Ministère de l'Enseignement Supérieur, de la Recherche Scientifique et de l'Innovation (MESRSI)

Université Nazi BONI (UNB)

Ecole Supérieure d'Informatique (ESI)

Master Système d'information Option : Système d'aide à la décision (SI/SAD)

Web mining for the detection of hints to terrorism

Fouille des données web pour la détection des relents du terrorisme

Author:
M. José Arthur OUEDRAOGO

Supervisor:
Dr Ferdinand GUINKO
Department of computer science
University Ouaga 1, Pr Joseph KI-ZERBO

Abstract

Abstract — Since the events of 11 September 2001, terrorism has spread throughout the world, threatening the safety and tranquility of humanity. As a potential contributor to the development of advanced technologies, data mining finds multiple application in predicting terrorist attacks or any trace of terrorism on the web. Web mining can be generally defined as the application of data mining techniques to extract useful knowledge from the Web Data. Web mining can be further categorized as web content that includes text, images, record etc, web structure which includes hyperlinks, tags etc, and web usage including http logs, app server logs. The use of the Web Mining technique that focuses on extracting web access patterns from usage data will allow early detection of terrorist attacks so that prevention can be done on time. However, the approach we propose in this study to help understand the behavior of Internet users has three phases: pre-treatment Logs files, classification of pages and classification of Internet users. In the pre-processing phase, the queries are organized into visits that represent the units of interaction between the web users and the web server. In the page classification phase, an internal representation of the website is created from the log files to retrieve navigation paths.

Keywords: Data mining, Data mining algorithms, weka, Facebook graph api, hints of terrorism and web usage mining.

Contents

List of Figures

List of Tables

List of Acronyms

ECD	*Extraction process from databases*
GTD	*Global Terrorism Databases*
NLP	*Natural Language Processing*
NLTK	*Natural Language Toolkit*
WCM	*Web Content Mining*
WSM	*Web Structure Mining*
WUM	*Web Usage Mining*

General Introduction

Web mining is the application of data mining techniques to discover constants, patterns or models in Internet resources or data about it. Nowadays, web data mining is a powerful software technology used to automatically obtain correlations between computerized data from different sources predict events or establish strategies. Thus, for example, "data mining" allows financial organizations to reduce credit card fraud by analyzing transactions in real time. Adapted to large-scale distribution, these software are able, after having analyzed the behavior of consumers through their purchases, to suggest the outline of a marketing campaign... In the security context, web data mining nowadays appears to be a program that uses queries, searches or analyzes in one or more databases to discover or locate predictive forms or indicator abnormalities in a terrorist database or criminal act. Aware of this reality, we decided to conduct studies on a current topic that is entitled: « Web mining for the detection of hints to terrorism». The present paper is organized as follows. The first part presents a motivation and background of our study, objects of the first three chapters. The second part consists of the last three chapters and is devoted to presenting the proposed methodology for extracting knowledge from social networks. The last chapter presents the results and the discussion related to our study.

Chapter I

MOTIVATION AND BACKGROUND

Introduction

Conducting a project on a given theme requires excellent mastery of the study theme. To do this, we let's first present the context of our study. Then it will be question for us to present a review of the literature. Finally, we will discuss the process of web usage mining (WUM).

i PRESENTATION OF CONTEXT OF THIS STUDY

The characterization of Internet users who frequent a website and the identification of their browsing patterns is a key issue for web designers who aim to assist the user, predict his behavior and personalize the consultation. These three considerations motivated significant efforts in the analysis of web user traces on Web sites and the adaptation of classification methods to web data in recent years. The present work is part of this research, proposing a Logs file processing methodology to study the behavior of users of a website by exploiting different classification methods.

i.1 Context of this study

With the exponential growth in the number of online documents and new pages each day, the Web has become the primary source of information. This development has resulted in a rapid growth of web-based activity, and an explosion of data resulting from this activity. To analyze these types of data, new methods of analysis have emerged, grouped under the term "Web Mining" whose three current development axes are the Web Content Mining (WCM) which is interested in the analysis of the contents of the pages. Web, the Web Structure Mining (WSM), which is interested in studying the links between Web sites and the Web Usage Mining (WUM) which is interested in the study of the use of the Web. This last branch of Web Mining, defined as the application of the Knowledge Extraction process from databases (ECD) to data. Creators of Web sites interested in retaining Internet users who visit their sites and seeking to attract new visitors need to analyze the behavior of Internet users in order to extract web access patterns for improvement and a customization of the sites.

i.2 Description of the problem

Today, many Web Usage Mining projects have been conducted by researchers. Some of this work has focused on the study of the first phase of the WUM process, namely the pre-processing of data, while others have focused on determining the behavioral patterns of Internet users who visit websites. This second axis is the focus of our research work, which is the subject of this paper. Indeed, assuming that there is some correlation between the different visitor practices on a website and their personal characteristics, our goal is to build navigation profiles enriched with user traits. In other words, we seek to identify and qualify groups of users in relation to their motives for browsing on a website or traits representing interests in particular those of terrorism.

i.3 Problematic

Data mining is a program that uses queries, searches, or analyzes in one or more databases to discover or locate predictive forms or anomalies indicative of a terrorist or criminal act. A definition certainly different from the traditional definition, but the document clearly outlines the specific issues of pattern search to identify and predict behavior. The problem of data growth and the difficulty of identifying the relevant information is traditional, but in the specific context of terrorism the speed of reaction is a differentiating element of marketing or risk where the time factor is not so critical. Data mining tools are set up to help human stakeholders in the detection and analysis of "suspicious forms". While in companies data mining is an element of competitiveness by allowing better understanding or analysis of data, in the field of anti-terrorist research, data mining is a way of linking new sources of information and to allow collaboration between international organizations.

i.4 Definition of concepts

This section will allow us to define some key concepts of our study.

