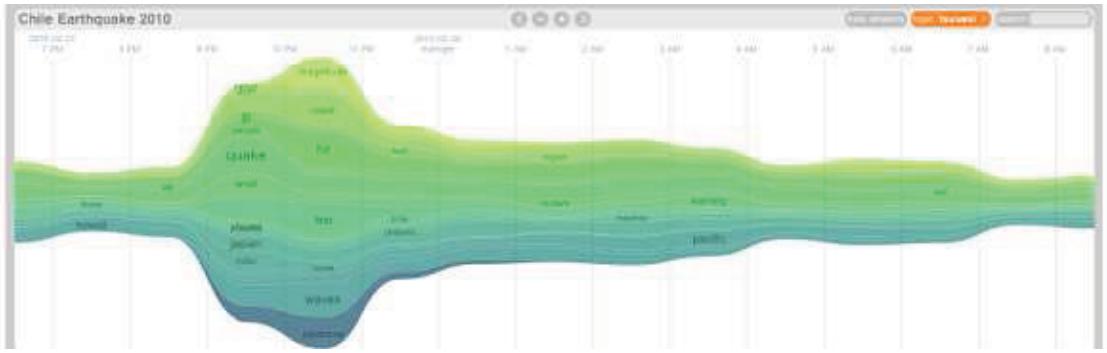


Figure 2.5: TwitInfo system with different individual interfaces that visualize and aggregate events and sub-events of a topic queried by a user (Marcus et al., 2011)

In the timeline series visualization (shown in Figure 2.5, at the left middle), the peak of each topic depends on the volume of related tweets, and each peak is labelled automatically (by a devised machine learning algorithm) by the most frequently used words retrieved in those tweets. Users can drill down more into the event by filtering out data using the timeline range and also by selecting each labelled topic to create a sub-event that corresponds to that topic. Selecting a specific topic in the timeline series (shown in Figure 2.5, at the left middle) visualization makes all other visualization interfaces update simultaneously to represent this new request. Tweets are sorted in the tweet panel (Figure 2.5, at the right top) according to their relevance to the currently selected topic. One of the most interesting features of TwitInfo is the sentiment analysis visualization, which appears in the form of a pie chart. This system segments each tweet of the selected topic as a variation of three factors: positive, negative, and neutral sentiments. In the pie chart (Figure 2.5, at the right bottom), blue represents positive, red

indicates negative, and white is used for neutral sentiments towards the selected topic. This sentiment analysis of tweets is also reflected on the map, where each tweet is plotted as a balloon icon coloured same as the sentiment of the tweet represents in the pie chart (Figure 2.5, at the left bottom).

Visual Backchannel is another application that visualizes Twitter data in real time using multiple views. Dork, Gruen, Williamson & Carpendale (2010) proposed a novel way of visualizing evolving Twitter conversations in real time. This visualization interface shows large scale data over time by visualizing topic streams from the data with a flexible time slider of different granularities, a visualization of user's activity around a spiral curve, and frequently shared images using image clouds. This visualization interface includes interesting features like topic stream as waves (see Figure 2.6, A) in which new topics are stacked on or above one another for a particular time period. The wave moves left to right to show topics over different time periods. This visualization distinguishes new topics from recent past topics by their position in the same vertical position of the wave along with a corresponding color encoding from blue to yellow green. It also separates recent topics from past topics by their horizontal position. Topic strengths, identified as the normalized frequency of words, are shown by the wave amplitude at that topic's position. All topics currently appearing in incoming tweets are highlighted in yellow (degree of opacity represents frequency of tweets) to notify users about the currently discussed topics.



A) Timeline with topic streams

B) User's activity C) Tweets D) Image clouds

Figure 2.6: Visual Backchannel's interfaces show tweets regarding the earthquake in Chile by using a timeline, topic streams, visualizing user's activity, tweets, and image clouds (Dork et al., 2010)

The people activity cloud (see Figure 2.6, B) represents how active people (users) are on Twitter over time by showing the frequency of their activities which is displayed in the system by increasing or decreasing the size of their label, and also the size of the dot placed for each user around a spiral curve. Interestingly, the saturation of the colour of each dot also changes with the frequency of the original tweets posted by the corresponding Twitter user. In the image cloud (see Figure 2.6, D), image size represents how frequently each image is shared in Twitter conversations over time.

Another similar tool is **Trendsmap** (Trendsmap, n.d.), a real-time Twitter data visualization tool that portrays trends of tweets over time by topics, and by geographic locations (see Figure 2.7). Most importantly, this visualization categorizes all tweets by different themes and topics based on their hash tags and most frequently used words. Trendsmap shows all the categories on Google Maps, and updates those in real-time. Users can see tweets from each category by clicking on the category labels on the map.

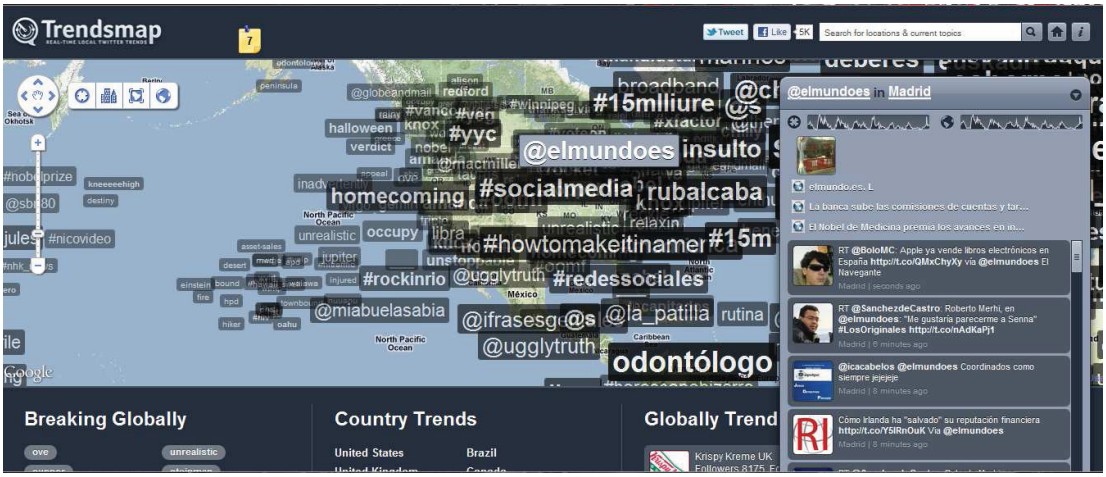


Figure 2.7: Trendsmap visualizes tweets categorized by topics and users on the map. When a topic or user is selected, it also shows all the associated tweets in the separate message window (Trendsmap, n.d.)

Search-based web applications such as **Tweetsters** visualize recorded Twitter data generated from different locations of the US (United States) in a coordinated view. In Tweetsters' visualization, the users' queried terms from tweets are visualized geographically on the US map (limited to 70 cities) (Kim, Jeong, Chew, Bonner, & Stasko, 2009). This visualization presents term(s) as node(s) by plotting them on a specific location of the map if they appear in any of the tweets generated from that particular location (see Figure 2.8). Tweesters provides a two way movable time slider to specify the time period in which users are interested to see how terms appear in tweets

over time. As each node (term) is visually encoded by the frequency of tweets in which it appears, users can gain insights on how a term in tweets is evolving over time. Interestingly, this visualization allows users to compare two or more terms, to understand the popularity of topics or products over time. To distinguish between multiple terms in the visualization, a unique color coding is used for each term queried by the users. The process of comparing multiple terms over time expedites the decision making process; in particular, it benefits users doing any kind of survey research work by allowing them to compare sets of data.

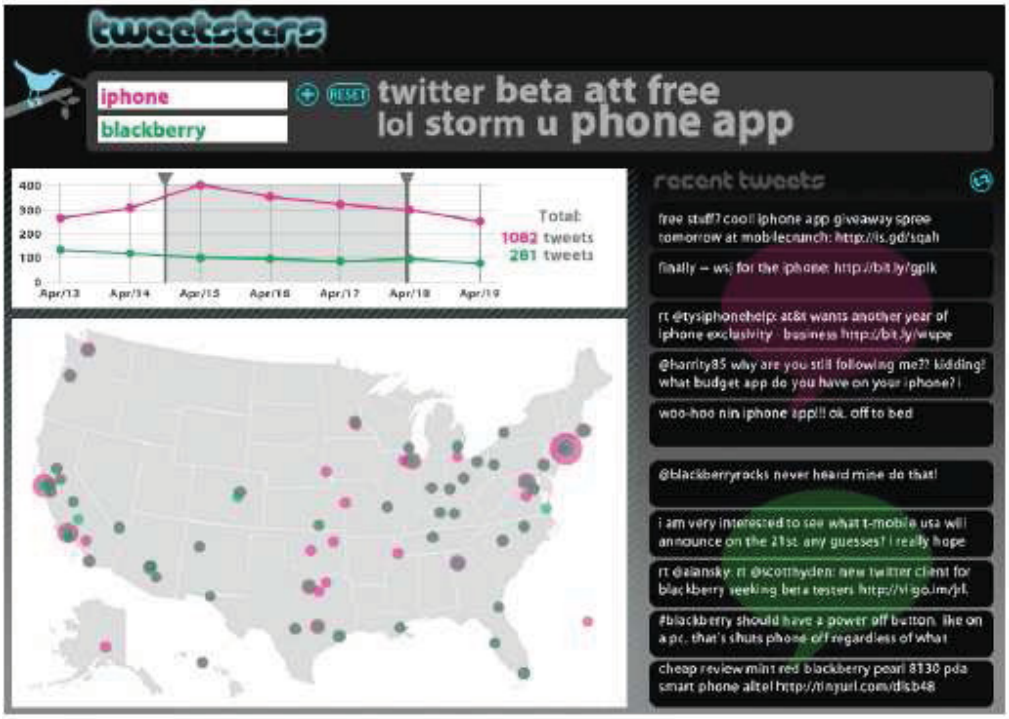


Figure 2.8: Tweetsters allows users to compare two or more terms, to understand the popularity of topics or products over time (Kim et al., 2009)

Another search-based web application, **Cartoview** (Cartoview, n.d.) visualizes data (see Figure 2.9) from Twitter, Wikipedia articles, Youtube videos, Flickr images, and other

data from various social media networks on Google Maps. Cartoview works as an information-retrieval interface by which users can search for an interesting term to get started, or can focus on a specific location to retrieve the data of interest. In this case, Cartoview coordinates all user interactions by providing several features to the users include zoom, pan, mouse-hover, selection, and search. Each kind of data is visualized as icons on the map, and is also presented in text format in a separate message panel.

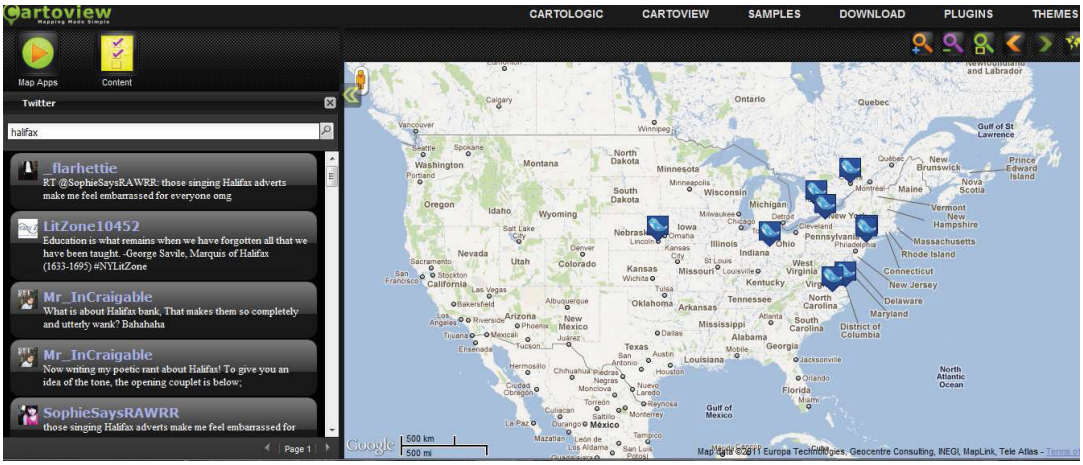


Figure 2.9: Cartoview shows tweets about Halifax by placing icons on the map and presenting related tweets in a separate message panel (Cartoview, n.d.)

A similar kind of visualization search-interface is the **Word on the Tweet** (Heatmap for Twitter-The Word on the Tweet, n.d.)**.** In this visualization, users can search for any word in the dialogue box to retrieve relevant tweets (see Figure 2.10). Icons are encoded into the heat map by the frequency of tweets on the map. When users click on heat map icons at the upper zoom level, this visualization provides all the tweets on the pop-up windows with tag-clouds generated using frequently appearing terms in those tweets. This visualization features tag clouds of retrieved tweets at the top of the pop-up window to help users focus on more important terms by their frequency of occurrences in tweets.

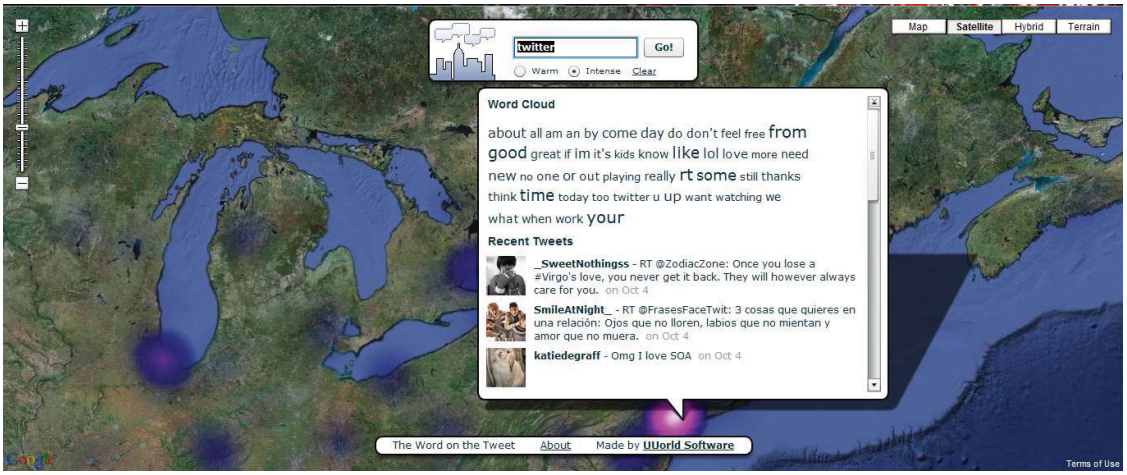


Figure 2.10: Word on the Tweet provides a tag-cloud of tweets at the top of a pop-up window (Heatmap for Twitter-The Word on the Tweet, n.d.)

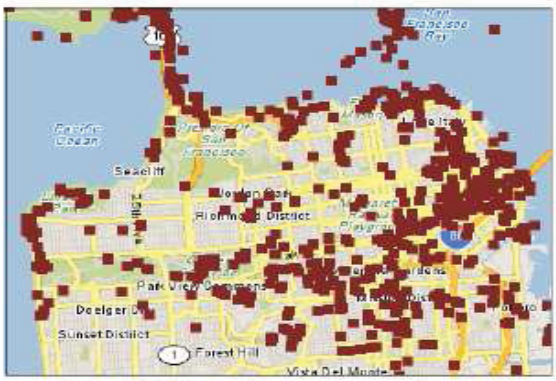
The resulting visualization interface includes different categories of topics to filter out unwanted information and help users focus on a specific task, but does not work as a search-based interface.

Oscar Twitter Map (Oscar Twitter Map, n.d.)**,** from Neoformix visualizes the 2009 Academy Awards on a map by using Twitter messages that were collected during the time the show aired. In this visualization, users get an insight into the event based on the tweets which mentioned specific names (movies, people) and associated qualitative words (adjectives). This visualization (see Figure 2.11) also provides users with the option to move the time slider manually or automatically to see the prominence of different subjects such as movies, people**,** and nominees over time. In addition, people are arranged in categories according to their roles in movies. Most importantly, it shows adjectives that are associated with a particular person or movie in the tweets, in order to classify audience views as positive or negative.

Figure 2.11: Oscar Twitter Map visualizes key terms in Twitter conversations that happened during the television broadcast of the 2009 Academy Awards (Oscar Twitter Map, n.d.)

Instead of summarizing Twitter conversations, **TagMaps** (see Figure 2.12) visualizes collections of photographs from a large multi-user collection using maps and based on the photo's geo-referenced meta-data collected from Flickr such as: location and time (where and when the photographs were taken), photographer's user information, tags (textual labels) of each photograph, and other externally derived parameters such as their quality and relevance (Jaffe, Naaman, Tassa, & Davis, 2006). The relevance of a photograph to any subset is defined by how important the photograph is to the collection based on other distinguishing parameters such as: how recent the photo is, the time of day and day of the week it was taken, the social network of the user, and user attributes. TagMaps uses a summarization algorithm that applies hierarchical clustering to a collection of photographs by receiving locations of the photographs as a parameter. The ranking score

of each cluster in the hierarchy is computed by using the meta-data of the photographs. Interestingly, this system uses the ranking score to find the most relevant photographs that define a particular location, so that it can show a subset of photographs at the current zoom level on the map. By this approach TagMaps reduces the visual clutter of having too many icons overlapped on the map.



(a) All San Francisco photos

(b) An automatic summary of San Francisco photos

Figure 2.12: The Tag Maps system summarizes photos by their features to show the most representative image at the higher zoom level (Jaffe et al., 2006)

2.4 DISCUSSION

This chapter described several areas where applications use geo-reference data to visualize data points based on their geographic location. Geo-visualizations are becoming more popular in visualizing social media data to better understand relations between different entities (people, organizations, topics) and their online activities. Below is a summary of some key features of the visualization interfaces discussed in this chapter.

Geographic maps to present data spatially: Maps are essential to understand the spatiality of the data in case of geo-visualization. Most of the applications discussed above incorporate some form of a map to plot their data on. Their geo-referenced meta-data is usually related to either the locations from where the data originated or to which it refers.

