APPROVAL SHEET

Title of Dissertation: REAL TIME BIG DATA ANALYTICS FOR
PREDICTING TERRORIST INCIDENTS

Name of Candidate: Ibrahim Toure
Doctor of Philosophy, 2017

Dissertation and Abstract Approved: Aryya G
(Dr. Aryya Gangopadhyay)
(Professor and Chair)
(Information Systems)

Date Approved: 7/27/2017

ABSTRACT

Title of dissertation:	REAL TIME BIG DATA ANALYTICS FOR PREDICTING TERRORIST INCIDENTS
	Ibrahim Toure, Doctor of Philosophy, 2017
Dissertation directed by:	Professor Aryya Gangopadhyay Department of Information Systems

Terrorism is a complex and evolving phenomenon. In the past few decades, we have witnessed an increase in the number of terrorist incidents in the world. The security and stability of many countries is threatened by terrorist groups. Perpetrators now use sophisticated weapons and the attacks are more and more lethal. Currently, terrorist incidents are highly unpredictable which allows terrorist groups to attack by surprise. The unpredictability of attacks is partly due to the lack of real time terrorism data collection systems, adequate risk models, and prediction methodologies. To bridge the gap between terrorist incidents and counter-terrorism measures, it is crucial to develop real time terrorism data collection systems along with novel and proven risk models and prediction methodologies. In this research, we developed a set of systems and methodologies to collect and analyze terrorism related data. The methodologies include terrorism data summarization for root cause analysis, cluster analysis of terrorist attacks into groups with similar patterns, a novel risk model that uses data collected via our data collection system, and a prediction method that uses our risk model and Markov Chains.

Our methodology for root cause analysis utilizes our novel algorithm, along with Latent Dirichlet Allocation and historical terrorism data of START. Our clustering method segregates terrorist groups based on their similarities in attack patterns. Our data collection system is an automated crawler engine that collects data from selected data sources via RSS Feeds and on demand. The result is real time data that gets preprocessed automatically using selected keywords. Our novel terrorism risk model utilizes the preprocessed data to calculate terrorism risk levels at different locations. Lastly our prediction method utilizes our terrorism risk model, and Markov Chain models to predict future terrorist incidents in different countries. We have implemented a fully automated system that does not require any manual interventions for collecting and calculating the risk values. The results obtained in this research show a promising terrorism prediction system which predicts future attacks up to three months prior to the occurrence of an attack with a maximum of 96.85% precision and 96.32% recall. Our software system and methodologies can be a useful tool for terrorism analysts to improve counter-terrorism measures, and potentially prevent future terrorist attacks.

REAL TIME BIG DATA ANALYTICS FOR PREDICTING TERRORIST INCIDENTS

by

Ibrahim Toure

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Dr. Aryya Gangopadhyay, Chair/Advisor
Dr. Jianwu Wang
Dr. Vandana Janeja
Dr. Yelena Yesha
Dr. Zhiyuan Chen

Dedication

I dedicate this dissertation to my father Mamadou Toure, my mother Mariam Fadiga, my grandparents Abdoulaye Toure and Bintou Toure, my wife Zeinabou Kouma, and my children: Mariam I. Toure and Mamadou I. Toure.

To my father, Mamadou Toure - you have been a role model, mentor, and of tremendous support throughout my life and my education. I would certainly not be able to complete my Ph.D. study without your continuous support. Thank you Papa!

To my mother, Mariam Fadiga - you have been supportive from day one of my Ph.D. study, and your continued acknowledgment of my work gave me the energy to move forward and tackle the obstacles encountered. Thank you Tamam!

To my late grandfather Abdoulaye Toure and my grandmother Bintou Toure - arguably, every successful person needs a strong foundation in terms of education, culture, and compassion. After spending most of my childhood with you, I found myself enriched with all the core values I needed to set high goals and work toward reaching them. This dissertation is the reward of your teachings. Thank you!

To my wife, Zeinabou Kouma - you are my companion and my best friend. I cannot thank you enough for your patience and your emotional support throughout my Ph.D. study. Thank you for understanding the countless nights I spent on my computer instead of listening to your conversations. But I promise I will be a better husband from now on. Thank you Neka Zei!

To my two year old daughter Mariam Ibrahim Toure, and my six month son Mamadou Ibrahim Toure, I started my Ph.D. study before you were born. However, your arrival gave me a reason for pushing forward with my studies hoping to set an example for you two. I am confident that one day, both of you will be writing your Ph.D. dissertation dedications like I am doing now. I believe in you and I know you will accomplish great goals in your life. Remember to stay happy and seek a stable life.

To my unborn children - you might not be born yet, but know that you already occupy a big place in my heart and that Mariam and Mamadou will set good examples for you like I am trying to do now.

A final note to all my children - remember the teachings, the way of life, and the values me and your mother show you. We have both worked hard hoping to set perfect examples for you. Help each other and be happy!

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisor Dr. Aryya Gangopadhyay for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this dissertation. I could not have imagined having a better advisor and mentor for my Ph.D. study.

I would like to thank the committee: Dr. Lina Zhou, Dr. Vandana Janeja, Dr. Yelena Yesha, and Dr. Zhiyuan Chen for their insightful comments and encouragement, but also for the hard questions which incented me to widen my research from various perspectives. My special thanks go to Dr. Jianwu Wang for agreeing to serve on the committee in such short notice.

My thanks go to Dr. Mavis Sanders for doing a thorough review of this dissertation document and providing valuable feedback. Dr. Sanders has been an advisor, an inspiration, and a role model for me, without whom, I could not produce a streamlined dissertation document.

I would like to thank my family: my parents, my wife, and my brothers and sisters for supporting me spiritually throughout writing this dissertation and my life in general.

Finally, my thanks go to whoever reads this dissertation and finds it interesting.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Terrorism is a complex and evolving phenomenon. In the past few decades, we witnessed an increase in the number of terrorist incidents in the world. The security and stability of many countries is threatened by terrorists groups. Even though the probability of high lethal terrorist incidents in the United States is low [121], foreign terrorist groups such as ISIS and Al-Qaeda are growing at an alarming rate. Perpetrators now use sophisticated weapons and the attacks are more and more lethal [122].

Currently, terrorist incidents are highly unpredictable which allows terrorist groups to attack by surprise. The unpredictability is partly due to the lack of real time data and adequate risk analysis methodologies. To narrow the gap between terrorist groups and counter-terrorism experts, it is imperative to develop novel and proven methodologies along with sophisticated systems specially designed to solve terrorism related issues.

1.1 Motivation

Many controversies arise around terrorism. One famous expression in the terrorism study domain is: *"One person's terrorist is another person's freedom fighter"* An example of this is *Nelson Mandela* who was a well renowned freedom fighter of South Africa, but was only removed from the U.S. terrorists watch list in 2008. Another example is Hezbollah, a U.S. designated terrorist group in Lebanon whose members occupy seats in the Lebanese parliament [124]. Therefore, while some countries consider specific perpetrators as terrorists, others consider them as liberators, and thus support them. To make the issue even more complex, the supporters of a designated terrorist group often label the countering group as the terrorists. However, the motivation of this research is not to find the right or wrong doer, but to develop solutions to improve counter-terrorism planning and potentially prevent terror attacks.

In fact, the first challenge encountered in terrorism research is the definition of the term *terrorism*. What does it actually mean? Surprisingly, there is no agreed upon or international definition of the term terrorism. There have been debates and discussions about the term but no common definition has been deducted. Since the 1990s, the United Nations [81, 82, 83] has attempted but has not been able to establish an international definition of terrorism, which makes it impossible for some nations to join forces to combat it. In addition, different organizations within the same countries have different definitions for the term. For instance, the U.S. Federal Bureau of Investigation (FBI) [79] and the Central Intelligence Agency (CIA) [78] have different definitions for the term terrorism. We follow the definition of the Study of Terrorism and Response to Terrorism (START) [65] as it is a government sponsored program and we use their data extensively in this research. START defines terrorism as *"the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear,* *coercion, or intimidation".*

Research studies have shown the reasons people turn to terrorism are multiple, inconclusive, and highly contested. The following are some common causes of terrorist incidents: ethnicity, nationalism, separatism, poverty and economic problems due to globalization, anti-democracy, dehumanization, and religion.

Terrorist groups attack for different reasons; while some attack civilians to intimidate governments, some attack facilities to secure income, and some seek power and wealth by controlling territories. Terrorist groups are highly dynamic and secretive, which makes it challenging to track their activities and prevent incidents. Part of the dynamism comes from the relationship between different terrorist groups as they can unite or separate to create sub-groups.

In general, all terrorist acts are motivated by two things:

First, social and political injustice - that is, people choose terrorism when they are trying to right what they perceive to be a social, political or historical wrong.

Second, the belief that violence or its threat will be effective and usher in change. According to Martin et. al [151], many terrorists *"chose violence after long deliberation, because they felt they had no choice".*

Our research sheds light on some of the reasons for terrorism based on data from historical and recent incidents. It is also important to note that the reasons behind terrorist incidents evolve with time in complex ways. For instance, a given terrorist group may claim to defend a religion, but attack the people of the same religion. Moreover, the same terrorist group can attack government infrastructures. Therefore, analyzing the reasons behind terrorist incidents requires special attention.

1.2 Proposed Work

For one to study terrorism, many parameters need to be considered including: the behavior and perspectives of individuals, the facts of the terrorist incidents, the world of political systems, the causes of terrorism, and the consequences of terrorism.

In this research, we develop a set of novel methodologies and software solutions to assist counter-terrorism experts to discover the reasons and motivations of terrorist groups, visualize and classify groups based on their similarities, view impacts, estimate terrorism risk values in real time, and predict terrorist incidents.

In our first solution, we create a novel algorithm that uses Latent Dirichlet Allocation (LDA) [57] to analyze historical terrorism data. The result is a summary of root causes of terrorist incidents.

In our second solution, we create a novel algorithm that uses the Latent Semantic Indexing (LSI) [23] and K-Means clustering method to extract and classify terrorist incidents that share similar attack patterns. The result is clusters of similar terrorist groups. This special grouping allows analyses based on attack patterns.

Our third solution addresses the following topics: real time collection of terrorist incidents data, real time terrorism risk of countries, terrorist incident prediction.

1.3 Research Contributions

The main contributions of the proposed research are as follows:

Terrorism Root Cause Analysis: Mine the reasons behind terrorist incidents using recent and historical terrorist incidents data.

Clustering of Terrorist Groups: Cluster terrorist groups based on similarities in attack patterns.

Real Time Terrorist Incidents Data: Develop a system to automatically collect real time terrorist incidents data from reliable sources.

Terrorism Risk Model: Develop a novel mathematical risk model to calculate the terrorism risk values of different locations. Moreover, estimate and plot on a dynamic map the impacts of terrorist groups on different locations (e.g. countries and cities).

Terrorist Incident Prediction: Use our system and Markov model to predict the risk of terrorist incidents.

Our system and methodologies can assist terrorism analysts to improvement counter-terrorism planning and hopefully save lives. In addition, our solutions can bring more awareness of and clarity about the scourge of terrorism, and can help overcome obstructions to its demise.

Finally, we want to enable other researchers in the counter-terrorism area or in other subject areas such as health care to use our methodologies.

1.4 Organization of Dissertation

The rest of the dissertation is organized as follows. Chapter 2 provides information on the background and related work on data mining and terrorism research. Chapter 3 provides details on the methodologies and software we developed to analyze terrorism data. Chapter 4 presents experimental results and evaluations.

Finally, we conclude this dissertation in Chapter 5 and describe areas for future work.

Chapter 2

Background and Literature Review

2.1 Background

With the exponential growth of data generated by companies and individuals through computers and hand held devices, data mining has become a necessary tool for businesses and government agencies in decision making. The information discovered in data can be used to make quick decisions in critical situations. Data mining can also assist governments in responding to or predicting incidents. As there are many data types including text, time-series, and geographic information system (GIS) data, it is difficult to identify a single methodology that is applicable to all data types. Even though many prediction methodologies exist, it is yet challenging to predict incidents related to terrorism. Furthermore, data mining becomes more challenging when researchers are dealing with noisy, unstructured, incomplete, and heterogeneous data.

Data mining covers the following areas: privacy preservation, classification, clustering, association rule mining, anomaly detection, spatial data mining, graph partitioning, and prediction.

2.1.1 Privacy Preserving

Many data sets contain identifiable information such as name, social security number, address, salary, medical information, etc. Most of the time, this information is confidential. Therefore, it should not be made available to the public, otherwise, malicious people can use the identifiable information for their own interests. However, useful information can be extracted from these confidential data and used to advance a wide array of benign organizational goals. Thus, companies often share their data with third parties. Consequently, many methods have been proposed to hide identifiable information in data sets so that it can be safely shared with data miners without disclosing any sensitive information or identifiable data. The following articles Verykios et al. [34], Dwork [35], Clifton et al. [36], Mukherjee et al. [37], and Sweeney [38] proposed methods to address privacy preserving issues including hiding, removing, distorting, or codifying the identifiable information without compromising the integrity or quality of the original data sets. Thus, the main goal of these methods is to generate a transformed data that can still be mined. One can measure the efficiency of a privacy preserving method based on the relevance of the results obtained from the modified data to the results of the original data. The results of the modified data should be as close as possible to the results mined from the original data. In my opinion, the most significant research problem in privacy preserving is how to hide confidential information in data without losing the quality of the original data.

2.1.2 Classification

Support Vector Machines (SVM) are usually trained and used to correctly categorize previously unseen data for a variety of purposes. Classification methods can be used, for example, to determine if a new customer is credible or not based on characteristics such as age and salary. Another application domain is fund distribution among Ph.D. students to cover conference fees. The criteria may include first time requester, year in the program, available funds etc. Some of these classification methods are proposed in Sebastiani [29] and Li [30].

Another category of classification is association rule mining, in which the goal is to mine co-occurrence of events. For instance, association rule mining can be used to find out the products that are mostly purchased together. Such information can be used by stores to place products that are purchased together on the same aisles. Some association rule mining methods are discussed in Chen et al. [31].

A summary is a set of attribute-pairs and an aggregation of values at a higher level. A summary method is presented in Gengm et al. [32]. An example of a summary is: the total number of students graduated from UMBC I.S. department since 2005. In this case, the summary is based on the count of graduated students since 2005 and from the I.S. department only. An example of a resulting summary is "UMBC I.S. graduated 2005-2017 — 78".

Among many classification methods, Decision Trees (DT) induction method [30] is one of the most popular. It is used in data mining as well as in artificial intelligence. SVMs or classifiers take an input data and map it to pre-defined categories (set of classes) to determine to which class the data belongs. For instance, given the credit score, education level, and age of a customer, a bank can determine if that customer is credible or not using a trained classifier. Unlike clustering, classification requires a supervised learning and labeled data. Finally, depending on the classification method, scalability on large data sets is problematic. To accommodate time and computation complexity, some classification methods use pruning to avoid unnecessary processing.

2.1.3 Clustering

Clustering is often used to detect outliers in data sets and consists of grouping similar data points in the same cluster. Measuring the interestingness of clusters is particularly challenging, because a cluster may be false positive, false negative, true negative or true positive making it difficult to differentiate between good and bad results. Another challenge related to that fact is detecting outliers, because it is only if the resulting clusters are true positive, that one can get true outliers. Moreover, it is difficult to set the boundaries of overlapping clusters. Another challenge is to determine the ideal number of clusters. Generating too few or too many clusters will result in missing information that is potentially interesting. Another challenge is that a cluster can be completely merged into another cluster if there are not enough parameters (i.e. if each data point is represented by only x and y). Therefore, it becomes interesting to combine user profiles with clustering methods as explored by Zhang, Karabatis, Chen et al. [44] so that clusters are not only generated based on data points, but also based on user profiles. One can compute precision and recall to assess the quality of resulting clusters.

Clustering consists of forming clusters by placing similar data points in the same cluster. Typically, clustering methods minimize the intra-cluster distances and maximize the inter-cluster distances. Thus, the formed clusters contain only objects belonging to the same category. The most challenging part of clustering is measuring its fitness or the quality of its results especially if it is performed on a large volume of data.

2.1.4 Association Rule Mining

Association rule mining is particularly interesting for grocery stores and retailers. Placing two products that are more frequently purchased together close to each one on a shelf can help businesses to make millions of dollars. Yet, distributing the association rule mining results obtained from a store to other stores is challenging, because the results may be generalized to stores at different locations. The reason for that is the behavior of customers changes based on locations (e.g. customers from different locations have different preferences), and other factors. Moreover, a scalability issue will arise when data size reaches a certain level. Furthermore, in my knowledge, there is no association rule mining method that attempts to predict the purchases made in different seasons (i.e. winter, summer, and fall) so that the planning is done in advance. Finally, another interesting feature in this category is data mining based on product profiles. For instance, a food profile may include calories, sugar, protein, sodium etc. In this case, similar products or new products may be suggested to be moved along with resulting products of the association rule mining tool. In my opinion, the most significant research problem in categorization is how to categorize unseen data, by using either supervised or un-supervised methods, to the proper class, cluster, or set.

2.1.5 Anomaly Detection

Detecting anomalies in data is challenging especially if they are unstructured and heterogeneous. Data may be in multimedia or text formats such as a Youtube, CBS videos or user feedbacks collected from eBay, Amazon etc. Typically, a data miner does not know in advance what he/she is looking for. Also, if we are to analyze text data, manually typed by customers, the first challenge is the choice of appropriate method in the variety of methods available. An inexperienced user (e.g. business owner) may not know the tool that works best to solve his/her problems. More challenging is to develop a novel approach to analyze unstructured text data.

In the survey of Chandola et al. [50], many anomaly detection methods are discussed with their application areas such as network intrusion, image processing, structural defect, medical and public health anomaly detection, etc. In my opinion, the most significant research problem in anomaly detection is how to generate a summary of problems that occur in data.

2.1.6 Spatial Data Mining

Data related to geographic locations may be difficult to mine especially due to the nature of data. A spatial data may have different dimensions and properties. Moreover, a defective device may generate erroneous data streams which lead to incorrect results. Also, different locations may be affected by weather or other natural effects. Some novel anomaly detection in spatial data have been introduced by Janeja et al. [52], Shi and Janeja [51] and by Ester et al. [49]. In my opinion, the most significant research problem in spatial data mining is how to manipulate geographic information system (GIS) data and analyze it with other properties.

2.1.7 Large data sets

Dealing with large data sets like Google's or Facebook's databases require a lot of computation resources and time. Thus, it is indispensable to develop highly scalable methods and techniques to efficiently address these issues. The following sub-areas can be classified under large data sets: data processing, data storage, information retrieval, graph partitioning, and data integration. I further detailed these methods in the following sub-sections.

2.1.7.1 Data Processing

Processing large data sets is challenging, especially if atomicity, consistency, isolation and durability (ACID) have to be satisfied. With the limitations of current computer resources, buffering terabytes of data into memory to be computed simultaneously across networks of thousands of computers is impossible without efficient methods. MapReduce was developed at Google and presented by Dean and Ghemawat [39]. This method is especially efficient due to its flexibility to process large data across hundreds of computers, and merge the final result. Furthermore, it is highly fault tolerant and can recover from computer crashes without affecting the final result, thus satisfying ACID. The only disadvantage I found in MapReduce is that if the machine that runs the master fails, the whole data processing stops, because there will be no way to control the workers.

2.1.7.2 Storage

I believe that storage is one of the most important components in large data sets. Storage of hundreds of terabytes of data across hundreds or thousands of computers requires efficient storing methods. When developing a method to efficiently store data, one should keep in mind the space limitations and to make sure that the stored data can be accessed faster. Big Table was developed at Google and presented by Chang et al. [41], a survey of data warehousing research issues are presented in Chadhuri and Dayal [40] and Chang et al. [42], and an overview of the evolution of data formats and storage is presented in Stonebraker and Hellerstein [27].

2.1.7.3 Information Retrieval

Data is stored under different formats, using different technologies resulting in heterogeneous formats. After storing terabytes of data in these conditions, retrieving them becomes a complex task. Moreover, because of the fast growth of data, highly flexible and scalable methods need to be developed to efficiently handle information retrieval. Some methods are proposed to address these issues in Karabatis et al. [47, 48]; other methods are also presented in Jing et al. [43] and Langville et al. [54].

2.1.7.4 Graph Partitioning

Most graph partitioning methods are to address community detection problems. With the globalization of social media, websites like Twitter and Linkedin are getting more and more attractive to a variety of audiences. One of the challenging parts of mining social networks is to accurately determine the interests of each member in the network, and be able to predict the interests of a member based on his past interests or profile. If there is not enough information in the profile of a member to predict his interests, how can we use his friends' profile information to predict his interests? At this point, it is important to detect communities, because this will facilitate targeting a group of people sharing the same interests. A detailed survey of graph partitioning methods was presented by Fortunato [55]. On the other hand, Tang and Liu [56] proposed a method to determine the interests of members based on their latent affiliations.

2.1.7.5 Data Integration

The heterogeneous nature of data makes it challenging to merge multitudinous data sources from different databases. Reviews in Halevy [45], Bleiholder [46] identified some important issues in data integration such as duplicate detection (effectiveness and efficiency) and conflict classification (schematic conflicts, identity conflicts, and data conflicts). Ontologies methods attempt to solve these issues, but are limited especially if there is no pre-determined labeled data to match the two data sources. I classified the work of Heinrichs and Lim [28] as a data integration category because it assesses the interests of adopting data mining tools in businesses.

2.2 Related Work

Terrorism is not new; it goes back to the beginning of recorded history where perpetrators use terror as a mean to intimidate and coerce. According to Rapoport [97] there are 4 waves of terrorism that were recorded since the 1880s. First, the “Anarchist Wave” appeared in the 1880s which lasted for about 40 years, followed by the “Anti-Colonial Wave” which began in the 1920s and continued until the 1960s. The “New Left Wave” was the third wave started in the late 1960s and dissipated largely in the 1990s leaving a few active groups in Sri Lanka, Spain, France, Peru, and Columbia. Fourth, the “Religious Wave” began in 1979 and is still continuing today. Based on observations of recent increases in terrorist incidents, the religious wave will be active for many years. During the leftist wave in 1975, Jenkins [145] concluded that “terrorists want a lot of people watching, not a lot of people dead.”

However, by 2006, he changed his original dictum to “many (although not all) terrorists want a lot of people watching and a lot of people dead.”

The religious wave brought along the highest rate of suicide bombings. One of the reasons for that increase can be explained by the religious ideologies lectured to the suicide bombers that make them willing to die for the cause. Such perpetrators are made to believe in ideologies like “I have nothing to lose, but I have everything to gain” [85], which refers to the fact that dying for the religion guarantees heaven. However, not only religious ideologies motivate suicide bombers. Riaz [86] in his article “What Motivates the Suicide Bombers?” stated that suicide bombers are mostly not motivated by religious purposes only, but also by “Humiliation, Revenge, and Altruism”. Other reasons, such as threats against the family of the suicide bomber and money may motivate the perpetrator. Riaz’s conclusions are based on the analysis of suicide bombers data collected from 1981 to 1996 in Iraq, Palestine-Israel, Afghanistan, Pakistan, and Sri Lanka.

Terrorist groups manage to transfer money using legal financial institutions and other illegal methods. Many efforts have been made to dismantle the financing of terrorists by making laws against money laundering and terrorism financing. Certain government laws require financial institutions to create software systems and develop methodologies to track transactions, halt suspicious activities, and potentially uncover terrorism financing networks.

In the 1990s, some efforts were made to control funds raised and sent to needy countries by NGOs. IEEPA (International Emergency Economic Act IEEPA-1977) was improved in the 1990s by making many laws such as: 1990 Anti-Terrorist Act which made funding, or providing resources to foreign terrorist organizations illegal. In 1995, Federal laws were made to authorize the deportation of terrorists and sanctions against fundraisers. In 1996, the Anti-Terrorism and Effective Death Penalty Act was made to criminalize providing support to terrorists, and also allowed civil suits against aiding in the commission of a terrorist act. In the same time period, Anti-Money laundering laws required banks to submit statements on Suspicious Activity Reports (SARs), and implement Know Your Customer (KYC) systems to prevent terrorist groups from using financial institutions to launder funds.

The efforts of controlling terrorism financing increased post 9/11, which resulted in the creation of departments specially focused on preventing terrorism financing. Such departments are: the FBI's Terrorist Financing Operations Section (TFOS); the US Customs' Operation Green Quest, the Justice Department's Unit for investigating financing; the CIA Interagency control of the gathering of Intel, tracking, and disrupting terrorist incidents; the National Security Council's Policy Coordination Committee (PCC); and finally the Office of Foreign Assets Control (OFAC). Working with the other organizations mentioned above, OFAC is used by the U.S to target and freeze the funds of alleged terrorist groups. Moreover, in 2007, the U.S. government established the Anti-Money Laundering and Counter-Terrorist Financing Unit (AML/CTFU) [93] by Law No. (46) as a financially and administratively independent Unit.

The U.S. designated foreign terrorist groups are in other countries and mostly beyond the country's legal reach. As of April 2015, a total of 59 Foreign Terrorist Organizations (FTOs) [120] were designated by the Secretary of State in accordance with section 219 of the Immigration and Nationality Act (INA). The United States allies with many countries to provide resources and adequate training to combat terrorist groups.

As there is no international law against terrorist groups, some countries do not restrict designated terrorist groups and such countries are sometimes the safe haven of terrorist groups. In addition, some countries support designated terrorist groups financially and provide them other resources.