Web mining

Web mining is the process of using data mining techniques and algorithms to extract information directly from the Web by extracting it from Web documents and services, Web content, hyperlinks and server logs. The goal of Web mining is to look for patterns in Web data by collecting and analyzing information in order to gain insight into trends, the industry and users in general. Web mining can also defined as a branch of data mining concentrating on the World Wide Web as the primary data source, including all of its components from Web content, server logs to everything in between. The contents of data mined from the Web may be a collection of facts that Web pages are meant to contain, and these may consist of text, structured data such as lists and tables, and even images, video and audio [52].

Categories of Web mining:

- **Web content mining** — This is the process of mining useful information from the contents of Web pages and Web documents, which are mostly text, images and audio/video files. techniques used in this discipline have been heavily drawn from natural language processing (NLP) and information retrieval.

- **Web structure mining** — This is the process of analyzing the nodes and connection structure of a website through the use of graph theory. There are two things that can be obtained from this: the structure of a website in terms of how it is connected to other sites and the document structure of the website itself, as to how each page is connected.

- **Web usage mining** — This is the process of extracting patterns and information from server logs to gain insight on user activity including where the users are from, how many clicked what item on the site and the types of activities being done on the site.

Data mining

Data Mining is a non-basic process of updating relationships, correlations, dependencies, associations, models, structures, trends, classes, factors obtained by navigating through large sets of data, usually recorded in databases (relational or not), navigation using methods mathematical, statistical or algorithmic [1, 51]. Also, Data mining is the process of analyzing hidden patterns of data according to different perspectives for categorization into useful information, which is collected and assembled in common areas, such as data warehouses, for efficient analysis, data mining algorithms, facilitating business decision making and other information requirements to ultimately cut costs and increase revenue. Data mining is also known as data discovery and knowledge discovery. **The major steps involved in a data mining process are:**

- Extract, transform and load data into a data warehouse;
- Store and manage data in a multidimensional databases;
- Provide data access to business analysts using application software;
- Present analyzed data in easily understandable forms, such as graphs.

The first step in data mining is gathering relevant data critical for business. Company data is either transactional, non-operational or metadata. Transactional data deals with day-to-day operations like sales, inventory and cost etc. Non-operational data is normally forecast, while metadata is concerned with logical database design. Patterns and relationships among data elements render relevant information, which may increase organizational revenue. Organizations with a strong consumer focus deal with data mining techniques providing clear pictures of products sold, price, competition and customer demographics.For instance, the retail giant Wal-Mart transmits all its relevant information to a data warehouse with terabytes of data. This data can easily be accessed by suppliers enabling them to identify customer buying patterns. They can generate patterns on shopping habits, most shopped days, most sought for products and other data utilizing data mining techniques [10, 11]. The second step in data mining is selecting a suitable algorithm - a mechanism producing a data mining model. The general working of the algorithm involves identifying trends in a set of data and using the output for parameter definition. The most popular algorithms used for data mining are classification algorithms and regression algorithms, which are used to identify relationships among data elements. Major database vendors like Oracle and SQL incorporate data mining algorithms, such as clustering and regression tress, to meet the demand for data mining.

Text mining

Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text. High-quality information

is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interest. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling. Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods [35, 36, 37]. A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted.

ii LITERATURE REVIEW

This literature review will guide us in the project's progress. It is divided into three main themes: similar projects already carried out in the literature, the mathematical classification of data mining algorithms and the tools used in data mining.

ii.1 Organization of hints of terrorism

Terrorism is calculated to create an atmosphere of fear and alarm through violence or threat of violence [21, 23]. Terrorist attacks can cause mass casualties, infrastructure damage, or public concern with great impact. The motive of terrorist activist mainly comes from

politics, religion, revenge, etc. The aim is not violence itself but to cause political and religious changes[24, 28]. To achieve their goals terrorists use several channels such as:

- **Cyber-terrorism** is one of the major terrorist threats posed to our nation today. As we have mentioned earlier, there is now so much of information available electronically and on the web. Attack on our computers as well as networks, databases and the Internet could be devastating to businesses. It is estimated that cyber-terrorism could cause billions of dollars to businesses. For example, consider a banking information system. If terrorists attack such a system and deplete accounts of the funds, then the bank could loose millions and perhaps billions of dollars. By crippling the computer system millions of hours of productivity could be lost and that equates to money in the end. Even a simple power outage at work through some accident could cause several hours of productively loss and as a result a major financial loss. Therefore it is critical that our information systems be secure [25, 26, 27].