Time slider to show data temporally: Most of the visualizations also incorporate a time slider to allow their users filter out data by using time parameters. Interestingly, applications that visualize real time data do not have any time slider, except for Visual Backchannel (Dork et al., 2010) and TwitInfo (Marcus et al., 2011). Both of these visualizations annotate their timelines with streams of topics to show real time data as time progresses.

Visual encodings of information: Visual encodings of information are essential in order to represent multiple dimensions of the data by size, colour, shape, and position. Some advanced visual encoding techniques such as heat maps have been used by few applications reviewed above such as AudioBoo (Disaster Response and Assistance, n.d.) and the Word On Tweet (Heatmap for Twitter-The Word on the Tweet, n.d.). Heat maps represent the intensity of data generated from a particular region by using the colour

intensity to encode the frequency of messages coming from a particular location. This is an especially useful technique when trying to visualize multiple data points that happen to be in the same location or within a small distance from each other.

Animation to represent the changing nature of online data over time: 2010-Twitter-candidates (Tracking Twitter Traffic About the 2010 Midterm Elections, n.d.) and Oscar Twitter Map (Oscar Twitter Map, n.d.) incorporate interesting animations of the live data from Twitter. Such animations enable users to see how data are changing over time. This feature can help users to find potential temporal patterns in data over time.

Multi-view coordinated interfaces: Most of the visualizations present multiple views based on the same underlying data to provide users with a platform to explore the data from different perspectives. This feature is especially prominent in applications such as TwitInfo (Marcus et al., 2011), Visual Backchannel (Dork et al., 2010), SensePlace2 (MacEachren et al., 2011), and Tweetsters (Kim et al., 2009). The multi-view interfaces commonly use such visual techniques as *brushing* and *linking* which connect and update all views of the interface to reflect any changes made by a user in one of the views.

Search interface: Some visualizations act as information retrieval systems, where users need to search for a query to get started with the visualization such as SenesePlace2 (MacEachren et al., 2011), TwitInfo (Marcus et al., 2011), Tweetsters (Kim et al., 2009). Although potentially useful, the limitation related to this kind of interface is that users would need to understand the underlying data in the first place to come up with a search query.

Tag-clouds to visualize popular topics: Tag-clouds are becoming a popular way to bring forward most frequently used topics appearing in the textual data. This kind of text

visualization provides users with an overview of the whole dataset. The size of each topic (word) often represents how often it was mentioned in text. Some visualizations discussed above that incorporate tag-clouds include SensePlace2 (MacEachren et al., 2011), Oscar Twitter Map (Oscar Twitter Map, n.d.), Tweetsters (Kim et al., 2009). and the Word On Tweet (Heatmap for Twitter-The Word on the Tweet, n.d.).

Some limitations of the existing visualizations:

Below is a brief discussion of some major limitations of the applications reviewed above.

Complexity of Data Representation: Complex data representations such as heat maps have aesthetic values and provide a good way to represent high-level trends; however, it may also make it difficult for user to access and review individual data points. For instance, in the case of AudioBoo (Disaster Response and Assistance, n.d.), the application uses a heat map feature; but users need to continuously zoom in to the map to be able to get into the higher zoom level to explore the actual data points. For this reason, it was decided not to include a heat map representation in AcademiaMap-GIV.

Lack of geographic reference: In 2010-Twitter-candidates (Tracking Twitter Traffic About the 2010 Midterm Elections, n.d.), the visualization does not incorporate any map, instead it displays geographical information of the data as texts (by directions) on the screen. This visualization could be even more informative and useful to its users if it incorporated a geographic map to show geographic areas of their influence online.

Missing filtering options: Some of the applications do not provide necessary features for the users to interact with the system. For example, in Trendsmap (Trendsmap, n.d.), different categories (place, person, event) of information are presented together on the map with labels to give an overview of the current trends on Twitter. However, this

visualization is overly cluttered with category labels as it keeps updated its visualization upon the arrival of new data (real time streaming). However, Trendsmap does not provide any data filtering option to show only user preferred category of information on the map. As a result, this visualization may cause confusion among users during their tasks.

Placement of a tag cloud: Sometime the wrong placement of interface components can exhibit certain problems in case of usability of the system. In SensePlace2 (MacEachren et al., 2011), placement of the tag cloud in a separate window rather than in the time slider window can distract users during their task. If it is placed near the time slider, users would be able to follow the evolving conversations on Twitter easily by relating those topics with the timeline changes. Interestingly, in the usability test, users did not find this tag cloud feature useful, although specific reasons for this are not stated.

The review of the existing visualizations is of great benefit to this thesis work because it allowed us to understand the different ways people usually visualize data geospatially while enabling us to identify the limitations of those visualizations. There are some features provided by the visualizations that we surveyed that have turned out to be most common and useful features to be considered for any geo-visual interface. These include time sliders for data filtering, animations of data over time, tag-clouds for visualizing popular topics, and multi-view coordinated interfaces. This thesis work will incorporate these features in AcademiaMap-GIV, while at the same time attempting to address some of the limitations of the existing visualizations. The next chapter will discuss the design process and user interface of AcademiaMap-GIV.

CHAPTER 3: METHODOLOGY AND IMPLEMENTATION

In this chapter, we focus on the design and implementation process of our web-based interactive GIV (Geo-based Information Visualization) system, AcademiaMap-GIV, which visualizes online conversational data from social media (using Twitter as a case study). The goal of this project is to make the rapidly-changing conversations that happen online among academics more perceivable and useful by visualizing them geographically and temporally.

Our intention is to visualize the underlying communicative patterns in these online groups, specifically between online peers and popular topics over time (as interactive tag-clouds) based on their posted tweets. Visualizing this information geographically (based on the location stated in the scholars' profile) and temporally helps us to better understand how scholarly information is being disseminated and shared over time across different regions in the world. To ensure that AcademiaMap-GIV has an effective visual interface, Ben Shneiderman's (1996) directive: "Overview first, zoom and filter, details on demand" is adopted here for designing the interface. The idea behind this approach is to provide users with a high level view of the data at first by providing necessary visualization cues. Afterwards, users can choose to drill down (through zooming and filtering) further on the data to focus on a specific segment of it in order to find interesting underlying patterns of the data. With each of user interaction, more details about the data are provided to help the users make sense of the task at hand. The

following sections of this chapter will discuss the methodology of designing the visualization interface, the system architecture, and the visualization techniques.

3.1 SYSTEM OVERVIEW-ACADEMIAMAP-GIV

AcademiaMap-GIV starts by loading all necessary data from the database collected via Twitter API (Application Programming Interface). Initially, an overview of the dataset is provided to the user by displaying each person from our dataset as a node/marker on the map (see Figure 3.1). Only users who posted at least one message (or tweet) within the last month are displayed. However, users can change the time period that is displayed and review data from the previous months as well, by using the time period sliders. More specifically, users can filter conversations (tweets) using one of the two sliders: one to only show messages posted on a particular day and the other for filtering by a broader time range. Users can also see how often people mentioned or forwarded (retweeted) each other's messages by clicking on any of the visible nodes. Popular topics (also known as *hash tags*) are shown as interactive tag clouds to provide an overview of the discussed topics over time. Users can click on any popular topic from the tag cloud to read all of the messages that mentioned this topic within the selected time range.

One of the interesting features of our system is that its users can observe how tweets are posted geographically and over time using a user-friendly animation. Users can pause the animation at any specific point to better view the data. This kind of animation may help users find trends in multi-dimensional data (Robertson, Fernandez, Fisher, Lee, & Stasko, 2008).

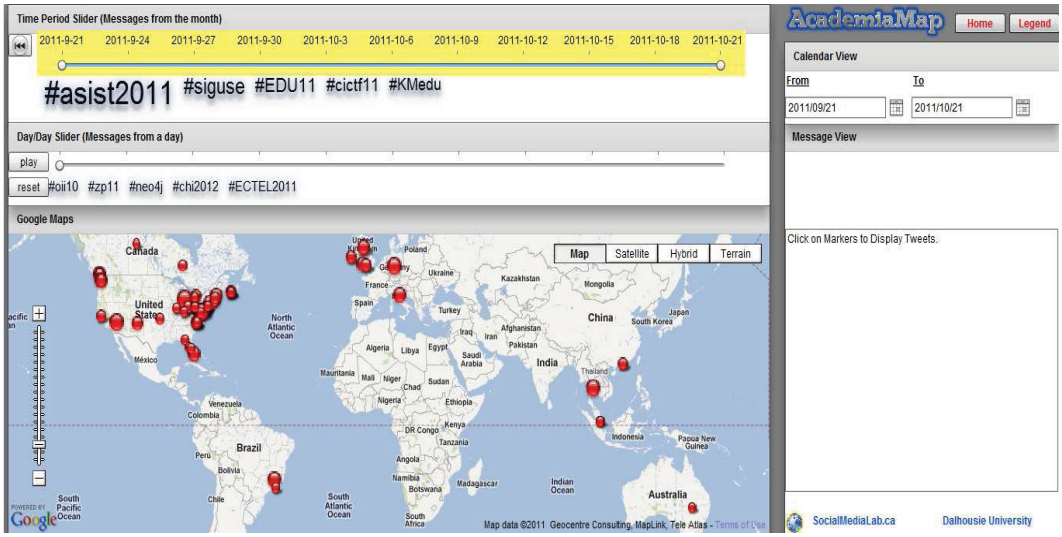


Figure 3.1: Main screen of AcademiaMap-GIV after loading the data

3.2 SYSTEM IMPLEMENTATION AND DESCRIPTION

AcademiaMap-GIV is a web-based mash-up comprised of several APIs and frameworks, as well as several kinds of programming languages (scripting, object-oriented, and static-typed). In this section, the details of the system implementation are discussed including: development tools, data retrieval methods, analysis, and interaction design.

Development Tools: The integrated development environment, Adobe Flash Builder 4, was used as the main development tool. Inside Flash Builder, the Flex framework and Google Maps API for Flash were used for developing the core interface components and user interactions on the map. Flex is a framework based on programming languages such as Action Script 3 (an object-oriented and strongly typed language) and MXML (Macromedia eXtensible Markup Language-XML based scripting language). The Flash platform was chosen for this web-based geo-visualization due to its compatibility with the other systems previously developed by the Social Media Lab. The developed GIV

application is built with the intention to be eventually integrated as a plug-in for other web-based applications, such as ICTA (Internet Community Text Analyzer), which also uses the Flash platform for visualizations.

Data Retrieval and Analysis: AcademiaMap-GIV follows only public Twitter accounts and collects only public Twitter messages. For AcademiaMap-GIV to be able to collect someone's public messages, it first has to be allowed to follow that person's Twitter account by the owner of the account. If a Twitter user decides to opt-out from the automated collection of their tweets, he or she can do so by simply removing our research Twitter account - @Research4SML- from their list of followers. This is clearly specified in the "bio" section of the profile page of the @Research4SML account. At the moment, the system's targeted audience includes scholars, students, and practitioners in the field of Library and Information Science who follow the American Society for Information Science & Technology (ASIS&T) annual conference 2011 on Twitter. We retrieve their public Twitter messages (called "tweets") on a regular basis using a web crawler (Twitter API). After the retrieval of data, each tweet is stored in the SQL (Structured Query Language) database with all its attached attributes: user name, tweet, time of tweet, location, mention, and hash tag(s). To retrieve data from this database, a set of API calls was developed by Sreejata Chatterjee of the Social Media Lab.

Data Loading and Parsing: Data loads into the visualization interface in several steps. First, a maximum and minimum time line is requested (latest and oldest dates of tweets) to the server over the HTTP service using API calls to load the data for the latest month.

For each one-month period, data is loaded in 5 (at most) API requests to the server, using a maximum and minimum date for tweets in each call. The data is retrieved in a form of a JSON (JavaScript Object Notation) array (See below):

{"**tweets**":"*RT @spmallette: Getting a few lines of #rexster code written this afternoon. #tinkerpop "*,
"**Twitter_username**":"twarko",
"**location**":"Santa Fe, New Mexico",
"**timeoftweet**":"2011-03-11 15:58:57",
"**retweet**":"spmallette",
"**mentions**":"spmallette",
"**hash_tags**":"rexster tinkerpop"}

Figure 3.2: Sample Tweet (in JSON) with the attached attributes received from the server

After loading and saving all the necessary data into the application, all data is arranged into collections of arrays by each unique user name. For each unique user, one object is created with multiple attributes and associated values such as “Twitter_username”, “location”, “tweets”, “mentions”, “retweet”, or “hash_tags”. For the “tweets”, “mentions”, “retweets”, and “hash_tags” attributes, arrays are used to store multiple values of the same attribute. The attribute “tweets” for each user consists of other attribute values such as “timeoftweet”, “location”, and “tweets” (=the actual message). The detailed structure of the data table is shown in Figure 3.3 below.

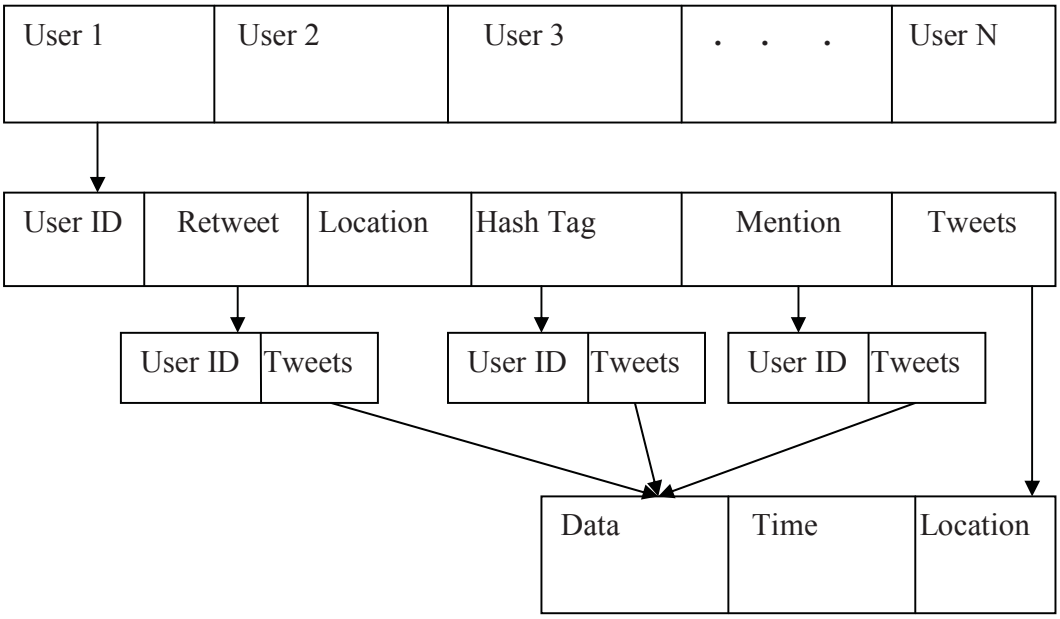


Figure 3.3: Internal Data Representation

User Interface (UI) Components (Views): AcademiaMap-GIV supports a multi-view approach (see Figure 3.1) to drilling down into the data to make sense of the underlying patterns that the data may contain. In the spatiotemporal analysis, it is necessary to provide users with so-called *coordinated* interfaces, where different views are focused on the multi-dimensional data from different perspectives. Moreover, it enables users to get insights on the data by focusing on it from multiple directions over time with respect to locations. Any changes in one view affects the data displayed in the other corresponding views.

In AcademiaMap-GIV, each node/marker on the map (see Figure 3.1) represents each Twitter user from the dataset. Each Twitter user is represented on the map by assigning a unique latitude and longitude address using the function described in Appendix D.

Hovering the mouse over a node shows more information about the corresponding user's profile. Clicking on a node shows all of the messages in the message panel at the right (see Figure 3.1). Lines are drawn on the map to show mutual connections between two persons (between two nodes on the map). Nodes can be filtered out by using the time sliders (top side of Figure 3.1) to see messages for a specific time period or from a specific day. Users who have not posted during that time will not appear. Underneath each slider, the five most frequently used hash tags (popular topics) mentioned in the tweets (messages) are displayed in the form of a tag cloud (see Figure 3.1). These hash tags change over time whenever users drag a slider or switch between sliders to display a new time period. See Appendix C for viewing high level block diagram of the system.

Below is a more detailed description of each interface element.

Nodes and Lines: Each node's size on the map is encoded by the frequency of messages/tweets posted by that person. The frequency encoding to the node allows interface users to get an overview on the dominant nodes (who posts most messages). Initially, all nodes are filled with the red colour (see Figure 3.1). If a node is selected, it turns from red to green (see Figure 3.4). If the selected Twitter user also mentioned someone else in his or her tweets, there will be a line(s) connecting the selected user to those whom he/she mentioned (see Figure 3.6). By clicking on the line, it will show only the messages that were exchanged between these two individuals (see Figure 3.12). If one of the top five hash tags is selected, the Twitter users who used this hash tag in their message(s) will also be indicated by changing the colour of their nodes from red to green (see Figure 3.5).