Ozgul et al. [102] developed a novel method called the Crime Prediction Model (CMP) to analyze and identify terrorist groups of unsolved attacks. CMP learns from terrorist attacks, matches them based on the similarities of their properties, and then clusters them into groups. This method was applied to a database of real life terrorist attacks that occurred in Turkey between 1970 and 2005. The predictions of CMP gave a good precision value for big terrorist groups and provided a good enough recall value for small terrorist groups. The average precision value for CPM is 76% and the average recall is 52%.

Shaikh and Wang [142] proposed the Investigative Data Mining (IDM) method to identify key nodes in terrorist networks. The method consists of mining networks of terrorists, identifying the most influential actors in the network (leaders), and the coordinators of transactions (gatekeepers) for activities such as passing weapons. The Social Network Analysis (SNA) methods were used on a dataset of U.S. embassy bombings in Tanzania, to identify the main actors in the network. First, an adjacency matrix was constructed with the dataset of 16 terrorists, where the value 1 means there is a connection between the two actors, and 0 otherwise. Second, the key actors are identified based on the computed results of the Degree of centrality, Betweenness, Closeness, and Eigenvector Centrality. Finally, the resulting connections of actors were plotted on a graph. The main actors had a higher number of connections. After identifying the main actors, appropriate actions can be carried out by either removing or isolating the key actors to destabilize the network, thus potentially disrupting the interactions and plans of the group.

Gay [111] examined the parameters of an underwater terrorism event. The research validated the parameters used by many in the intelligence and risk communities as applicable to underwater terrorism and verified the anecdotal expectation that the underwater environment impacts an adversary's planning cycle. It provided operational definitions for 17 distinct parameters and provided the relationships between the parameters to assist the modeling and simulation community in developing robust threat or risk tools.

Sun et al. [113] have proposed a new event driven extraction task and four pattern-based document selection strategies. The method is applied to the terrorism event information extraction. The objective is to select as few documents as possible to construct the event related entity and relation instances. Performance metrics were defined to compare the proposed document selection strategies and several experiments were conducted on two datasets. Experimental results conclude that the proposed strategies perform well on the extraction task. Among the proposed strategies, PartialMatch shows the best performance. For a dataset containing 1000 documents, the PartialMatch extracted 95% of the required event related information by selecting only 5% of the documents.

Hu et al. [114] developed a method to analyze biomedical literature for potential bioterrorism weapons. The method used Geissler's 13 search key words criteria to identify all the possible properties of viruses. Then, based on the judgment of a domain expert, the properties are assigned a MeSH term and an importance weight. Then, the biomedical literature is searched to retrieve all the relevant documents. The method uses virus names extracted from the biomedical literature to discover novel connections between viruses and potential bioterrorism viruses.

Khalsa [22] developed a forecasting terrorism tool using indicators and analytic techniques to guard against the common pitfalls of terrorism analysts. His tool has 3 primary types of warning picture views: country list view, target list view, and indicator list view. His methodology consists of 23 tasks and 6 phases of warning analysis. The 23 tasks include: 4 annual tasks, 3 monthly tasks, 14 daily tasks, and 2 as-required tasks. This methodology requires three types of analysts to operate it: Raw Reporting Profilers, Indicator Specialists, and Senior Warning Officers. The data collection in this research is purely manual and requires time and resources.

Croushore [132] in his article Frontiers of Real-Time Data Analysis reported many data revision methods used in finance to forecast financial recession. However, these methods cannot be applied to terrorism studies due to the uncertainty of human behavior.

Kengpol et al. [149] proposed a risk assessment method to predict the distribution range radius of terrorist incidents. The authors proposed a Risk Assessment Radar Chart to prevent Improvised Explosive Device (IED) terrorist incidents. The IED incidents data from 2007 to 2013 in the capital district of Yala province, the southern part of Thailand, were collected and used in the proposed methodology as a case study. The results are distance based incident predictions which specify in kilometer radius and number of days, the possible insurgencies. The model was tested against real life events and has a precision of prediction rate of approximately 50.41% (distance in kilometer) and 43.90% (in number of days). Our methods however, provide higher precision and recall values for terrorist incident predictions.

Karl Roberts and John Horgan [147] analyzed the risk factors to be considered when developing risk level methodologies to assess terrorism. Some of the suggestions include historical factors, clinical risk, and risk management factors. However, no risk model is developed in this research.

Ezell et al. [148] used probability tree and decision tree to simulate the risk level of terrorist incidents. The data used are based on assumptions of the occurrence of different events. Probabilistic risk analysis (PRA) and event trees have been shown to be useful approaches for assessing terrorism risks, especially for creating a baseline comparison of these risks. Decision trees, like PRA have limitations and are not complete solutions. Moreover, the limitations of the decision tree approach may be difficult to surmount upon implementation given the fact that adversaries' objective functions and level of ability to predict tree outcomes are unknown and difficult to estimate. The results were not tested against real life events.

Henry et al. [146] provides a definition of risk and discusses the relationships among threats, vulnerabilities, consequences, and risk. In addition, the research suggests a method for constructing a single dimensional estimate of city risks, which is designed to perform well across a wide range of threat scenarios, risk types, and other sources of uncertainty. The method provides a framework for comparing the performance of alternative risk estimates given uncertainty in terrorists' intentions and capabilities, target vulnerabilities, and the likely consequences of successful terrorist attacks. An estimate of the annual probability of terrorism risk based on the threats to a target, vulnerability of the target to the threats, and consequences if the attack is successful are reported. The risk simulations results of the method show that the annual risk of New York ranges from 3.04 to 30,400. The risk level in the methodology does not have boundaries. Moreover, the method is not event based, but rather based on assumptions of event and population density of cities. The results were not tested against real life events.

Darby J. L. [150] developed a software tool based on fuzzy sets, approximate reasoning, and the belief/plausibility measure of uncertainty. The input data are based on the reasoning of the user of the system. Similar to the research conducted by Henry et al. [146], the risk was partly based on threat, vulnerability, and consequence.

Chapter 3

Research Methodology

In this Chapter, we will provide details on our methodologies.

We developed a set of solutions to reach our objectives below:

Terrorism Root Cause Analysis: Analyze historical and real time data to extract the root causes of terrorist incidents.

Clustering of Terrorist Groups: Cluster terrorist groups based on similarities in attack patterns.

Real Time Terrorist Incidents Data: Develop a system to automatically collect real time terrorist incidents data from reliable sources.

Terrorism Risk Model: Develop a new mathematical risk model to calculate the terrorism risk of locations (i.e. countries and cities). Moreover, calculate and plot on a dynamic map the impacts of terrorist groups on different countries.

Terrorist Incident Prediction: Predict terrorist incidents in different countries.

3.1 Terrorism Root Cause Analysis

Toure and Gangopadhyay [64] create a novel methodology based on the Latent Dirichlet Allocation (LDA) [130]. Then, we apply the method to the historical terrorist incidents data collected over several decades [80]. The result is a text summary that provides insights into the reasons and motivations of terrorist groups at the time of the incidents. More details are provided below on the methodology and results.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a mixed membership model that has been used to discover hidden themes or ”topics”' in a document corpus. In the context of a corpus of text documents, a topic model captures the underlying themes or topics that exhibit themselves in different proportions in the documents. The topics themselves are distributions over words or terms that appear in the corpus. Given that the only observable parameters are the words that appear in the documents, the challenge is to estimate the hidden parameters such as the word distributions in topics, the topic proportions in documents, and the word assignments to generate the documents. The LDA is a mixed membership model that generates each document in a corpus as a bag of words given the hidden parameters. The challenge is to estimate the hidden parameters given the observable data. Several methods have been proposed for the parameter estimation of the LDA, and we follow the mean variational methods [18] in this work. The LDA model estimation and inference were done using the lda-c implementation [130].

Method

The input data to our method is a set of texts describing each attack. These texts can be obtained from the news media including the Internet, print media, as well as other sources such as agencies tracking information on terrorist attacks. Our goal in this methodology is to discover the “hidden” themes in these texts. The first step is to generate a *term-document matrix* ($\mathcal{D}$), which consists of the *terms* that occur in the rows and the *documents* in the columns. Each document corresponds to the textual information on each attack. Each entry in $\mathcal{D}$ corresponds to the frequency of occurrence of each term corresponding to each attack document. Prior to creating the matrix $\mathcal{D}$, we remove common stop words such as article, prepositions, etc. The matrix $\mathcal{D}$ is used to create a number of topic models with different numbers of topics.

Our proposed method is presented in Algorithm 1. The first input parameter to the algorithm is the term-document matrix, which is described above. The second input parameter, $\mathcal{N}$, is simply a list of numbers that represent the number of topics to be created for each topic model. In our case this was {50, 100, 300, 350, 400}. These numbers are somewhat arbitrary as there is no known method for determining the "ideal" number of topics. However, having too few topics will capture the broad-brush patterns and too many will overfit the model to the data. The third input parameter k is the number of top terms for each topic. Each topic model $\mathcal{T}_i$ consists of n_i number of topics. Each topic is a list of terms where the j^{th} topic in $\mathcal{T}_i$ is denoted by T_{ij}. Each T_{ij} is truncated to only the top-k terms (line 4 in Algorithm 1). Next we take the intersection of the truncated topics to create the truncated top-k terms T_i (line 6 in Algorithm 1). The final ordered list of terms T is created by taking the intersection of all truncated ordered sets (line 8 in Algorithm 1). The summary, shown in lines 9 and 10 in Algorithm 1, can be generated manually or by using natural language generation, which is not discussed in this paper.

The Big O notation is used to analyze the performance or complexity of algorithms in this dissertation.

Algorithm 1 Algorithm for Analyzing Terrorist Attacks

Input: $\mathcal{D}$: an $m \times n$ Term-Document matrix; $\{n_1, n_2, \ldots, n_l\} \in \mathcal{N}$: set of number of topics; k: number of significant terms in a topic

Output: Summary of attack patterns $\mathcal{S}$.

```
for $i = 1$ to $l$ do
    Create a topic model $\mathcal{T}_i$ corresponding to the number of topics $n_i \in \mathcal{N}$
    for $j = 1$ to $n_i$ do
        Truncate topic $T_{ij} \in \mathcal{T}_i$ to the top-$k$ topics
    end for
    $T_i = \bigcap_{j=1}^{n_i} T_{ij}$
end for
$T = \bigcap_{i=1}^{l} T_i$
Generate summary $\mathcal{S}$ from $T$
Return $\mathcal{S}$
```

For $t \in [T]$, let n_t^* be the number of words assigned to topic t in the maximum a posteriori (MAP) assignment. Then, the MAP assignment x is found by solving the following optimization problem:

$$\max_{x_{it} \in \{0,1\}} \sum_{i,t} x_{it} l_{it}$$
$$\text{subject to } \sum_t x_{it} = 1, \sum_i x_{it} = n_t^*,$$

which is equivalent to weighted b-matching in a bipartite graph (the words are on one side, the topics on the other) and can be optimally solved in time $O(bm^3)$, where $b = max_t n_t^* = O(N)$ and $m = N + T$ [161].

As we truncated the topic $T_{ij} \in \mathcal{T}_i$ to the top-k topics where k equals 20; the inner loop can be ignored as it is constant time.

Therefore, the complexity of our algorithm is $O((lN)(bm^3))$ where l is the number of topic models and N represents the size of the input data.

The results of this method are provided in the next chapter.

We published this methodology [64] in the proceedings of the IEEE Conference on Technologies for Homeland Security (HST'2012), Boston, November 2012.

3.2 Clustering of Terrorist Groups

Terrorism activities occur in many parts of the world. Such activities seemingly occur randomly in different locations, at different times, and are caused by different perpetrators. Thus, it is a challenging task to find patterns in activities related to terrorism.

3.2.1 Analyzing Terror Attacks using Latent Semantic Indexing

Clustering of perpetrators into similar groups can provide valuable information such as common characteristics among the various groups, the types of targets typically attacked, and weapons used in such attacks. In this research, we use the terror data collected by START (Study of Terrorism and Responses to Terrorism) from 2000-2011. We developed a method to classify terrorist groups based on their attack patterns by analyzing textual descriptions of such attacks using latent semantic indexing and clustering. The resulting information can be used for counter-terrorism globally. In addition, the results can help develop security measures to protect infrastructures such as hospitals, schools, and government agencies as well as saving the lives of innocent people.

3.2.2 Methodology

3.2.2.1 Latent Semantic Indexing

Latent semantic indexing (LSI) [33] is an indexing and information retrieval method that uses the singular value decomposition (SVD) [34] technique to identify the semantic meaning of words based on their patterns, and the relationships between the terms and concepts contained in a text collection. LSI follows the logic that words used in the same context tend to have similar semantic meanings. One of the primary properties of LSI is its ability to establish associations between terms that appear in similar contexts.

3.2.2.2 Method

In this research, we mine the textual summary information of terror events. Our method uses the Latent Semantic Indexing [33] to cluster attacks based on their similarities. We first extract the data over a period of 11 years from 2000 to 2011 from START. We remove common English words using a stop word list [35] to avoid bias with non-content terms. We segment the extracted data into three segments (2000 - 2005, 2006 - 2011, and 2000 - 2011) to perform different analysis. We then generate a term-document matrix of these segments. Next, we compute the Singular Value Decomposition (SVD) of the term-document matrix. Then, we cluster the first two vectors with the highest energy by using the simple Kmeans clustering method. The vectors with the highest energy represent the data that contain the most semantically connected information of the summaries. Finally, we plot the results in a 2D graph and analyze them.

Our proposed method is presented in Algorithm 2. First, we extract the summary information of the attacks. Next, we attribute a unique ID to each attack for later identification of the clusters, and remove all the stop words like $the, a, etc.$ using a stop word dictionary [35]. Next, we use macros to insert a blank line after each row so that a term-document matrix can be processed properly. Using the identifiable summary data processed so far, we generate the term-document matrix. Next, we generate the first 6 singular value decomposition vectors. Then, we choose the 2 vectors (v_x, v_y) that have the highest energy. Next, we form 2 clusters using Kmeans and assign each pair (v_x, v_y) to either the first cluster $\mathcal{C}_1$ or to the second

Algorithm 2 Algorithm for Clustering Terror Events

```
Input: Textual Summary of Attacks
Output: Clusters of Similar Attacks;
  Extract the Textual Summary of Attacks;
  Attribute a Unique Identifier to each Attack;
  Remove Stop Words;
  Generate $\mathcal{D}$: an $m \times n$ Term-Document Matrix;
  Generate $(v_1, v_2, \ldots, v_z)$ = LSI $(\mathcal{D})$;
  Extract $(v_x, v_y)$ with the highest energy;
  for $i = 1$ to 2 do
    $\mathcal{C}i$ = Kmeans $(v_x, v_y)$;
    $\mathcal{G} = \mathcal{C}i$; // Plot the Cluster $\mathcal{C}i$ on Graph $\mathcal{G}$.
  end for
  Return Final Graph $\mathcal{G}$;
```

cluster $\mathcal{C}_2$. The resulting clusters are displayed in a 2D graph. Finally, the resulting clusters are groups of attacks that are similar to one another and have similar patterns.

For computational complexity, finding the optimal solution to the Kmeans clustering problem for observations in d dimensions is:

If k and d (the dimension) are fixed, the problem can be exactly solved in time $O(n^{dk+1})$, where n is the number of entities to be clustered [160]. In our context, k and d are very small and can be ignored.

Thus, the complexity of Algorithm 2 is $O(Nn^{dk+1})$ where N represents the size of the input data.

We name this methodology *Analyzing Terror Attacks using Latent Semantic Indexing*. We present its results in Chapter IV.

We published this methodology [65] in the proceedings of the IEEE Conference on Technologies for Homeland Security (HST'2013), Boston, November 2013.

3.3 Real Time Terrorist Incidents Data

Real time data are needed to develop a real time terrorism risk model. Currently, there is no real time data collection system available to researchers. Therefore, we design and implement a data collection system.

3.3.1 Data Collection System

The data collection system has multiple web crawler instances connected to our database. The architecture of the systems is illustrated in *Figure 3.1*.

The news media is the number one source for getting real time incidents information. An advantage is that the news media data is often delivered through RSS feeds, therefore, we use RSS feeds to collect data from news sources that offer that capability. We consider the information reliable as it is usually collected at the scene of the incident. Moreover, the title and the short description of the news has enough information for our method. Thus, to avoid unnecessary data collection efforts, we collect only the title and the short description of news. However, the media data do not include the exact location information. For instance, the media companies do not provide the longitude, latitude, zip codes of the incidents. When such data sources become available, this research can be extended to cover granular location data. Our system supports data in languages other than English such as French, Spanish, and Arabic. However, this research covers data in English only.

As we are using RSS feeds, our system can collect data from foreign countries not blocked by the US government. For example, Al-Jazeera Africa version is not accessible from the US. One solution for that is to setup a data collection server abroad.

Our second data source is twitter. We collect live streaming data from twitter which consists of tweets posted in real time. Some tweets contain location information that can be used to pinpoint where the incidents occurred. However, many tweets do not include location information.

After long monitoring, all incidents that make it to the news are collected by our system, but it does not collect incidents that do not reach the news media or twitter.

Currently, our system stores data on a UMBC local server. For scalability, prevention of data lost, and worst case data collection scenarios, we may use Amazon Web Services to store our data in the future.

3.3.1.1 Web Crawler

As illustrated on *Figure 3.1*, the web crawlers periodically download news data from reliable sources such as New York Times, Washington Post, and Al Jazeera. The number of news sources that can be fetched is virtually unlimited. The execution time to check for news updates can be set to as short as one second. Our time settings criterion is to set the recheck period to five minutes to assure all news of interest are fetched on a regular basis. A web crawler can be scheduled to fetch one or many news sources. Similarly, the same news source can be fetched by many crawlers, which is significantly beneficial as when a crawler fails, another one can pick up the task. In addition, the information retrieval is more reliable when at least two or three crawlers are placed on each news source. Our system is distributed which allows the crawlers and other components of the system to communicate and exchange messages.

The algorithm 3 presents the functions crawlers perform. The inputs of the

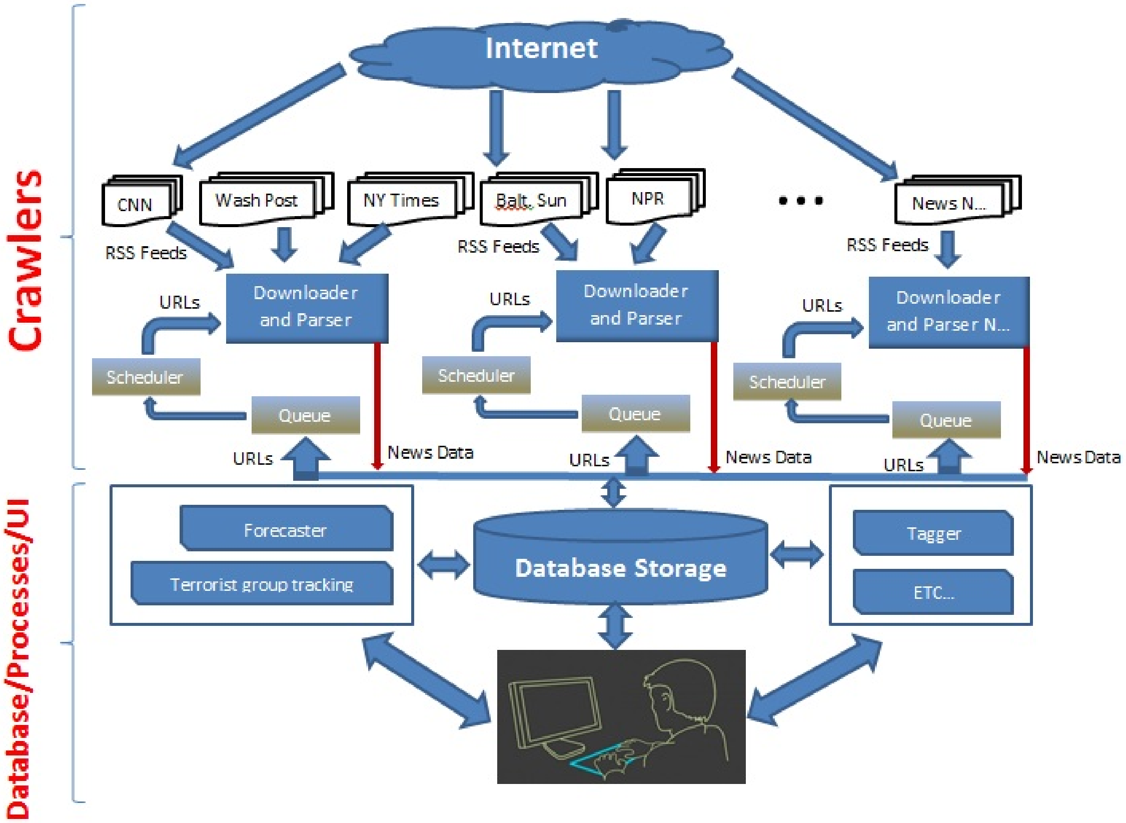


Figure 3.1: System Architecture

crawlers are: $\mathcal{I}$ the unique identifier of the crawler; $\{a_1, a_2, \ldots, a_n\} \in \mathcal{A}$ the list of news sources assigned to the crawler $\mathcal{I}$; T the wait time before re-fetching the news source; and $\mathcal{N}$ the number of news to download during each fetch. The input parameters are dynamic and can be configured in the settings for each crawler. The output of our algorithm is $\mathcal{D}$ the downloaded news data.

The first step of the crawler is to open an infinite execution loop that never ends. At each run of the algorithm, the crawler resolves the path of the news source. Then, it loads the number of $\mathcal{N}$ news specified in the input, from the news source a_i assigned to the crawler $\mathcal{I}$. The downloaded data a_i get stored in the variable

$\mathcal{D}_i$ for further processing. The system cleanses the data in $\mathcal{D}_i$ to remove special characters, *HTML* tags, etc. Start a loop on the news data loaded in $\mathcal{D}_i$. For each news data in $\mathcal{D}_{ij}$, if the news data $\mathcal{D}_{ij}$ already exist in the database, skip the record, otherwise, store the new record $\mathcal{D}_{ij}$ in the table $\mathcal{G}$. The system groups the news downloaded $\mathcal{D}_{ij}$ into $\mathcal{D}$ and returns it. Then, wait for T time before re-fetching the news source. After T time passes, re-run the algorithm from step one. The complexity of Algorithm 3 is $O(N^2)$ where N represents the size of the input data (the number of news).

Once the data are downloaded and stored in our database, other components of the system such as the country and city scanners filter the data and classify the contents for further processing. We use the data filtered in our terrorism risk level and incident prediction models.

Currently, we have three web crawlers deployed to fetch data from 20 news sources. The data collection of those 20 sources cover enough data for our research. Our topics of interest in this research are international incident news, and crime reports. The number of news sources and web crawlers can be multiplied to cover more topics. For other research subjects such as health care, the crawler can be used to fetch data from other sources that provide RSS feeds capabilities.

3.3.1.2 Country and City Scanners

The crawlers regularly add new data to our database. We develop two automatic scanners: the country level and the city level scanners. There are two options

Algorithm 3 Crawler Algorithm

Input: $\mathcal{I}$: ID of the web crawler; $\{a_1, a_2, \ldots, a_n\} \in \mathcal{A}$: List of news sources assigned to the crawler $\mathcal{I}$; T: Wait time before re-fetch; $\mathcal{N}$: the number of news to download during each fetch

Output: Downloaded News Data $\mathcal{D}$.

```
for $i = 1$ to $\infty$ do
    Resolve Source Paths
    Load $\mathcal{N}$ of $a_i$ from source of $\mathcal{I}$
    $\mathcal{D}_i = a_i$
    Cleanse $\mathcal{D}_i$ remove special characters
    for $j = 1$ to $n_i$ do
        if $\mathcal{D}_{ij}$ $\exists$ in Database then
            Skip $\mathcal{D}_{ij}$
        else
            $\mathcal{G} = \mathcal{D}_{ij}$
        end if
    end for
    $\mathcal{D} = \bigcap \mathcal{D}_{ij}$
    Return $\mathcal{D}$
    Wait $T$ time before re fetch
    After $T$ time, re run the algorithm from step one
end for
```

to implement the scanners. First, set a rescan time T and the system automatically triggers the scan at every T time (e.g. every 5 minutes). Second, develop a real time messaging system that informs the scanners of new database updates. We chose the first option as it is more efficient to trigger the scanners based on a preset time than having messages going through the system which can create unnecessarily long message queues. Thus, similar to the web crawlers, the rescan process is automatic and it triggers on specific timings.

The algorithm 4 represents the tasks the Country scanner performs. The input parameters are: $\mathcal{N}$ the number of records to be stored before the scan process starts; and T the wait time before rescan automatically triggers. The output is: $\mathcal{R}$ the list of data scanned. The algorithm starts by counting the number k of data not yet scanned in $\mathcal{G}$. If k is higher or equal to the preset number $\mathcal{N}$, load the list of active countries into the variable $\mathcal{C}$. For each country $\mathcal{C}_i$, load the list of active terms $\mathcal{S}$ (i.e. kill, bomb, explosion, suicide, etc.). For each country C_i and term S_j, check if the C_i and S_j exist in the data $\mathcal{R}_l$ and store the total count in the variable x. If the resulting x is not equal to zero (0), check if the data $\mathcal{R}_l$ already exists in the destination table $\mathcal{O}$, store the total count in the variable y. If the new data does not exist in the destination table $\mathcal{O}$ ($y = 0$), insert the value of $\mathcal{R}_l$ into $\mathcal{O}$. Mark $\mathcal{R}_l$ as the latest record for the news id. In case the news already exists in the destination table $\mathcal{O}$, display the message ”News already added”. Group the values of $\mathcal{R}_l$ into $\mathcal{R}$. At this point, we can optionally set the data $\mathcal{R}_l$ as scanned to prevent rescans with other terms $\mathcal{S}_j$. Then, for tracking purposes, the system updates the table $\mathcal{H}$ with the scan id and current date time. After the above steps, set all the loaded data $\mathcal{R}$

as scanned. In case the data count is below the preset $\mathcal{N}$, display the message "No data to scan". Return the list of data scanned $\mathcal{R}$. The system waits for the preset T time before rescan. Finally, when the wait time T expires, the rescan automatically triggers, and the algorithm restarts to step one. The complexity of Algorithm 4 is $O(N^4)$ where N represents the size of the input data.