- **Network intrusions.** What happens here is that intruders try to tap into the networks and get the information that is being transmitted. These intruders may be human intruders or Trojan horses set up by humans. Intrusions could also happen on files. For example, one can masquerade as some one else and log into someone else's computer system and access the files. Intrusions can also occur on databases. Intruders posing as legitimate users can pose queries such as SQL queries and access the data that they are not authorized to know. Essentially cyber terrorism includes malicious intrusions as well as sabotage through malicious intrusions or otherwise. Cyber security consists of security mechanisms that attempt to provide solutions to cyber attacks or cyber terrorism. When we discuss malicious intrusions or cyber attacks, we need to think about the non cyber world, that is non information related terrorism and then translate those attacks to attacks on computers and networks. For example, a thief could enter a building through a trap door. In the same way, a computer intruder could enter the computer or network through some sort of a trap door that has been intentionally built by a malicious insider and left unattended

through perhaps careless design. Another example is a thief entering the bank with a mask and stealing the money [22, 32, 33, 34].

Most terrorist attacks plan on the internet can often be traced to the sources of the organizers of these attacks through WUM techniques [29, 30, 31].

ii.2 Similar projects in the literature

In depth research related to the hints of terrorism and web mining was carried out. However, we have not been able to find articles that deal directly with the search of web mining for the detection of terrorism. But so we did research on projects similar to the one realized in this memoir.The articles selected for this literature review are those that we believe to be the most relevant. They all follow a general approach of Web mining or data mining techniques by detailing all steps. Through our reading, web mining is a powerful software technology used to automatically obtain correlations between computerized data from different sources in order to predict events or strategies in many domains of daily life. Web mining through the use of data mining techniques is widely used in domain medical for the prevention of certain pathologies including for the prediction of the presence or the absence of a problem of heart type, [3]. The second most important issue according to our research is the prediction of all types of cancer, especially breast cancer [1, 2, 4]. Having seen the impact of data mining in the health field, we also contemplate the influence of data mining on the security aspect. Web mining, through the use of data mining techniques is much in the field of the national security of a country collecting the sensitive data to threaten the security to propose new measures of security. It is in this sense that during our reading articles dealing with the classification of terrorist threats, [5]. Also articles dealing with the comparison of jihad, homicides and crimes were met, the purpose of these readings is to apply data mining techniques to predict terrorist attacks or to find measures to predict terrorism,[6, 7, 8]. Finally,these techniques prove to be an important tool for countering terrorism, [9].

ii.3 Classification of data mining algorithms and Data mining techniques

For this project, we decided to use data mining to reach our goal.For many years, data mining has been used by several marketing companies around the world [10]. Data mining combines statistical analysis, intelligence artificial as well as technologies to extract relationships and patterns in huge databases [15]. Data mining algorithms can follow three different learning approaches: supervised, unsupervised, or semi-supervised. In supervised learning, the algorithm works with a set of examples whose labels are known. The labels can be nominal values in the case of the classification task, or numerical values in the case of the regression task. In unsupervised learning, in contrast, the labels of the examples in the dataset are unknown, and the algorithm typically aims at grouping examples according to the similarity of their attribute values, characterizing a clustering task. Finally, semi-supervised learning is usually used when a small subset of labelled examples is available, together with a large number of unlabeled examples. The algorithms resulting from this comparison are as follows: Decision Tree, K-Nearest Neighbor, Support Vector Machines, Naive Bayesian Classification and Neural Networks [11, 12].

- **Decision Tree** A Decision Tree Classifier consists of a decision tree generated on the basis of instances. A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces a certain discrete function of the input attributes values . The estimation criterion in the decision tree algorithm is the selection of an attribute to test at each decision node in the tree. The goal is to select the attribute that is most useful for classifying examples. A good quantitative measure of the

worth of an attribute is a statistical property called information gain that measures how well a given attribute separates the training examples according to their target classification. This measure is used to select among the candidate attributes at each step while growing the tree.The result of the experiment shows that the best splitting attribute in each case was found to be outlook with the same order of splitting attributes for both indices. The decision tree techniques has the restriction that the training tuples should reside in memory, so, in the case of very large data, decision tree construct therefore becomes inefficient due to swapping of the training tuples in and out of the main and cache memories. ID3 and C4.5 Algorithm are used as decision classifier model.

$$G(x,y) = H(x) - \sum_{i \in value(y)} \frac{|\Delta y_i|}{|\Delta y|} H(y_i)$$

Figure I.1: Mathematical method of Decision Tree algorithm

- **K-Nearest Neighbor Classifiers (KNN)** KNN (K-nearest neighbors) algorithm is a supervised learning method. It can be used for both regression and classification. To make a prediction, the K-NN algorithm will not compute a predictive model from a Training Set as is the case for logistic regression or linear regression. Indeed, KNN does not need to build a predictive model. Thus, for KNN there is no learning phase proper. This is why it is sometimes categorized in Lazy Learning. To make a prediction, K-NN relies on the dataset to produce a result. K nearest neighbors (KNN) is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (distance function).KNN has been used in statistical estimation and pattern recognition since 1970's. A case is classified by a majority voting of its neighbors, with the case being assigned to the class most common among

its K nearest neighbors measured by a distance function. If K=1, then the case is simply assigned to the class of its nearest neighbor.