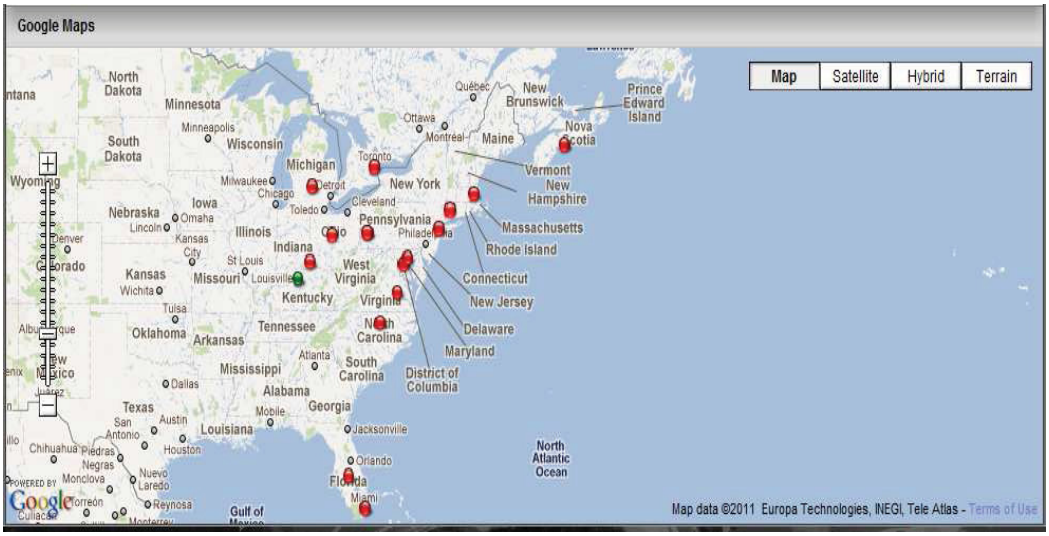


Figure 3.4: A green node is overlaid when a red node is selected

Figure 3.5: Green nodes double in size when a hash-tag is selected that appears in its messages

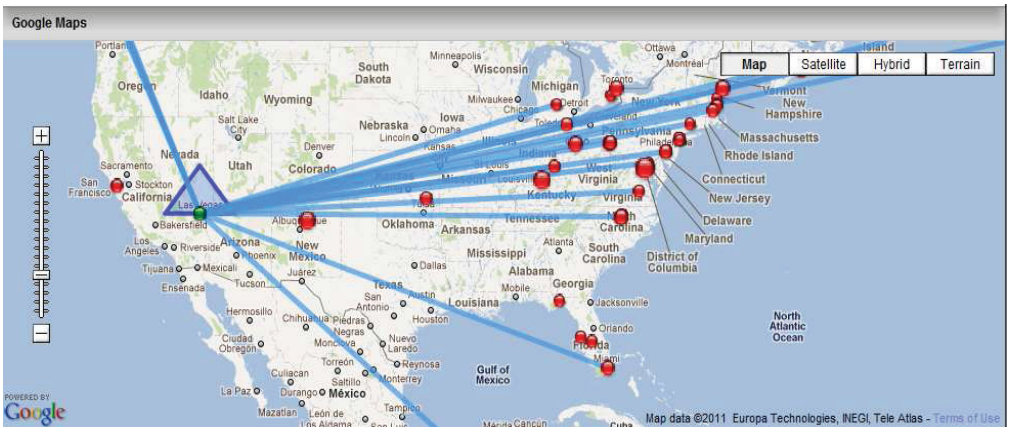


Figure 3.6: Each line joins two nodes if one node (person) mentions and/or retweets another.

Note: Triangles are drawn around a node to represent self-reference messages of the selected node on the map

Polygons (triangles): Polygons (triangles) are drawn around a selected node (as shown in Figure 3.6) if the node mentioned himself or herself in any of the tweets within the selected time period. If users click on the polygon, it shows those messages in the message box.

Circles: If a message is selected in the message box, a circle will appear around the node on the map to indicate who posted this message. The map will also simultaneously adjust to center the node in the middle of the screen (See Figure 3.7).

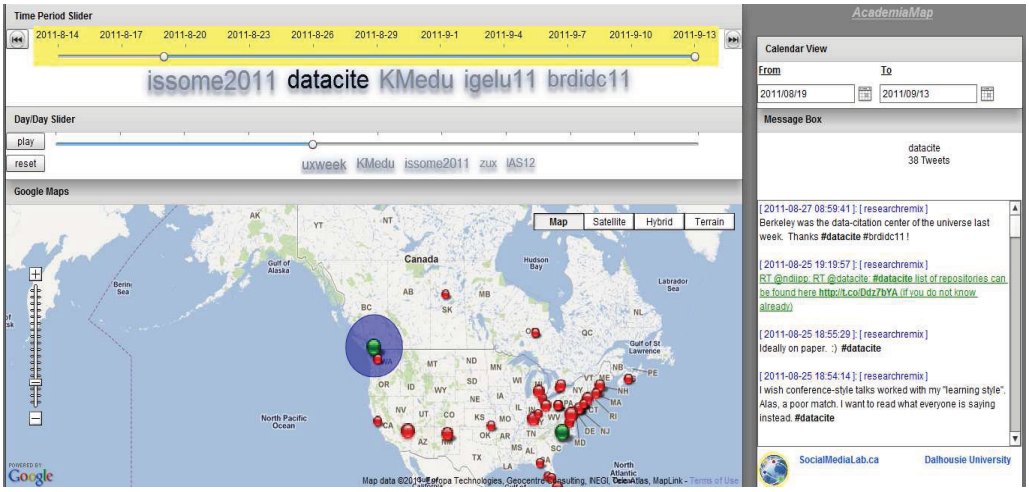


Figure 3.7: Circle around a node helps interface users to identify Twitter users' corresponding nodes geographically on the map by selecting their posted messages in the message panel

Timeline View: The timeline view consists of the three interface components: "Time period" slider, "Day/day" slider, and the Calendar display. The two different time sliders were used to enable in the visualization to delineate and compare the evolving activities by the Twitter users over a broader range of time (weeks) and over a single day. Users can easily switch between these two sliders (the time period slider and the day/day slider) by clicking on the corresponding slider.

Time Period Slider: Time period slider is set initially to allow a 30 day time period to be displayed in order to provide an overview of the segmented dataset. The timeline of the currently loaded segment is shown at the top of this slider in form of dates. By default, the time period slider is enabled (highlighted in yellow) at start-up to show the minimum and maximum time range of the selected dataset. Users can change a date/time range by dragging the two thumbs provided in this slider (see Figure 3.8).

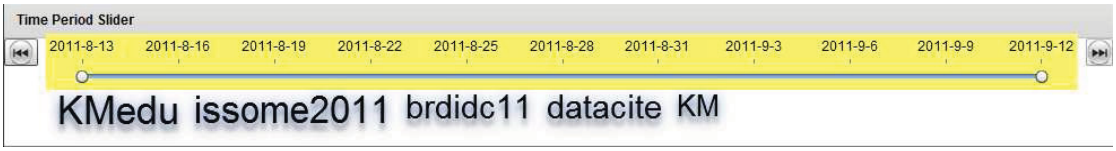


Figure 3.8: The time period slider shows and works on minimum and maximum time ranges from the selected segment of the data

Day/Day Slider: When the Day/Day slider is enabled (highlighted in yellow) it displays messages/nodes from a single day. Users can set this slider to any point (any day) by dragging the single thumb provided in this slider. The most interesting functionality related to this slider is the activity animation. The button labeled (as shown in Figure 3.9) "play" is provided on the left side of this slider to automatically run an animation of how these online geographic conversations evolve over a single day time period. The "reset" button reverts the whole interface back to the initial state (the current month with no nodes selected). Some basic user-interactions provided in this feature are described underneath in few points:

1. When users press the play button, the current timeline switches to the timeline associated with the day/day slider.
2. During the period of animation, the slider point of the day/day slider moves by one day as the animation goes forward.
3. During the period of animation, the current day (as dates) is shown in the calendar display to visualize the system current timeline.
4. Tag-clouds are generated for the current timeline and shown underneath of the day/day slider to visualize the popular topics (hash tags) for a specific day.
5. Users can pause the animation at any time by dragging the day/day slider, clicking anywhere on the slider, or by hitting the pause button.

6. Users can select a node before starting or even during the animation to see the activity just for that particular Twitter user over time.

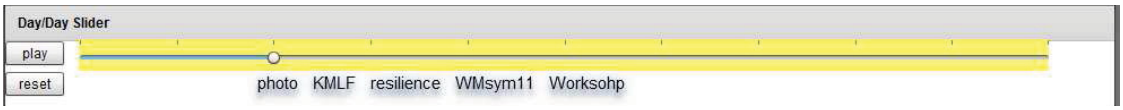


Figure 3.9: In the Day/Day slider, each time the slider is dragged a single day of the selected data is displayed

Calendar display: In AcademiaMap-GIV, a pop-up calendar is also provided to users as an additional option to view and select a time period (as shown in Figure 3.10). The user's selected time range as dates are shown in the text fields labelled as "From" and "To". However, when the Day/Day slider is active in the visualization, the "From" text field becomes disabled (as it is not needed in this case).

Figure 3.10: Calendar display allows the user to choose specific dates

Message Box: The message box contains all of the tweets posted by a selected node, as well as the profile information of the person (see Figure 3.11). If this person mentions someone (an arrow indicates mentions), then the mentioned person's profile and picture are shown at the right corner of the top portion (shown in Figure 3.12). Messages/tweets posted by a selected person (node) are shown in the middle of the message box. Messages

are sorted in chronological order; the most recent messages during the selected time period come at the top. In each message, mentioned names and all hyperlinks are made bold to highlight them in the text. Also, hyperlinks are enabled by using regular expressions so that users can click and open the link in a browser.

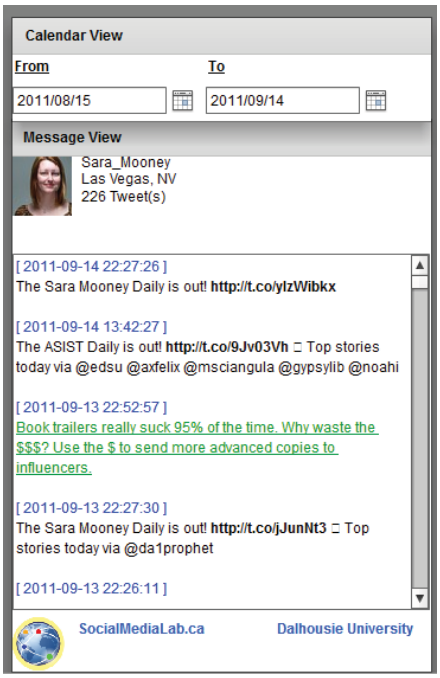


Figure 3.11: Message Box contains all the tweets posted by the selected node during any particular time period and also the profile information

Figure 3.12: If the selected node (person) mentions someone in their message and the connecting line between them is clicked on, then the messages are shown in the panel with both profiles information at the top

Tag-cloud View: In the tag-cloud view, the five most frequent hash tags (popular topics) from the dataset are shown underneath each slider (shown in Figure 3.13) to provide an overview of the popular topics. The position and size of each hash tag indicate the topic's popularity. The Flex glow filtering and shadowing built-in functions are applied to the hash tags to make the tag cloud more attractive and also more perceivable to users.

Hash tags move during one of the following three interactions:

1. whenever a user drags any of the sliders;
2. when the "play" button is clicked to run the animation; and
3. when a date is chosen from the calendar.

All hash tags are click-able. Clicking on a hash tag shows green nodes (double in size) on the map (see Figure 3.7). These green nodes represent the Twitter users who used the selected hash tag in their messages.

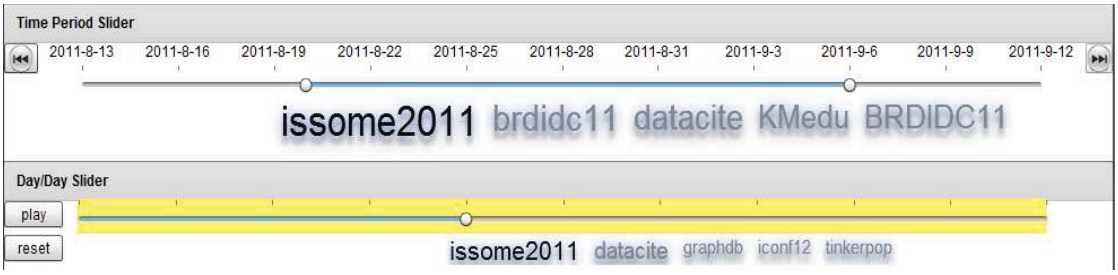


Figure 3.13: Parallel tag-clouds of different time periods help users understand and compare the evolving nature of popular topics (same topics are highlighted)

Zoom and pan: Google Maps Flash API default zoom and pan features are used alongside with other interface components to allow users to better explore the map visualization. Google default zoom and pan scroll bars are positioned on the map and also zooming feature by scrolling mouse-wheel is loaded to provide users with better flexibility in performing this operation.

Tooltip: More details are provided to users in a form of 'tooltips' whenever they mouse-over an interactive object in the visualization. This kind of interaction cues help users perform their tasks at hand. For instance (see Figure 3.14), when the user mouse-overs a node in the visualization, the profile information (Twitter account name and location) of the Twitter user (represented by the node) is provided in a pop-up hint box.

Figure 3.14: Whenever users mouse-over on an interactive object, tooltips are provided to help users understand about the possible interactions that the object leads to

User Interaction Design: As mentioned above, all visualization objects are clickable, each click alternates between their "on" and "off" states. In Figure 3.15, the interaction hierarchy is shown from a high-level view. For example, users can get started with the system exploration by interacting with the tag clouds or nodes representing Twitter users on the map, and then following this interaction hierarchy, users can interact with lines connecting different nodes or with messages. All interaction steps are saved to allow the users to get back to the previous state. For instance, when a user clicks on a line (if it exists) after selecting a node, the message box will show all the messages in which selected node mentioned or/and retweeted someone else's messages. Now if the user deselects the line by clicking on it again, all the messages from the previous state will be loaded and shown in the message box once more. This kind of cascading interaction is used to allow users to select and deselect objects and to see the corresponding changes

during the interaction with a particular object. Sample scenarios of basic user interactions with the system are provided in Appendix B along with their work flow diagrams.

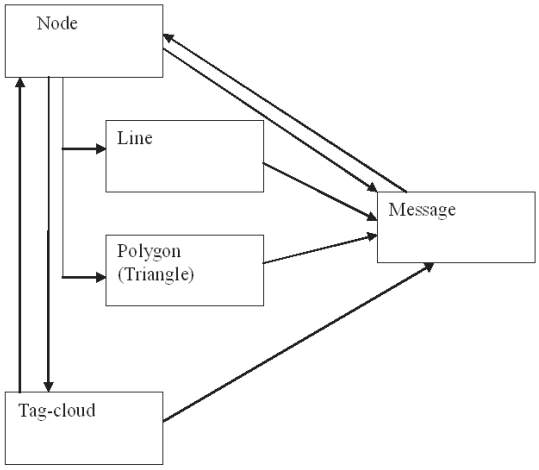


Figure 3.15: Interaction Hierarchy: the arrow is used here to show the direction of generations of interactive objects during user interactions

3.3 VISUAL ENCODING TECHNIQUES: Representing Information through Size, Colour, Position and Shape

Four different types of visual encoding techniques were used in the system to represent multiple dimensions of data points: encoding by size, colour, position, and shape. Each is discussed in detail below.

Encoding by size: First, the size of each node on the map represents the number of messages posted by each Twitter user. (In AcademiaMap-GIV, an image of a dot in the PNG format is used to represent the node.) Although there is no standard function to encode the frequency of messages posted by each user, two commonly used functions are: linear and square root (Viegas, Wattenberg, & Feinberg, 2009). After conducting a number of empirical tests, we have decided to use a simple linear function to calculate the

size of each node when visualizing data for a single day and a square root function for longer time periods. This is to account for the differences in the number of messages posted in a single day versus during in a longer time period. As a result, depending on which time slider (a broader range or a single day) is active, one of the following size encoding formulas is applied:

The node size (*N*) for a single day is calculated as:

$$N = D + Freq$$

The node size (*N*) when the time range is longer than a single day is calculated as:

$$N = D + \sqrt{Freq}$$

where

D = the default size of the node

Freq = the number of messages posted by the user within the specified time period

In both cases, the same threshold value is used to specify the upper bound value for the size of the node. The main reason to have the two different formulas is to indicate that the frequency of messages based on a longer time period outweighs the frequency of messages from a single day.

To calculate the size of each word in the tag cloud (*F*), a different frequency encoding function was used (also determined empirically):

$$F = Fm + \frac{FreqT}{FreqMt} \times Factor$$

where

Fm = the minimum font size

FreqT = the word frequency

FreqMt = the frequency of the most frequent word

Factor = scaling factor

In the above equation, the frequency of each term is normalized based on the most frequent term. Here, the most frequent term is displayed with the largest font size, which is defined by the factor of scaling (constant), and the minimum font size (constant). The rest of the terms are drawn in comparable font sizes, with respect to the most frequent term. Because both of the tag clouds (for a broader range and single day timeline) are visible at the same time, the same frequency encoding equation is applied to them. This allows users to visually compare the relative popularity of topics over a broader and single day time period.

Encoding by colour: In the visualization, to differentiate between the selected and non-selected actions by the users, colour plays a significant role. Nodes turn from *red* to *green* and lines between nodes turn from *blue* to *red* once they are selected. In case of the tag clouds, all tags are initially drawn in black; however, if a tag is selected by a user, then non-selected tags are drawn in grey, and the selected one remains black. All messages in the panel are rendered in black with the timeline headings in blue. When the user selects a

message from the panel, the selected message is coloured in *green*. The green colour is used here to comply with the same method used for the node.

Encoding by position: The location information of Twitter users are retrieved from their profiles and encoded on the map. A function provided by Google Maps for geo-coding is used to parse the location text to get the actual coordinates of the user location (latitude and longitude). Because some of the Twitter users may be located at the same location, we would need to specify a unique location (in latitude and longitude) for each user to avoid the overlap between the nodes. (For details on the Geo Address Resolution function, see Appendix D).
The tag clouds are also positioned in a way so that they can easily correspond to the current timeline and they also move along with the time slider.

Encoding by shape: Different shapes of objects are used in the visualization to represent different forms of information: circular nodes are used to represent Twitter users; lines and triangles - connections between nodes.

3.4 ITERATIVE PROTOTYPING

The user interface of AcademiaMap-GIV is developed in several iterations by soliciting and incorporating feedback from potential users of the visualization system, members of the Social Media Lab (SML) at Dalhousie University and Dr. Anatoliy Gruzd, a domain expert in designing and evaluating information visualizations. Expert feedback is really valuable and important for designing human centered interfaces for visualization systems before evaluating it on the target users. It works as a supplementary phase for designing and evaluating information visualization interfaces (Tory & Moeller, 2005). In different

phases of this design process, visualization interfaces were modified and also new interfaces were introduced and evaluated in accordance with the collected user and expert feedback.