The algorithm 5 represents the operations performed by the City scanner. The City scanner uses the results of the Country scanner in $\mathcal{O}$ as its input data source. In addition, the city scanner uses the following input parameters: the number of records $\mathcal{N}$ to be stored before scan starts, and the wait time T before rescan automatically re-triggers. The output value is the list of data $\mathcal{D}$ scanned.

The first step of the algorithm is to count the number of new incidents data in $\mathcal{O}$ not yet scanned and store the total count in the variable z. If the value in z is greater or equal to the preset value $\mathcal{N}$, load the list of active Countries in $\mathcal{O}$ into the variable $\mathcal{C}$. For each country name in $\mathcal{C}$, load the list of active cities into $\mathcal{A}$ for country $\mathcal{C}_i$. Thus, the variable $\mathcal{A}$ contains both the list of active countries and cities. Load the incidents data from $\mathcal{O}$ to scan, and store it in the variable $\mathcal{D}$. For each record in $\mathcal{D}$, load the data $\mathcal{D}_l$ in the queue. Count the number of records in $\mathcal{D}_l$ that contain both the country name C_i and the city name A_j, store the total count in the variable x. If the value of the x is not equal to zero (0), store the incident information of $\mathcal{D}_l$ in the destination table for city incidents $\mathcal{W}$. The risk model of cities uses the incidents data in $\mathcal{W}$ after validation. Next, the system sets the value of $\mathcal{D}_l$ as the latest record in $\mathcal{W}$ for the incident id. Group the values of $\mathcal{D}_l$ into the variable $\mathcal{D}$. Optionally set the value of $\mathcal{D}_l$ as scanned to prevent rescan with other

Algorithm 4 Country Scanner

Input: $\mathcal{N}$: Number of records to be stored before scan start; T: Wait time before rescan

Output: List of Data Scanned $\mathcal{R}$.

```
for $q = 1$ to $\infty$ do
    $k = \sum$ Data in $\mathcal{G}$ not yet scanned
    if $k >= N$ then
        $\mathcal{C}$ = List of active countries
        for $i = 1$ to $m_i$ do
            $\mathcal{S}$ = List of active terms
            for $j = 1$ to $n_j$ do
                $\mathcal{R}$ = Load data not yet scanned
                for $l = 1$ to $k_l$ do
                    $\mathcal{R}_l$ = Load data in the queue
                    $x = \sum \exists\ (C_i$ and $S_j) \in \mathcal{R}_l$
                    if $x \neq 0$ then
                        $y = \sum \mathcal{R}_l \exists$ in $\mathcal{O}$
                        if $y = 0$ then
                            $\mathcal{O} = \mathcal{R}_l$
                            Set $\mathcal{R}_l$ as latest record $\in \mathcal{O}$
                        else
                            print  ”News already added”
                        end if
                    end if
                end for
                $\mathcal{R} = \bigcap \mathcal{R}_l$
            end for
            Set $\mathcal{R}_l$ as scanned to prevent rescan (Optional)
        end for
        Update table $\mathcal{H}$ with scan id and current date time to track scans
        Set all $\mathcal{R}$ as scanned
    end if
    Return $\mathcal{R}$
    Wait for $T$ time before rescan
    When $T$ time expires, automatically trigger rescan. Re-run the algorithm to Step One
end for
```

cities in $\mathcal{A}_j$. Update the scan tracker table $\mathcal{H}$ with the scan id, the id of $\mathcal{D}_l$ and the current time stamp. The data in $\mathcal{H}$ is for reporting and monitoring the system performance. After scanning all the values in $\mathcal{D}$, set all the values in $\mathcal{D}$ as scanned. Otherwise, if the value in z is less than the preset number of records $\mathcal{N}$, print the message "No data to scan". Return the values $\mathcal{D}$ of incidents scanned. Wait for the time T before re-triggering another scan. Finally, when the set time T expires, automatically trigger the scan and re-run the algorithm from step one.

Note that the scan processes for both countries and cities in an infinite loop; that is, when the scan process starts, the system continues to automatically scan at each T time until the process is manually terminated. The complexity of Algorithm 5 is $O(N^4)$ where N represents the size of the input data.

3.3.2 Data Duplicates

Our system collects hundreds of news data on a daily basis. The country and city scanners automatically filter incidents data. However, news data collected from media companies come with many duplicates that we need to depict. For instance: different news media (e.g. New York Times and CNN) can report the same incident; the same news media can provide many updates on the same incident; and the news media can discuss incidents that occurred recently (e.g. 2 days ago).

Example of Similar News:

Search key word: **Ferry**

Source: News York Times

Ferry Capsizes Off Philippines, Leaving Dozens Dead: [Thu, 02 Jul 2015 13:45:31 GMT]

Algorithm 5 City Scanner

Input: $\mathcal{N}$: Number of records to be stored before scan start; T: Wait time before rescan; $\mathcal{O}$: List of Country incidents $o_1, o_2, o_3, \ldots, o_n$

Output: List of Data Scanned $\mathcal{D}$.

```
for $q = 1$ to $\infty$ do
    $z = \sum \mathcal{O}$ not scanned
    if $z >= N$ then
        $\mathcal{C}$ = List of active Countries $\in \mathcal{O}$
        for $i = 1$ to $m_i$ do
            $\mathcal{A}$ = List of active Cities of Country $\mathcal{C}_i$
            for $j = 1$ to $n_j$ do
                $\mathcal{D}$ = Load data not scanned $\in \mathcal{O}$
                for $l = 1$ to $k_l$ do
                    $\mathcal{D}_l$ = Load data in the queue
                    $x = \sum \exists\ (C_i$ and $A_j) \in \mathcal{D}_l$
                    if $x \neq 0$ then
                        $\mathcal{W} = \mathcal{D}_l$
                        Set $\mathcal{D}_l$ as latest record in $\mathcal{W}$
                    end if
                end for
                $\mathcal{D} = \bigcap \mathcal{D}_l$
            end for
            Set $\mathcal{D}_l$ as scanned to prevent rescan (Optional)
        end for
        Update table $\mathcal{H}$ with scan id and current date time to track scans
        Set all $\mathcal{D}$ as scanned
    else
        Print ”No data to scan”
    end if
    Return $\mathcal{D}$
    Wait for $T$ time before rescan
    When $T$ time expires, automatically trigger rescan. Re-run the algorithm to Step One
end for
```

A ferry carrying 189 people overturned near a port in the central Philippines on Thursday, killing at least 36 people.

Source: News York Times

Dozens Killed as Ferry Capsizes Off Philippines: [Thu, 02 Jul 2015 09:13:00 GMT]

At least 34 people aboard the vessel, which was carrying 173, died, and by Thursday afternoon, 118 passengers were said to have been rescued.

Source: Baltimore Sun

Ferry capsizes in Philippines; 35 dead, many missing: [Thu, 02 Jul 2015 13:39:00 GMT]

A ferry carrying 189 passengers and crew capsized Thursday minutes after it left a central Philippine port in choppy waters, leaving at least 35 dead and 20 others missing, coast guard officials said.

Source: Baltimore Sun

Ferry capsizes in Philippines; 36 dead, 26 missing: [Thu, 02 Jul 2015 10:36:00 GMT]

A ferry carrying 189 passengers and crew capsized Thursday as it left a central Philippine port in choppy waters, leaving at least 36 dead and 26 others missing, coast guard officials said.

Source: CNN

Dozens die as ferry capsizes: [Thu, 02 Jul 2015 08:21:28 EDT]

Source: Reuters

Philippine ferry sinks, killing at least 36, but most passengers survive: [Thu, 02 Jul 2015 10:58:08 GMT]

Source: BBC

Deadly ferry sinking in Philippines: [Thu, 02 Jul 2015 08:00:59 GMT]

To minimize and potentially eliminate duplicates, we develop and integrate a software package (similarity analyzer) to our system to analyze the similarity between two or more text data using the cosine similarity formula. The similarity analyzer uses the output data of the country and city scanners O and W respectively as its input data. The results are the similarity scores between each input text data. This process is automatic; the similarity analyzer runs one time in every 24 hours and it covers a period of 48 hours of data at each run.

Algorithm 6 illustrates the tasks the similarity analyzer performs. The inputs of the algorithm are incidents data from Country O and City W scanners. The output of the algorithm is a matrix of Cosine ($\theta_{m,n}$) scores. The first step is to Cleanse the input data O and W, and store the values in the variable $\mathcal{A}$. Transpose the documents in $\mathcal{A}$ ($\mathcal{A}^T$). Create an $m \times n$ term-document matrix (tdm) $\mathcal{D}$ with the documents in $\mathcal{A}$. Create an $m \times n$ mapping matrix $\mathcal{M}$ which we use to map the values of the original documents $\mathcal{A}$ to the resulting matrix $\mathcal{M}$. Calculate the cosine similarity using the term-document matrix $\mathcal{D}$. Set the similarity threshold to $s = 0.35$. The similarity threshold is set to $s = 0.35$ as after analysis of multiple cosine results, 99% of similar news have a score above 0.35. Next, the system counts the number of rows in matrix ($\theta_{m,n}$) and stores the value in the variable m $= \theta_i$. Similarly, count the number of columns in matrix ($\theta_{m,n}$) and store the value in the variable n $= \theta_j$. For each row value $i = 1$ to m_i and for each column value $j = 1$ to n_j in $\theta_{m,n}$, assign the cosine value $\theta_{i,j}$ to κ. If the value of κ is greater or equal to the threshold s, print to screen "Similar: $\kappa >= s$". If the row number in M_i is not equal to the column number in M_j, return M_i, M_j, and κ. We do this verification ($M_i <> M_j$) to exclude data that maps to itself in the matrix where the values are always equal to one. Next, the system updates the similarity values in the destination table $\mathcal{S}$ with M_i, M_j, and κ. If κ is below the threshold s, print "Not similar: $\kappa < s$". Finally, return the mapping matrix $M_{m,n}$, the cosine similarity matrix $\theta_{m,n}$, and the information updated $\mathcal{S}$ in the system.

We store the results of this algorithm in our system ($\mathcal{S}$) for user validation. When the analyst validates the data, the system automatically groups news based on the similarity values. The similarity grouping considerably facilitates the data validation process. The complexity of Algorithm 6 is $O(N^2)$ where N represents the

Algorithm 6 Text Similarity Using Cosine

Input: Incidents data from Country O and City W scanners

Output: Matrix of Cosine ($\theta_{m,n}$) scores; Mapping matrix $M_{m,n}$; System Updates $\mathcal{S}$

```
$\mathcal{A}$= Cleansed $O + W$
$\mathcal{A} = \mathcal{A}^T$
Create an $m \times n$ term-document matrix ($tdm$) $\mathcal{D}$ with ($\mathcal{A}$)
Create an $m \times n$ mapping matrix $\mathcal{M}$ for ($\mathcal{A}$)
Calculate the similarity = Cos($\theta$) = $\frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2 \times} \sqrt{\sum_{i=1}^{n} (B_i)^2}}$, with the $tdm$ ($\mathcal{D}$)
$s = 0.35$; the similarity threshold
m = $\theta_i$; matrix row number count
n = $\theta_j$; matrix column number count
for $i = 1$ to $m_i$ do
    for $j = 1$ to $n_j$ do
        $\kappa = \theta_{i,j}$
        if $\kappa >= s$ then
            print "Similar: $\kappa >= s$"
            if $M_i <> M_j$ then
                return $M_i$
                return $M_j$
                return $\kappa$
                Update system: set similarity values $\mathcal{S}$ with $M_i$, $M_j$, and $\kappa$
            end if
        else
            print "Not similar: $\kappa < s$"
        end if
    end for
end for
return $M_{m,n}$
return $\theta_{m,n}$
return $\mathcal{S}$
```

size of the input data.

3.3.3 Data Validation User Interface

Our system automatically calculates the risk level values without manual intervention. However, we provide the option to users or experts to confirm incidents as terrorist attacks. We design our system to provide user data validation which gives control and flexibility to analysts to decide on incidents that qualify as terrorist incidents as per their definition. This flexibility makes our system usable by a wide range of risk analysts. Our previous algorithms [4, 5] namely country and city scanners, and duplicates identification [6] do tremendous tasks to alleviate the manual validation efforts. The analyst can access the data validation user interface any time to perform different tasks.

We develop the algorithm 7 to present the filtered incidents data to the analyst for validation. The algorithm performs four main tasks: **apply** user updates for non similar news data; apply user updates to **confirm** news similarity; conduct regular searches (**doSearch**); and conduct searches using **similar** news data. The inputs of the algorithm are: $\mathcal{O}$: the country incidents data; $\mathcal{S}$: the output values of the similarity algorithm [6]; α: is one of these key values: $apply, doSearch, confirm, Similar$; $dateStart, dateEnd, cName$: search filters; o: option category. The output of the algorithm is the data validated $\mathcal{O}$ and $\mathcal{S}$. Both tables $\mathcal{O}$ and $\mathcal{S}$ receive the updates.

The first step of the algorithm is to check the values of α against the following values: $apply, doSearch, confirm$. If one of those values is set, it indicates the screen is being reloaded (e.g. it is not the first display of the user interface), and thus captures the search interval start date *dateStart*, end date *dateEnd*, the country name *cName*, and category option o loaded on the screen. This verification helps the analyst not to re-enter the same values at each reload. Next, on submit of the data form, the system provides one of the three values of α: *apply*, *doSearch*, or $confirm$. If α is equal to *apply*, count the number of records loaded and store the value in the variable r. For each record in r_i, store the values of o_i and s_i. The variable o_i contains the information of the option the analyst selects (e.g. false alarm, confirmed, etc.), and s_i is the unique incident identifier in our system. Next, the system updates the table $\mathcal{O}$ with the option value O_o to o_i that has a record identifier $O_s = s_i$. Print to screen "Record Id: $s_i->$ Updated to: $s_i = o_i$". After looping through all the values of r_i and updating $\mathcal{O}$, display "Update Completed Successfully".

Else, if the value of α is equal to **confirm**, count the total number of records and store the result in the variable r. For each record in r_j, count the number of records in $\mathcal{O}$ for which the option is equal to 4 (i.e. similar), and the data unique id O_s is equal to s_j. Store the count value in the variable c. If c is not equal to zero, update the incidents data value in $\mathcal{O}$, set O_o to o_j that has the unique identifier O_s equal to s_j. If the value of the option o_j is equal to 4, update the values in the similarity table $\mathcal{S}$, set S_o to o_j that has the unique record identifier S_s equal to s_j. Print to screen "Record Id: $s_j - >$ Updated to: $s = o_j$". When the updates are complete, print to screen "Update Completed Successfully".

If α is equal to **doSearch**, and if the option o is not equal to 4 (i.e. similar news), load in the variable Z the values country names $cName$, options o, titles, links, and descriptions from incident table $\mathcal{O}$ satisfying the following search conditions: dates between *dateStart* and *dateEnd*. For each value in Z, load Z_k on the screen.

If the option o is equal to 4, load in the variable Z the values country names $cName$, options o, titles, links, and descriptions from incident table $\mathcal{O}$ and similarity table $\mathcal{S}$ satisfying the following search conditions: dates between *dateStart* and *dateEnd*. For each value in Z, load Z_k on the screen. The complexity of Algorithm 7 is $O(N)$ where N represents the size of the input data.

Note that the analyst needs to validate incidents data only one time, then the system automatically includes it in the risk calculation for both countries and cities.

3.3.3.1 Data Validation Options

When the system automatically classifies incoming news data as terrorist incidents, the counter-terrorism analyst has the following five options (categories) to classify incidents:

1. Unconfirmed: News data automatically classified by the system as terrorist incident, but not yet confirmed by the analyst.

2. Confirmed: Incidents data validated by the analyst as terrorist incident.

3. False Alarm: News data automatically classified by the system as terrorist incident, but confirmed by the analyst as not an incident.

4. Similar: News data automatically classified by the system as terrorist incident,

Algorithm 7 Custom Incidents Data Validation

Input: $\mathcal{O}$: Incidents Data; $\mathcal{S}$: Filtered Similarity Data; α: apply OR doSearch OR Confirm; dateStart; dateEnd; cName; Option o

Output: Data Validated $\mathcal{O}$ and $\mathcal{S}$

```
if ∃ apply or doSearch or Confirm then
    capture dateStart; dateEnd; cName; Option $o$ on current User Interface
end if
if $\alpha$ == apply then
    $r$ = total records on screen
    for $i = 1$ to $r_i$ do
        Update $\mathcal{O}$ set $O_o = o_i$ for $O_s = s_i$
    end for
else if $\alpha$ == confirm then
    $r$ = total records on screen
    for $j = 1$ to $r_j$ do
        $c = \sum$ (data $\mathcal{O}$ for $O_o = 4$ and $O_s = s_j$); 4 = Similar
        if $c <> 0$ then
            Update $\mathcal{O}$ set $O_o = o_j$ for $O_s = s_j$
            if $o_j == 4$ then
                Update $\mathcal{S}$ set $S_o = o_j$ for $S_s = s_j$
            end if
        end if
    end for
else if $\alpha$ == doSearch then
    if $o <> 4$ then
        $Z$ = cName, $o$, title, description ∃ between $dateStart$ and $dateEnd \in \mathcal{O}$
        for $k = 1$ to $l_k$ do
            Load on Screen: $Z_k$
        end for
    end if
else if $o == 4$ then
    $Z$ = cName, $o$, title, description ∃ between $dateStart$ and $dateEnd \in \mathcal{O}$ and $\mathcal{S}$
    for $k = 1$ to $l_k$ do
        Load on Screen: $Z_k$
    end for
end if
```

but validated by the analyst as a duplicate.

5. Watch: News data automatically classified by the system as terrorist incident, but the analyst needs to verify with other analysts if the incident qualifies as terrorist incident. This option can be updated to any of the four options after verification.

3.3.3.2 Notifications

The analyst can setup alerts to notify her of new incoming incidents. Our notification module currently sends alerts to the incidents data validation user interface and the search page. The alerts can be redirected to emails, mobile devices, or other systems.

The algorithm 8 illustrates the tasks our notification module performs. The inputs of the algorithm are: the potential terrorist incidents data $\mathcal{O}$ (automatically classified by the country scanner algorithm 4); and the time t to re-trigger the notification job. The output values are: the list of notifications sent $\mathcal{Z}$; and the list of notification destinations $\mathcal{Q}$. The first step of the algorithm is to count the number of new incoming news in $\mathcal{O}$, and store the count result in the variable κ. If the value of κ is not equal to zero, load incoming terrorist incidents data from $\mathcal{O}$ and store the results in the variable $\mathcal{Z}$. For each incidents data in Z_i, send notification for it to the destination Q. Update the incident table $\mathcal{O}$ set the notification flag of the incident record O_i to one (1) indicating the notification is sent for O_i. Return the list of notifications Z along with the corresponding destinations Q. Print to screen "Notifications sent successfully". Then the system waits for the preset t time before automatically re-triggering the notification job. When t time expires, re-trigger the notification job and re-run the algorithm from step one.

The notification process runs in an infinite loop and can be manually terminated by the analyst or the system administrator.

The complexity of Algorithm 8 is $O(N^2)$ where N represents the size of the input data.

Algorithm 8 Notifications

Input: Potential Incidents Data $\mathcal{O}$; Time t to re-trigger notification job.

Output: Notifications Sent $\mathcal{Z}$; Notifications Destinations Q

```
 1: for $\beta = 1$ to $\infty$ do
 2:     $\kappa = \sum$ Number of new incoming terrorist incident news in $\mathcal{O}$
 3:     if $\kappa <> 0$ then
 4:         $Z$ = Load incoming terrorist incident in $\mathcal{O}$
 5:         for $i = 1$ to $n_i$ do
 6:             Send notification $Z_i$ to destination $Q$
 7:             Update $\mathcal{O}$ set notification flag = 1 for $O_i$
 8:         end for
 9:         return Notification list $Z$ along with $Q$
10:         print Notifications sent successfully
11:     end if
12:     Wait for $t$ time before automatically re-triggering notification job
13:     When $t$ time expires, re-trigger notification job
14: end for
```

3.4 Real Time Terrorism Risk Model

To calculate the real time terrorism risk values, we develop a terrorism risk model and apply it our real time terrorist incidents data.

The risk levels of countries and cities partly depend on their terrorist incidents frequency. The risk of a country or particular location is higher when it undergoes many criminal and terrorist incidents. For our research purpose, we consider incidents that qualify as terrorist incidents and other types of data such as political issues.

Furthermore, time is an important factor in determining the risk of a location as it allows one to represent the current situation. For instance, a location that undergoes frequent criminal incidents in 2000 is considered high risk in that year. However, if no incidents occur at the same location in 2015, it is considered safer.

Based on those two assumptions, we develop the *theories* below to derive a terrorism risk model to calculate the real time risks of countries and cities.

3.4.1 Theories for Terrorism Risk Model

Theory 1: News Filtering and Incident Identification.

Terrorist incidents news include terms like: bomb, explosion, suicide, etc. To retrieve all the terms that occurred frequently in terrorist incidents data, we applied our first method that uses LDA, and TensorFlow - Word2Vec, and Swivel [159] on our data to extract the best terms.

Using LDA, Word2Vec and Swivel [159] on the historical terrorist attacks data of START and our real time data, we identified highly frequent key words in terrorist attacks data. Our system utilizes the keywords to scan incoming data to filter and classify potential terrorist attacks or events that can lead to terrorist attacks. The automatic filtering using the keywords eliminates the need for manual processing.

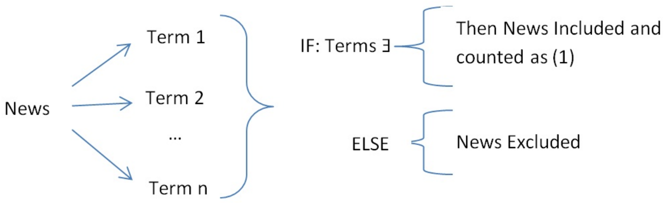


Figure 3.2: News Inclusion Criteria

A few of the resulting keyword terms are: Suicide, Terror, Islamic, Bomb, Shoot, Explosion, Blast, Kill, Qaeda, Hezbollah, Jihad, and Boko Haram.

Theory 2: Time Factor

Recent terrorist incidents have higher impact on the risk than old incidents.

Theory 3: Frequency Factor

The terrorism risk level of countries and cities is based on the frequency of terrorist incidents occurring in those particular locations.

As mentioned above in *theory 1*, not all the news data that the system downloads are included in the risk level calculation. We define two main inclusion criteria listed below.

3.4.2 Inclusion Criteria

3.4.2.1 Criteria 1:

As illustrated on *Figure 3.2*, only news data that contain our predetermined terms such as suicide or explosion are included.

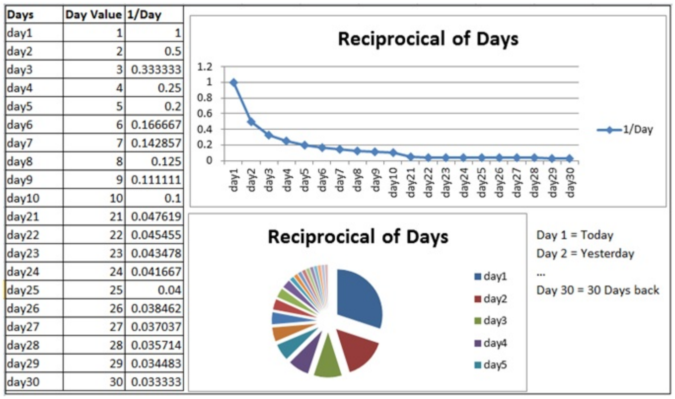
3.4.2.2 Criteria 2:

Given that we collect hundreds of news data on a daily basis, while calculating the risk for a location, we limit the data usage by setting the inclusion period to 30, 60 days data, etc. In fact, we decide to make our model flexible as experts may have different arguments on what is the ideal period of data to consider to calculate an accurate risk level. The inclusion period has an impact on the risk level values and how the risks should be interpreted. For example, the risk level value 20% using 60 days incidents data has a different meaning than a risk level value 20% using 90 days incidents data. Therefore, the results of the terrorism risk using 60 days incidents data should be interpreted differently than the terrorism risk using 90 days incidents data. For this research, we decide to use 60 days incidents data to calculate the risk level as we believe 60 days data encompass enough information to determine the real time terrorism risk level.

Considering the theories and inclusion criteria, we use the time and the frequency factors to derive the real time terrorism risk model.