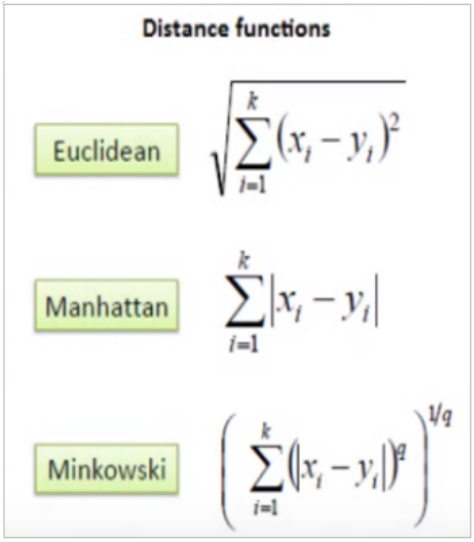

Figure I.2: Mathematical method of KNN algorithm

- **Support Vector Machine (SVM)** SVM was first introduced by Vapnik and has been very effective method for regression, classification and general pattern recognition. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra. This is where the kernel plays role. For linear kernel the equation for prediction for a new input using the dot product between the input (x) and each support vector (xi) is calculated as follows:

$$f(x) = B(0) + \text{sum}(ai * (x,xi))$$

Figure I.3: Mathematical mathod of SVM algorithm

Data mining is highly effective, so long as it draws upon one or more of these techniques:

- **Tracking patterns.** One of the most basic techniques in data mining is learning to recognize patterns in your data sets. This is usually a recognition of some aberration in your data happening at regular intervals, or an ebb and flow of a certain variable over time. For example, you might see that your sales of a certain product seem to spike just before the holidays, or notice that warmer weather drives more people to your website.

- **Classification.** Classification is a more complex data mining technique that forces you to collect various attributes together into discernible categories, which you can then use to draw further conclusions, or serve some function. For example, if you're evaluating data on individual customers' financial backgrounds and purchase histories, you might be able to classify them as "low," "medium," or "high" credit risks. You could then use these classifications to learn even more about those customers.

- **Association.** Association is related to tracking patterns, but is more specific to dependently linked variables. In this case, you'll look for specific events or attributes that are highly correlated with another event or attribute; for example, you might notice that when your customers buy a specific item, they also often buy a second, related item. This is usually what's used to populate "people also bought" sections of online stores.

- **Outlier detection.** In many cases, simply recognizing the overarching pattern can't give you a clear understanding of your data set. You also need to be able to identify anomalies, or outliers in your data. For example, if your purchasers are almost exclusively male, but during one strange week in July, there's a huge spike in female purchasers, you'll want to investigate the spike and see what drove it, so you can either replicate it or better understand your audience in the process.

- **Clustering.** Clustering is very similar to classification, but involves grouping get her based on their similarities. For example, you might choose to cluster different

demographics of your audience into different packets based on how much disposable income they have, or how often they tend to shop at your store.

- **Regression.** Regression, used primarily as a form of planning and modeling, is used to identify the likelihood of a certain variable, given the presence of other variables. For example, you could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression's main focus is to help you uncover the exact relationship between two (or more) variables in a given data set.

- **Prediction.** Prediction is one of the most valuable data mining techniques, since it's used to project the types of data you'll see in the future. In many cases, just recognizing and understanding historical trends is enough to chart a somewhat accurate prediction of what will happen in the future. For example, you might review consumers' credit histories and past purchases to predict whether they'll be a credit risk in the future.

ii.4 Data mining tools

Data mining has a wide number of applications ranging from marketing and advertising of goods, services or products, artificial intelligence research, biological sciences, crime investigations to high-level government intelligence. Due to its widespread use and complexity involved in building data mining applications, a large number of Data mining tools have been developed over decades. Every tool has its own advantages and disadvantages. Data mining provides many mining techniques to extract data from databases. Data mining tools predict future trends, behaviors, allowing business to make proactive, knowledge driven decisions. The development and application of data mining algorithms requires use of very powerful software tools. As the number of available tools continues to grow the choice of most suitable tool becomes increasingly difficult. The top six open source tools available for data mining are briefed as below [13, 14]. The data mining tools we have been

paying attention to during this review of the literature are: Rapidminer, Orange, Weka, Knime, kell and R.

- **RAPIDMINER**

  RAPIDMINER is a software platform developed by the company of the same name

Figure I.4: Rapidminer for Data mining
Download Rapidminer

that provides an integrated environment for machine learning, data mining, text mining, predictive analytic and business analytic. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process. Rapid Miner uses a client/server model with the server offered as Software as a Service or on cloud infrastructures.