In the iterative design process, rapid prototyping is a really useful and workable method to get feedback from potential users at different phases of the development of an application. There are two main types of software prototyping methods, often adopted by developers (Stasko, 1997): formative and summative. A formative approach is dependent on testing the usability of the system on the potential users at different stages of the development. All the recorded feedback from the users are considered and incorporated into the next cycle of the development. The objective of this prototyping method is to maximize and ensure the usefulness of the User Interface (UI) design by involving users at different cycles of the development. On the other hand, a summative design process deals with only one usability test of the system after the end of its iterative design. A summative design process is easy to deal with and less time consuming but in the long run it might appear too expensive and actually be too time consuming if the later conducted usability test results in the need for critical changes to the core UI design.

Considering major trade-offs of both prototype design processes, a hybrid of formative and summative approach was adopted for developing the AcademiaMap-GIV visualization UI. A formative process will allow us to incorporate the necessary changes required as the iteration of prototype development goes on, by using an informal user study on the potential users. The intention was to gradually change the design and features as the system evolves over time with feedback from the informal user study. It is

quite similar to evolutionary prototyping, in which an initial prototype is evaluated as it evolves over time until it satisfies the user's need to appear as a final product (Stasko, 1997). A formal exploratory user study (discussed in the next chapter) is also conducted on the target users at the final iteration, which is considered as summative approach in relation to the current prototype. Major usability decisions regarding the interface design are discussed below.

Decision #1: Two time sliders VS. one time slider for filtering out data

Filtering out data with respect to time is one of the most important features of AcademiaMap-GIV because it enables users to visualize the data for a specific time period. Initially, only one time slider was available (shown in Figure 3.16). However, during the earlier informal user study, users were confused by the possible interactions with the slider. Specifically, the confusion was invoked during the 'playing' of the animation, which shows the activity of Twitter users over time by automatically moving the left slider point by a single day. When the animation starts, the left thumb (point) of the slider moves towards the right thumb but only shows tweets from a single day indicated by the left thumb. From the user's point of view, it should show messages between the ranges specified by both thumbs. A possible solution could be providing users with the third (movable) thumb in the same slider, but three points in one slider could make users more confused. Finally, the decision was made to provide the second, "dedicated" slider for filtering data on the basis of daily activities (see Figure 3.17).

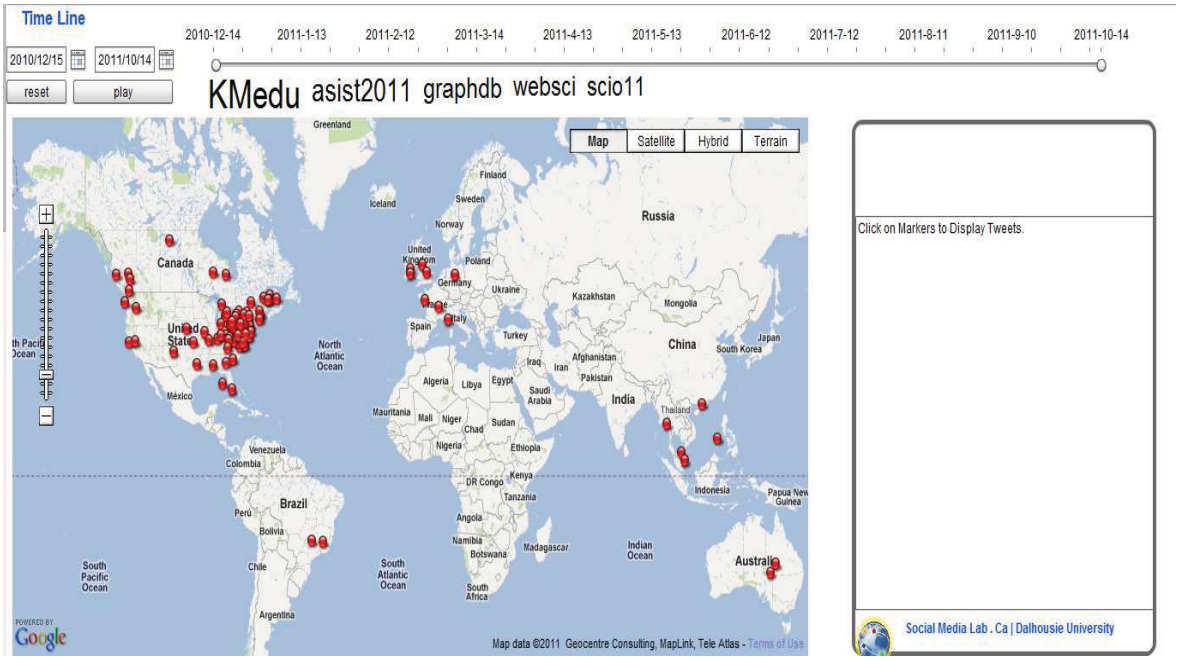


Figure 3.16: In the earliest version, only single time slider was positioned to reflect two types of timelines

Decision #2: Tag cloud arrangement with the corresponding sliders

Previously, the tag clouds were not positioned closely to the corresponding sliders (see Figure 3.17), unlike shown in Figure 3.18. The users found this confusing. In the next iteration of the interface, both tag clouds were positioned just underneath their corresponding sliders (see Figure 3.18) to eliminate this confusion.

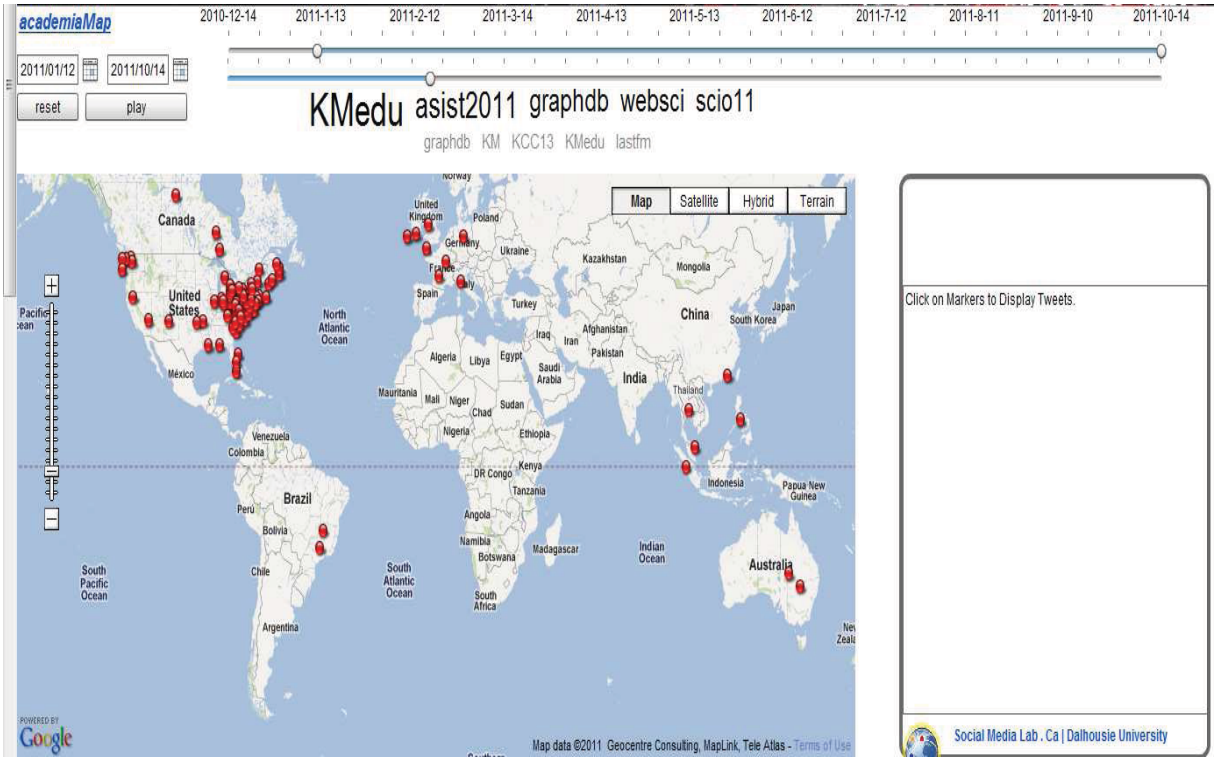


Figure 3.17: In the earlier version, tag-clouds were not positioned with the corresponding time sliders closely

Decision #3: Highlighting active sliders

In an early iteration shown in Figure 3.17, it was not clear which slider was currently active. That is unless they noticed the colour of the tag clouds (which when active are black, and when inactive are grey) corresponding to the sliders, or if they noticed the date fields in the calendar display to understand the currently active time line. However, it is time consuming and distracting for users to find this out, which could mislead their understanding of the data to some extent. To overcome this situation, the currently active slider is highlighted in yellow (shown in Figure 3.18) to help users disambiguate the active slider from the other one.

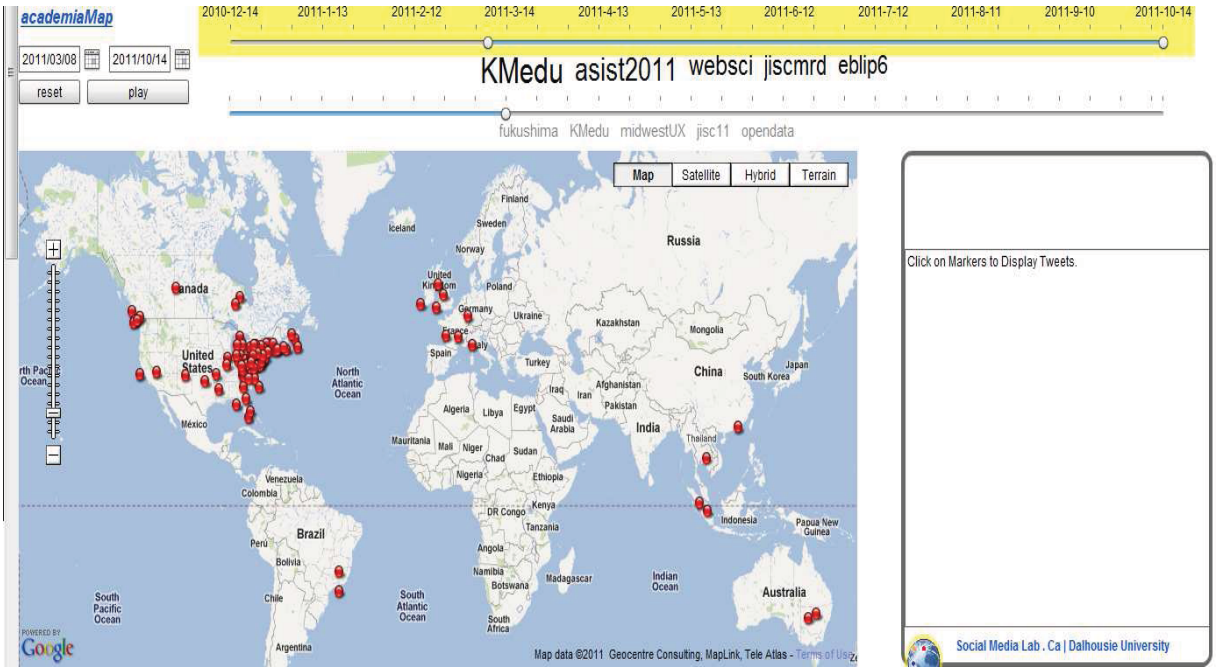


Figure 3.18: The active time slider is being highlighted in yellow colour to help users understand the current timeline

Decision #4: Loading the whole dataset at once VS. Data segmentation

In one of the earlier iterations of the interface, the whole dataset was loaded at once (from the oldest tweets to the most recent tweets). However, because of the large number of data points, the system responses to users' requests were relatively slow. To rectify this, only data from one month is loaded at a time. Users have the flexibility to load data from a different time period using the Pervious and Next buttons provided at the two ends of the time slider (as shown in Figure 3.21).

Decision #5: Font coloring, shading, and shadowing in the Tag Cloud

In the prior iterations, the tag clouds positioned underneath both sliders were without any font shading or effects as shown in Figure 3.18. However, it was not clear to the users

that you can actually click on tags in the tag clouds. To make the interface more intuitive, shadowing functions were used on the fonts (shown in Figure 3.19). Moreover, the HTML anchor tag was added to the font to make those terms appear as hyperlinks whenever users mouse-over a tag. In a later version, shown in Figure 3.1, the "#" sign is appended in front of each hash tag in the tag clouds to follow Twitter convention.

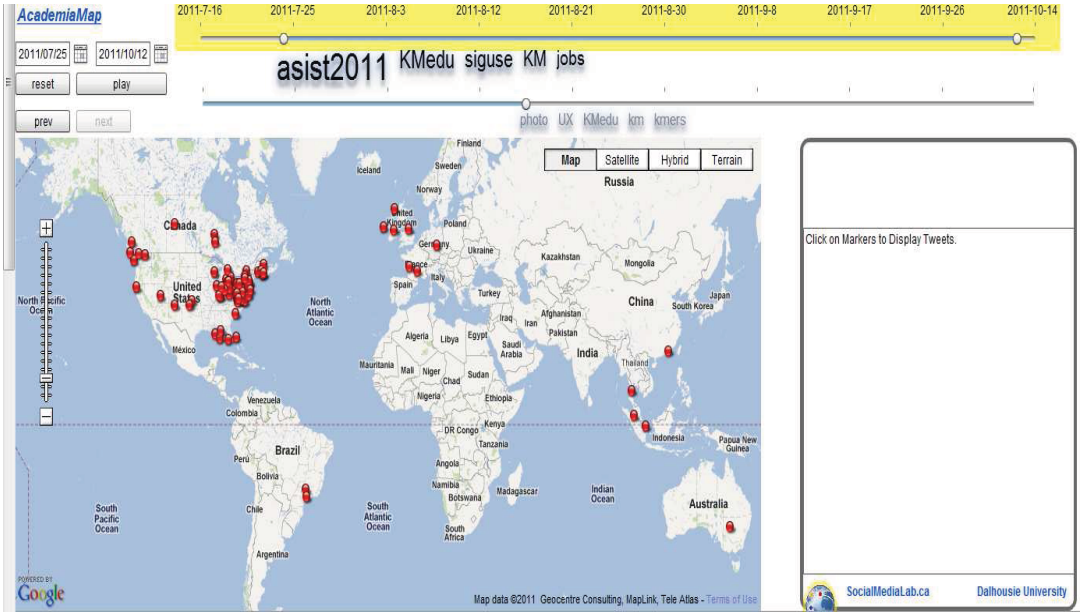


Figure 3.19: Shadowing effect is used to make the tag clouds more attractive and intuitive to the users

Decision #6: Tag cloud positioning within the slider panel

The tag cloud moves with the slider point (thumb) along the slider. However, if it is allowed to move with the slider point without any interruption, then it can exceed the boundary of the interface, as shown in Figure 3.20. In this case, users found it difficult to use the tag cloud with the default scroll bar. To address this problem, the movement of the tag cloud is restricted so it is not allowed to move with the slider point beyond the

boundary of the component which it is attached to (shown in Figure 3.21, in the Day/Day slider).

Figure 3.20: Tag-cloud (the lower one) exceeds the boundary of the component in which it is positioned

Decision #7: Grouping and labelling interface components

In an earlier iteration, the components of the visualization did not have any individual panels (shown in Figure 3.20) which made it difficult for users to distinguish between different tasks and different areas. To make the interface more intuitive, the related components were grouped, labelled and organized with the individual panels (shown in Figure 3.21).

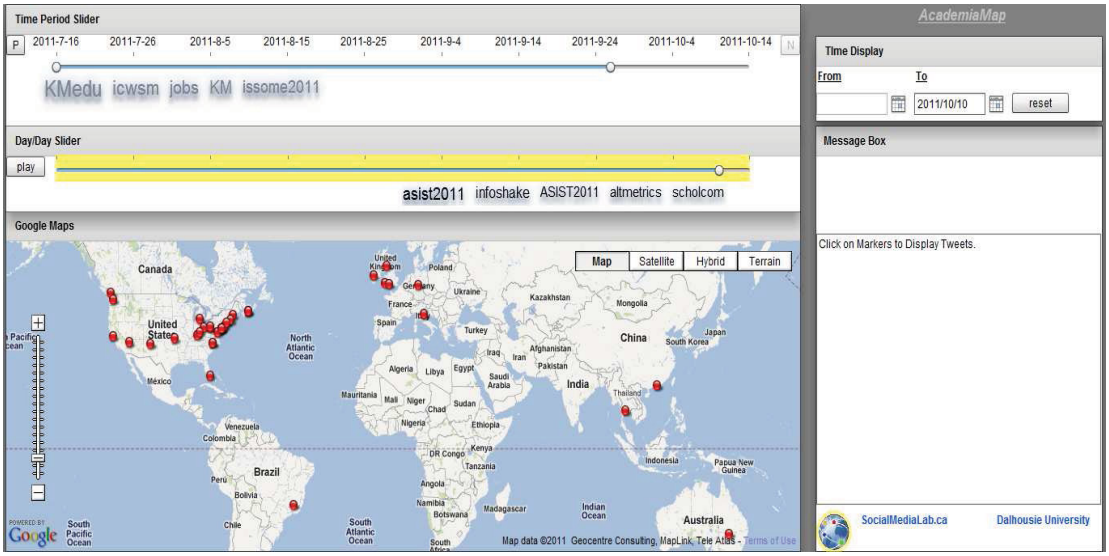


Figure 3.21: Flex interface panels are used to re-group different components of the interface

Decision #8: Encoding of message frequency

In an earlier version, nodes on the map were visualized without encoding the frequency of messages (as seen in Figure 3.21). Based on the initial user feedback, the interface was modified to represent the frequency of messages by changing the size of each node (see Figure 3.1). The encoding functions are described in Section 3.3.

Decision #9: Loading users' profile photos

In an earlier iteration, profile photos of Twitter users were preloaded before the start of AcademiaMap-GIV main interface (as shown in Figure 3.22). However, users found it time consuming for all photos to load (even for about 100 Twitter users). This approach would become even more severe if a larger number of Twitter users are added to the dataset in the future. To address this issue, it was decided to load profile images on demand, only when a node is selected by a user.