3.4.3 Time Factor

Time is an important determinant of our risk model and it is represented by θ, where θ is the number of days past the incident. The risk will only include incidents that occurred in the preset t time period (e.g. previous 30 days, 60 days incidents data). According to our *theory 2*, recent incidents have higher impact on the risk level than older incidents. We use the reciprocal of θ to attribute higher weight



Days	Day Value	1/Day
day1	1	1
day2	2	0.5
day3	3	0.333333
day4	4	0.25
day5	5	0.2
day6	6	0.166667
day7	7	0.142857
day8	8	0.125
day9	9	0.111111
day10	10	0.1
day21	21	0.047619
day22	22	0.045455
day23	23	0.043478
day24	24	0.041667
day25	25	0.04
day26	26	0.038462
day27	27	0.037037
day28	28	0.035714
day29	29	0.034483
day30	30	0.033333

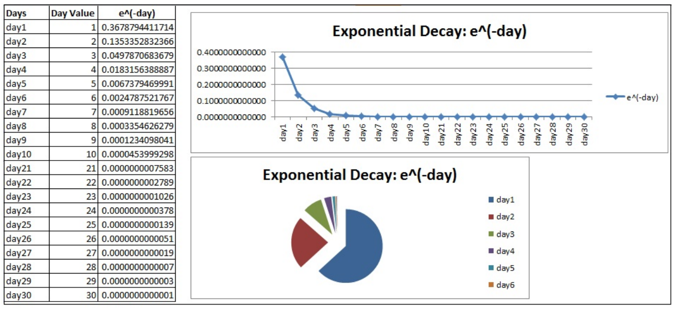
Figure 3.3: Reciprocal graph

values to recent incidents and lower weight values to older ones. *Figure 3.3* shows how the values are attributed based on the time the incidents occur.

We represent time by θ, the reciprocal of θ is $\frac{1}{\theta}$. The days are counted backward that is the current date is $\frac{1}{1}$, one day in the past (current day plus one day) is $\frac{1}{2}$, two days in the past is $\frac{1}{3}$, three days in the past is $\frac{1}{4}$ and so on.

Given θ is the start date and t the end date with n number of days between the two dates. The events from day θ to day t days is as follows: $\sum\limits_{i=t}^{\theta} \frac{1}{\theta-i+1}$

The exponential decay formula in *Figure 3.4* is a potential substitute for the reciprocal in the time factor, however we decide to use the reciprocal instead as its values decrease more slowly. As opposed to exponential decay, with the reciprocal, the past incidents do not lose their values too quickly, therefore, incidents lose their values at a steady pace.



Days	Day Value	e^(-day)
day1	1	0.3678794411714
day2	2	0.1353352832366
day3	3	0.0497870683679
day4	4	0.0183156388887
day5	5	0.0067379469991
day6	6	0.0024787521767
day7	7	0.0009118819656
day8	8	0.0003354626279
day9	9	0.0001234098041
day10	10	0.0000453999298
day21	21	0.0000000007583
day22	22	0.0000000002789
day23	23	0.0000000001026
day24	24	0.0000000000378
day25	25	0.0000000000139
day26	26	0.0000000000051
day27	27	0.0000000000019
day28	28	0.0000000000007
day29	29	0.0000000000003
day30	30	0.0000000000001

Figure 3.4: Exponential Decay

3.4.4 Frequency Factor

The system captures the incidents (terrorist attacks or major events) in real time for each country and city. Each incident is counted only once. The risk score on a given day is measured based on the number of incidents the have occurred since n days prior to the present day. Further we use a decay function to attach more weight to the recency of the events. Hence, the total number of events that have occurred t days prior to the present day is: $\sum_{i=t}^{\theta} \lambda_i$

To get the real time risk level of a particular location, we take the product of the frequency and the time factors for that location.

Therefore, the terrorism risk level based on t days is:

$$\sum_{i=t}^{\theta} \lambda_i \cdot \sum_{i=t}^{\theta} \frac{1}{\theta - i + 1}$$

Finally, the Terrorism Risk Model is:

$$\text{Risk} = \sum_{i=t}^{\theta} \left[\lambda_i \left(\frac{1}{\theta - i + 1} \right) \right]$$

The value of t can be set to include the desired range of days.

We apply this risk model to our real time data to calculate the real time terrorism risk levels of countries and cities.

For instance, given 60 days incidents of a location as: day 60, day 59, day 58, day 57, . . . , day 5, day 4, day 3, day 2, day 1 (current day). And 5 incidents occurred during the 60 days:

- 1 incident on day 41
- 1 incident on day 20
- 2 incidents on day 14
- 1 incident on day 9

The risk for the current day of that location is calculated as:

$1(\frac{1}{41}) + 1(\frac{1}{20}) + 2(\frac{1}{14}) + 1(\frac{1}{9}) = 32.84$

We present the country and city level terrorism risk values using this risk model in Chapter IV.

We published this methodology [66] in the proceedings of the IEEE Conference on Technologies for Homeland Security (HST'2015), Boston, April 2015.

3.5 Enhanced Risk Model

Before stepping into developing a terrorist incident prediction model, we first need to describe how we addressed limitations of our terrorism risk model.

Frequency Factor: Currently, our risk model sums incidents on a daily basis, however, when we project the risk level with such frequency calculation, our results get skewed. Thus we need to improve it to best fit our terrorist incident prediction model.

Time Factor: We currently use the reciprocal of the time θ ($\frac{1}{\theta}$). However, the time curve of the reciprocal decreases rapidly, thus we decided to update it to make it decrease at a slower pace.

Normalization: Our risk model lacks upper and lower boundaries. This poses a major limitation as we cannot make a difference between a risk level value 10 and 50. It is clear that 50 is higher risk in this comparison, but without maximum and minimum boundaries, one cannot situate himself in terms of the meaning of the risk value. Therefore, we normalized our terrorism risk model to include upper and lower boundaries.

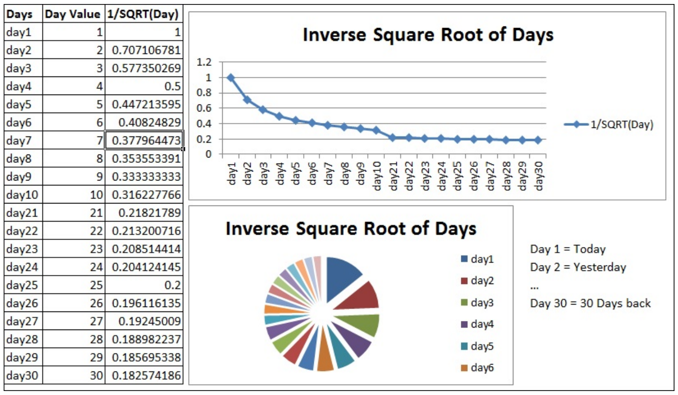
We provide details on these improvements and the enhanced terrorism risk model in this section.

3.5.1 Frequency Factor

For the purpose of this research, we are only interested in knowing if an incident occurs or not in a given day. Therefore, instead of summing up incidents on a daily basis which can result in more than one incident per day per location, we change the frequency logic to the following:

- A location can have a minimum of zero (0) incident per day.
- A location can have a maximum of one (1) incident per day.

Therefore, even when more than one (1) incident occurs in a country or city in the same day, we evaluate the combined incidents values as one (1).



Days	Day Value	1/SQRT(Day)
day1	1	1
day2	2	0.707106781
day3	3	0.577350269
day4	4	0.5
day5	5	0.447213595
day6	6	0.40824829
day7	7	0.377964473
day8	8	0.353553391
day9	9	0.333333333
day10	10	0.316227766
day21	21	0.21821789
day22	22	0.213200716
day23	23	0.208514414
day24	24	0.204124145
day25	25	0.2
day26	26	0.196116135
day27	27	0.19245009
day28	28	0.188982237
day29	29	0.185695338
day30	30	0.182574186

Day 1 = Today
Day 2 = Yesterday
...
Day 30 = 30 Days back

Figure 3.5: Inverse Square Root Graph

3.5.2 Time Factor

We use the reciprocal in our risk model to incorporate the time factor. However, similar to the exponential decay, the reciprocal values converge toward zero (0) too quickly. For instance, on the second day of an incident, it loses 50% of its value $\left(\frac{1}{2} = 0.5\right)$

Therefore, we replace the reciprocal of θ $\left(\frac{1}{\theta}\right)$ with the inverse of the square root of θ $\left(\frac{1}{\sqrt{\theta}}\right)$. The inverse of the square root of θ is shown on *Figure 3.5*, where the values converge toward zero (0) at a slower pace which indicates that past incidents lose their values at a slower pace. The inverse of the square root allows us not to over evaluate day 1 (current day) incidents over past but recent incidents.

3.5.3 Normalization

There are no upper and lower boundaries to our terrorism risk level values. In consequence, the risk level values can increase indefinitely. Furthermore, without boundaries, we cannot associate meanings to the risk values.

We add a normalization factor to the risk model to control the risk values. The normalized risk is always between 0 and 100.

To normalize the risk value, we divide our risk model by the sum of the preset time t values, that is the sum of $\left(\frac{1}{\sqrt{\theta}}\right)$ for the period t time *whether an incident occurs or not.*

Finally, we multiply the risk value by 100, to raise its value up to one hundred.

Combining the enhancements above, we rewrite our terrorism risk model as follows:

$$\text{Enhanced Risk} = 100 \times \sum_{i=t}^{\theta} \left(\frac{\lambda_i \left(\frac{1}{\sqrt{\theta - i + 1}} \right)}{\frac{1}{\sqrt{\theta - i + 1}}} \right)$$

3.6 Weighted Risk Model

We developed a weighted risk model that includes the following:

1. Instead of using 0 and 1 as daily incident values, allow multiple incidents in the same location in a day.

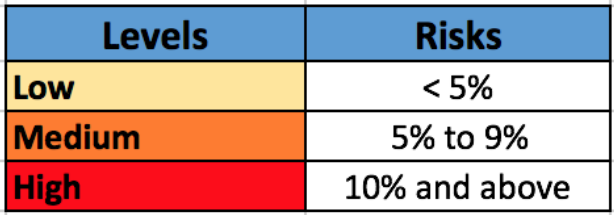
2. Assign different weights (W) to each incident based on the number of frequency, and the location that was attacked. The value of W is one, if only one incident occurred in a given day, and the value of W is two where there are two or more incidents at a location.

Our Weighted Risk Model is:

$$\text{Weighted Risk} = 100 \times \sum_{i=t}^{\theta} \left(\frac{W_i \lambda_i \left(\frac{1}{\sqrt{\theta - i + 1}} \right)}{W_i \left(\frac{1}{\sqrt{\theta - i + 1}} \right)} \right)$$

We discretized the risk levels after doing multiple evaluations of the risk values and the number of attacks that occur afterward. Our risk model is discretized into three classes: Low, Medium, and High where risk values below 0.5 is Low, 0.5 to 9 is Medium, and 10 and above is High.

The levels are assigned to the risk values as shown in figure 3.6.



Levels	Risks
Low	< 5%
Medium	5% to 9%
High	10% and above

Figure 3.6: Risk Level Standards

We published this methodology [67] in the proceedings of the IEEE Conference on Technologies for Homeland Security (HST'2016), Boston, April 2016.

3.7 Terrorist Incident Prediction

Predicting terrorist incidents is challenging and requires considering multiple factors such as recent and historical incidents data, time, and political incidents.

3.7.1 Markov Chain

Markov chains represent a class of stochastic processes of great interest for the wide spectrum of practical applications such as molecular biology, economics, pollutant dispersion models, PageRank, Web server HTTP requests, sales, and weather prediction. In particular, discrete time Markov chains (DTMC) permit one to model the transition probabilities between discrete states by the aid of matrices [155].

3.7.1.1 Markov Chain - Mathematical Concepts

A DTMC is a sequence of random variables $X_1, X_2, ..., X_n, ...$ characterized by the Markov property. The Markov property states that the distribution of the forthcoming state X_{n+1} depends only on the current state X_n and does not depend on the previous ones $X_{n-1}, X_{n-2}, \ldots, X_1$.

$P_r(X_{n+1} = x_{n+1}|X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P_r(X_{n+1} = x_{n+1}|X_n = x_n).$

The set of possible states $S = \{s_1, s_2, ..., s_r\}$ of X_n can be finite or countable and it is named the state space of the chain.

The chain moves from one state to another (this change is named either 'transition' or 'step') and the probability p_{ij} to move from state s_i to state s_j in one step is named transition probability:

$p_{ij} = P_r(X_1 = s_j|X_0 = s_i).$

The probability of moving from state i to j in n steps is denoted by:

$p_{ij}^{(n)} = P_r(X_n = s_j|X_0 = s_i).$

The probability distribution of transitions from one state to another can be represented into a transition matrix $P = (p_{ij})_{i,j}$, where each element of position (i, j) represents the transition probability p_{ij} . E.g., if $r = 3$ the transition matrix P is shown below:

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$

The distribution over the states can be written in the form of a stochastic row vector x.

We generated Markov transition matrix using algorithm 9. The first input of the algorithm is $\mathcal{R}$ the list of countries and their corresponding risk levels $\{Low, Medium, High\}$. The second input parameter of the algorithm is $\mathcal{M}$ the number of months to be included. In this study, we choose three months attacks as that represents enough data for our prediction. The third input parameter is $\mathcal{A}$ the list of attacks for all countries. The output of the algorithm is an $m \times n$ Markov transition matrix $\mathcal{T}$. The first step of the algorithm is to create an $m \times n$ matrix $\mathcal{X}_{ij}$ that stores the sum of attacks for the given risk level $\mathcal{R}_i$ and the month $\mathcal{M}_j$. The second step is to calculate the sum of each row $\mathcal{X}_i$ and store the values in Z_i. The attack matrix $\mathcal{X}_{ij}$ gets stored in $\mathcal{Y}$. Next, the Markov transition β_{ij} gets constructed with the attack matrix $\mathcal{Y}_{ij}$ divided by total of attack rows $\mathcal{Z}_i$. Then, the matrix β_{ij} gets stored in the transition matrix $\mathcal{T}$. The Markov transition matrix $\mathcal{T}$ is returned as output of the algorithm.

The complexity of Algorithm 9 is $O(N^4)$ where N represents the size of the input data.

We use the values of the attacks of the previous month as the initial state of the Markov prediction.

Algorithm 9 Algorithm for Generating Markov Transition Matrix

```
Input:
$\mathcal{R}$: List of countries and their corresponding risk levels $\{Low, Medium, High\}$
$\mathcal{M}$: The number of months $\{x_1, \ldots, x_m\}$
$\mathcal{A}$: Attacks
Output: An $m \times n$ Markov Transition Matrix $\mathcal{T}$.
for $i = 1$ to $m$ do
    for $j = 1$ to $n$ do
        $\mathcal{X}_{ij} = \sum \mathcal{A} \in [\mathcal{R}_i][\mathcal{M}_j]$
    end for
    $Z_i = \sum \mathcal{X}_i$
end for
$\mathcal{Y} = \mathcal{X}_{ij}$
for $i = 1$ to $m$ do
    for $j = 1$ to $n$ do
        $\beta_{ij} = \frac{\mathcal{Y}_{ij}}{\mathcal{Z}_i}$
    end for
    $\mathcal{T} = \beta_{ij}$
end for
Return Markov Transition Matrix $\mathcal{T}$
```

The next chapter IV presents experiments and results found using our systems and methodologies.

Chapter 4

Experiments and Evaluations

In this Chapter, we present the results of our methodologies developed to address our objectives below:

Terrorism Root Cause Analysis: Analyze historical and real time data to extract the root causes of terrorist incidents.

Clustering of Terrorist Groups: Cluster terrorist groups based on similarities in attack patterns.

Real Time Terrorist Incidents Data: Develop a system to automatically collect real time terrorist incidents data from reliable sources.

Terrorism Risk Model: Develop a new mathematical risk model to calculate the terrorism risk of locations. Moreover, calculate and plot on a dynamic map the impacts of terrorist groups on different locations (e.g. countries and cities).

Terrorist Incident Prediction: Predict terrorist incidents of different countries.

4.1 Terrorism Root Cause Analysis

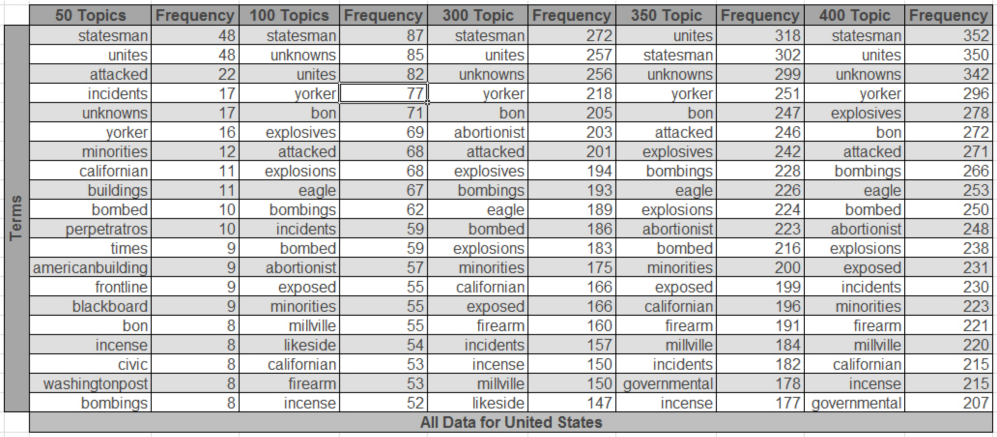
The objective of this method is to analyze historical and real time data to extract the root causes of terrorist incidents.

4.1.1 Experimental Results

Much work has been conducted in data mining and there is still a lot to do. Researchers developed methodologies in areas like unstructured text mining, privacy, classification, clustering, anomaly detection, graph partitioning, and spatial data mining to analyze different types of data. We developed novel methodologies in this research to analyze terrorism related data.

4.1.1.1 Analysis

In our analysis, the documents are the number of attacks and the terms are the number of words. In order to decide on the best number of topics, we generated 50, 100, 300, 350 and 400 topics on the same dataset. Fifty (50) topics generated a greater variety of words and its 12 most frequent words were similar to the 20 most frequent words in the other topics (100, 300, 350 and 400). The results are shown in Figure 4.1.



	50 Topics	Frequency	100 Topics	Frequency	300 Topic	Frequency	350 Topic	Frequency	400 Topic	Frequency
Terms	statesman	48	statesman	87	statesman	272	unites	318	statesman	352
	unites	48	unknowns	85	unites	257	statesman	302	unites	350
	attacked	22	unites	82	unknowns	256	unknowns	299	unknowns	342
	incidents	17	yorker	77	yorker	218	yorker	251	yorker	296
	unknowns	17	bon	71	bon	205	bon	247	explosives	278
	yorker	16	explosives	69	abortionist	203	attacked	246	bon	272
	minorities	12	attacked	68	attacked	201	explosives	242	attacked	271
	californian	11	explosions	68	explosives	194	bombings	228	bombings	266
	buildings	11	eagle	67	bombings	193	eagle	226	eagle	253
	bombed	10	bombings	62	eagle	189	explosions	224	bombed	250
	perpetratros	10	incidents	59	bombed	186	abortionist	223	abortionist	248
	times	9	bombed	59	explosions	183	bombed	216	explosions	238
	americanbuilding	9	abortionist	57	minorities	175	minorities	200	exposed	231
	frontline	9	exposed	55	californian	166	exposed	199	incidents	230
	blackboard	9	minorities	55	exposed	166	californian	196	minorities	223
	bon	8	millville	55	firearm	160	firearm	191	firearm	221
	incense	8	likeside	54	incidents	157	millville	184	millville	220
	civic	8	californian	53	incense	150	incidents	182	californian	215
	washingtonpost	8	firearm	53	millville	150	governmental	178	incense	215
	bombings	8	incense	52	likeside	147	incense	177	governmental	207
	All Data for United States									

Figure 4.1: Term Frequencies across Topics for United States

The 20 most frequent resulting topics were similar for the 100, 300, 350 and 400 topics. We used a MySQL database to count the frequency of terms and exported the top 20 most frequent terms.

Figure 4.2 shows the results we obtained using our method on terrorist attacks that occurred in United States from 1970 to 2010.

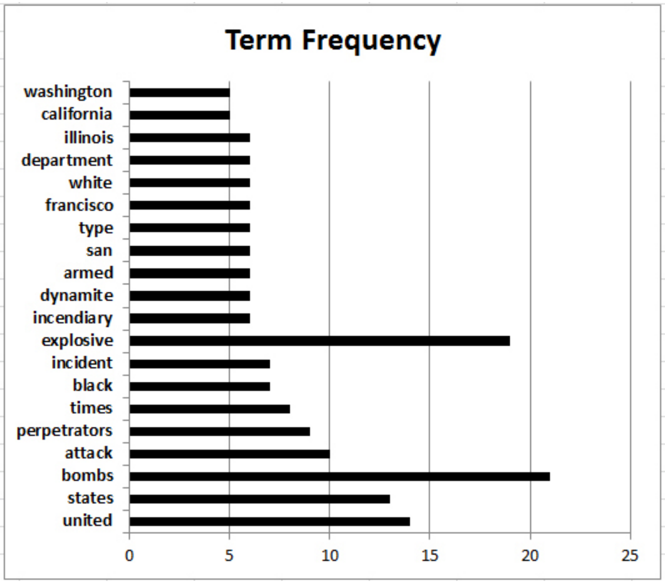


Figure 4.2: All Dataset for United States

After analyzing the top 20 most frequent terms in the topics, we mined useful information about the problems, reasons, motivations, effects, and weapons used in terrorist attacks. The following text shows the information gleaned from the dataset.

Terrorist attacks occurred in all the five continents, but more frequently in the following countries: United States, Iraq, Armenia, Colombia, India, and Baghdad. Moreover, many attacks took place in Europe and in the Caribbean. The Military and Police are involved in many attacks. Bombs, dynamites, firearms are mostly used in terrorist attacks, and some weapons remain unknown. Asiacell was victim of many terrorist attacks. In addition to damages caused to innocent civilians, businesses suffered the most in terrorist attacks. Some perpetrators remained unknown. In the 1980s, many assassinations occurred. In the 2000s, the group hizb was responsible of many attacks.

In the United States, abortion was one of the major causes of terrorist attacks especially in the 1980s and 1990s. Most of these attacks occurred in New York and California. Perpetrators used arson on clinics and sabotaged equipment. In the 1970s, in Washington, San Francisco, Illinois and California, ethnicity issues based on black and white was a major problem and caused many attacks. In the 1990s and 2000s, terrorists attacked mostly for respect of their rights and for liberation. The ALF (Animal Liberation Front) was responsible for many attacks in the 1990s; Jews figured in many attacks as well. Moreover, in general in the United States, minorities struggled for their rights.

Finally, bombs, dynamites, firearms, and incendiary devices were the most frequent weapons used in terrorist attacks.

4.1.1.2 Summary of the Methodology using LDA

In this study, we have described a novel method based on the Latent Dirichlet Allocation (LDA) to analyze text data on terrorist attacks that occurred between 1970 and 2010. We segmented the data set into decadal sub-sets, and then performed LDA analysis on each decade as well as the entire data. The results of this study have importance for homeland security, because they can be used to make decisions in dealing with future terrorist attacks. Moreover, it provides information for homeland security on the reasons behind these attacks. Our proposed method can be applied to any text data to generate topic models as patterns and subsequently terms that occur frequently in the topic models. A simple term-frequency based method using the raw text will not provide this information. Such information can be useful in other domains such as patents and product documentations for agencies such as the Food and Drug Administration (FDA) and patent offices that have to deal with big data. Methods such as these are also useful in domains such as healthcare where an abundance of text data exists.

The research discussed above provides a good introduction to the analysis of terrorist incidents. However, these methods are limited when it comes to predicting terrorist incidents in real time. Our methodologies and systems provide further means to analyze Big data and get higher precision and recall values on predicting incidents. In addition, our methodologies can be used to analyze and extract useful information from historical data on terrorism.

4.2 Clustering of Terrorist Groups

The objective of this methodology is to cluster terrorist groups based on similarities in attack patterns.

4.2.1 Experimental Results

Our method uses the Latent Semantic Indexing [33] to cluster attacks based on their similarities. We use 11 years terrorist incidents data collected by START from 2000 to 2011.

4.2.2 Analysis

In our experiment, we preprocess the data and then apply our method to it. The START [28] database contains 124 attributes. We retrieved the summary attribute only for our experiment. The summary contains the detailed information about who, when, where and how the attacks occurred. Also, we focused on more recent data which is more likely to have more useful information about current terrorism. As the nature and the techniques used by terrorist groups change over time, focusing on more recent attacks allows us to identify more up to date patterns. We choose the sample data for the United States only and for the periods of 2000 - 2005, 2006 - 2011, and 2000 - 2011.

The figures below are plots of the two vectors with the highest energy. The reference numbers are the unique identifier of each attack in the database. The year 2000 to 2011 has a total of 240 records, 2000 to 2005 has 169 records, and 2006 to 2011 has 71 records. As we grouped the figures in two clusters, the blue points and red points represent cluster $C1$ and cluster $C2$ respectively.

Figure 4.3 is the summary of the attacks that occurred from 2000 to 2011. Our method clustered together the attacks with similar patterns. That is, the attacks

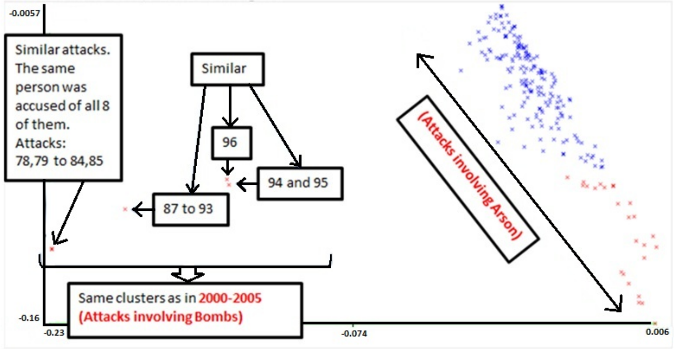


Figure 4.3: Clusters 2000 to 2011

that involved bombs are clustered on the same side and the attacks involving arson are clustered together on the other side. As mentioned in figure 4.3, the attacks with reference 78 to 85 are all clustered together and the same person was accused of all 8 attacks. The attacks with references 87 to 93 represent attacks with similar scenarios and with similar weapons. The attacks with references 94 and 95 are similar to each other as well. The attack with reference 96 is closely similar in weapon usage from reference 87 to 95. Overall, the attacks clustered together are more closely related in weapon usage or scenario. The position of the clustered references 87 to 93, 94 and 95, and 96 do not change even after changing the number of records. That makes our method robust to the dynamism of the English language where different people describe the same scenario using different words. However, the key words remain the same such as *bomb*, *arson*, *explosion*, *etc*. That is, even if a shorter or longer summary data were selected to be analyzed, it would not have an effect on the structure of the clusters. Similar attacks will always remain clustered together independent of changing the data size. As you may have noticed already, the similarity of the attacks is based on the similarity in the summary description.