- **WEKA (Waikato Environment for Knowledge Analysis)**

  Weka is a collection of machine learning algorithms for data mining tasks. These algorithms can either be applied directly to a data set or can be called from your own Java code. The Weka (pronounced Weh-Kuh) workbench contains a collection of several tools for visualization and algorithms for analytics of data and predictive modeling, together with graphical user interfaces for easy access to this functionality.

Figure I.5: Weka Data mining
Download Weka

- **ORANGE**

  Orange is a component-based data mining and machine learning software suite, fea-

Figure I.6: Orange Data mining
Download Orange

turing a visual programming frontend for explorative data analysis and visualization, and Python bindings and libraries for scripting. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It is implemented in C++ and Python. Its graphical user interface builds upon the cross-platform framework.

ii.5 Methods used in data mining projects

Before starting such a project, it is necessary to determine which method we should privilege to have a rigorous and complete approach. Various methods are used in data mining projects. Indeed, several authors have stated methodologies to follow. In this literature

review, two methods are stated. First, there are the phases of a data mining project according to the *Cross industry model standard process for data mining*. According to them, a project of this kind is an iterative process and adaptive [17]. It was developed in 1996 by a group of experts [18]. Figure 2.1 illustrates this method. Each of the steps presented

Figure I.7: Cross-Industry Standard Process for Data Mining (CRISP-DM)
More info

in Figure I.7 has specific objectives and we can turn to a previous step as needed. It is not a linear process. In this translation, the author speaks of understanding the craft, but in other articles he name this step understanding the problem. It is necessary, among others in this step, determine our goals. Then there are the stages of understanding and preparation of data where one works on the present data. The preparation of the data is one of the most important aspects in a data mining project [17]. Then it is the model at this stage, the analyst chooses the models and will test them with several times by changing the

settings to determine which is the best model. Always at this stage, he must determine the evaluation criteria to compare the models between them. Then comes the evaluation stage. At this point, we must ensure that our results are comply with the criteria of commercial success. Finally, the last step according to this method is deployment, that is, using new knowledge to make improvements in the processes [17]. Second, Berry and Linoff have also developed a method called the virtuous circle of data mining [19]. As the name suggests, it is still not a process linear, but in this case a circle.

Note that in the virtuous circle of data mining, shown in Figure I.8, there is less in the

Figure I.8: Virtuous Cycle of Data Mining
More info cycle

model presented in Figure I.7. The first step is the identification of business opportunities is to identify the field of study. In addition, it is necessary to determine the objective of the project at this stage. Subsequently, the analyst must analyze the data present and group them together. In the third step, the analyst implements the techniques of selected data mining. Finally, just like the CRISP evaluation step, it measure the results of actions [20].

iii WEB USAGE MINING PROCESS

The Web Usage Mining (WUM) process is commonly divided into three main stages: data preprocessing, data mining, and results analysis. A preliminary step is to collect the data from the Web to analyze but as part of our study we will use a social network. We present in this chapter each of these stages as well as a detailed description of the data and the treatments necessary for its realization.

iii.1 Web Usage Mining Process

WUM is the application of data mining techniques to discover patterns of use from web data in order to better understand and serve the needs of web applications. The first step in the WUM process, once the data is collected, is the pretreatment of the Logs files which consists of cleaning and transforming the data. The second step is data mining to discover association rules, a sequence of web pages often appearing in visits and "clusters" of users with similar behaviors in terms of content visited. The analysis and interpretation stage closes the WUM process. It requires the use of a set of tools to keep only the most relevant results.

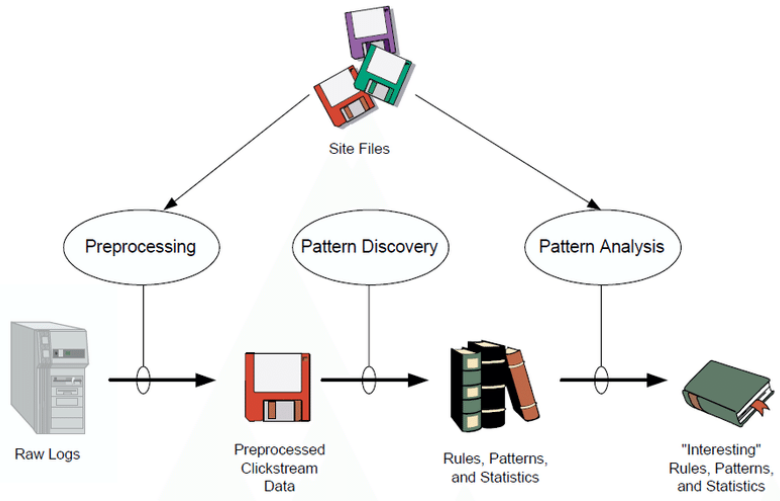

Figure I.9: Web mining process
More info

iii.2 Collection of data

The first step in the WUM process is to collect the web data to be analyzed. The two main sources of data collected are the data recorded at the server level and the data recorded at the client level. Another source is data stored at the proxy server, which is intermediate in client-server communication. Nowadays new technologies of data collection namely APIs are available thanks to the social networks that store multiple data.