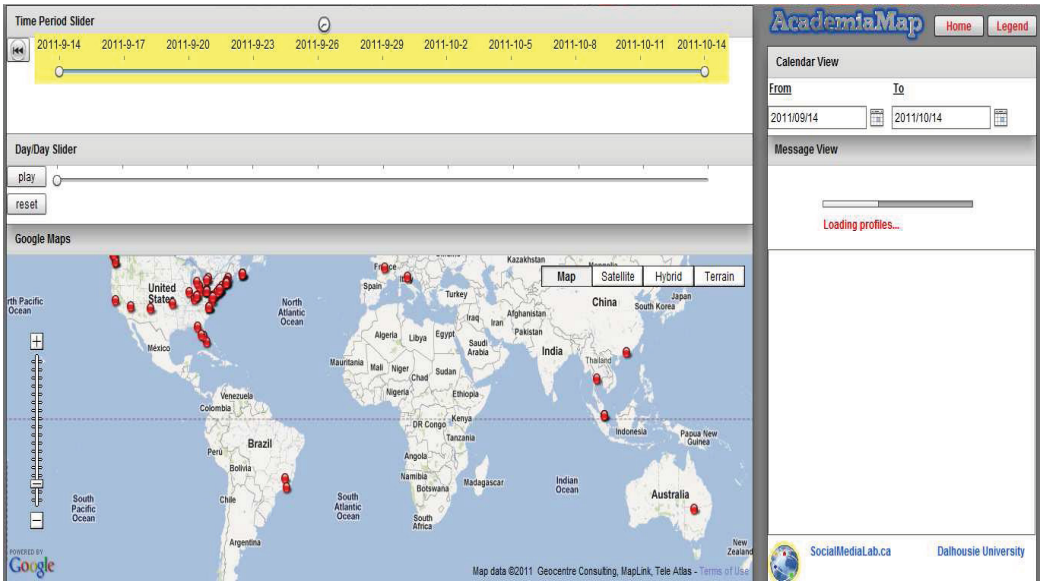


Figure 3.22: Date loading in the earlier iteration of the prototype

Decision #10: Limiting Map Zooming Options

Originally, users were allowed to zoom out from the map up to the highest level of the zoom, allowed by the map. However, at the highest level of the zoom, multiple world maps were shown to the users (see Figure 3.23), which is an artifact of Google API. This feature often distracts users in performing their tasks by showing multiples maps with corresponding nodes overlaid on them. To overcome this interface deficiency, users are now only allowed to perform zoom-out option up to a certain level (zoom level=2).

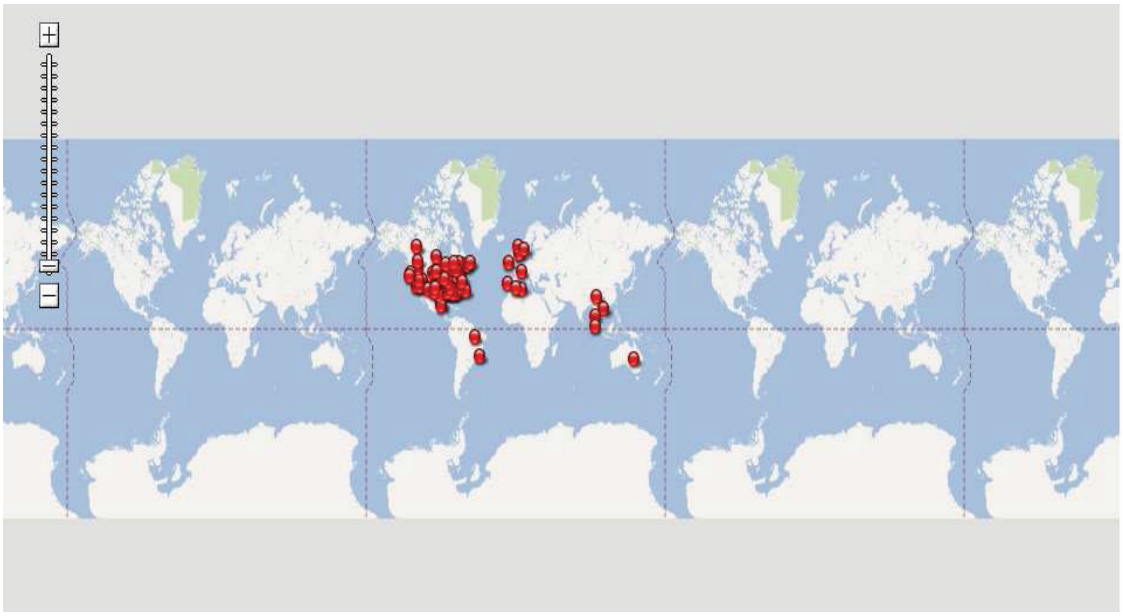


Figure 3.23: Default zoom option shows multiple maps at the lowest zoom level

The next chapter describes our exploratory user study conducted to see how the current version of AcademiaMap-GIV can be improved.

CHAPTER 4: EXPLORATORY USER STUDY

4.1 STUDY DESIGN

The exploratory user study was conducted to evaluate the current prototype of AcademiaMap-GIV and to help us comprehend the complexity of the system when using it for the first time, as well as to explore the design space in order to discover future improvements. Before conducting the study, the Research Ethics approval (Ethics number: 2011-2498) was obtained from the Social Sciences and Humanities Human Research Ethics Board of Dalhousie University.

Study Participants: To evaluate AcademiaMap-GIV, we invited potential users to test the system: Twitter users who follow the Twitter account for the 2011 ASIS&T (American Society for Information Science & Technology) annual conference. Out of the 111 Twitter users who followed the 2011 ASIS&T Twitter account, 74 individuals had public email addresses and were invited via a direct email to participate in the study. Additionally, the study participants were recruited via a public Twitter message. No compensation was offered to the participants for their participation in this study.

Study Description: The study was designed as a 30-minute, online usability testing session. During the session, participants were asked to use the system (after going through a short online video tutorial) and then gave us feedback by completing a short online survey. The survey questions can be found in Appendix E. The survey was created using a survey system called Limesurvey.

The study was conducted remotely on the participant's computer and during the time of their choice. The session was structured as follows. First, each participant was asked to view a 3-minute video tutorial on how to use the system. The goal of this tutorial was to get users acquainted with the system and its functionalities. It was useful for the users to understand the available features of the system and to experiment with them for the first time. After finishing the introductory tutorial, the participants were asked to use the system for 15 minutes. Due its novel nature, there could be many different ways to interact with AcademiaMap-GIV. Therefore, in this exploratory study, we did not want to restrict user's interaction with the interface by assigning a specific task(s) to potential users. Instead, we wanted potential users tried the system in the context of their broader activities.

Finally, after trying the system, each participant was asked to fill out a brief online questionnaire to gage their experience with AcademiaMap-GIV and solicit future improvements. The survey contained several questions about users' experience with the system in general and with the main features in particular. The total of 24 individuals completed the survey. The survey responses were used to measure the users' attitude towards the system and to determine features of the system that users like or dislike. In addition, it helped us to discover possible improvements to be implemented in the next iteration of the interface design. Below are the summary of the study results.

4.2 STUDY RESULTS

The following results are based on the ratings provided by the study respondents regarding different features of the AcademiaMap-GIV interface.

Learnability of AcademiaMap-GIV: Out of 24 respondents, only 8% (2) found that it was difficult to learn how to use the system (see Figure 4.1). (No one found it "very difficult" to learn.) "very. Whereas most respondents, 71% (17), rated the difficulty level of learning between "easy" and "very easy"; as one respondent stated: "*The interface is very visual. After clicking on a person, the visual change is also very relevant and directly connected to what idea it is conveying.*"

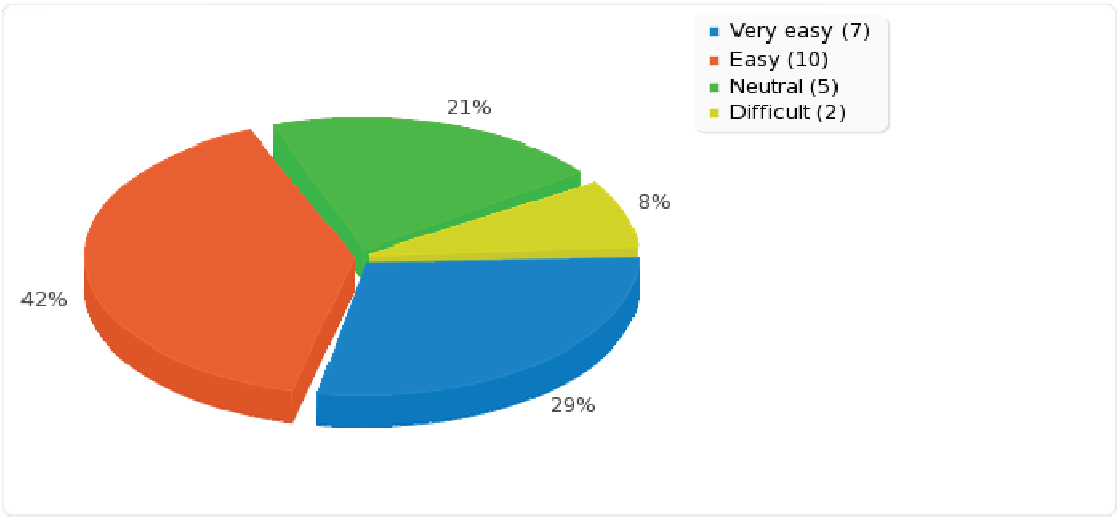


Figure 4.1: Most of the respondents found learning AcademiaMap-GIV is easy

One of the main reasons why most of the participants found it "easy" or "very easy" to learn how to use the system is likely because of their familiarity with some other similar visualizations. This is supported by the following quotes left by the study participants:

- *"Very intuitive. Resembles similar types of mashups already in existence."*
- *"Familiarity with most of the tools and widgets makes learning how to use this a fairly painless task."*

Also another user noted that the 3-minute video demo was helpful in learning the main features and functionalities of the system:

- *"...your video demo is perfect - not too long and not too short and covers everything you needed to know to use the tool."*

Next the participants were asked to evaluate the following five features of AcademiaMap-GIV according to their usefulness on the scale from 1 to 5: 1 – *Least Useful*; 2 – *Not Useful*, 3 – *Neutral*, 4 – *Useful*, 5 – *Most Useful*.

1) **Date/Time Range Filter:** None of the respondents found this feature "least useful", and only 1 respondent found it "not useful" (see Figure 4.2). Whereas most of the respondents, 87% (21) rated this feature between "useful" and "most useful". Moreover, one-third (7) among them found this feature "most useful"; as one study respondent noted: "*I found the ability to filter the tweets based on date/time range to be the most useful feature. The reason I selected this feature over the others is that I can foresee a specific research use case for this functionality.*"

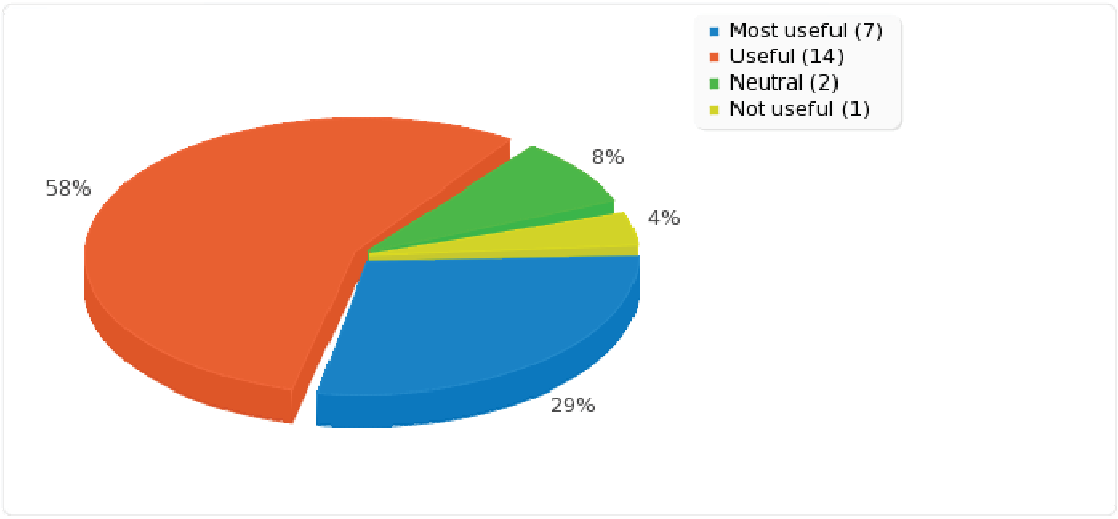


Figure 4.2: Most of the respondents found date/time range filter useful

2) **Day-by-day Animation:** Only 8% (2) of total respondents (24) found this feature "least useful", (none of them selected "not useful"); whereas more than half 58% (14) of the total respondents rated this feature between "useful" and "most useful" (Figure 4.3). Interestingly, comparing to any other feature, the day-by-day animation received the most "neutral" responses (8 out of 24).

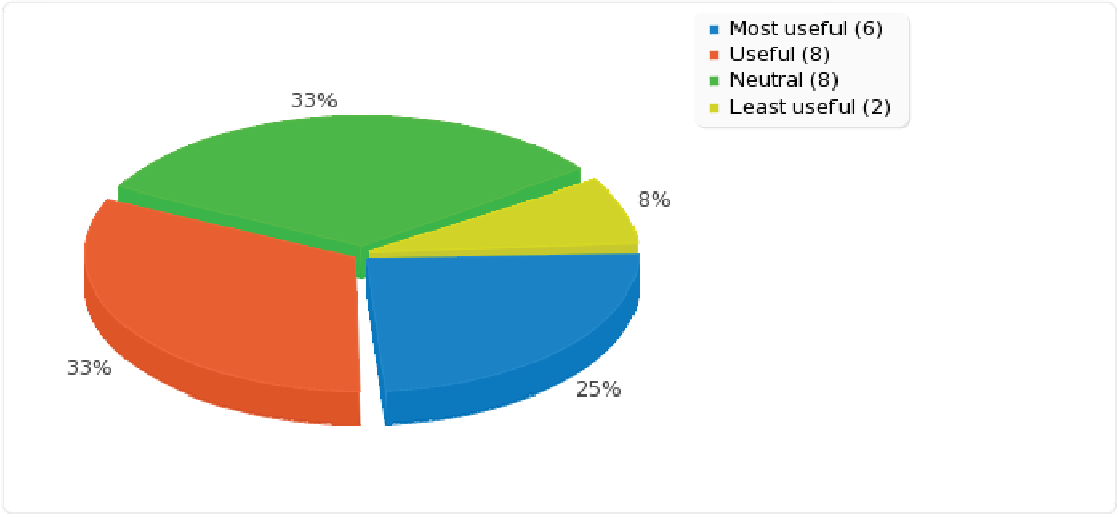


Figure 4.3: Most of the respondents found day-by-day animation useful

3) **Connections between Users:** Once again, only 1 respondent found this feature "least useful", and none of the respondents found it "not useful". Whereas most respondents, 88% (21), rated this feature between "useful" and "most useful" (see Figure 4.4). For example, one respondent found this feature especially useful because it helped him to find who is connected to whom in the communication network on Twitter. And another respondent found this feature useful because "*in a matter of seconds, I was visually able to see the correspondence made by scholars from different parts of the world.*"

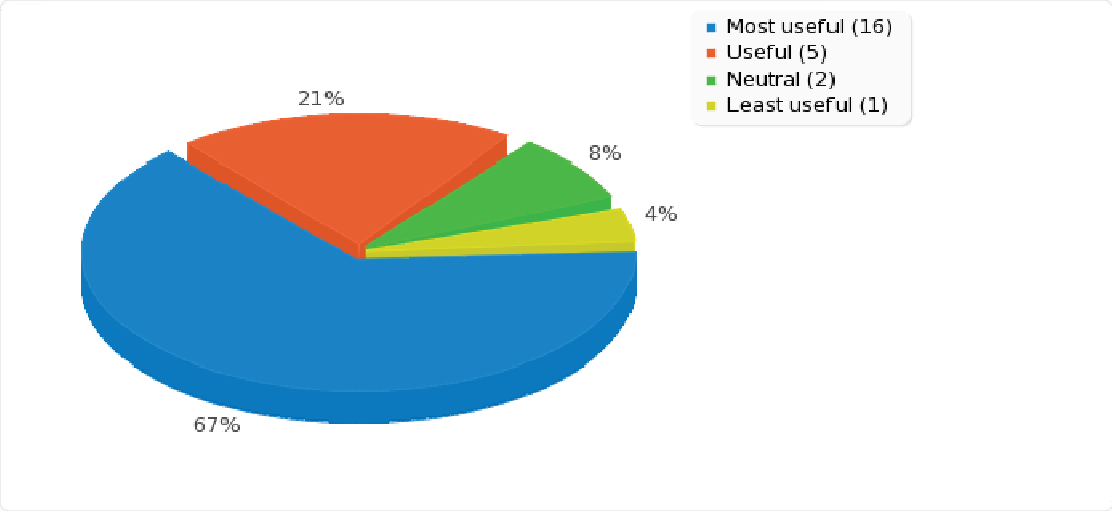


Figure 4.4: Most of the respondents found connections between users most useful

4) **Popular Topics (hash-tags):** None of the respondents found this feature "least useful" and only 1 respondent found it "not useful". Whereas almost all of them, 91% (22) of the total respondents rated this feature between "useful" and "most useful" (see Figure 4.5). The following comments left by the respondents explain why they found it useful: *"The hashtags: They shows what kind of topic the people are*

following or interested in.", "*playing day by day I really liked how you could see the hashtags an individual had used*"