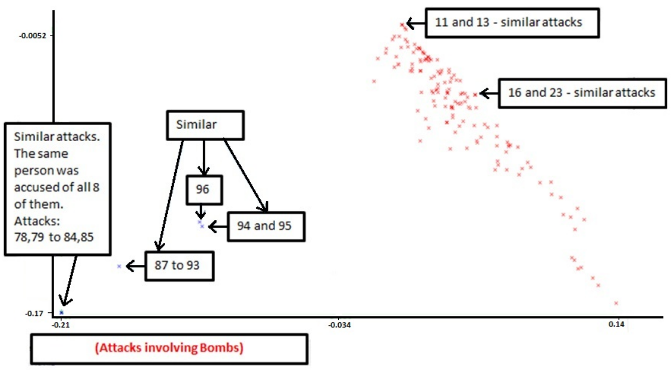


Figure 4.4: Clusters of 2000 to 2005

The left angle of figure 4.4 is the same as the one of figure 4.3. This supports our argument mentioned above that the number of records analyzed does not affect the performance of our method. We randomly picked two data points in the figure, the first point is 11 and the second point is 13. Both include arson and were caused by the same perpetrators. Similarly, 16 and 23 both include arson and were caused by the same terrorist group.

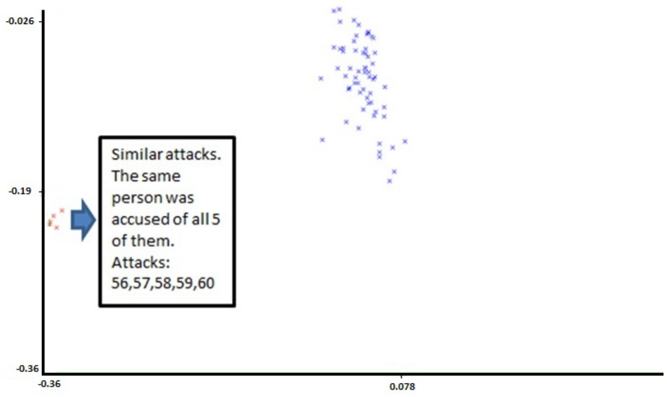


Figure 4.5: Clusters of 2006 to 2011

Figure 4.5 represent the data of 2006 to 2011. The cluster $C1$ (in red) contains 5 different attacks. The same person was accused of all these attacks.

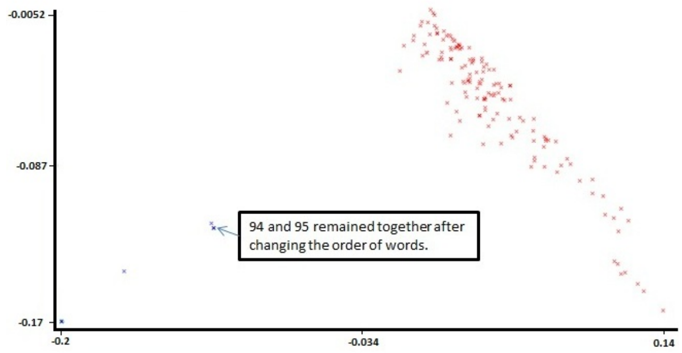
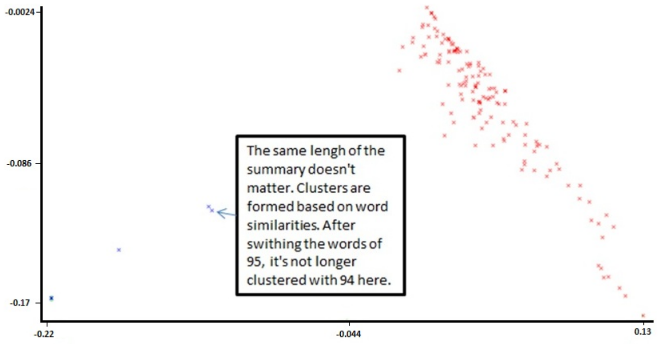


Figure 4.6: Clusters of 2000 2005 Word Order Switched

In the next two experiments, we add noise to the original data to test the stability of our method. In the first case, we changed the order of the words in figure 4.6, and in the second scenario in figure 4.7, we kept the same length of words but changed the words.

In figure 4.6, we test whether our method is robust to changing the positions of words or breaking the grammar rules. As we know that the attacks 94 and 95 are clustered together, we shuffled the positions of the words in the summary by making them random and not respecting the grammar logics. After applying our method to this mixed data, the clusters remained the same. The attacks 94 and 95 remained clustered together as they were before switching the positions of the words. Thus, the position of the words in the sentence does not affect the results of our method.

To test the robustness of our method, we kept the same length of words but changed the words (e.g. attacks to home). In this example, the length of the referenced summary 95 was 104 words. We replaced these 104 words by other 104 words randomly picked from the Internet. We plotted these results on figure 4.7. As a result, the attack with reference 95 was no longer grouped with 94 as they were



The same lengh of the summary doesn't matter. Clusters are formed based on word similarities. After swithing the words of 95, it's not longer clustered with 94 here.

Figure 4.7: Summary of 2000 to 2005 Words Changed but keeps the same Length

before. Thus, our method clusters only based on the similarities in the description of attack summaries.

4.2.3 Summary of the Methodology using LSI

Preventing terror attacks is a challenging task for the Department of Homeland Security (Homeland Security). Terrorist groups are very dynamic. They change locations, targets, weapons, and their names. Tracking the activities of such groups and countering or preventing their attacks is tedious. Past records can be thoroughly studied using technology to extract useful information that can be used to assist in counter-terrorism measures. In this research, we used the data collected by START [65] over the last decade to cluster the summaries of the attacks. Our method incorporates Latent Semantic Indexing (LSI) and successfully clustered attacks based on their similar patterns independent of the length of attack summaries and of the order or positions of words in attack descriptions. Our method can be used by Homeland Security to quickly cluster and categorize terrorist groups based on their attack patterns and weapon usage similarities. Our method can increase the efficiency of counter-terrorism measures and reduce the response time in categorizing a terrorist group in order to take the appropriate action.

4.3 Real Time Terrorist Incidents Data and Terrorism Risk Model

In this section, we present the results of the terrorism risk model applied to the real time terrorist incidents data.

4.3.1 Experimental Results

When the analyst (optionally) confirms terrorist incidents, the data are updated in the table $\mathcal{O}$. In case there is no manual interventions, the system automatically filters and stores relevant data in the table $\mathcal{O}$. In this experiment, we apply the terrorism risk model to the incidents data in $\mathcal{O}$. For each country, we select the incidents data for a preset time period (60 days), then apply the risk model to it. The system calculates the risk level for the selected country and it includes the most recent incidents data. Including recent incidents data guarantees the risk level is real time. Furthermore, the system displays the risk levels for each country on a dynamic map.

We implement the risk model using the algorithm 10. The inputs of the algorithm are: the list of terrorist incidents data from the table $\mathcal{O}$, the list of countries $\mathcal{C}$, the time period t (e.g. 60 days data), and the current date c. The outputs of the algorithm are: the list of countries $\mathcal{C}$ along with their risk levels $\mathcal{R}$. The first step of the algorithm is to load the terrorist incidents data from $\mathcal{O}$ into λ, include only incidents that have dates above or equal to the current date c minus the period t. This selection method includes the most recent incidents data up to $c - t$ date. Next, the system loads the list of active countries from $\mathcal{C}$. For each country in $\mathcal{C}_i$ calculate its risk level, and store the risk value in the variable $\mathcal{R}_i$. Group the country $\mathcal{C}_i$ and their risk values $\mathcal{R}_i$, and store the values in the variable β.

Return β containing the list of countries $\mathcal{C}$ and the risk values $\mathcal{R}$. Plot the countries and their corresponding risk values on a dynamic map. Highlight the country map with red gradients, use lighter red for low risk values and darker red for high risk values. The performance of Algorithm 10 is $O(N)$.

Algorithm 10 Real Time Country Risk Levels

Input: $\mathcal{O}$: Incidents data; $\mathcal{C}$: List of countries; t: Period to include; c: Current Date

Output: $\mathcal{C}$: List of countries; $\mathcal{R}$: Country risk levels

```
$\lambda$ = Terrorist incidents in $\mathcal{O} >= c - t$
$\mathcal{C}$ = Active countries in $\mathcal{C}$
for $i = 1$ to $n_i$ do
    $\mathcal{R}_i = \sum_{\theta=1}^{t} \left[ \sum_{\lambda=1}^{n} \lambda_\theta \cdot \left(\frac{1}{\theta}\right) \right]$ for country $\mathcal{C}_i \in \mathcal{O}_t$
    $\beta = \sum \mathcal{C}_i \bigcup \mathcal{R}_i$
end for
return $\beta$ containing the list of countries $\mathcal{C}$ along with their risk values $\mathcal{R}$
Plot $\mathcal{C}$ and $\mathcal{R}$ on dynamic map
Assign red gradients to countries $\mathcal{C}$ based on the risk values $\mathcal{R}$
```

Moreover, our system allows analysts to zoom into countries to view their city level terrorism risk. To implement the city level terrorism risk, we developed the algorithm 11 to calculate and plot the city level terrorism risk levels.

The inputs of the algorithm [11] are: the list of terrorist incidents data from the table $\mathcal{O}$, the list of countries $\mathcal{C}$, the period t incidents data to include, and c the current date. The output of the algorithm is the list of countries $\mathcal{C}$ with their cities $\mathcal{S}$ along with the corresponding risk levels $\mathcal{R}$.

The first step of the algorithm is to load the terrorist incidents data from $\mathcal{O}$ into the variable λ, include only incidents that have dates above or equal to the current date c minus the period t. This includes the most current incidents data up to $c - t$ date. Next, the system loads the list of countries from $\mathcal{C}$. For each country in $\mathcal{C}_i$, load its active cities $\mathcal{S}_j$. For each city in $\mathcal{S}_j$, compute its risk level, and store the result in the variable $\mathcal{R}_j$. Group the country $\mathcal{C}_i$, city $\mathcal{S}_j$ along with their risk values $\mathcal{R}_j$, and store the results in the variable β.

Return β containing the list of countries $\mathcal{C}$, cities $\mathcal{S}$, and the risk values $\mathcal{R}$. Plot the cities and their corresponding risk values on a dynamic map. Add markers to the city map with their risk values. The system displays the risk levels on mouse hover. The performance of Algorithm 11 is $O(N^2)$.

Our next objective is to determine the impact level of different terrorist groups on countries. We develop the algorithm [12] to calculate the impact level of terrorist groups on different countries. We apply our risk model on the incidents data of terrorism groups. The data validation criteria of countries also apply to the terrorist groups. In addition, we include only incidents that contain the name of the terrorist groups. Therefore, if the analyst (optionally) confirms news data as a terrorist

Algorithm 11 Real Time City Risk Levels

Input: $\mathcal{O}$: Incidents data; $\mathcal{C}$: List of countries; t: Period to include; c: Current Date

Output: $\mathcal{C}$: List of countries; $\mathcal{R}$: Country risk levels

```
$\lambda$ = Terrorist incidents in $\mathcal{O} >= c - t$
$\mathcal{C}$ = Active countries in $\mathcal{C}$
for $i = 1$ to $n_i$ do
    $\mathcal{S}$ = Active cities $\mathcal{S}$ in $\mathcal{C}_i$
    for $j = 1$ to $m_j$ do
        $\mathcal{R}_j = \sum_{\theta=1}^{t} \left[ \sum_{\lambda=1}^{n} \lambda_\theta \cdot \left(\frac{1}{\theta}\right) \right]$ for city $\mathcal{S}_j \in (\mathcal{C}_i$ and $\mathcal{O}_t)$
        $\beta = \sum \mathcal{C}_i \bigcup \mathcal{S}_j \bigcup \mathcal{R}_i$
    end for
end for
return $\beta$ containing the list of countries $\mathcal{C}$, cities $\mathcal{S}$, and risk values $\mathcal{R}$
Plot $\mathcal{S}$ and $\mathcal{R}$ on dynamic map
Add markers to cities $\mathcal{S}$ with the risk values $\mathcal{R}$
```

incident, and it does not contain the name of a terrorist group, we exclude that incident from the impact level data. Note that if there is no manual interventions from the analyst, the system automatically filters the data, and retrieves incidents data that contain the names of terrorist groups.

The inputs of the algorithm [12] are: the terrorist group name $\mathcal{G}$ we want to calculate the impact level for, the list of terrorist incidents data from the table $\mathcal{O}$, the list of countries $\mathcal{C}$, the period t incidents data to include, and c the current date. The outputs of the algorithm are: the terrorist group name $\mathcal{G}$, the list of countries $\mathcal{C}$ with the corresponding impact levels $\mathcal{R}$.

The first step of the algorithm is to load the terrorist incidents data from $\mathcal{O}$ into the variable λ, include only incidents that have dates above or equal to the current date c minus the period t $(c - t)$. Next, the system loads the list of active countries from $\mathcal{C}$. For each country in $\mathcal{C}_i$, calculate its impact level, and store the result in the variable $\mathcal{R}_i$. Regroup the terrorist group name $\mathcal{G}$, country $\mathcal{C}_i$, along with their impact values $\mathcal{R}_i$, and store the results in the variable β.

Return β containing the terrorist group name $\mathcal{G}$, the list of countries $\mathcal{C}$ and the impact level values $\mathcal{R}$. Plot the countries and their corresponding impact values on a dynamic map. Highlight the country map with red gradients, use lighter red for low impact values and darker red for high impact values. The performance of Algorithm 12 is $O(N)$.

Algorithm 12 Terrorist Groups Impact Levels

Input: $\mathcal{G}$: Terrorist group; $\mathcal{O}$: Incidents; $\mathcal{C}$: Countries; t: Period; c: Current date

Output: $\mathcal{G}$: Terrorist group; $\mathcal{C}$: Countries; $\mathcal{R}$: Impact levels

```
$\lambda =$ Terrorist incidents $\in \mathcal{O} >= c - t$
$\mathcal{C} =$ Active countries $\mathcal{C}$
**for** $i = 1$ to $n_i$ **do**
    $\mathcal{R}_i = \sum_{\theta=1}^{t} \left[ \sum_{\lambda=1}^{n} \lambda_\theta \cdot \left(\frac{1}{\theta}\right) \right]$ for group $\mathcal{G} \in (\mathcal{C}_i$ and $\mathcal{O}_t)$
    $\beta = \sum \mathcal{G} \bigcup \mathcal{C}_i \bigcup \mathcal{R}_i$
**end for**
**return** $\beta$ containing the group name $\mathcal{G}$, countries $\mathcal{C}$, and the impact values $\mathcal{R}$
Plot the countries $\mathcal{C}$ on a dynamic map along with impact levels $\mathcal{R}$
Assign red gradients to countries $\mathcal{C}$ based on the impact values $\mathcal{R}$
```

4.3.2 Analysis

We illustrate in *Figure 4.8*, the incident scanning process for countries and cities.

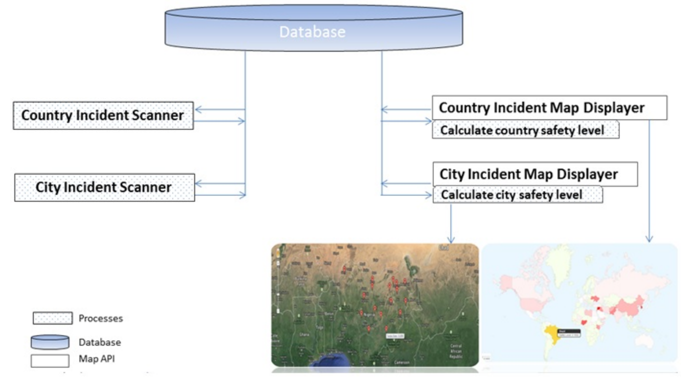

Database
Country Incident Scanner
City Incident Scanner
Country Incident Map Displayer
Calculate country safety level
City Incident Map Displayer
Calculate city safety level
Processes
Database
Map API

Figure 4.8: Incidents Scanners

4.3.2.1 Country Level Terrorism Risks

We illustrate the results of our country level risk algorithm [10] in *Figure 4.9*. The map represents the real time risk level of all countries undergoing terrorist incidents. Countries with darker risk colors have higher terrorism risk.

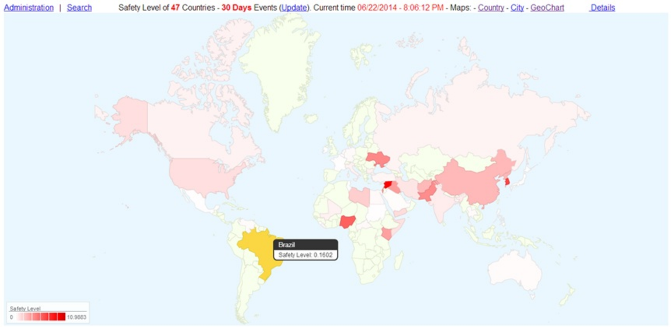
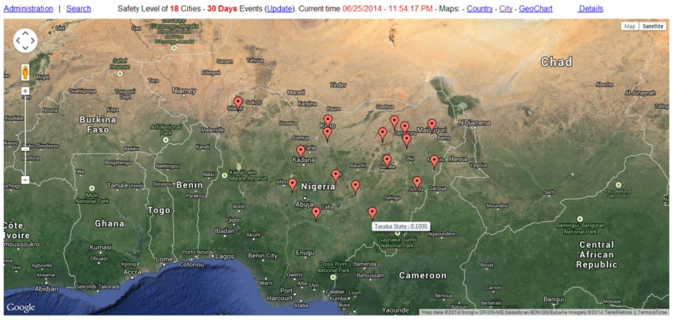

Administration | Search   Safety Level of 47 Countries - 30 Days Events (Update). Current time 06/22/2014 - 8:06:12 PM - Maps: - Country - City - GeoChart   Details

Figure 4.9: Country Level Terrorism Risks

We generated the geo-chart in *Figure 4.9* in June 2014. After analyzing the results of the country level terrorism risk on *Figure 4.9*, we indexed the following countries with higher terrorism risks: *Syria, South Korea, Nigeria, Ukraine, Pakistan, Israel, Iraq, Kenya, Afghanistan, and North Korea.* These countries are among the most dangerous places in the world in terms of terrorism risks.

4.3.2.2 City Level Terrorism Risks

The map in *Figure 4.10* represents the risk levels of cities of Nigeria in June 2014. These risk levels are calculated based on 30 days incidents data. A close analysis of the data shows that a total of 18 cities are affected by terrorists and most of the incidents are caused by the insurgency group Boko Haram.



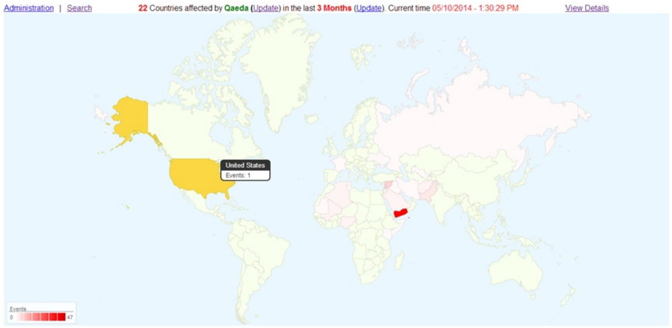
Administration | Search Safety Level of 18 Cities - 30 Days Events (Update). Current time 06/25/2014 - 11:54:17 PM - Maps: - Country - City - GeoChart Details

Figure 4.10: City Level Terrorism Risks

4.3.2.3 Terrorist Group Impact Levels

Using the real time data and the risk model, our system calculates the impact level of terrorist groups on countries. We use the name of the terrorist group Al-Qaeda as input of our algorithm [12], and three months terrorist incidents data. Our system returns the results in *Figure 4.11*, which represents the impact levels

Al-Qaeda had on different countries in the three-month time period. The darkness of the red color indicates the impact level of Al-Qaeda, and darker red represents higher impact. The country *Yemen* is the most affected by Al-Qaeda.



22 Countries affected by Qaeda (Update) in the last 3 Months (Update). Current time 05/10/2014 - 1:30:29 PM

Figure 4.11: Al-Qaeda Impact Levels on Countries

Note that so far, we apply our risk model on three different data: country, city, and terrorist groups. This suggests our risk model can be applied in other areas as well.

Additionally, we verified if the US terrorist alert [152] coincides with the risk indication of our method. However, after a year of monitoring, the US terrorist alert was not raised. After tracing back, we found that the last time the US terrorist alert was raised was in 2009. On its terror alert page [153], the Department of Homeland Security (DHS) noted that *"Raising the threat condition has economic, physical, and psychological effects on the nation."* Therefore, the terrorism alert is raised by DHS only under careful considerations.

4.4 Weighted Risk Model:

We applied the weighted risk model to Baltimore shooting data [154] that contain longitude and latitude information as follows:

- Calculate the risk level of different locations
- Each risk level was calculated using 30 days incidents data
- The risk levels were calculated for 01/01/2016 to 02/23/2016 for different locations
- Finally, we took the locations and perform the following:

  1. Calculate the distance radius between the locations and the incidents that occurred in the next 24 hours.

  2. The distance to the closest risk location from the incident is retained.

We analyzed the distance radius in miles between the locations marked as high risk and the locations where incidents occur in the next 24 hours.

4.4.1 Risk of Crime Prediction Statement

A crime will occur near the high risk areas in Baltimore within the next 24 hours.

4.4.2 Results

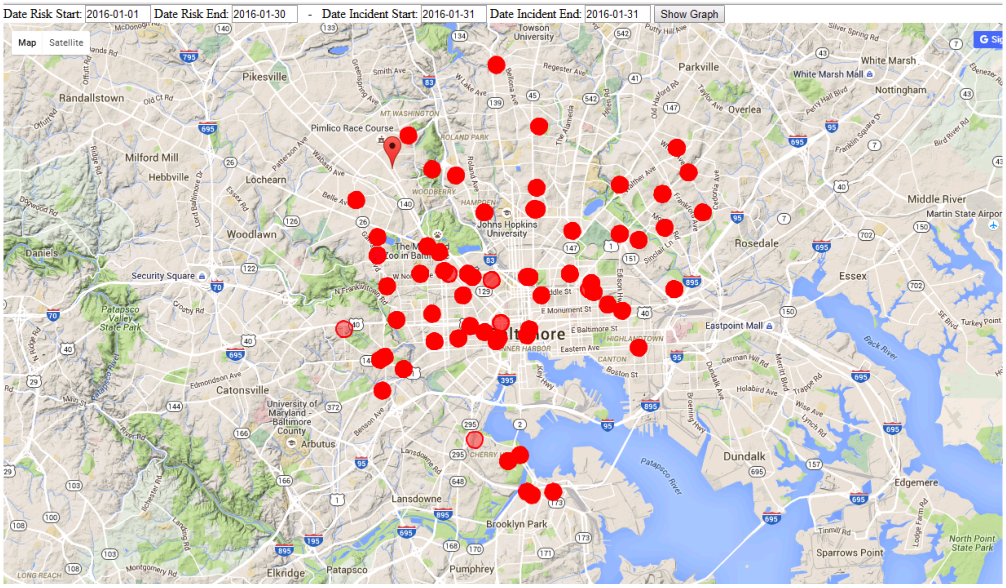
The results show that out of 84 incidents, 90% of them occurred within 0.5 miles from the closest high risk location.

Weighted Risk Model Results		
Incident Distance from Closest High Risk Location	Incidents	Percentage
Distance < 0.50 mile	75	**90%**
Distance >= 0.50 mile and < 1.00 mile	6	7%
Distance >= 1.00 mile and <=1.50 miles	3	3%
Total	**84**	**100%**

Table 4.1: Weighted Risk Model Results

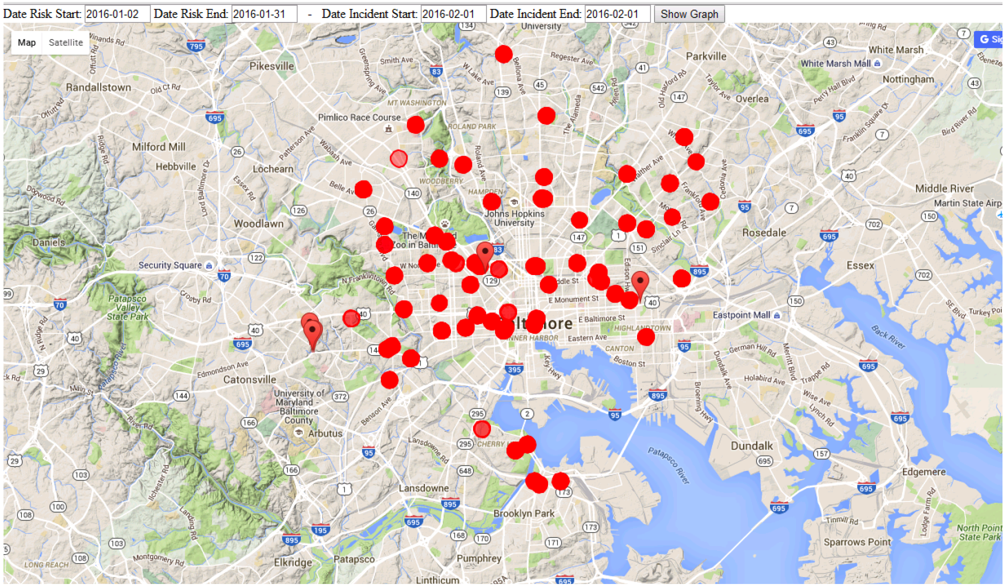
The maximum recorded distance between a new incident and the nearest risk location is 1.17 miles; and the shortest recorded distance is 0.011 mile - about 17 meters.