Data recorded at the server level

Each request to display a web page from a user can generate several requests. Information about these queries (including the names of the requested resources and the Web server responses) are stored in the Log files of the Web server. Logging data into Server-side Logs (Logs) is used to identify the set of users accessing the Web site. In addition, server logs provide content data, structure information, and meta-information about web pages (file size, date of last change). However, the Log files of the Web servers present major problems as reported in the previous chapter.

Data recorded at the client level

Data is collected at the client node through agents implemented in Java or Java script. These agents are embedded in the web pages (in the form of java applets, for example) and used for a direct collection of information from the client station (examples of information: the time of access and abandonment of the site, browsing history). Another data collection technique is to use a modified version of the browser. This technique allows you to save the web pages visited by a user as well as access time and response time and send them to the server. The first method is to collect data on a user navigating on a single website. On the other hand, a modified browser allows the collection of data on a user navigating on several Web sites. The problem that arises in the second case is how to convince Internet users to use this modified browser in their browsing knowing that it can be considered a

threat to their privacy. The information recorded at the client station is more reliable than the information recorded at the server since it solves the problem of caching and session identification.

Twitter and social media as sources

Twitter as a data source has gained lot of prominence in recent years. It is ranked as one of the top 10 popular websites, having 400 million registered users and over 500 million tweets generated everyday [38]. Additionally, information about disasters can be extracted from news channels and blogs through APIs, RSS feed or web scraping. Sentiment analysis [39], [40], stock market [41], public health [42], general public mood and finding political alignments [43, 44] are some of the areas where twitter data have been used.
Some of the advantages of tweets are as follows:

- Although it is unstructured, it has some structure by the limitation of 140 characters;
- It can use hashtags, which give semantic annotations of the tweets;
- The tweets have geocodes, which can help us spatially map the sources of the tweets.

Following are the fields that are available from twitter:

- archive source: API source of the tweet (twitter-search or twitter-stream);
- text: contents of the tweet itself, in 140 characters or less;
- to user id: numerical ID of the tweet recipient;
- from user: screen name of the tweet sender,
- id: numerical ID of the tweet itsel;,
- from user id: numerical ID of the tweet sender,

- iso language code: code (e.g. en, de, fr, ...) of the sender's default language (not necessarily matching the language of the tweet itself);
- source: name or URL of the tool used for tweeting (e.g., tweet deck, ...);
- profile image url: URL of the tweet sender's profile picture,
- geo type: form in which the sender's geographical coordinates are provided;
- geo coordinates 0: first element of the geographical coordinates;
- geo coordinates 1: second element of the geographical coordinates;
- created at: tweet timestamp in human–readable format (set by the tweeting client inconsistent formatting);
- time: tweet timestamp as a numerical unix timestamp.

iii.3 Pre-treatment of data and Data transformation

Data pre-processing is divided into two main phases: a data cleansing phase and a transformation phase.

After collecting the data, it will be necessary first to clean and prepare the data in order to be able to analyze these data. For this we used some text mining technique described in the figure below : In order to improve the performance and data quality, all the irrelevant characteristics are debarred while the data is being uploaded into RapidMiner tool. The major steps involve the separation of the document into tokens; this task is called Tokenization [16]. The next step is concerned with the transformation process of all the characters where each document title is created in a lower case. Stop words filtering is involved in the third step, where English is filtered through this operator. A single English word is required to be signified by each token. All tokens that were similar to stop words were eradicated from the provided document by an operator. The document must have only one stop word per line. The last step is concerned with the text processing phase

Figure I.10: Text mining process

that involves filtering the tokens according to the length. The minimum number of the characters that the token should have is 4, while the maximum number is 25 characters.

iv THE HELD SOLUTION

Several interesting solutions for solving our problem have been encountered in our review of the literature. It is with this in mind that we will consider taking inspiration from the data collection technique proposed by twitter [38, 45]. As well as the J48 algorithm of the weka tool [46] for analyzing the data collected.

Conclusion

In this chapter, we presented the context of our study, the review of the literature and the different phases of the WUM process. Our study is centered on the exploration of data, the following chapter is dedicated to the methodology of solving our problematic.

Chapter II

METHODOLOGY

Introduction

Any data mining project follows a precise approach as set out in *Figure I.7.* For this project, we decided to follow the *Cross Industry Standard model method Process for Data Mining* [17].In this section, a clear and transferable description of the project is produced. This is the actual collection of data, the preprocessing of data and finally the use of techniques with the Weka tool [46].

i Data Set Collector

The "Terrorist Attacks Dataset' where taken from the Global Terrorism Database. The GTD data set is an open source database, most comprehensive and world's most significant data available on terrorism incidents used for the experiment, taken from the National Consortium for the study of terrorism and Responses of Terrorism (START) initiative at the University of Maryland, which broadcasts the terrorism incidents report about the world from 1970 to 2016, and includes information about more than 170,000 terrorist events as well as the vast information more than 120 variables, and include information on more

than 83,000 bombings, 18,000 assassinations, and 11,000 kidnappings since 1970. Also, includes information on at least 45 variables for each case.But in the case of our study we used Facebook's Graph API to collect data that is related to the hype of terrorism in posts as well as in communication groups [46].