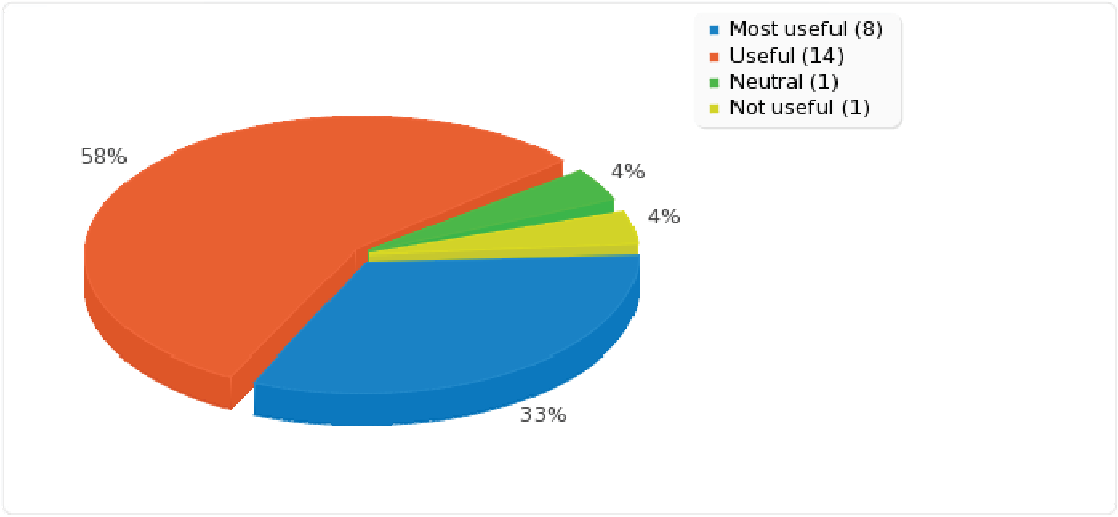


Figure 4.5: Most of the respondents found popular topics feature useful

5) **The Message Panel:** Finally, almost all of the respondents (except 1), 96% (23) rated this feature between "useful" and "most useful" (see Figure 4.6). It is not surprising that none of the study respondents found this feature "least useful" or "not useful" since people usually prefer to get the context of the data that they are exploring. In this case, the context is represented by the messages posted by a user (or a group of users) and displayed in the message panel. For example, one study respondent commented regarding the usefulness of this feature stating that "*Seeing a link on the map between users and then seeing their tweets on the right hand column was most useful because I could see the conversation.*"

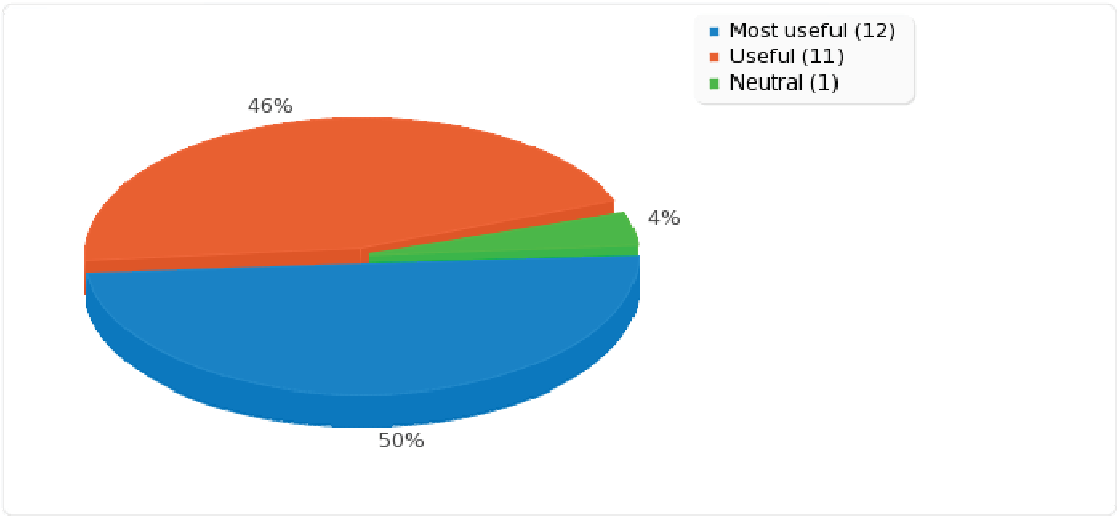


Figure 4.6: Almost all of the respondents found the message panel feature useful or most useful

In sum, based on the number of "useful" and "most useful" responses, the *Message Panel* and the *Popular Topics* are the two most useful features as determined by the study participants, followed by the *Connections between Users* and the *Date/Time Range Filter*. The *Day-by-day Animation* feature received the least number of positive responses and it also received the lowest average rating scores based on the 5-point scale (from 1 – "Least Useful" to 5 – "Most Useful"). See Figure 4.7 below. Based on the examination of the comments left by the respondents, many found the *Day-by-day Animation* interesting, but not as useful; as one participant stated "*Day-by-day animation, it's cool not sure it's useful.*" This observation is in line with the previous research by Robertson et al. (2008) who also found that the trend-type visualizations are good for presentation but not very useful for the analysis of data.

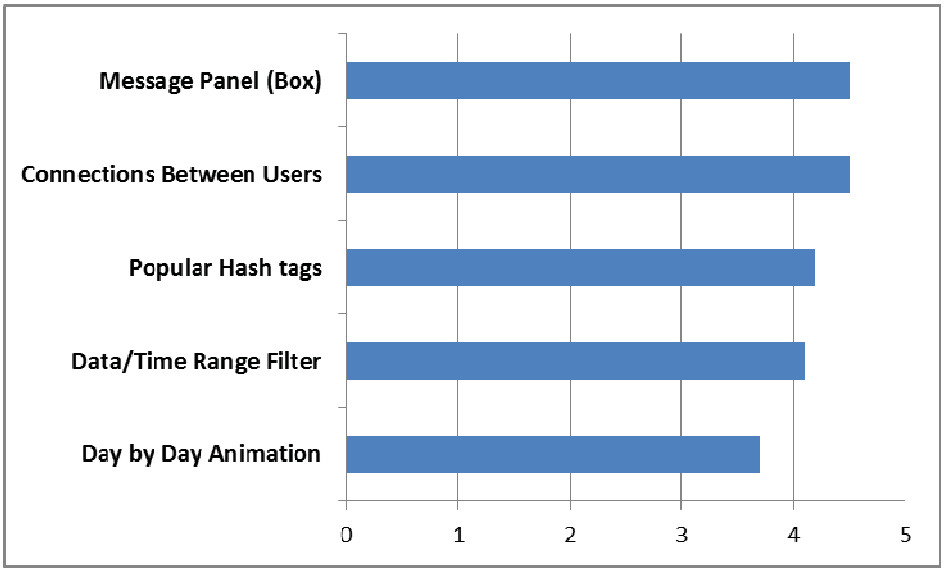


Figure 4.7: The average rating scores on the 5-point scale (from 1 – "Least Useful" to 5 – "Most Useful") based on the 24 responses (X axis) for the five main features (Y axis)

The survey was also helpful in identifying opportunities for improvement. For example, 5 respondents found it difficult to understand the difference between the two time sliders in regards to their functionality, leaving comments such as *"Is it necessary to have both sliders and the calendars to do this?"* And at least 2 users stated that it was problematic in their view to have three different places to select a date/time period. Other difficulties with the interface included trying to select a node in a dense area with many nodes in a close proximity to each other. Another feature that was mentioned in this category was the *triangle* shape that if clicked, displays all of the tweets with self-references. For instance, one respondent noted that "*Clicking on the triangle can be a little annoying because it is very close to the node symbol",* and another person even questioned its usefulness in general.

In conclusion, the survey results helped us to identify the most and least useful features of the interface as well as to determine how AcademiaMap-GIV can be improved in the future. In short, most of the respondents of the study found AcademiaMap-GIV easy to use, and also found most of the available features useful. However, as indicated above, a few respondents also found some difficulties with the interface. The next chapter will outline our future work to address the interface issues identified in the study.

CHAPTER 5: CONCLUSION AND FUTURE WORK

5.1 SUMMARY

Geo-based Information Visualizations (GIVs) allow people to analyze data points based on their related geographic locations. This approach is usually adopted where a large-scale geo-referenced dataset is present, and users are trying to find a way to examine the non-obvious relationship and hidden patterns within the dataset. GIVs are widely used in different fields such as: investigative analysis, social network analysis, health situation awareness, business analysis, traffic management, tourism, environmental science, weather forecast, news updates, geo-information retrieval, and so on. One of the emerging trends in GIV is to visualize social media data to show how information flows between users of popular social networking sites such as Facebook, Twitter, and Flickr who are also located in different cities and countries. Due to its public nature and the large number of users (who collectively generate the enormous amount of data), most of the visualizations in this area rely on conversational data from Twitter (Twitter.com), a popular microblogging service for exchanging short messages.

Based on the existing work, this thesis proposed a web-based interactive GIV system, AcademiaMap-GIV, to visualize online conversational data from social media (using Twitter as a case study). To ensure that AcademiaMap-GIV has an effective visual interface, a high level view of the data is provided at first by providing necessary visualization cues. Afterwards, users are allowed to drill down (through zooming and filtering) further on the data to focus on a specific segment of it. A rapid prototyping approach was adopted to develop AcademiaMap-GIV over multiple iterations. In each iteration, an informal user study was conducted with a group of potential users.

A formal exploratory user study was also conducted on the target users to improve the design of the interface and also to understand if the visualization incorporates necessary features to satisfy users' needs. The study result was very satisfactory as most of the study respondents found AcademiaMap-GIV useful and effective in regards to visualizing scholarly conversations on Twitter. Most importantly almost all of the users gave positive feedback regarding the overall learnability of AcademiaMap-GIV. The users stated that the user interface was easy to learn, very intuitive and visual. Furthermore, they mentioned that the overall layout of the interface as well as legitimate positioning and labelling of each component within the interface made it easy to use the system.

In the following sections, some potential uses (5.2), and future directions (5.3) of the research with AcademiaMap-GIV are discussed in detail.

5.2 POTENTIAL USES OF ACADEMIAMAP-GIV

The main goal of this research was to make rapidly-changing conversations happening online among academics more visible and useful by visualizing them geographically and temporally. AcademiaMap-GIV achieves this by enabling its users to find and focus on popular topics, on tweets posted by a specified user(s) or only on tweets from a particular time period. Based on the exploratory user study, the three major uses of AcademiaMap-GIV were identified. First, the system may be useful to follow real-time or recorded conversations happening around various scholarly events such as conferences. Second, the system may also be useful as an information gathering tool to locate and follow research news. In recognition of these two possible uses of AcademiaMap-GIV, the

Social Media Lab at Dalhousie University, for example, has already started using the system to identify popular topics discussed by scholars in the ASIS&T community. The lab is now posting their monthly findings with regard to this at http://AcademiaMap.com/blog.

Finally, the system was found to be especially useful in identifying connections among scholars in the Twitterverse. By examining who tends to retweet and mention whom, it makes it easy to identify scholars with similar research interests. This feature can also help AcademiaMap-GIV users to find other scholars with similar research interests and start following them on Twitter. And since the system visualizes scholars from all over the world, AcademiaMap-GIV users can benefit by expanding their research connections internationally.

5.3 FUTURE WORK

Besides getting positive feedback from the users, some limitations of AcademiaMap-GIV have been discovered through the formal user study. And more importantly, many respondents also provided some suggestions regarding the possible future improvements to the interface. In this section, some of the future directions of the research are discussed.

Modifying the existing interface: First, a few respondents (3) commented about the calendar view as the least useful feature in the interface because it has the same functionality as the time sliders do. Therefore, in the next iteration of AcademiaMap-GIV, the calendar view will be removed from the interface. Another feature, that shows self-referencing in tweets using triangles, was also commented by two respondents as not

very useful and will be removed in the next design iteration. Third, it was determined that showing the two time-related sliders at the same time on the screen confused some of the users during the usability study. For example, some thought that the lower slider (day/day slider) should represent hours of the day as the other one shows the timeline in days. One approach to address this misconception would be by separating the two sliders and putting them in different tabs as shown in Figure 5.1, so that users can only see one slider at the time. As a side benefit, such approach would also provide more real estate on the screen for the map visualization.

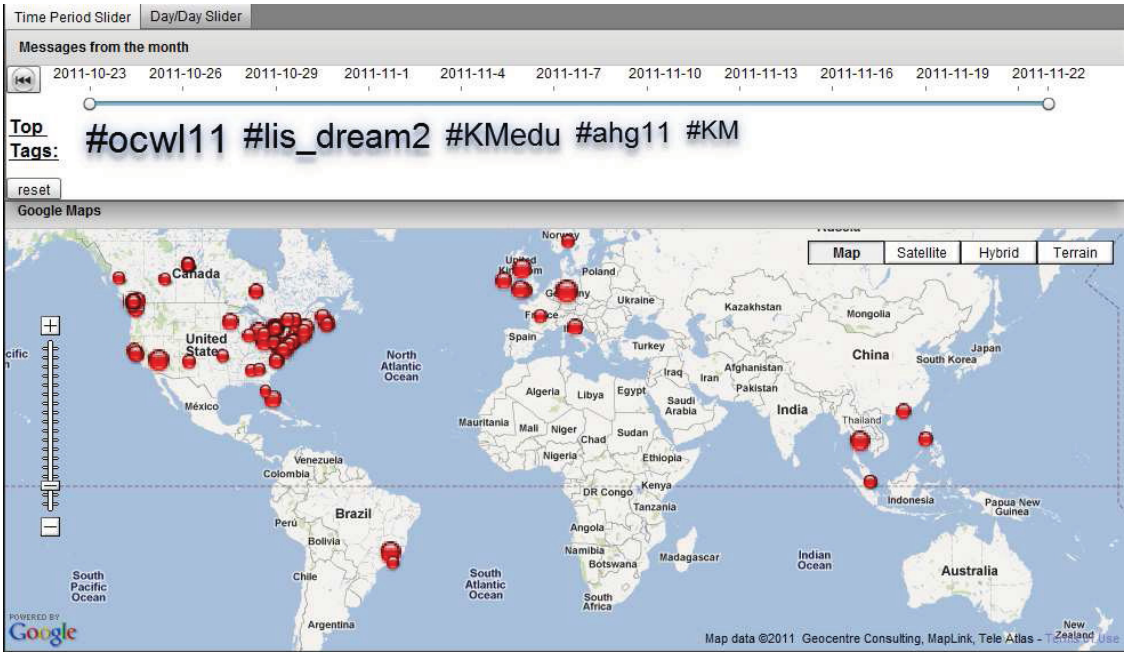


Figure 5.1: The layout of having time sliders in different tabs provides more space to the map as well as may alleviate some confusion regarding the active timeline

Incorporating new features:

1) *New encoding approach to address scalability*: To visualize a massive, dynamic dataset, scalability is the key issue. In the future, we plan to add more Twitter users to follow by AcademiaMap-GIV. However, additions of more Twitter users can make

the visualization more cluttered because more nodes are going to be plotted on the map to represent them. Furthermore, it will take more time to load and display new data points. To address this challenge, the following steps will be considered:

A. Encoding Twitter users from a specific region to one bigger node (community node), and showing that node as a representative of them at higher zoom levels; similarly to the **Mappa** (Mappa, n.d.) application in Figure 5.2. This visualization focuses only on outdoor activities within New Zealand to visualize popular spots as icons on Google Maps. Mappa incorporates marker (icon) clustering as a visual encoding feature such that clicking on an encoded bigger icon at higher zoom levels visualizes more icons, or more encoded icons, on the map.

B. Rendering nodes on demand with respect to the current zoom level. As the AcademiaMap-GIV user selects a certain zoom level, all nodes encoded by the community node are rendered.

Figure 5.2: Mappa application visualizes popular spots using marker clustering and colour encoding (Mappa, n.d.)

2) *Overview of connections between people*: In the current version of AcademiaMap-GIV, users need to select each and every node to see the entire network of Twitter users. However, some of the study respondents are interested to see an overview of the entire network for all Twitter users at the same time. One possible approach to incorporating this feature is to use the saturation of colour of each node to encode the information regarding the frequency of links to other nodes. For instance, nodes in darker colour will represent those Twitter users, who have mentioned more people in their tweets than others. This approach will provide a better overview by visualizing the most "communicative" Twitter users based on their posting patterns.
3) *Selection of multiple nodes*: At present, users can select only one node on the map but in the future, users will be allowed to select multiple nodes to see and compare their activities all together on the map.
4) *Statistical analysis with graphs*: In the future, the activity of each Twitter user will be presented in different kinds of graphs using statistical analysis of the data at hand. It will help users to review a summary of each Twitter user's activity over time.
5) *Incorporating data from other sources:* In the future, other social media data as well as data from other sources relevant to scholars (e.g., bibliographic data) will be incorporated as well as information about Twitter user's academic or professional affiliation (if applicable) to expand the existing visualization.

The endeavour of visualizing scholarly online conversations has received numerous encouraging words from the study participants who commented by saying that *"It's interesting to see how the information changes and moves. Good job", "This is a useful*

app.","good tool", "Excellent work! I really liked the interface and the concept.", "great job! "Nice work, Keep it up", "This is a really interesting project!". So, although the thesis part of AcademiaMap-GIV is completed, the work on its next release has already begun.

REFERENCES

Audioboo / The BooMap. (n.d.). Retrieved February 1, 2011, from http://audioboo.fm/boos/map

Cartoview - Google maps example. (n.d.). Retrieved February 1, 2011, from http://www.cartologic.com/products/cartoview-1.0.0beta/samples/gm-black/index.html

Charleston SC Trip Planner. (n.d.). Retrieved March 12, 2011, from http://www.ota.cofc.edu/tripplanner/

CND Election 2011 · Social Media Lab.Ca. (2011). Retrieved November 18, 2011, from http://socialmedialab.ca/?cat=30

Data Visualization: Journalism's Voyage West | Rural West Initiative. (n.d.). Retrieved July 13, 2011, from http://www.stanford.edu/group/ruralwest/cgi-bin/drupal/visualizations/us_newspapers

Dewar, M. (2010). Visualisation of Activity in Afghanistan using the Wikileaks data on Vimeo. Retrieved April 10, 2011, from http://vimeo.com/14200191

Disaster Response and Assistance. (n.d.). Retrieved February 12, 2011, from http://tmapps.esri.com/egypt_unrest/index.html

Dork, M., Gruen, D., Williamson, C., & Carpendale, S. (2010). A visual backchannel for large-scale events. *Visualization and Computer Graphics, IEEE Transactions on*, *16*(6), 1129–1138.