Figures 4.12 , 4.13 , and 4.14 show the high risk locations in Baltimore in which we predicted incidents will occur in the following 24 hours. The incidents are represented by a marker while the risk levels are represented by red circles.



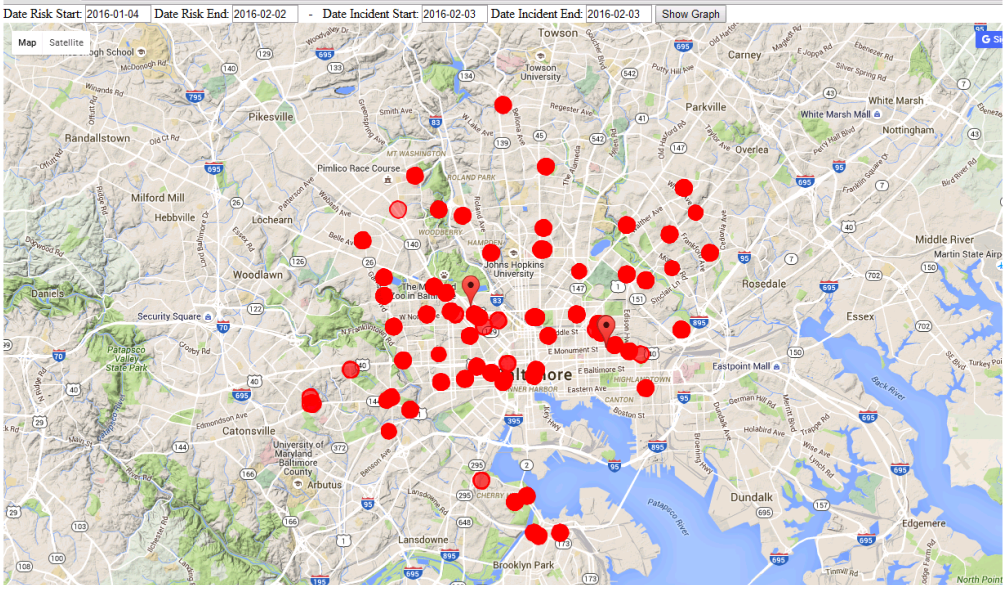
Date Risk Start: 2016-01-01 Date Risk End: 2016-01-30 - Date Incident Start: 2016-01-31 Date Incident End: 2016-01-31 Show Graph

Figure 4.12: Incidents on 01/31/2016



Date Risk Start: 2016-01-02 Date Risk End: 2016-01-31 - Date Incident Start: 2016-02-01 Date Incident End: 2016-02-01 Show Graph

Figure 4.13: Incidents on 02/01/2016



Date Risk Start: 2016-01-04 Date Risk End: 2016-02-02 - Date Incident Start: 2016-02-03 Date Incident End: 2016-02-03 Show Graph

Figure 4.14: Incidents on 02/03/2016

Out of 84 incidents and given that the weighted risk values are between 0 and 100:

- 63% of incidents occurred around locations that have risk values >= 50
- 37% of incidents occurred around locations that have risk values < 50

Those results suggest that high risk level locations undergo higher number of incidents.

Furthermore, we calculated the confidence interval for the distance between incident locations and the closest high risk area as follows:

Sample Data	0.029423251, 0.030039611, 0.065404006, . . .
Mean	0.299344523
Standard Deviation - Sample	0.237144419
Sample Size	84
Alpha	0.05
Confidence Interval - Based on sample SD	0.051463494
Upper Limit	0.351
Lower Limit	0.248

Table 4.2: Distance - Confidence Interval

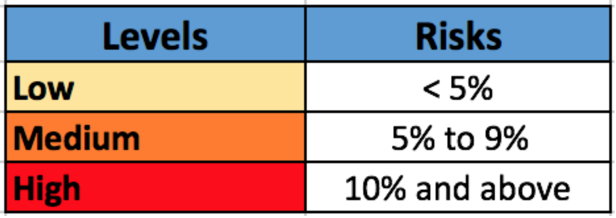
There is a **95%** chance the incident distance from our nearest high risk point is between **0.248** and **0.351 mile**.

After setting **Alpha = 0.01**, the confidence interval values changed to: There is a **99%** chance the incident distance from our nearest high risk point is between **0.231** and **0.368 mile**.

4.5 Terrorist Incident Prediction

Predicting terrorist incidents is challenging and requires considering multiple factors such as recent and historical incidents data, time, and political situations.

We defined standards to our risk model as Low, Medium, and High. The levels are assigned to the risk values as shown in figure 4.15.



Levels	Risks
Low	< 5%
Medium	5% to 9%
High	10% and above

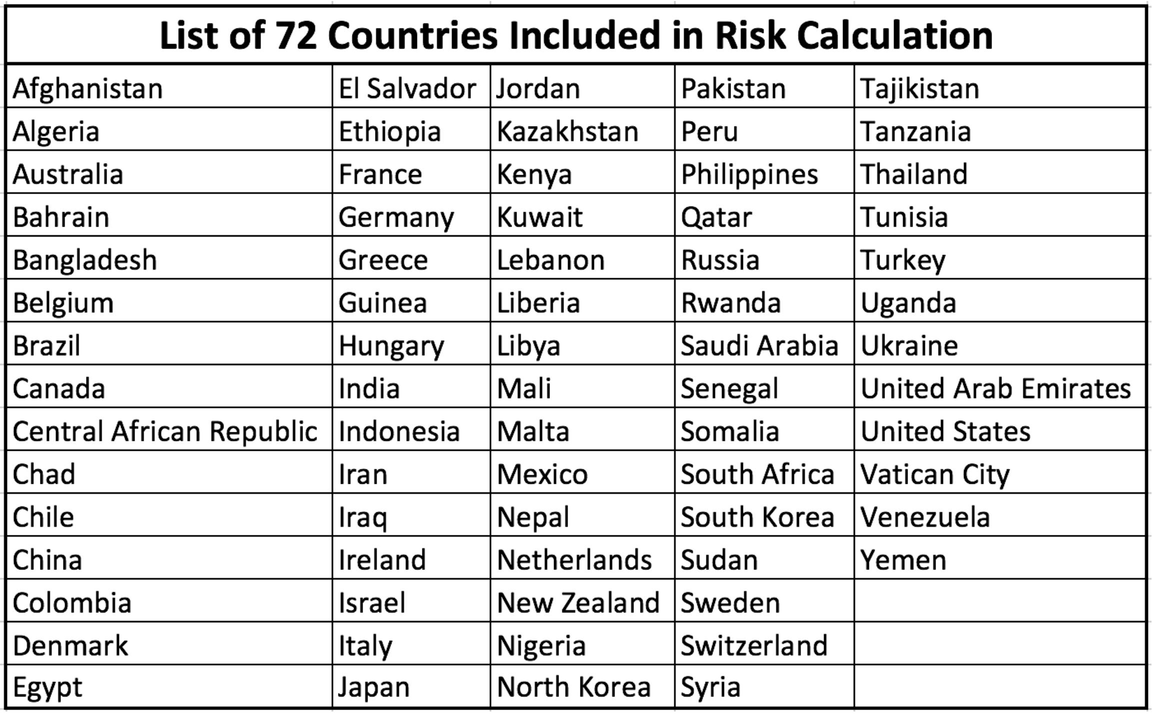
Figure 4.15: Risk Level Standards

4.5.1 Prediction - Statements

- When the risk is **High**, an incident will occur
- When the risk is **Low or Medium**, no incident will occur

4.5.1.1 Risk Prediction Calculation

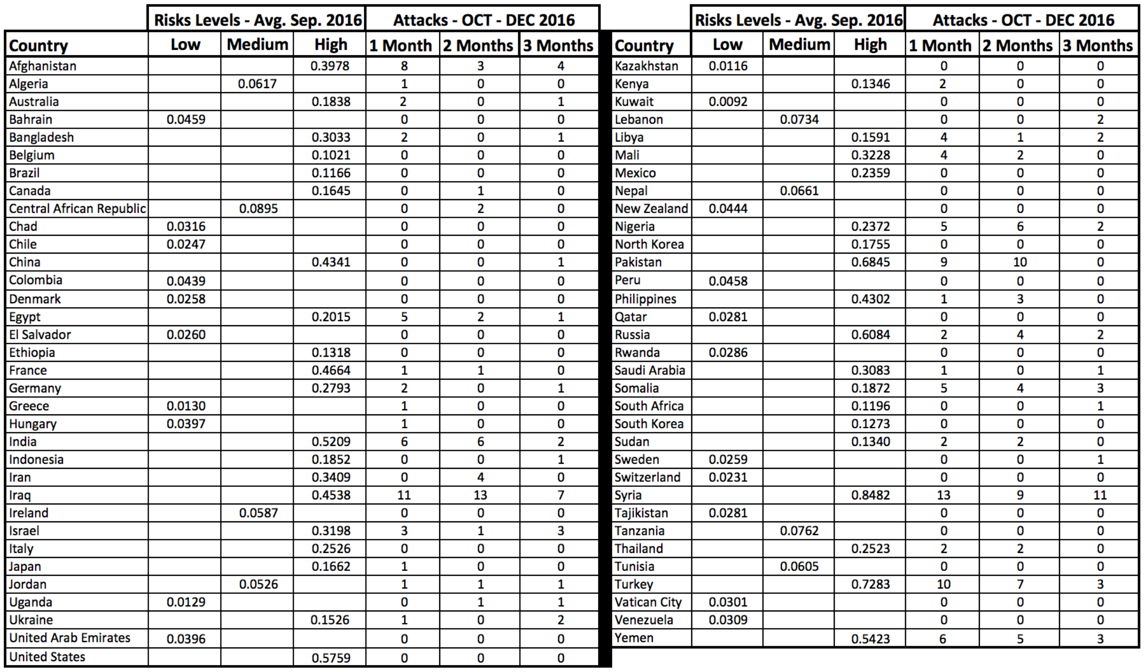
We process an equivalent of 10 months data for the list of countries in table 4.16 to validate our prediction methodology. The data included in the process was retrieved from our real time data collection system.



List of 72 Countries Included in Risk Calculation				
Afghanistan	El Salvador	Jordan	Pakistan	Tajikistan
Algeria	Ethiopia	Kazakhstan	Peru	Tanzania
Australia	France	Kenya	Philippines	Thailand
Bahrain	Germany	Kuwait	Qatar	Tunisia
Bangladesh	Greece	Lebanon	Russia	Turkey
Belgium	Guinea	Liberia	Rwanda	Uganda
Brazil	Hungary	Libya	Saudi Arabia	Ukraine
Canada	India	Mali	Senegal	United Arab Emirates
Central African Republic	Indonesia	Malta	Somalia	United States
Chad	Iran	Mexico	South Africa	Vatican City
Chile	Iraq	Nepal	South Korea	Venezuela
China	Ireland	Netherlands	Sudan	Yemen
Colombia	Israel	New Zealand	Sweden	
Denmark	Italy	Nigeria	Switzerland	
Egypt	Japan	North Korea	Syria	

Figure 4.16: Countries in Risk Prediction Calculation

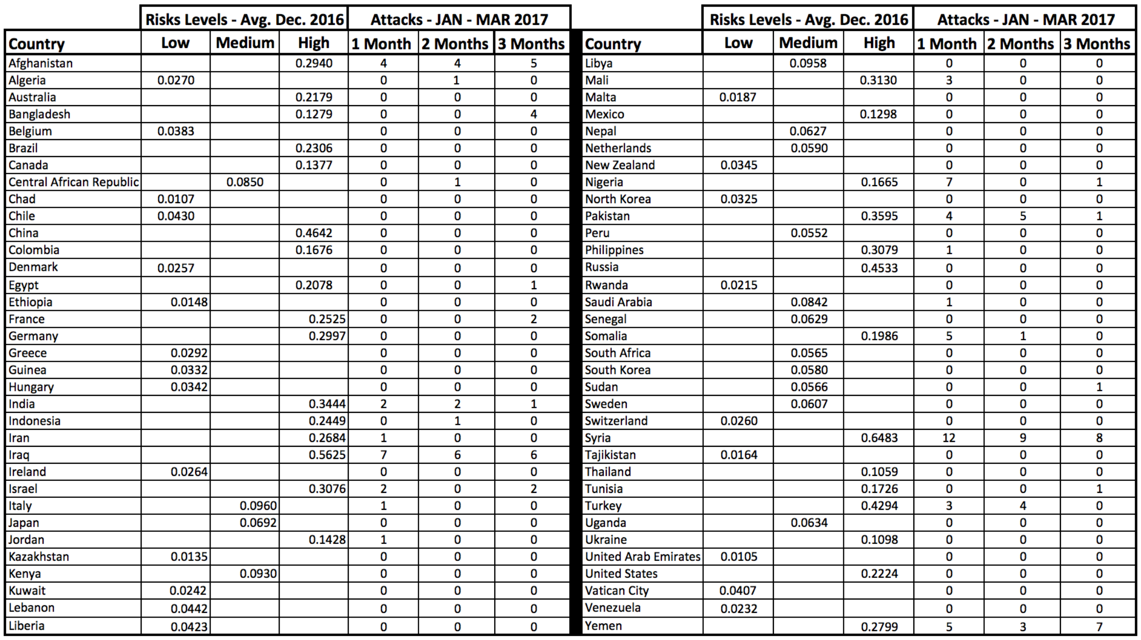
The risk values (low, medium, and high) for the countries in the table 4.17 are the average risk values from 09/01/2016 to 09/30/2016. These risk values are used to export the attacks occurred between October 2016 and December 2016. Those values from October 2016 and December 2016 are then used to calculate the precision and recall values of the prediction statements above.



	Risks Levels - Avg. Sep. 2016			Attacks - OCT - DEC 2016		
Country	Low	Medium	High	1 Month	2 Months	3 Months
Afghanistan			0.3978	8	3	4
Algeria		0.0617		1	0	0
Australia			0.1838	2	0	1
Bahrain	0.0459			0	0	0
Bangladesh			0.3033	2	0	1
Belgium			0.1021	0	0	0
Brazil			0.1166	0	0	0
Canada			0.1645	0	1	0
Central African Republic		0.0895		0	2	0
Chad	0.0316			0	0	0
Chile	0.0247			0	0	0
China			0.4341	0	0	1
Colombia	0.0439			0	0	0
Denmark	0.0258			0	0	0
Egypt			0.2015	5	2	1
El Salvador	0.0260			0	0	0
Ethiopia			0.1318	0	0	0
France			0.4664	1	1	0
Germany			0.2793	2	0	1
Greece	0.0130			1	0	0
Hungary	0.0397			1	0	0
India			0.5209	6	6	2
Indonesia			0.1852	0	0	1
Iran			0.3409	0	4	0
Iraq			0.4538	11	13	7
Ireland		0.0587		0	0	0
Israel			0.3198	3	1	3
Italy			0.2526	0	0	0
Japan			0.1662	1	0	0
Jordan		0.0526		1	1	1
Uganda	0.0129			0	1	1
Ukraine			0.1526	1	0	2
United Arab Emirates	0.0396			0	0	0
United States			0.5759	0	0	0
Kazakhstan	0.0116			0	0	0
Kenya			0.1346	2	0	0
Kuwait	0.0092			0	0	0
Lebanon		0.0734		0	0	2
Libya			0.1591	4	1	2
Mali			0.3228	4	2	0
Mexico			0.2359	0	0	0
Nepal		0.0661		0	0	0
New Zealand	0.0444			0	0	0
Nigeria			0.2372	5	6	2
North Korea			0.1755	0	0	0
Pakistan			0.6845	9	10	0
Peru	0.0458			0	0	0
Philippines			0.4302	1	3	0
Qatar	0.0281			0	0	0
Russia			0.6084	2	4	2
Rwanda	0.0286			0	0	0
Saudi Arabia			0.3083	1	0	1
Somalia			0.1872	5	4	3
South Africa			0.1196	0	0	1
South Korea			0.1273	0	0	0
Sudan			0.1340	2	2	0
Sweden	0.0259			0	0	1
Switzerland	0.0231			0	0	0
Syria			0.8482	13	9	11
Tajikistan	0.0281			0	0	0
Tanzania		0.0762		0	0	0
Thailand			0.2523	2	2	0
Tunisia		0.0605		0	0	0
Turkey			0.7283	10	7	3
Vatican City	0.0301			0	0	0
Venezuela	0.0309			0	0	0
Yemen			0.5423	6	5	3

Figure 4.17: Attacks Statistics - Oct. 16 - Dec. 16 - Using Sep. 16 Average Risks

The risk values (low, medium, and high) for the countries in the table 4.18 are the average risk values from 12/01/2016 - 12/31/2016. These risk values are used to export the attacks occurred between January 2017 and March 2017. Those values from January 2017 and March 2017 are then used to calculate the precision and recall values of the prediction statements above.



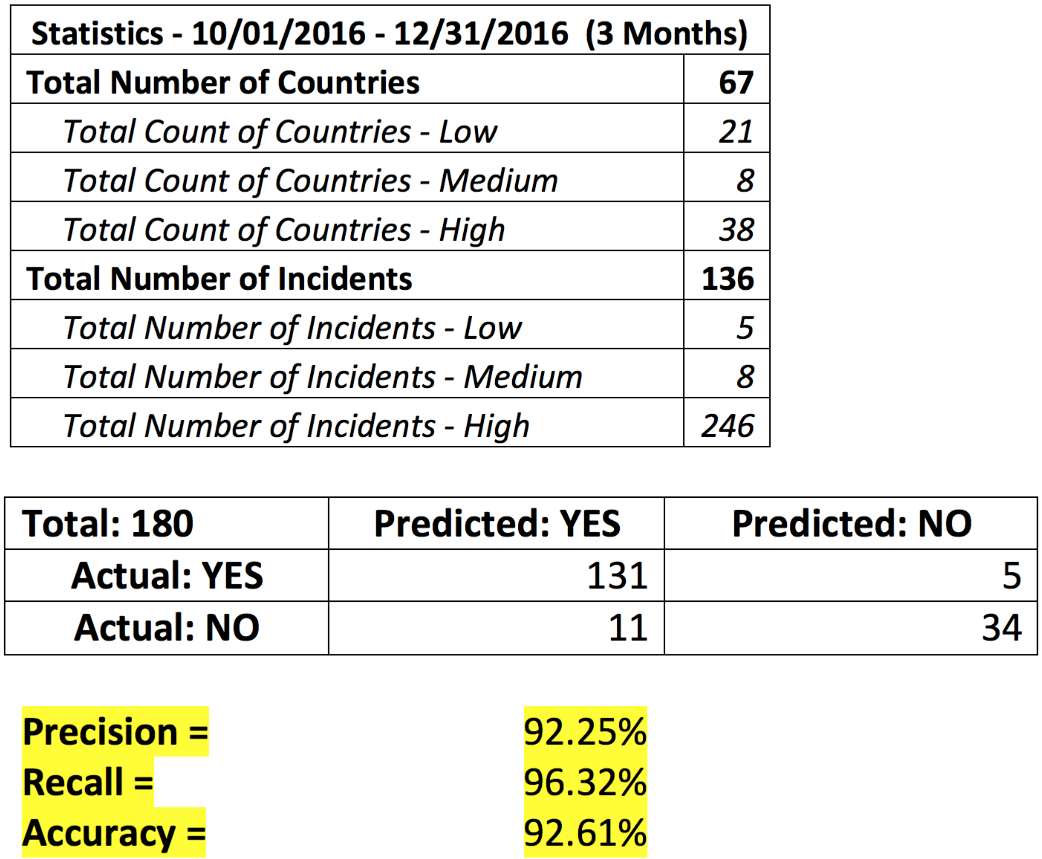
	Risks Levels - Avg. Dec. 2016			Attacks - JAN - MAR 2017				Risks Levels - Avg. Dec. 2016			Attacks - JAN - MAR 2017		
Country	Low	Medium	High	1 Month	2 Months	3 Months	Country	Low	Medium	High	1 Month	2 Months	3 Months
Afghanistan			0.2940	4	4	5	Libya		0.0958		0	0	0
Algeria	0.0270			0	1	0	Mali			0.3130	3	0	0
Australia			0.2179	0	0	0	Malta	0.0187			0	0	0
Bangladesh			0.1279	0	0	4	Mexico			0.1298	0	0	0
Belgium	0.0383			0	0	0	Nepal		0.0627		0	0	0
Brazil			0.2306	0	0	0	Netherlands		0.0590		0	0	0
Canada			0.1377	0	0	0	New Zealand	0.0345			0	0	0
Central African Republic		0.0850		0	1	0	Nigeria			0.1665	7	0	1
Chad	0.0107			0	0	0	North Korea	0.0325			0	0	0
Chile	0.0430			0	0	0	Pakistan			0.3595	4	5	1
China			0.4642	0	0	0	Peru		0.0552		0	0	0
Colombia			0.1676	0	0	0	Philippines			0.3079	1	0	0
Denmark	0.0257			0	0	0	Russia			0.4533	0	0	0
Egypt			0.2078	0	0	1	Rwanda	0.0215			0	0	0
Ethiopia	0.0148			0	0	0	Saudi Arabia		0.0842		1	0	0
France			0.2525	0	0	2	Senegal		0.0629		0	0	0
Germany			0.2997	0	0	0	Somalia			0.1986	5	1	0
Greece	0.0292			0	0	0	South Africa		0.0565		0	0	0
Guinea	0.0332			0	0	0	South Korea		0.0580		0	0	0
Hungary	0.0342			0	0	0	Sudan		0.0566		0	0	1
India			0.3444	2	2	1	Sweden		0.0607		0	0	0
Indonesia			0.2449	0	1	0	Switzerland	0.0260			0	0	0
Iran			0.2684	1	0	0	Syria			0.6483	12	9	8
Iraq			0.5625	7	6	6	Tajikistan	0.0164			0	0	0
Ireland	0.0264			0	0	0	Thailand			0.1059	0	0	0
Israel			0.3076	2	0	2	Tunisia			0.1726	0	0	1
Italy		0.0960		1	0	0	Turkey			0.4294	3	4	0
Japan		0.0692		0	0	0	Uganda		0.0634		0	0	0
Jordan			0.1428	1	0	0	Ukraine			0.1098	0	0	0
Kazakhstan	0.0135			0	0	0	United Arab Emirates	0.0105			0	0	0
Kenya		0.0930		0	0	0	United States			0.2224	0	0	0
Kuwait	0.0242			0	0	0	Vatican City	0.0407			0	0	0
Lebanon	0.0442			0	0	0	Venezuela	0.0232			0	0	0
Liberia	0.0423			0	0	0	Yemen			0.2799	5	3	7

Figure 4.18: Attacks Statistics - Jan. 17 - Mar. 17 - Using Dec. 16 Average Risks

4.5.1.2 Experiment 1: Prediction using Risk Values

The Three (3) Experiments below were run using the data of 72 unique countries with 6 months data.