ii Data Set collector with using Facebook's GRAPH API

Facebook as a data source has gained a lot of importance in recent years. It is ranked among the 5 most popular websites, with around 2.5 billion registered users and over 3 billion posts generated each day. Additionally, information about APIs, RSS feed or web scraping. Sentiment analysis, stock market, public health, general public mood and finding political alignments are some of the areas where Facebook data has been used.Facebook has several APIs to facilitate the development of solution interacting with its data. These APIs including Graph API. The Graph API is the primary way to get data into and out of the Facebook platform. It's an HTTP-based API that apps can use to programmatically query data, post new stories, manage ads, upload photos, and perform a wide variety of other tasks. The Graph API is named after the idea of a "social graph" — a representation of the information on Facebook. It's composed of :

- **nodes** — basically individual objects, such as a User, a Photo, a Page, or a Comment;
- **edges** — connections between a collection of objects and a single object, such as Photos on a Page or Comments on a Photo;
- **fields** — data about an object, such as a User's birthday, or a Page's name.

Typically you use nodes to get data about a specific object, use edges to get collections of objects on a single object, and use fields to get data about a single object or each object in a collection.

Note : Thanks to the many features offered by the *Graph API* we will use it for data collection because its implementation is very easy. Once this data is collected we will move on to the next step of processing the data.

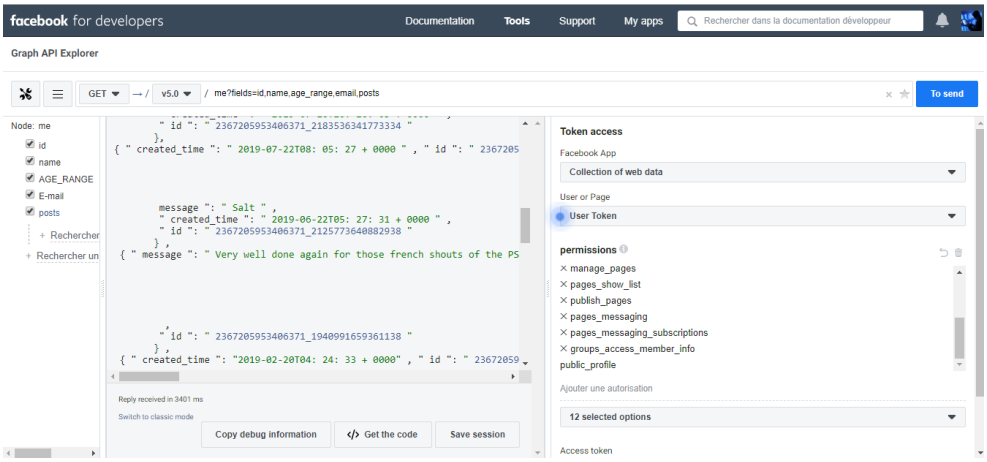

Figure II.1: Example of data with Graph API

iii Data Pre-Processing

Before applying techniques (algorithms) usually some pre- processing is performed on the dataset. It is necessary to improve the data quality to accomplish data processing. There are a few number of techniques used for data processing as data aggregation, data sampling, data discretization, variable transformation, and dealing with missing values.It is in this sense that for the data cleaning, we'll leverage Python's Pandas and NumPy libraries to clean data. For the preparation of the data we will base ourselves on the techniques of the Natural Language Processing (NLP) is a part of computer science and artificial intelligence which deals with human languages. In order to achieve this goal we will use the **NLTK** library of python.

Natural language toolkit (NLTK) is the most popular library for natural language processing (NLP) which was written in Python and has a big community behind it. NLTK also is very easy to learn, actually, it's the easiest natural language processing (NLP)

library that you'll use.

iv Using techniques with Weka Tool

The algorithms in this research are implemented based on WEKA. WEKA is an open source tool created in Java, a collection of machine learning algorithms allows the researcher to mine his own data for trends and patterns and the algorithms can have applied directly to a dataset. In this paper, you will discover the various ways that you can estimate the performance of your machine learning model in Weka tool. How to Evaluating your model using the training dataset, evaluating your model using a random train and test split, assess your model using k-fold cross-validation.There are some model evaluation techniques that you can choose from, and the Weka machine learning workbench others four of them: Training Dataset: prepare your model within the entire training dataset, after that, evaluate the model on the same dataset. Supplied Test Set: divided your dataset is manually utilizing another program. Percentage Split: randomly split your dataset into training and a testing partitions each time you evaluate a model. Cross-Validation: split the data into k-partitions or folds. Train a model on all of the partitions except one that is usually held away as test set after that calculate the average performance of all k models. And you can see these techniques in the Weka tool explorer on the classify tab after you have loaded a dataset. Also, each test option has a time. Evaluation options are concerned with determining the performance of a model on unseen data. Predictive modeling aims to build a model that performs best in a situation that we do not entirely understand, the future with new unknown data. We must use these types of robust statistical techniques to best estimate the performance of the model in this situation and the performance summary is provided in Weka when you evaluate a model. Whenever evaluating a machine learning algorithm on a classification issue, you are given a massive total of performance information to digest because of classification could be the most analyzed type of predictive modeling issue, and there are a wide variety of ways to consider the performance of classification