Earthquake: Twitter Users Learned of Tremors Seconds Before Feeling Them - The Hollywood Reporter. (2011). Retrieved November 17, 2011, from http://www.hollywoodreporter.com/news/earthquake-Twitter-users-learned-

tremors-226481

Eccles, R., Kapler, T., Harper, R., & Wright, W. (2008). Stories in geotime. *Information Visualization*, *7*(1), pp. 3–17.

GeoVISTA CrimeViz | GeoVISTA Center. (n.d.). Retrieved March 28, 2011, from http://www.geovista.psu.edu/CrimeViz/

Heatmap for Twitter - The Word on the Tweet. (n.d.). Retrieved March 8, 2011, from http://www.uuorld.com/wordontweet/#

Jaffe, A., Naaman, M., Tassa, T., & Davis, M. (2006). Generating summaries and visualization for large collections of geo-referenced photographs. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval - MIR '06*. Santa Barbara, California, USA, doi:10.1145/1178677.1178692

Junninen H., Lauri A., Keronen P., Aalto P., Hiltunen V., Hari P. & Kulmala M. (2009). Smart-SMEAR: online data exploration and visualization tool for SMEAR stations. *Boreal Env. Res. 14: pp. 447-457.*

Kapler, T., & Wright, W. (2005). Geotime information visualization. *Information Visualization*, *4*(2), pp. 136–146.

Kim, T., Jeong, H. Y., Chew, Y. C., Bonner, M., & Stasko, J. (2009). SocialVisualization for Micro-Blogging Analysis. IEEE InfoVis 2009. Atlantic City, NJ.

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., et al. (2011). SensePlace2: GeoTwitter Analytics Support for Situational Awareness. *IEEE Conference on Visual Analytics Science and Technology, Providence, RI, IEEE.*

Mappa. (n.d.). Retrieved April 2, 2011, from http://www.mappa.co.nz/

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011). TwitInfo: Aggregating and visualizing microblogs for event exploration. *Proceedings of the 2011 annual conference on Human factors in computing systems* (pp. 227–236).

Monitoring Swine Flu using Twitter. (n.d.). Retrieved March 22, 2011, from http://compepi.cs.uiowa.edu/~alessio/Twitter-monitor-swine-flu/

Oscar Twitter Map. (n.d.). Retrieved April 3, 2011, from http://www.neoformix.com/Projects/OscarTwitterMap/

Pan, C. C., & Mitra, P. (2007). FemaRepViz: automatic extraction and geo-temporal visualization of FEMA national situation updates. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on* (pp. 11–18).

Pan, C. C., Mitra, P., & Ganguly, A. R. (2007). *Spatio-Temporal Analysis on FEMA Situation Updates with Automated Information Extraction*. The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). San Jose, CA..

Robertson, G., Fernandez, R., Fisher, D., Lee, B., & Stasko, J. (2008). Effectiveness of Animation in Trend Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1325-1332. doi:10.1109/TVCG.2008.125

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations (pp. 336-343). IEEE Comput. Soc. Press. doi:10.1109/VL.1996.545307

Stasko, J. (1997). Rapid Prototyping. Retrieved November 8, 2011, from http://www.cc.gatech.edu/classes/cs6751_97_winter/Topics/rapid-proto/

Stryker, M., Turton, I., & MacEachren, A. M. (2008). Health GeoJunction: Geovisualization of news and scientific publications to support situation awareness. In *Geospatial Visual Analytics Workshop, GIScience*.

Sullivan, B. L., Kelling, S. T., Wood, C. L., Iliff, M. J., Fink, D., Herzog, M., Moody, D., et al. (2009). Data Exploration Through Visualization Tools. *In the* Proceedings of the Fourth International Partners in Flight Conference: Tundra to Tropics (pp. 415–418).

Tory, M., & Moller, T. (2005). Evaluating Visualizations: Do Expert Reviews Work? *IEEE Computer Graphics and Applications*, *25*, 8-11. doi:10.1109/MCG.2005.102

Tracking Twitter Traffic About the 2010 Midterm Elections - Interactive Feature - NYTimes.com. (n.d.). Retrieved April 5, 2011, from http://www.nytimes.com/interactive/us/politics/2010-Twitter-candidates.html

Trendsmap. (n.d.). Retrieved May 2, 2011, from http://trendsmap.com/

Twitter Blog: #numbers. (2011). Retrieved November 17, 2011, from http://blog.Twitter.com/2011/03/numbers.html

Twitter, Facebook and YouTube's role in Arab Spring (Middle East uprisings). (2011). Retrieved November 23, 2011, from http://socialcapital.wordpress.com/2011/01/26/Twitter-facebook-and-youtubes-role-in-tunisia-uprising/

Visualizing social media | VizWorld.com. (2009). Retrieved November 23, 2011, from http://www.vizworld.com/2009/06/visualizing-social-media/

Viegas, F. B., Wattenberg, M., & Feinberg, J. (2009). Participatory Visualization with Wordle. IEEE Transactions on Visualization and Computer Graphics, 15(6), 1137-1144. doi:10.1109/TVCG.2009.171

Water Conflict Chronology Map. (n.d.). Retrieved March 31, 2011, from http://www.worldwater.org/conflict/map/

Wu, Y. J., Wang, Y., & Qian, D. (2007). A google-map-based arterial Traffic Information System. In *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE* (pp. 968–973).

Zero Geography: Map of Wikileaks US Embassy Cables. (2010). Retrieved April 10, 2011, from http://www.zerogeography.net/2010/11/map-of-wikileaks-us-embassy-cables.html

APPENDIX A. SOCIAL MEDIA TERMINOLOGY

Source: The Twitter Glossary (https://support.Twitter.com/articles/166337-the-Twitter-glossary)

Follower

A follower is another Twitter user who has followed you.

Hashtag

The # symbol is used to mark keywords or topics in a Tweet. It was created organically by Twitter users.

Mention

Mentioning another user in your Tweet by including the @ sign followed directly by their username is called a "mention". Also refers to Tweets in which your username was included.

Retweet (noun)

A Tweet by another user, forwarded to you by someone you follow. Often used to spread news or share valuable findings on Twitter.

Retweet (verb)

To retweet, retweeting, retweeted. The act of forwarding another user's Tweet to all of your followers.

RT

Abbreviated version of "retweet." Placed before the retweeted text when users manually retweet a message. See also Retweet.

Tweet (verb)

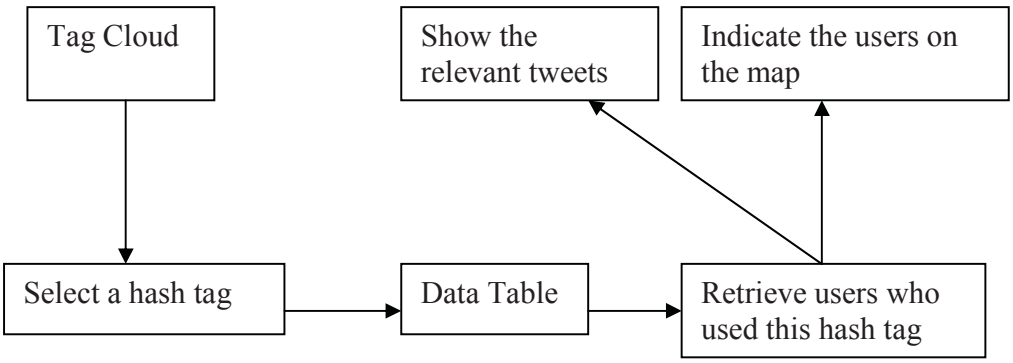
Tweet, tweeting, tweeted. The act of posting a message, often called a "Tweet", on Twitter.

Tweet (noun)

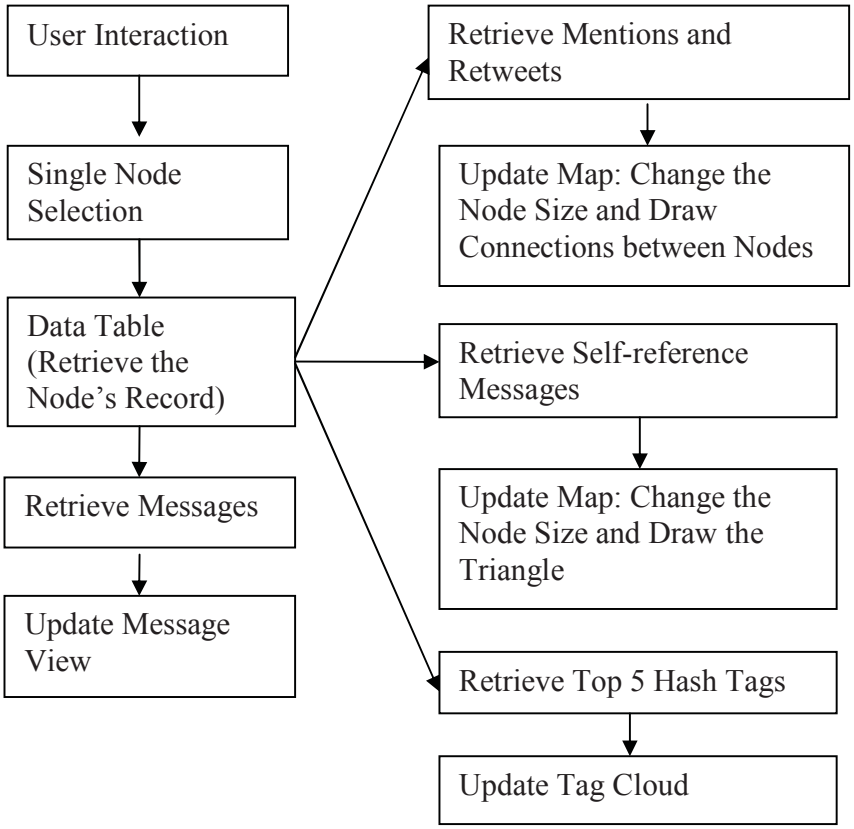
A message posted via Twitter containing 140 characters or fewer.

APPENDIX B. WORK FLOW DIAGRAMS

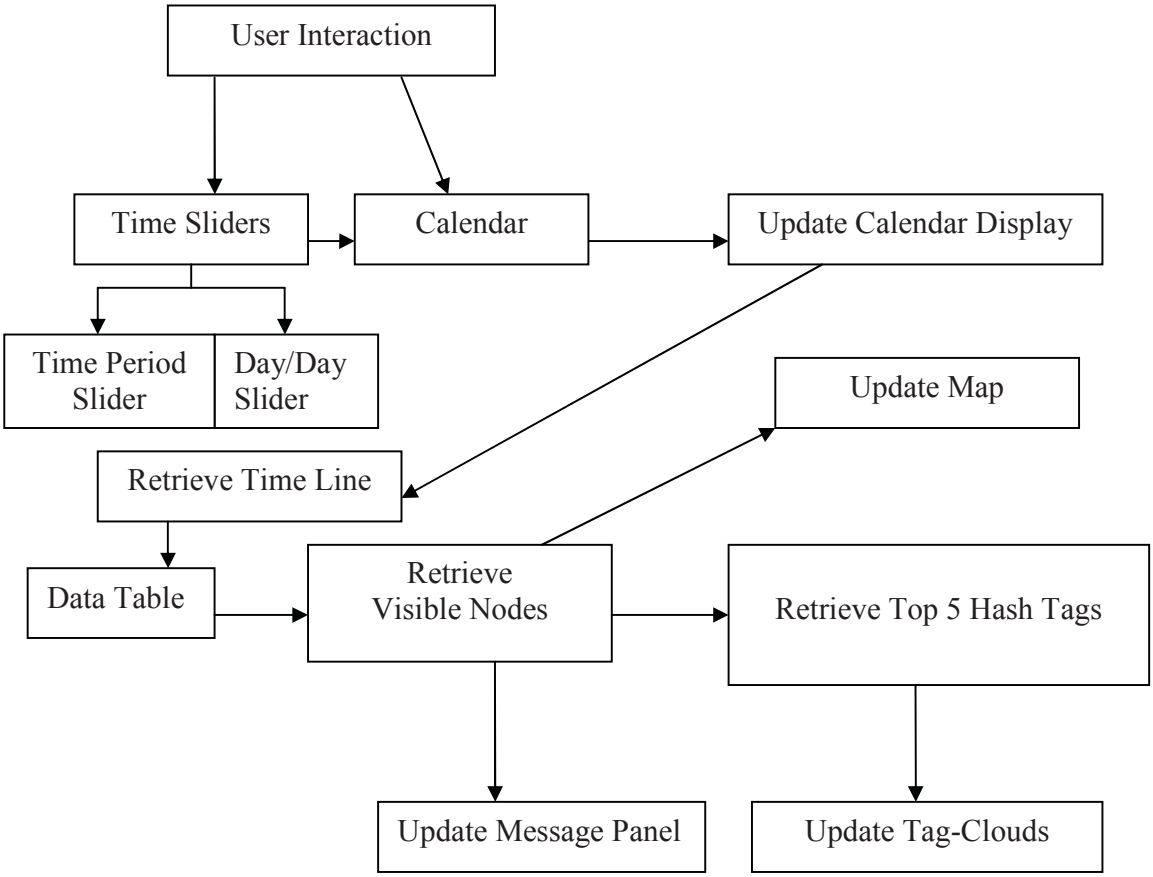
1) Work flow diagram of user-system interactions with the tag-clouds



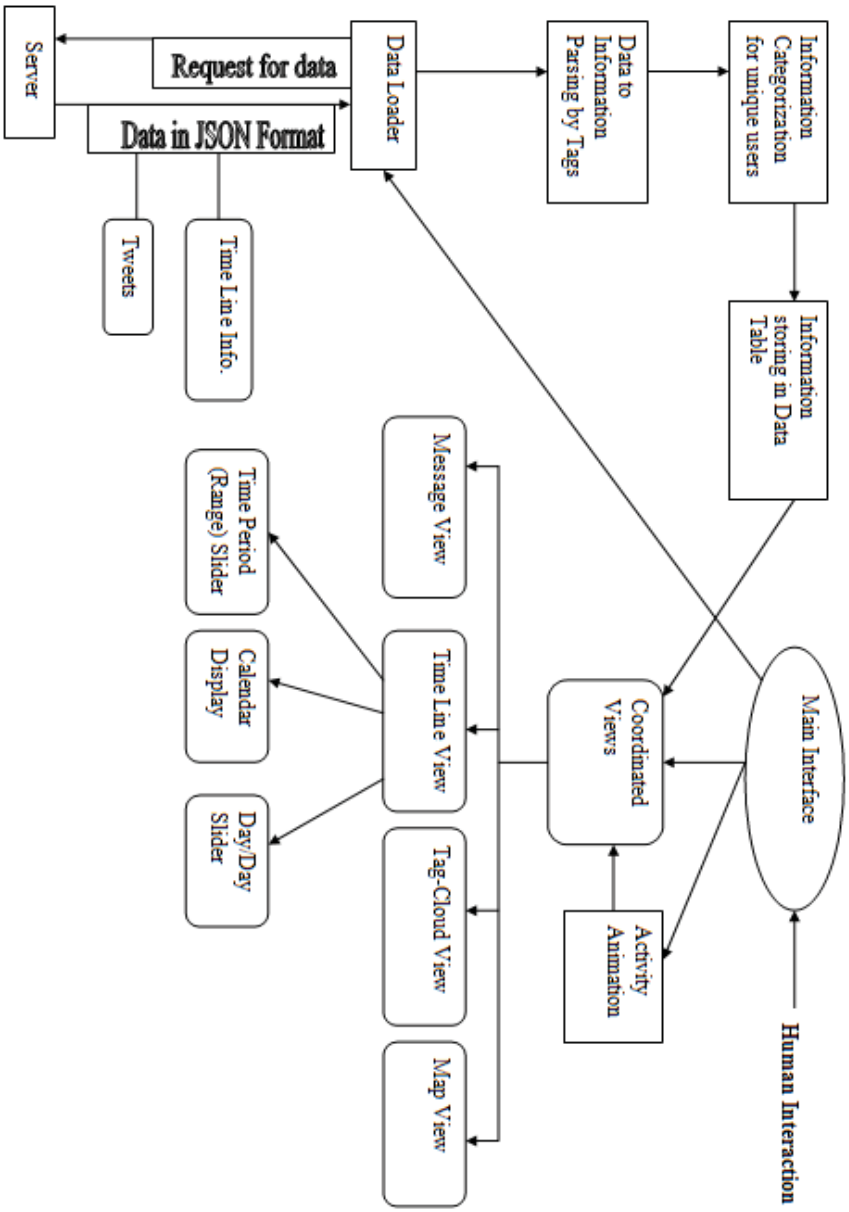
2) Work flow diagram of user-system interactions with the nodes (Twitter users)



3) Work flow diagram of user-system interactions with the timeline sliders



APPENDIX C. HIGH LEVEL SYSTEM VIEW



APPENDIX D. GEO ADDRESS RESOLUTION FUNCTION

When the data is loaded into the data structure of the application, the *NodeRenderer* function is called to render circular nodes (markers) to represent each unique user from the dataset on the map by assigning unique latitude and longitude addresses to them. The location information of each user is retrieved from the specific location field of the data table. Ensuring unique location of each node is necessary to overcome the problem of visual clutter in the visualization by those overlapped nodes (if their location points are the same on the map). The random numbers are generated from 0 to 0.01 range to slightly shift a conflicting location. This range was selected through empirical experimentations. The NodeRenderer function is described in pseudo code below:

```
NodeRenderer (username, location, Array of rendered points): returns Point with unique latitude, longitude
If Point:GeoCode (location) exists in the Array of rendered points:
Begin:
        newLat= latitude of Point + random number (range: 0~0.01)
        newLng= longitude of Point+ random number (range: 0~0.01)
        Point=Make new point of (newLat, newLng)
        If (Point exists in the Array of rendered points)
        Begin:
            NodeRenderer(username, location, Array of rendered points)
       End:
       Else
            Begin:
            Add Point in the array of rendered points
            End:
```

APPENDIX E. ONLINE SURVEY QUESTIONNAIRE

1. How would you rate the **learnability** of AcademiaMap-GIV:

Very Difficult	Difficult	Neutral	Easy	Very Easy
1	2	3	4	5

   - Why?