Risk Values from: 10/01/2016 12/31/2016



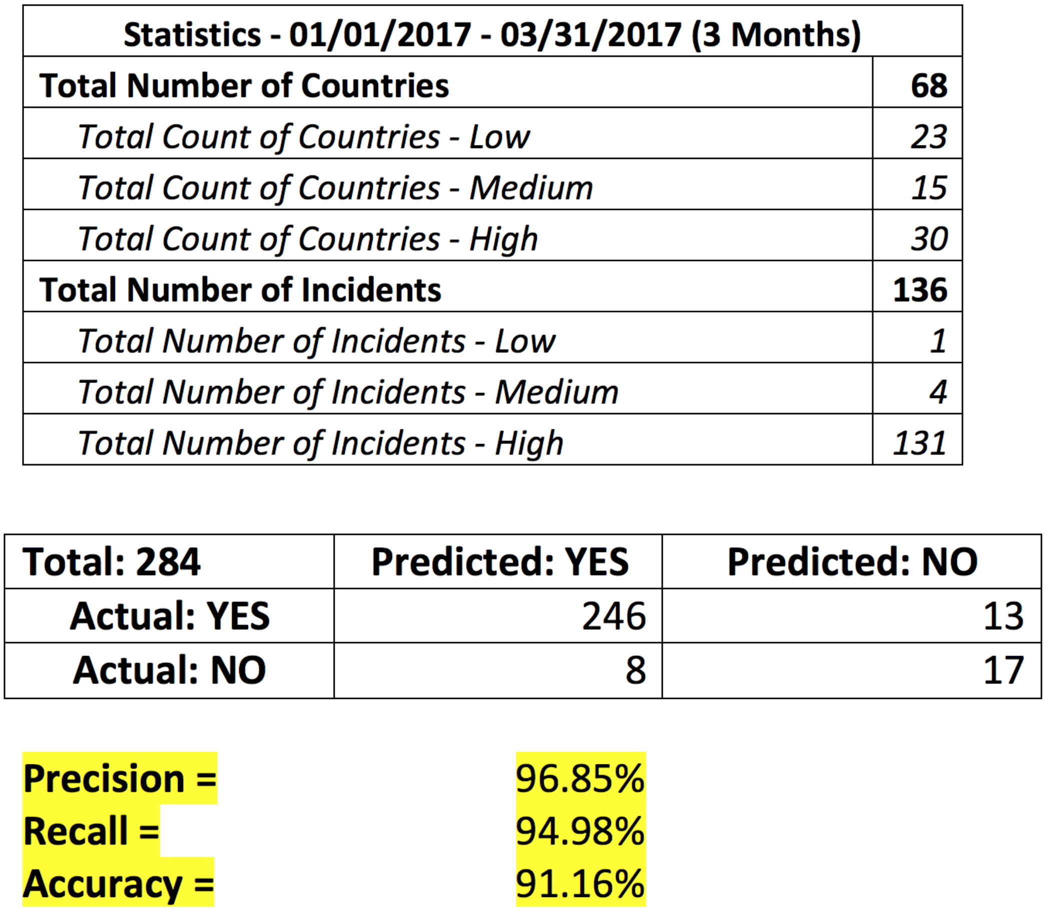
Statistics - 10/01/2016 - 12/31/2016 (3 Months)	
Total Number of Countries	**67**
Total Count of Countries - Low	*21*
Total Count of Countries - Medium	*8*
Total Count of Countries - High	*38*
Total Number of Incidents	**136**
Total Number of Incidents - Low	*5*
Total Number of Incidents - Medium	*8*
Total Number of Incidents - High	*246*

Total: 180	Predicted: YES	Predicted: NO
Actual: YES	131	5
Actual: NO	11	34

Precision =	92.25%
Recall =	96.32%
Accuracy =	92.61%

Figure 4.19: Experiment 1: Prediction using Risk Values

4.5.1.3 Experiment 2: Prediction using Risk Values



Statistics - 01/01/2017 - 03/31/2017 (3 Months)	
Total Number of Countries	**68**
Total Count of Countries - Low	*23*
Total Count of Countries - Medium	*15*
Total Count of Countries - High	*30*
Total Number of Incidents	**136**
Total Number of Incidents - Low	*1*
Total Number of Incidents - Medium	*4*
Total Number of Incidents - High	*131*

Total: 284	Predicted: YES	Predicted: NO
Actual: YES	246	13
Actual: NO	8	17

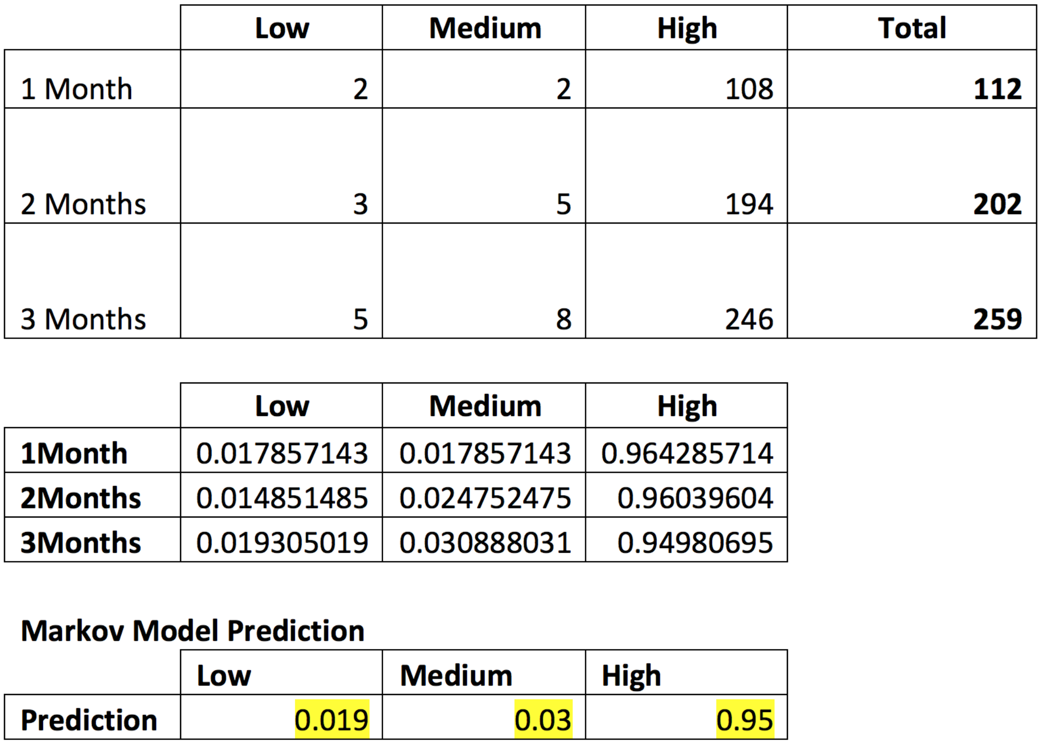
Precision =	96.85%
Recall =	94.98%
Accuracy =	91.16%

Figure 4.20: Experiment 2: Prediction using Risk Values

4.5.1.4 Experiment 3: Prediction using Markov Chain

Markov Chain [156] is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. The Markov Chain model is a proven prediction formula and has been used extensively in prediction systems in different domains such as Economics, Finance and Natural Sciences [155].

We construct the Markov Model using the 3 months data 10/01/2016 - 12/31/2016.



	Low	Medium	High	Total
1 Month	2	2	108	**112**
2 Months	3	5	194	**202**
3 Months	5	8	246	**259**

	Low	Medium	High
1Month	0.017857143	0.017857143	0.964285714
2Months	0.014851485	0.024752475	0.96039604
3Months	0.019305019	0.030888031	0.94980695

Markov Model Prediction

	Low	Medium	High
Prediction	0.019	0.03	0.95

Figure 4.21: Experiment 3: Prediction using Markov Chain

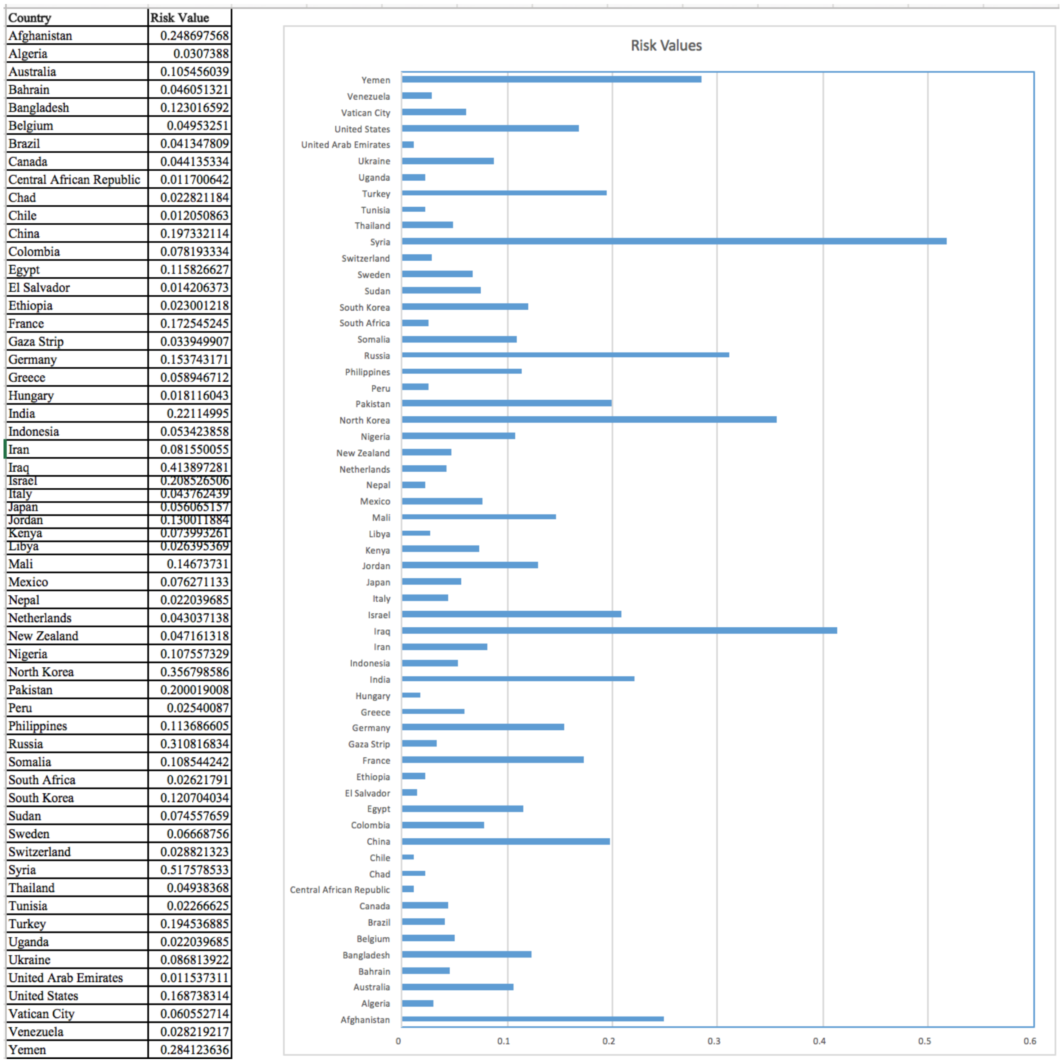
The results of the Markov Chain model show that there is a 95% chance for an incident to occur when the risk is high, 3% chance when the risk is medium, and 1.9% chance when the risk is low. The prediction results of Markov Chain model support our incident predictions using the values of our risk model in experiment 1 and experiment 2; where the prediction statement is: an incident will occur in high risk countries, and no incident will occur in medium and low risk countries.

4.5.1.5 France Incident Prediction

France April 2017 Incident:

France had a high risk of a terror attack on 4/10/2017; and on 4/20/2017 there was an attack in Paris.

The figure 4.22 shows the high risk level of France on 4/10/2017 along with other countries.

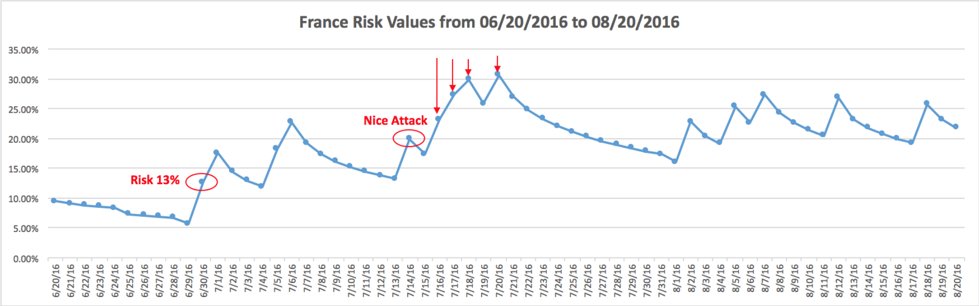


Country	Risk Value
Afghanistan	0.248697568
Algeria	0.0307388
Australia	0.105456039
Bahrain	0.046051321
Bangladesh	0.123016592
Belgium	0.04953251
Brazil	0.041347809
Canada	0.044135334
Central African Republic	0.011700642
Chad	0.022821184
Chile	0.012050863
China	0.197332114
Colombia	0.078193334
Egypt	0.115826627
El Salvador	0.014206373
Ethiopia	0.023001218
France	0.172545245
Gaza Strip	0.033949907
Germany	0.153743171
Greece	0.058946712
Hungary	0.018116043
India	0.22114995
Indonesia	0.053423858
Iran	0.081550055
Iraq	0.413897281
Israel	0.208526506
Italy	0.043762439
Japan	0.056065157
Jordan	0.130011884
Kenya	0.073993261
Libya	0.026395369
Mali	0.14673731
Mexico	0.076271133
Nepal	0.022039685
Netherlands	0.043037138
New Zealand	0.047161318
Nigeria	0.107557329
North Korea	0.356798586
Pakistan	0.200019008
Peru	0.02540087
Philippines	0.113686605
Russia	0.310816834
Somalia	0.108544242
South Africa	0.02621791
South Korea	0.120704034
Sudan	0.074557659
Sweden	0.06668756
Switzerland	0.028821323
Syria	0.517578533
Thailand	0.04938368
Tunisia	0.02266625
Turkey	0.194536885
Uganda	0.022039685
Ukraine	0.086813922
United Arab Emirates	0.011537311
United States	0.168738314
Vatican City	0.060552714
Venezuela	0.028219217
Yemen	0.284123636

Figure 4.22: France and Other Countries Risk Level

France June 2016 Incident:

A historical risk analysis shows that France reached a high risk level alert on 06/30/2016 and the Nice attack occurred on 07/14/2016 - two weeks after the high risk alert. The high risk alert remained high until the last results on 08/20/2016. In the same time period, France had a higher terrorist attack risk than its surrounding countries and Nice was the only one attacked in the week of 07/14/2016.



→ **Spikes on Dates:** 6/16/2016, 6/17/2016, 6/18/2016, 6/20/2016

Are due to continues reports on:

1. The nice Attack
2. The connection between the perpetrator with ISIS
3. ISIS claim of the attack
4. The extension of France's state of emergency
5. Stories of people who died during the attack

Figure 4.23: France Risk History

The arrows represent the spikes that occurred after the Nice attack on 07/14/2016. Hundreds of news were processed to generate the risk values causing the spikes. The spikes on the following dates: 6/16/2016, 6/17/2016, 6/18/2016, 6/20/2016 are due to many factors and continues reports on:

1. The Nice Attack

2. The connection between the perpetrator with Islamic State of Iraq and Syria (ISIS)

3. ISIS claim of the attack

4. The extension of France's state of emergency

5. Stories of people who died during the attack

France Risk History vs its Surrounding Countries:

The comparison of the risk level of France versus its surrounding countries shows that France had a higher risk terrorist incident. Only Nice, France was attacked during this time period.

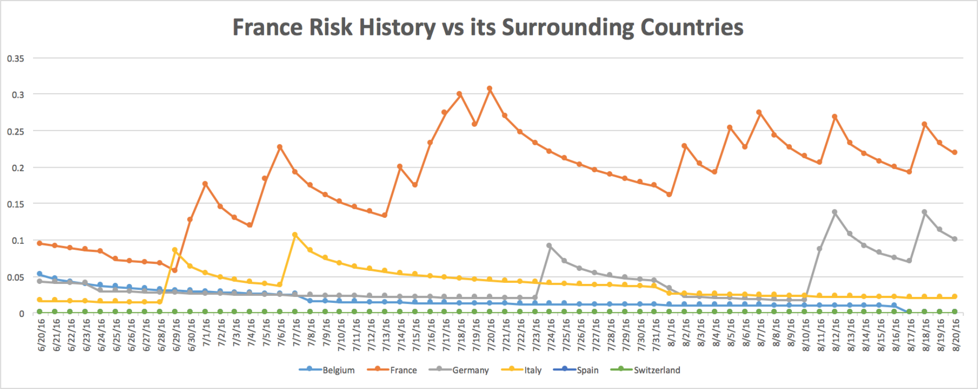


Figure 4.24: France Risk History vs its Surrounding Countries

4.6 Research Comparisons

We compare our prediction method (experiment 1 and experiment 2) to four other terrorism prediction systems. The comparison with the first method is qualitative as we use the same data to make the predictions. The comparison with the other method is quantitative as we did not use the same data to make the predictions. Our metrics - precision, recall, and accuracy values are significantly higher than the metrics of the other methods.

The details of each method are provided below:

Method 1 - IED:

There are not many methods that predict future terrorist incidents. Kengpol et al. [147] proposed a risk assessment method to predict the distribution range radius of terrorist incidents. The authors proposed a Risk Assessment Radar Chart to prevent Improvised Explosive Device (IED) terrorist incidents. The IED incidents data from 2007 to 2011 in the capital district of Yala province, the southern part of Thailand, were collected and used in the proposed methodology as a case study. The results are distance based incident predictions which specify in kilometer radius and number of days, the possible insurgencies. The model was tested against real life events and has a precision of prediction rate of 50.41%.

We applied our prediction method on the same data used by Kengpol et al. [147] to predict terrorist incidents of Yala province, Thailand. The data is collected by START team consisting of terrorist attacks in Yala province, Thailand from January 2007 to December 2011. The results of our method showed a prediction rate of 95% precision and 71% recall. Our prediction method provides higher precision and recall values for terrorist incident predictions. Furthermore, our prediction method is automated while the method proposed by Kengpol et al. [147] is manual and all the input parameters are select before processing.

Method 2 - CMP:

Fatih Ozgul et al. [100] developed a novel method called the Crime Prediction Model (CMP) to analyze and identify terrorist groups of unsolved attacks. CMP learns from terrorist attacks, matches them based on the similarities of their properties, and then clusters them into groups. This method was applied to a database of real life terrorist attacks that occurred in Turkey between 1970 and 2005. The predictions of CMP gave a good precision value for big terrorist groups and provided a good enough recall value for small terrorist groups. The average precision value for CPM is 76% and the average recall is 52%. Furthermore, our prediction method is automated while the method proposed by Fatih Ozgul et al. [100] is manual and all the input parameters are select before processing.

Method 3 - TGPM:

Abhishek Sachan and Devshri Roy [157] introduced a novel method named Terrorist Group Prediction Model for Counter Terrorism (TGPM). In this paper the authors have shown a terrorist group prediction model (TGPM) to predict the terrorist group involved in a given attack. This model initially learns similarities of terrorist incidents from various terrorist attacks to predict the responsible group. The model was validated with experimental results. The performance of the model shows the predictive accuracy of TGPM is 80.41%. Furthermore, our prediction method is automated while the method proposed by Abhishek Sachan and Devshri Roy [157] is manual and all the input parameters are select before processing.

Method 4 E-TGPS:

Sachan, A. [158] proposed an enhanced method named ”Enhanced Terrorist Group Prediction System for Counter Terrorism - E-TGPS”. This method presents an enhanced system that helps to predict the terrorist groups involved in attacks under investigation. The system initially learns similarities of terrorist activities from various past terrorist incidents to predict the responsible group. The method was validated with experimental results. The performance of the method shows the predictive accuracy of E-TGPS is 81.12%. Furthermore, our prediction method is automated while the method proposed by Sachan, A. [158] is manual and all the input parameters are select before processing.

Finally, our method outperforms all four methods with higher precision, recall, and accuracy. In addition, our method is based on real time data which allows one to predict future incidents in real life. On the counterpart, the methods we are comparing with use historical data for both their method development and prediction, which means that the incidents are already in the past. Thus, our systems and methods provide better solutions for predicting future terrorist incidents.

4.7 Applications and Advantages

4.7.1 Terrorism Root Cause Analysis

Our root cause analysis methods can be used to analyze historical and real time data to extract the reasons and consequences of terrorist incidents. This method will allow counter-terrorism experts to identify and find mitigations to the root causes that lead to the incident.

4.7.2 Clustering of Terrorist Groups

Our clustering method allows counter-terrorism experts to cluster terrorist groups based on similarities in attack patterns. The advantages of clustering terrorist groups are multiple. Clustering terrorist groups allows to potentially find ties between different terrorist groups. The identification of similar groups allows accelerate the understanding of a new group and allows making the right decisions. Furthermore, a law or method that work on a group will more likely work on similar groups. Thus, a successful method (e.g. dismantle financial sources) on one group can be applied to similar groups.

4.7.3 Real Time Terrorist Incidents Data

Data is key when it comes to analyzing terrorist attacks and finding the behavior of terrorist groups. Moreover, for counter-terrorism experts to quickly adapt and be informed of new threats of terrorist groups, real time data is a must and is a high priority requirement. Our data collection system provide solutions to automatically collect real time terrorist incidents data from reliable sources. Our data collection solution can be a useful tool to close a gap between terrorist groups and counter-terrorism experts.

4.7.4 Terrorism Risk Model

Assessing the risks of different locations is a critical task for counter-terrorism experts and any law enforcement officer. Our novel mathematical risk model is designed specifically to calculate the risk levels of different locations. Moreover, our system implements a functionality to plot on a dynamic map the impacts of terrorist groups on different locations (e.g. countries and cities). Our novel risk model is one of the key components for successful counter-terrorism measures and operations.

4.7.5 Terrorist Incident Prediction

It is always challenging to estimate the risks and predict future terrorist incidents. Our real time data collection system, risk models, and prediction systems provide excellent analytical tools to risk analysts and counter-terrorism experts to effectively analyze past and recent incidents data, and predict future terrorist incidents of different locations with high precision, recall, and accuracy.

Our risk and prediction methods can assist counter-terrorism experts to make quick and accurate decisions especially in uncertain situations. Furthermore, crime prevention units can use our systems and methods to locate high risk areas in a city and focus their patrols on those areas.

Our risk maps adhere to an easy localization of high risk areas which allows a quick turnaround for decision making.

4.8 Discussion of Findings

Predicting terrorist attacks is a challenging task and requires many resources and technology solutions. The methods we propose in this dissertation serve as a novel way for analyzing and predicting terrorist and crime related incidents. Our novel methods can play key roles in counter-terrorism efforts.

In this dissertation, we developed different methodologies such as Clustering of Terrorist Groups, Real Time Terrorist Incidents Data Collection system, different Terrorism Risk Model, and Terrorist Incident Prediction. Each of the mentioned methods plays its own important role in analyzing Big Data and discovering risk levels.

Our methodologies are at an initial but mature state. They can be applied in the real world for analyzing terrorism related data. Future research will address larger data profiling, and we plan to conduct further research on our methodologies in order to extend its applicability to other domains such as health care, machine learning, and stock market analysis.

4.9 System Usage

The System and Methodologies:

The System Allows Root Cause Analysis

The System Collects Real Time Data

The System Calculates Terrorism Risk Levels

The System Predicts Terrorist Incidents

Whats Next?:

Use the system on a daily basis

Monitor for Root Causes and Similarities

Monitor for (Medium and High Risks) for both Risk Levels and Incidents Predictions

Increase Security Measures on High Risks

Plan Counter-Terrorism Measures

Share Reports with Experts for Analysis and Planning

4.10 Technologies

In this research, we use technologies on a need basis. Whenever, we implement a new functionality, we first evaluate different programming languages and select the one that best fits our project. Some of our technology selection criteria are latency, scalability, and preferably open source.

Some of the technologies we use are: Java, HTML5, PHP, Python, Web2py, D3js, JavaScript, Bootstrap, Google Map APIs, MySQL, Couchbase on Amazon Web Services. And the tools are: MATLAB, R, Weka, MySQL Workbench, SQL Developer, XAMPP.

Chapter 5

Conclusion and Future Work

5.1 Summary of Contributions

In this research, we developed different methodologies to analyze terrorism related data. Our research contributes to counter-terrorism and data mining domains as described below:

Contribution to Terrorism Domain:

A real time terrorism data collection system that collects data on a periodic basis.

A real time terrorism data scanner to automatically discover potential incidents.

A terrorism risk model to estimate terrorist threats based on recent terrorism data.

A terrorism incident prediction model.

Contribution to Data Mining Domain:

Unstructured text data summarization method, using our novel algorithm and the latent dirichlet allocation.

Unstructured text data classification method, using our novel algorithm in addition to Kmeans clustering and the latent semantic indexing.

A novel risk model that combines the frequency and time factors. This model can be applied to other problems that deal with frequency and time factors. For example, the risk estimation of a disease outbreak per location.

Real time data made available for analysis by other researchers.

The different conditions our methodologies best fit in:

We developed three main methodologies in this research, and each is best fit to address different aspects of terrorism as described below:

1. Analyze Terrorism data using Latent Dirichlet Allocation:

This methodology is used for text summarization. It is best when the analyst has a large amount of data collected over a long time period. For instance, this methodology can be utilized to summarize historical health records. In our research, we used START terrorism data collected over the last 40 years. This methodology is not well suited for predicting future incidents as it does not have time parameters.

2. Analyze Terrorism data using Kmeans clustering and Latent Semantic Indexing:

This methodology is used for text classification and graphical view of corrected data. This methodology is best used with historical data; but preferably the data should be segmented based on different time periods in order to capture incidents that occurred on the same time frames.

3. Terrorism Incident Prediction:

This methodology is used for real time data analyses. As the method uses real time data, it goes along with a real time data collection system such as our news data collection system. It is best used to capture terrorism risk levels of different locations in real time. Moreover, used with the Markov Chain model, it allows to predict terrorist incidents. The results are based on recent incidents therefore they reflect the risk values of the current situation.

An important note is: what if no incident occurred in the last 60 days (number of days to be considered for the risk level calculation)? Can we still estimate the risk level and predict incidents?

Incidents can still be estimated, even if no incident occurred in recent days. In this scenario where no incidents occurred in the recent months, we can use the data collected based on warning signs. The warning signs include: threat made against the country, and the political situations. If there are no warning signs, the country terrorism risk level is classified as Low and no incident is expected to occur.

5.2 Future Work

Our system collects data in real time from different news media. However, news media companies do not provide exact location information of the incident such as the longitude, latitude, street address, and zip code. Therefore, when new data sources become available that provide exact location information, our methodologies can be used to calculate the terrorism risk levels up to the street level. Moreover, we focused on the minimal information (threats and incidents) to create the building block of terrorism risks, and incident predictions. While improving our methodologies, we plan to include other factors such as population density and infrastructure data.

During this dissertation research, we experimented, learned, and were enthusiastic about the results we discovered. We hope this dissertation becomes the beginning of a novel prediction system applicable to multiple domains.

In conclusion, we will continue this research and we will improve our methodologies based on the valuable comments and suggestions made by the committee members.

Bibliography

[1] S. L. Tomassen, "Conceptual ontology enrichment for web information retrieval," PhD, NTNU, 2011.

[2] P. Cimiano, *Ontology Learning and Population from Text - Algorithms, Evaluation and Applications.* Springer, 2006.

[3] J. Brank, M. Grobelnik, and D. Mladenić, "A survey of ontology evaluation techniques," in *Proceedings of the Conference on Data Mining and Data Warehouses SiKDD 2005*, no. a. Citeseer, 2005.

[4] A. Maedche and S. Staab, "Measuring similarity between ontologies," *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, pp. 15–21, 2002.

[5] J. Brank, D. Mladenic, and M. Grobelnik, "Gold standard based ontology evaluation using instance assignment," in *Proc. of the EON 2006 Workshop*, 2006.

[6] K. Dellschaft and S. Staab, "On how to perform a gold standard based evaluation of ontology learning," in *Proceedings of the 5th International Semantic Web Conference (ISWC)*, I. C. et al., Ed. Springer Verlag, 2006, pp. 228–241.

[7] R. Porzel and R. Malaka, "A task-based approach for ontology evaluation," in *ECAI Workshop on Ontology Learning and Population, Valencia, Spain.* Citeseer, 2004.

[8] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks, "Data driven ontology evaluation," in *Proceedings of Internation Conference on Language Resources and Evaluation*, 2004.

[9] M. Sabou, J. Gracia, S. Angeletou, M. d'Aquin, and E. Motta, "Evaluating the semantic web: A task-based approach," in *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference.* Springer-Verlag, 2007, pp. 423–437.

[10] P. Spyns and M. Reinberger, "Lexically evaluating ontology triples generated automatically from texts," *The Semantic Web: Research and Applications*, pp. 85–97, 2005.

[11] M. Sabou, V. Lopez, E. Motta, and V. Uren, "Ontology selection: Ontology evaluation on the real semantic web," in *Proceedings of the EON'2006 Workshop, "Evaluation of Ontologies on the Web"*, 2006.

[12] H. Alani, S. Harris, and B. O'Neil, "Winnowing ontologies based on application use," in *ESWC*, 2006, pp. 185–199.

[13] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *JMLR*, vol. 9, pp. 1981–2014, 2008.

[14] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[15] D. M. Blei, "Introduction to probabilistic topic models," *Communications of the ACM*, 2011.

[16] V. Spiliopoulos, G. Vouros, and V. Karkaletsis, "Mapping ontologies elements using features in a latent space," in *Web Intelligence, IEEE/WIC/ACM International Conference on.* IEEE, 2007, pp. 457–460.

[17] J. S. e. a. Luciano, "The translational medicine ontology and knowledge base: driving personalized medicine by bridging the gap between bench and bedside," *Journal of Biomed Semantics*, vol. 2(Suppl 2), 2011.

[18] Andrej Cherkaev, "Variational Methods for Structural Optimization", *Springer-Verlag*, 2000, pp. xii + 627

[19] D. Zeimpekis and E. Gallopoulos, TMG: A MATLAB toolbox for generating term-document matrices from text collections. Springer, 2006, pp. 187210.

[20] Sun Zhen, Lim, Ee-Peng, Chang Kuiyu, Ong Teng-Kwee, Gunaratna RohanKumar, Event-Driven Document Selection for Terrorism Information Extraction, Springer Berlin Heidelberg, 2005-01-01, Information extraction, 37-48

[21] Hu, Xiaohua and Yoo, Illhoi and Rumm, Peter and Atwood, Michael, Mining Candidate Viruses as Potential Bio-terrorism Weapons from Biomedical Literature, Springer Berlin Heidelberg, 2005, p 60-71.

[22] Khalsa SundriK., Forecasting Terrorism: Indicators and Proven Analytic Techniques, Springer Berlin Heidelberg, 2005, p. 561-566.

[23] Papadimitriou, Christos H. and Tamaki, Hisao and Raghavan, Prabhakar and Vempala, Santosh, Latent semantic indexing: a probabilistic analysis, Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, year 1998, ACM, p. 159–168.

[24] J. Shlens. A Tutorial on Principal Component Analysis. available at http://www.snl.salk.edu/ shlens/pca.pdf

[25] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11, 1 (November 2009), 10-18. DOI=10.1145/1656274.1656278 http://doi.acm.org/10.1145/1656274.1656278

[26] Wikipedia, Frequentist_probability - http://en.wikipedia.org/wiki/Frequentist_probability

[27] Michael Stonebraker Joseph M. Hellerstein:What Goes Around Comes Around Readings in Database Systems, Fourth Edition Morgan Kaufmann 2005.

[28] Heinrichs, J. H. and J. Lim (2003). "Integrating Web-Based Data Mining Tools With Business Models For Knowledge Management." Decision Support Systems 35(1): 103-112.

[29] Fabrizio Sebastiani. Machine learning in automated text categorization. ACM Comput. Surv. 34(1): 1-47 (2002).

[30] Li, X-B. A scalable decision tree system and its application in pattern recognition and intrusion detection, Decision Support Systems 41 (2005) 112-130.

[31] Chen, M.-S., Han, J. and Yu, P. S. Data Mining: An Overview from a Database Perspective, IEEE Transactions on Knowledge and Data Engineering, 8(6):866-883, 1996.

[32] Gengm L. and Hamilton, H. J. "Interestingness Measures for Data Mining: A Survey", ACM Computing Surveys, 38(2), 2006.

[33] Ulrike Luxburg. 2007. A tutorial on spectral clustering. Statistics and Computing 17, 4 (December 2007), 395-416.

[34] V. Verykios and E. Bertino and I. Fovino and L. Provenza and Y. Saygin and Y. Theodoridis. State-of-the-art in privacy preserving data mining, SIGMOD Record 33 (2004) 50–57", 2004, available online at http://citeseer.csail.mit.edu/713858.html.

[35] Cynthia Dwork. 2006. Differential privacy. In Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II (ICALP'06), Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.), Vol. Part II. Springer-Verlag, Berlin, Heidelberg, 1-12.

[36] Clifton, C., Kantarcioglu, M., and Vaidya, J. Defining Privacy for Data Mining, in National Science Foundation Workshop on Next Generation Data Mining, pages 126-133, 2002. Available online at http://cimic.rutgers.edu/ jsvaidya/pub-papers/ngdm-privacy.pdf.

[37] Mukherjee, S., Chen, Z. and Gangopadhyay, A. A Privacy Preserving Technique for Euclidean Distance-Based Mining Algorithms Using Fourier-Related Transforms, The VLDB Journal, 15:293-315, 2006.

[38] L. Sweeney, k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002, 557-570

[39] Jeffrey Dean and Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters". Communications of the ACM, vol. 51, no. 1 (2008), pp. 107-113

[40] S. Chaudhuri and U. Dayal. An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record, 26(1):65–74, March 1997.

[41] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber. "Bigtable: A Distributed Storage System for Structured Data". In proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2006, Seattle, WA

[42] Wu, M.-C. and Buchmann, A. "Research Issues in Data Warehousing", in Datenbanksysteme in Buro, Technik und Wissenschaft, pp. 61-82, 1997.

[43] Jing, J. and Helal, S. and Elmagarmid, A. Client Server Computing in Mobile Environments, ACM Computing Surveys, Vol. 31, No. 2, June 1999.

[44] D. Zhang, G. Karabatis, Z. Chen, B. Adipat, L. Dai, Z. Zhang Personalization and Visualization on Handheld Devices, ACM Symposium on Applied Computing, Dijon, France, April 2006.

[45] A. Halevy, A. Rajaraman, J. Ordille. "Data Integration: The Teenage Years". Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB), September 2006, Seoul, Korea, pp. 9-16.

[46] J. Bleiholder, and F. Naumann. "Data fusion". /ACM Computing Surveys/ 41, 1 (Dec. 2008), 1-41. DOI= http://doi.acm.org/10.1145/1456650.1456651

[47] G. Karabatis. "Using Context in Semantic Data Integration", Journal of Interoperability in Business Information Systems, Volume 1, Number 3, December 2006, pp. 9-21

[48] G. Karabatis, Z. Chen, V. Janeja, T. Lobo, M. Advani, M. Lindvall, R.L. Feldmann Using Semantic Networks and Context in Search for Relevant Software Engineering Artifacts Journal on Data Semantics, Volume XIV, Lecture Notes in Computer Science LNCS 5880, Springer-Verlag, pp. 74-104, 2009

[49] Martin Ester, Hans-Peter Kriegel, Jrg Sander (University of Munich), Algorithms and Applications for Spatial Data Mining, Published in Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, Taylor and Francis, 2001.http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/Chapter7.revised.pdf

[50] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages.

[51] Lei Shi, V. Janeja : Anomalous Window Discovery for Linear Intersecting Paths. IEEE Trans. Knowl. Data Eng. 23(12): 1857-1871 (2011)

[52] V. Janeja, Nabil R. Adam, Vijayalakshmi Atluri, Jaideep Vaidya: Spatial neighborhood based anomaly detection in sensor datasets. Data Min. Knowl. Discov. 20(2): 221-258 (2010)

[53] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze. "Web Search Basics". Chapter 19 in Introduction to Information Retrieval, Cambridge University Press, 2008. (also available online http://nlp.stanford.edu/IR-book/pdf/19web.pdf)

[54] A. Langville and C. Meyer, A Survey of Eigenvector Methods for Web Information Retrieval, SIAM Review, vol. 47, no. 1, pp. 135161, 2005.

[55] S. Fortunato. "Community detection in graphs", Physics Reports 486 (2010) 75174.

[56] Lei Tang, and Huan Liu. Leveraging Social Media Networks for Classification. Journal of Data Mining and Knowledge Discovery (DMKD), 2011

[57] D. M. Blei, A. Ng, and M. Jordan, Latent dirichlet allocation, JMLR, vol. 3, pp. 9931022, 2003.

[58] Eamonn Keogh, Stefano Lonardi, and Chotirat Ann Ratanamahatana. 2004. Towards parameter-free data mining. In Proceedings of the tenth ACM

SIGKDD international conference on Knowledge discovery and data mining (KDD '04). ACM, New York, NY, USA, 206-215.

[59] Ma, J. & Perkins, S. Online Novelty Detection on Temporal Sequences. Proc. International Conference on Knowledge Discovery and Data Mining, August 24-27, 2003.

[60] Dasgupta, D. & Forrest,S. Novelty Detection in Time Series Data using Ideas from Immunology." In Proc. of the International Conference on Intelligent Systems (1999).

[61] Yairi, T., Kato, Y., & Hori, K. Fault Detection by Mining Association Rules from House-keeping Data, Proc. of Intl Sym. on AI, Robotics and Automation in Space, 2001.

[62] Shahabi, C., Tian, X., & Zhao, W. TSA-tree: A Wavelet-Based Approach to Improve the Efficiency of Multi-Level Surprise and Trend Queries The 12th Intl Conf on Scientific and Statistical Database Management (SSDBM 2000)

[63] Steve Kramer. 2010. Anomaly detection in extremist web forums using a dynamical systems approach. In ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD '10). ACM, New York, NY, USA, , Article 8 , 10 pages.

[64] Ibrahim Toure, Aryya Gangopadhyay, A Method for Analyzing Terrorist Attacks, IEEE Technologies for Homeland Security, November 2012.

[65] Ibrahim Toure, Aryya Gangopadhyay, "Analyzing Terror Attacks using Latent Semantic Indexing", IEEE Technologies for Homeland Security, November 2013.

[66] Ibrahim Toure, Aryya Gangopadhyay, "Analyzing Real Time Terrorism Data", IEEE Technologies for Homeland Security, November 2015.

[67] Ibrahim Toure, Aryya Gangopadhyay, "Big Data Analytics For Predicting Incidents", IEEE Technologies for Homeland Security, November 2016.

[68] Tianxia Gong, Chew Lim Tan, Tze Yun Leong, Cheng Kiang Lee, Boon Chuan Pang, Tchoyoson Lim, C.C., Qi Tian, Suisheng Tang, Zhuo Zhang, , "Text Mining in Radiology Reports," Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on , vol., no., pp.815-820, 15-19 Dec. 2008.

[69] Blatak, J., , "First-order Frequent Patterns in Text Mining," Artificial intelligence, 2005. epia 2005. portuguese conference on , vol., no., pp.344-350, 5-8 Dec. 2005.

[70] Tekiner, F., Tsuruoka, Y., Tsujii, J., Ananiadou, S., , "Highly scalable Text Mining - parallel tagging application," Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, 2009. ICSCCW 2009. Fifth International Conference on , vol., no., pp.1-4, 2-4 Sept. 2009

[71] Ning Zhong, Yuefeng Li, Sheng-Tang Wu, , "Effective Pattern Discovery for Text Mining," Knowledge and Data Engineering, IEEE Transactions on , vol.24, no.1, pp.30-44, Jan. 2012

[72] Raymond Y. K. Lau, S. Y. Liao, Ron Chi-Wai Kwok, Kaiquan Xu, Yunqing Xia, and Yuefeng Li. 2012. Text mining and probabilistic language modeling for online review spam detection. ACM Trans. Manage. Inf. Syst.

[73] Report Problems if Vaccines form http://vaers.hhs.gov/resources/vaers form.pdf

[74] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[75] SentiWordnet from http://sentiwordnet.isti.cnr.it/

[76] http://www.cdc.gov/az/a.html

[77] Food and Drug Administration http://www.fda.gov/MedicalDevices/default.htm

[78] CIA https://www.cia.gov/News-information/cia-the-war-on-terrorism/terrorism-faqs.html

[79] FBI http://www.fbi.gov/stats-services/publications/terrorism-2002-2005

[80] Study of Terrorism and Response to Terrorism (START) Codebook http://www.start.umd.edu/gtd/downloads/Codebook.pdf

[81] Definition of Terrorism United Nations http://www.un.org/News/Press/docs/2012/gal3433.do

[82] Definition of Terrorism United Nations http://www.un.org/News/Press/docs/2005/gal3276.do

[83] Various Definition of Terrorism http://www.azdema.gov/museum/famousbattles/pdf/Terroris

[84] Research on Terrorism http://www.terrorism-research.com/

[85] Chen H., Reid E., Sinai J., Silke A., Ganor B., Research on Terrorism, Integrated Series In Information Systems, 2008, Springer US, P 27-50

[86] Riaz Hassan, What Motivates the Suicide Bombers?, YaleGlobal, Sept. 2009, http://yaleglobal.yale.edu/content/what-motivates-suicide-bombers-0

[87] START, Annex of Statistical Information, Country Reports on Terrorism 2012 http://www.state.gov/documents/organization/210288.pdf

[88] Google Maps APIs https://developers.google.com/maps/

[89] Foreign Terrorism Organization Designation, Boko Haram and Ansaru, http://www.start.umd.edu/start/publications/

[90] Sequential Analysis http://www.hsph.harvard.edu/betensky/bio276.html

[91] Sampling Methods: http://en.wikipedia.org/wiki/Sampling

[92] Noam Chomsky, The Culture of Terrorism.

[93] Terrorism Financing: http://www.amlu.gov

[94] Sequential Methods "Wikipedia.org Sequential Methods"

[95] Radiological Fact Sheet: http://www.dhs.gov

[96] Terrorism Financing by Center for Homeland Defense and Security Naval Postgraduate School (http://www.chds.us), Module 1-6 : https://www.youtube.com/watch?v=ITJomMgTKLI

[97] Rapoport, David, C. Four Waves or Rebel Terror and September 11, Antropoethics, vol. 8, no. 1 (Spring/Summer 2002).

[98] analysis-of-digital-financial-data: http://www.fbi.gov

[99] http://micastore.com/Vanguard/PastIssues/2010April.pdf

[100] Boko Haram Attack on U.N Building in Nigeria, August 2011, http://www.nytimes.com/2011/08/27/world/africa/27nigeria.html?pagewanted=all

[101] Foreign Terrorist Organizations, US Department of States, from 1997-2013 http://www.state.gov/j/ct/rls/other/des/123085.htm

[102] F. Ozgul, Z. Erdem, and C. Bowerman, Prediction of past unsolved terrorist attacks, in International Conference on Intelligence and Security Informatics, ISI 09. IEEE, 2009, pp. 3742.

[103] M. Shaikh and W. J., Investigative data mining: Identifying key nodes in terrorist networks, in Proceedings of the Multitopic Conference, 2006. INMIC 06. IEEE, 2006, pp. 2324.

[104] D. M. Blei, A. Ng, and M. Jordan, Latent dirichlet allocation, JMLR, vol. 3, pp. 9931022, 2003.

[105] D. M. Blei, Introduction to probabilistic topic models, Communications of the ACM, 2011. [Online]. Available: http://www.cs.princeton.edu/blei/papers/Blei2011.pdf

[106] M. Steyvers and T. Griffiths, Probabilistic topic models. Springer, 2006, pp. 187210.

[107] T. Griffiths, M. Steyvers, and J. Tenenbaum, Topics in semantic representation, Psychological Review, vol. 114, no. 2, pp. 211244, 2007.

[108] A. Cherkaev, Variational Methods for Structural Optimization, ser. Applied mathematical sciences, 2000, vol. 140. [Online]. Available: http://www.math.utah.edu/book/vmso

[109] D. Galanis, G. Karakatsiotis, G. Lampouras, and I. Androutsopoulos, An open-source natural language generator for owl ontologies and its use in protege and second life, in Proceedings of the EACL 2009 Demonstrations Session, 2009, pp. 1720.

[110] D. Zeimpekis and E. Gallopoulos, TMG: A MATLAB toolbox for generating term-document matrices from text collections. Springer, 2006, pp. 187210.

[111] Richard Gay. 2012. Key factors in managing and understanding the risk of underwater terrorism. In Proceedings of the 2012 Symposium on Emerging Applications of M/S in Industry and Academia Symposium (EAIA 12). Society for Computer Simulation International, San Diego, CA, USA, Article 12, 8 pages.

[112] D. Zeimpekis and E. Gallopoulos, TMG: A MATLAB toolbox for generating term-document matrices from text collections. Springer, 2006, pp. 187210.

[113] Sun Zhen, Lim, Ee-Peng, Chang Kuiyu, Ong Teng-Kwee, Gunaratna RohanKumar, Event-Driven Document Selection for Terrorism Information Extraction, Springer Berlin Heidelberg, 2005-01-01, Information extraction, 37-48

[114] Hu, Xiaohua and Yoo, Illhoi and Rumm, Peter and Atwood, Michael, Mining Candidate Viruses as Potential Bio-terrorism Weapons from Biomedical Literature, Springer Berlin Heidelberg, 2005, p 60-71.

[115] Papadimitriou, Christos H. and Tamaki, Hisao and Raghavan, Prabhakar and Vempala, Santosh, Latent semantic indexing: a probabilistic analysis, Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, year 1998, ACM, p. 159168.

[116] J. Shlens. A Tutorial on Principal Component Analysis. available at http://www.snl.salk.edu/ shlens/pca.pdf

[117] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11, 1 (November 2009), 10-18. DOI=10.1145/1656274.1656278 http://doi.acm.org/10.1145/1656274.1656278

[118] Wikipedia: "Al-Qaeda in the Islamic Maghreb"

[119] Wikipedia: "Taliban"

[120] Country Reports on Terrorism 2015: http://www.state.gov/j/ct/rls/crt/2014/

[121] Miller, Erin. 2014. "Terrorist Attacks in the U.S. between 1970 and 2012: Data from the Global Terrorism Database GTD)." http://www.start.umd.edu/

[122] Asal, Victor H., and R. Karl Rethemeyer. 2008. "The Nature of the Beast: Terrorist: The Organizational and Network Characteristics of Organizational Lethality." Journal of Politics (January): 437-449. http://sitemason.vanderbilt.edu/files/iVnJS0/Rethemeyer.pdf

[123] New York Metro News, http://nymag.com/nymetro/News/features/10558/

[124] Casey L. Addis, Christopher M. Blanchard, Hezbollah: Background and Issues for Congress, Congressional Research Center, Jan. 2011. http://www.fas.org/sgp/crs/mideast/R41446.pdf

[125] Justin J. Kemptona and Jonathan P. Stewart,b M.EERI. Prediction Equations for Signicant Duration of Earthquake Ground Motions Considering Site and

Near-Source Effects, Earthquake Spectra, Volume 22, No. 4, pages 9851013, November 2006. http://peer.berkeley.edu/

[126] Seifert, Katherine R. and Clark McCauley (2014) Suicide Bombers in Iraq, 2003-2010:Disaggregating Targets Can Reveal Insurgent Motives and Priorities(Jan) http://www.tandfonline.com/doi/abs/10.1080/09546553.2013.778198(April 2, 2014)

[127] Gruenewald, Jeff, and Joshua D. Freilich, Steven M. Chermak, William S. Parkin. Research Highlight: Violence Perpetrated by Supporters of al-Qa'ida and Affiliated Movements (AQAM): Fatal Attacks and Violent Plots in the United States, Research Brief to the Resilient Systems Division, Science and Technology Directorate, U.S. Department of Homeland Security. College Park, MD: START, 2014. http://www.start.umd.edu/

[128] Brachman, Jarret. 2014. "Transcending Organization: Individuals and 'The Islamic State.'" START Analytical Brief. College Park, Maryland. June. http://www.start.umd.edu/

[129] Simonelli, Corina. 2014. "The Evolution of the Islamic State of Iraq and the Levant (ISIL): Relationships 2004-2014." START Fact Sheet. College Park, Maryland. June. http://www.start.umd.edu

[130] http://www.cs.princeton.edu/ blei/lda-c/index.html

[131] RAND Database of Worldwide Terrorism Incidents (2009), Retrieved from http://www.rand.org/nsrd/projects/terrorism-incidents/download.html

[132] Ogle, Virginia E., and Michael Stonebraker. "Chabot: Retrieval from a relational database of images." Computer 28.9 (1995): 40-48.

[133] Google Map API https://developers.google.com/maps

[134] List of Countries, US Department of States http://www.state.gov/misc/list/

[135] MySQL Database http://www.mysql.com/

[136] PHP Documentation http://php.net/docs.php

[137] HTML, CSS, Javascript, XML Schema for Developer source http://www.w3schools.com/

[138] JQuery documentation http://api.jquery.com/

[139] Really Simple Syndication format - http://validator.w3.org/feed/docs/rss2.html

[140] Bosc, Patrick, and Olivier Pivert. "SQLf: a relational database language for fuzzy querying." Fuzzy Systems, IEEE Transactions on 3.1 (1995): 1-17.

[141] F. Ozgul, Z. Erdem, and C. Bowerman, Prediction of past unsolved terrorist attacks, in International Conference on Intelligence and Security Informatics, ISI '09. IEEE, 2009, pp. 37-42.

[142] M. Shaikh and W. J., Investigative data mining: Identifying key nodes in terrorist networks, in Proceedings of the Multitopic Conference, 2006. INMIC '06. IEEE, 2006, pp. 23-24.

[143] MySQL, Stopwords. full-text stopwords, mysql documentation.

[144] MySQL Developers site https://dev.mysql.com/doc/refman/5.0/en/c-api.html

[145] Brian Michael Jenkins, Stray Dogs and Virtual Armies: Radicalization and Recruitment to Jihadist Terrorism in the United States Since 9/11, RAND Corporation (OP-343), 2011.

[146] Henry H. Willis, et. al. "Estimating Terrorism Risk", RAND Corporation 2005.

[147] Karl Roberts, John Horgan , Risk Assessment and the Terrorist, Terrorism Research Initiative 2008. http://www.terrorismanalysts.com/pt/index.php/pot/article/view/38/html

[148] Ezell, B, Bennett, S, Von Winterfeldt, D, Sokolowski, J, & Collins, A 2010, 'Probabilistic Risk Analysis and Terrorism Risk', Risk Analysis: An International Journal, 30, 4, pp. 575-589, Business Source Premier, EBSCOhost, viewed 17 July 2015. http://www.dhs.gov/xlibrary/assets/rma-risk-assessment-technical-publication.pdf

[149] Kengpol, Athakorn, Neungrit, Pakorn A decision support methodology with risk assessment on prediction of terrorism insurgency distribution range radius and elapsing time: An empirical case study in Thailand. Elsevier Ltd, Computers & Industrial Engineering September 2014 75:55-67. Link to article

[150] Darby, J.L., Tools for evaluating risk of terrorist acts using fuzzy sets and belief/plausibility, Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American Year: 2009 Pages: 1 - 5, IEEE Conference Publications DOI: 10.1109/NAFIPS.2009.5156446

[151] Martin Gill, Ona Ekhomu, "Pratical Security strategies and Emerging Trends" Elsevier 2010.

[152] https://www.dhs.gov/national-terrorism-advisory-system

[153] "Homeland Security Advisory System". September 27, 2007. Retrieved 2007-09-27.

[154] Baltimore Call For Service Data, https://www.spotcrime.com/md/baltimore

[155] Giorgio Alfredo Spedicato, Tae Seung Kang, Sai Bhargav Yalamanchi and Deepak Yadav, "The markovchain Package: A Package for Easily Handling Discrete Markov Chains in R", 2016.

[156] Bremaud P. Discrete-Time Markov Models. In Markov Chains, pp. 5393. Springer (1999).

[157] Abhishek Sachan, Devshri roy, TGPM: Terrorist Group Prediction Model for Counter Terrorism, In: International Journal of Computer Applications, Vol 44 (10), 2012, pp. 49-52.

[158] Sachan, A. "E-TGPS: Enhanced Terrorist Group Prediction System for Counter Terrorism" International Journal of Computer Applications (0975 8887). Volume 117 No. 24, May 2015

[159] Google TensortFlow "A Deep Learning Tool" https://www.tensorflow.org/

[160] Inaba, M.; Katoh, N.; Imai, H. (1994). Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. Proceedings of 10th ACM Symposium on Computational Geometry. pp. 332339. doi:10.1145/177424.178042.

[161] Schrijver. Combinatorial optimization. Polyhedra and eciency. Vol. A, volume 24 of Algorithms and Combinatorics. Springer-Verlag, Berlin, 2003. ISBN 3-540-44389-4. Paths, flows, matchings, Chapters 138.