     _____
     _____
     _____

2. Can you tell us about actually **using** AcademiaMap-GIV?

   - What was easy

     _____
     _____
     _____

   - What was difficult

     _____
     _____
     _____

3. Please rate the **usefulness** of the following features on the scale from 1 to 5:

Least Useful	Not Useful	Neutral	Useful	Most Useful
1	2	3	4	5

   - Date/time range filter
   - Day-by-day animation
   - Connections between users
   - Popular topics (hash-tags)
   - The messages panel to the right

4. Which AcademiaMap-GIV features did you found **most** useful? And why?
5. Which AcademiaMap-GIV features did you found **least** useful? And why?
6. Can you suggest any improvements to the interface design?
7. Any other comments?
8. ☐ I consent that any of my comments in the survey can be quoted **anonymously** in any reports or publications

APPENDIX F. GEO-VISUALIZATIONS OF NON SOCIAL MEDIA DATA

In this section, different GIV applications are discussed that use data sources other than from social media.

1) Investigative analysis

CrimeViz (GeoVISTA CrimeViz, 2009), developed by Pennsylvania State University GeoVISTA Center, produces visualizations that support analysts in their sense making and investigative activities. This web-based application runs on a near real time dateset contributed by the District of Columbia Data Catalog. The visualization (see Figure F.1) provides an interactive map (using the Google Maps service) with multiple features such as the crime type filter, the animation of data over time, and a set of bar charts to show crime rates over time.

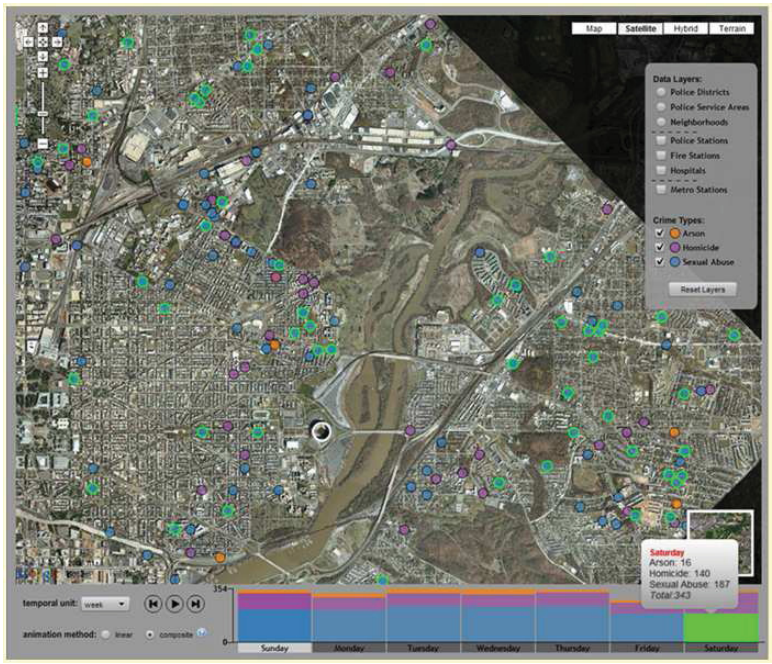


Figure F.1: CrimeViz visualizes different categories of criminal activities as nodes on the map over time (GeoVISTA CrimeViz, 2009)

Similarly, **GeoTime Information Visualization** is a 3D visualization system that is specifically designed for the crime and investigative analysis (Kapler, & Wright, 2005; Eccles, Kapler, Harper, & Wright, 2008). Though this application is developed for military personnel, it is also applicable to any kind of analytic work, for instance, business analysis tasks, in which analysts may wish to connect multiple events with respect to location and time.

2) Situation Awareness

Health GeoJunction demonstrates how visualization can be applied to make people aware of current health-related epidemics. This web-based visualization extracts information from the scientific database such as PubMed, the World Health Organization (WHO) database, and reports about outbreak incidents from the World Animal Health Organization (OIE) (Stryker, Turton, & MacEachren, 2008). Then, Health GeoJunction uses maps, timelines, and tag clouds to inform users about any outbreaks (see Figure F.2a). Each icon side represents the frequency of PubMed abstracts of each case. If a document is selected in the visualization, it relates to the original location of the document, along with the referred places in the text, by drawing arcs between them as showed in Figure F.2b.

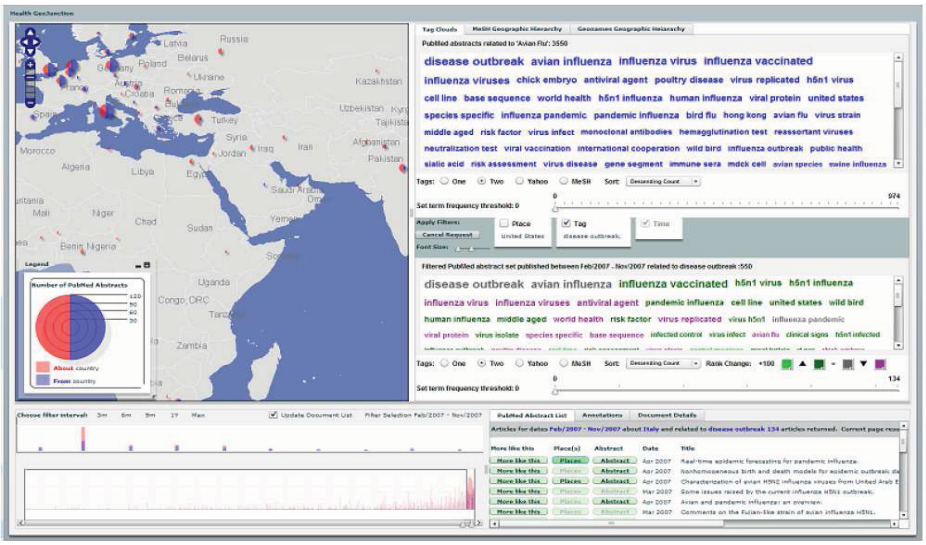


Figure F.2a: Health GeoJunction interface includes tag clouds, a map, and a timeline to visualize data in a coordinated view (Stryker, Turton, & MacEachren, 2008)

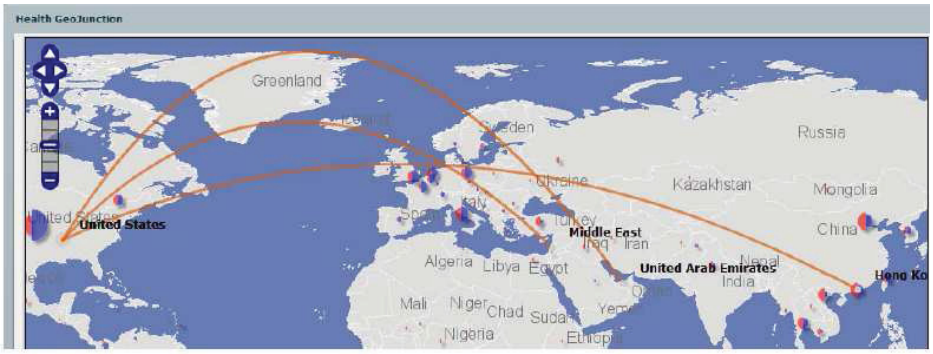


Figure F.2b: Arcs are used to relate places that are described in the abstract of a selected article to the publication site of the document (Stryker, Turton, & MacEachren, 2008)

Weather forecasting is another domain where situation awareness GIVs are commonly used. For example, **FemaRepViz** is a web-based application that visualizes weather updates on Google Earth and Google map by extracting data from the news reports using text extraction algorithms (Pan, & Mitra, 2007; Pan, Mitra, & Ganguly, 2007).

3) Visualizing Historical Data

Water Conflict Chronology Map (Water Conflict Chronology Map, n.d.) visualizes historical data based on different incidents that include conflicts for water in different regions around the world since 0 BC. This application (see Figure F.3) puts a marker for each incident on the map, and also shows the incident snippet in the right panel of the window. If any incident is selected on the panel, it points to the associated marker with the information window placed on the marker. Users can also filter the results by selecting any particular region, date range, and conflict types.

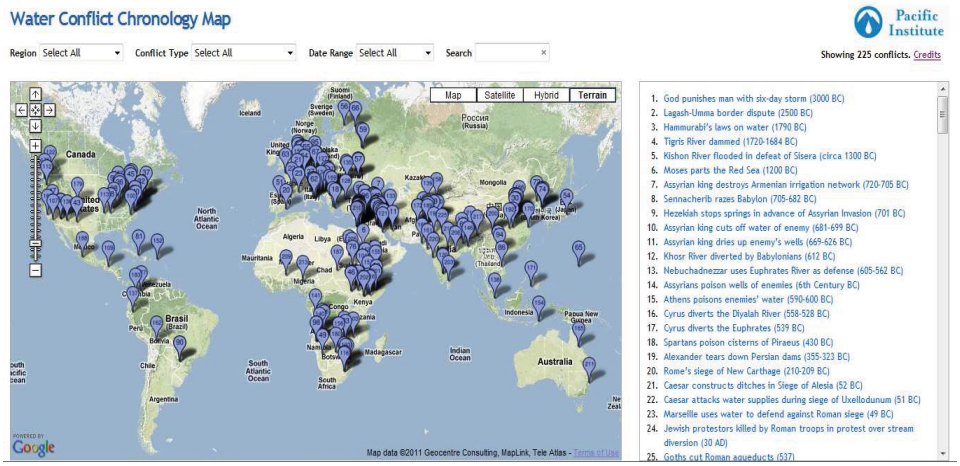


Figure F.3: Water Conflict Chronology Map: all of the conflicts for water in different regions around the globe since 0 BC (Water Conflict Chronology Map, n.d.)

Journalism's Voyage West uses another kind of historical data - the growth of newspapers across the U.S. during the period from 1690 to 2011 (Data Visualization: Journalism's Voyage West, n.d.). This application presents different newspapers as nodes on the map from where they were published during their active periods (see Figure F.4). Each node size represents the frequency of publications in a particular city. This visualization includes a time slider to show newspaper publications over time. This

newspaper application also incorporates details-on-demand, filtering features by time and also by different categories of newspapers. After clicking on a node on the map, more details are presented about the selected newspaper.

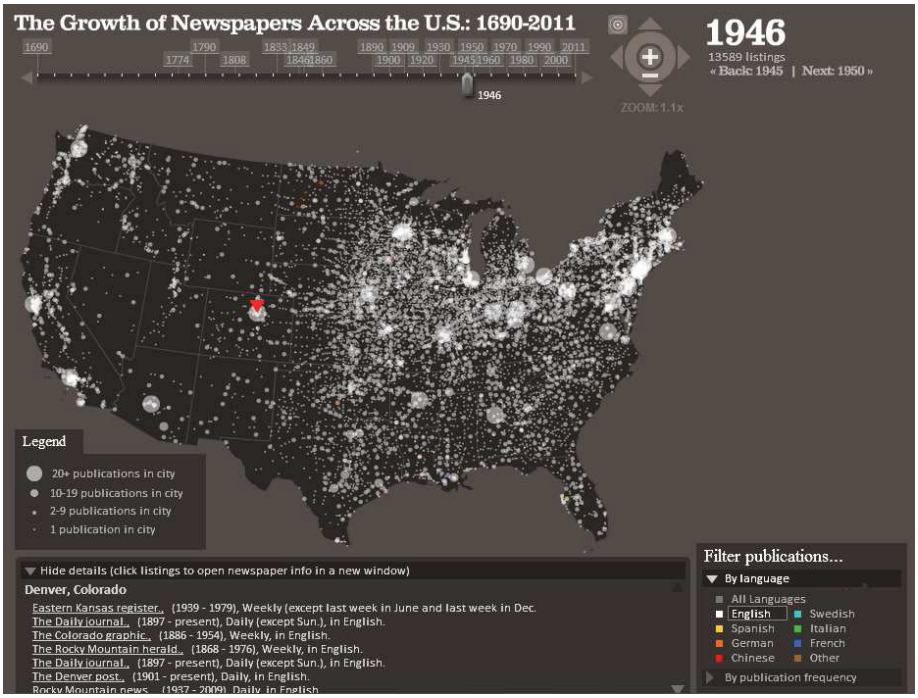


Figure F.4: Journalism's Voyage West: this application presents different newspapers as nodes on the map from where they originated (Data Visualization: Journalism's Voyage West, n.d.)

Another set of examples of historical data visualizations are based on data about and from the US embassy cables released by Wikileaks. **Map of Wikileaks US Embassy Cables** (Zero Geography: Map of Wikileaks US Embassy Cables, 2010) is one such visualization, which displays and categorizes all of the US embassy cables leaked by Wikileaks based on their geographic points of origin. Circular markers are used to indicate how many cables were leaked from a particular location (see Figure F.5).

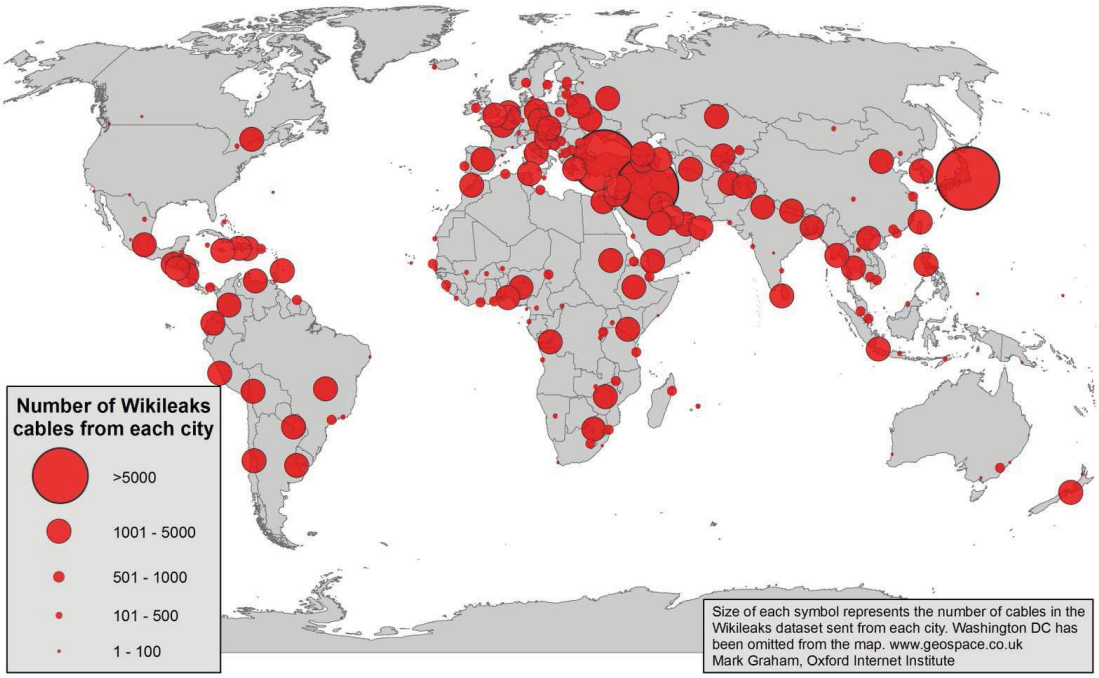


Figure F.5: In this visualization, size of each icon represents the frequency of Wikileaks cables from each city on the map (Zero Geography: Map of Wikileaks US Embassy Cables, 2010)

The next GIV**, Visualisation of Activity in Afghanistan using the Wikileaks data**, (Dewar, 2010) by Mike Dewar is an animated visualization of activities that occurred in Afghanistan from 2004 to 2009. The data was collected from Wikileaks based on the time and location where the incident occurred. A heat map is used here to show the intensity of different types of events taking place at any particular region over a one-month time window. The animation made from this visualization runs at 10 days per second and the heat map is generated for every day during the period from 2004 until 2009.

4) Tourism

In tourism, geographic visualization systems are often for trip planning and for getting acquainted with popular places. **Interactive Trip Planner** is one such GIV application that helps tourists to plan their future trips (Charleston SC Trip Planner, n.d.). In particular, this application visualizes different interesting places in the city of Charleston, South Carolina, USA on Google Maps. This web based application clusters all important places into different categories. These categories include antiques, apparel, aquariums, art galleries, entertainment and venues, bed and breakfasts, dining, and many more. Users can plan trips using this application by adding different places from these categories to their itinerary. Users can also make any point as the starting point of the intended trip based on the current itinerary list. This action connects all the itinerary items to show the entire trip path on the map. Another GIV in this category is **Mappa** (Mappa, n.d.). Mappa focuses on outdoor activities within New Zealand. Tourists who intend to go to New Zealand can explore popular places in New Zealand using a Google Map interface (see Figure F.6).

Figure F.6: Mappa application visualizes popular places for outdoor activities by category on the map using marker clustering and colour encoding (Mappa, n.d.)

5) Environmental Effects on Animal Life

Environmental issues like pollution and weather change severely affect animals' lives and habitat. Data visualization tools can greatly facilitate research in this area such as visualizing changes in birds' habitats and migration patterns. Data visualization tools accomplish this by visualizing the data on maps and also connecting different environmental incidents that may be responsible for such changes, such as excessive air pollution, and leakage of oil into the sea water (Sullivan et al., 2009). For example, the **smart-SMEAR II** (**S**tations for **M**easuring the forest **E**cosystem-**A**tmosphere **R**elationships) provides a unique platform for researchers to study cross-disciplinary environmental problems based on data from a boreal pine forest. This web-based system visualizes a holistic view of aerosol particles, atmospheric chemistry, fluxes, and tree and soil processes on Google maps that facilitates a better understanding of possible connections between different atmospheric elements with the forest itself (Junninen et al., 2009).

6) Traffic management

A real-time traffic management system is another area where geo-based visualizations are often used to display streets along with updates on the recent traffic congestion based on available traffic flow data from a traffic agency or reports from other drivers. An example of such GIV is **Google-map-based Arterial Traffic Information (GATI)**. GATI visualizes and organizes arterial networks of urban streets in the City of Bellevue, Washington (Wu, Wang, & Qian, 2007).