algorithms. Therefore, the first thing to note in the performance for classification algorithms is classification accuracy the ratio of the number of correct predictions away from all forecasts made frequently presented as a percentage where 100% is the foremost an algorithm can perform. The second one is accuracy by class take note of the real positive and false positive rates to get the predictions for each class which may be instructive from the class break down to get the problem is uneven, or you will find more than two classes. As well as the last one is confusion matrix a table showing the number of predictions for each class in comparison to the number of instances that actually participate in each class. The terrorism data set of world wide is splitting into two main sets: Training dataset with 66% percentage and Testing dataset with 34% percentage from the whole data set, and that is applied by using the default setting of Weka tool [47, 48, 49, 50].

Conclusion

This chapter allows us to show the methodology of our solution namely in present our data collection tool, then the techniques we use for the processing of collected data and finally the algorithm and tools for analyzing this data. This chapter gives us the basics to start the next one or we present the results of our solution.

Chapter III

RESULTS AND DISCUSSIONS

In this chapter, we will present the results obtained during our experiments on the hints of terrorism. The other part summarizes analysis and discussion related to the results obtained.

i Results and discussions

In our experience, we have applied different algorithms to our database dealing with the hints of terrorism. The pre-processed dataset consists of 286 instances of data is converted to *. ARFF which is a file to use by the Weka tool. Each test option has a time. Therefore, the algorithms result obtained according to two test options which are:

- Evaluation on Test split that divides the input data set into 66% for the training data and 34% for the test set.
- 10 Fold Cross-Validation.

The results from the applied classification algorithms in the two approaches will be evaluated according to four performance measures which are the classification Accuracy, Recall, F-measure also called F-Score. In case of dividing the input data set into 66% for the

training data and the remaining 34% for testing, the results are shown in Table I which provide a clear comparison among the selected classifiers according to accuracy, precision, recall and F-measure which shows that:

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Precision	Recall	F-Measure
Decision Tree	209	77	0.777	0.866	0.819
KNN	203	83	0.748	0.886	0.811
Naive Bayes	204	82	0.769	0.846	0.819
J48	216	70	0.757	0.96	0.846
SVM	196	90	0.742	0.846	0.791

Table III.1: Test split and accuracy results

From the accuracy point of view, J48 correctly classified about 75.7% of the data it means 118 items out of 184 in the 34% test split of the data SVM is outperformed Bayes Net which correctly approximately 61.413% of the data. It is evident that the accuracy of KNN achieved the lower accuracy 51.087% among the other classifiers although it has the lowest precision, recall, and f-measure from other classifiers. Bayes Net classifier produces higher precision, f-measure values than other classifiers, and J48 classifier has the higher recall from different classifiers. The overall performance of NB is near from SVM classifier. It is obvious that a comparison is applied on our five classifiers due to precision, recall, and F-measure which shows us that J48 has the highest accuracy than other classifiers, it does not mean that it performs well in other results. Bayes Net classifier performs also well in all results. The overall performance of NB is near from SVM results. KNN has lower accuracy than all other classifiers although it performs well in other measures.

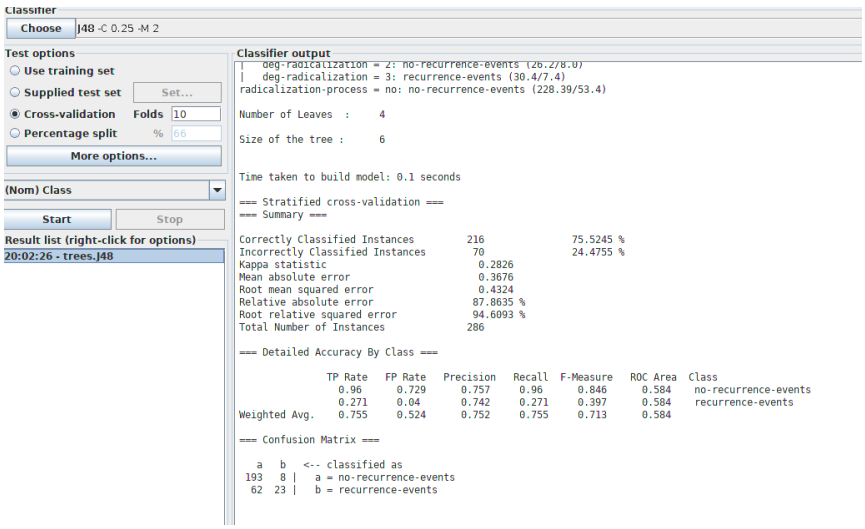

Figure III.1: Performance measures results of used classifiers j48 algorithm.

These pictures above shows the performance measures in case of using classification based on 10-fold cross-validation where J48 has lower precision, F-measure values than SVM, but it could not consider more accurate than J48. NB classifier performs well and very near from Bayes Net especially in recall and F- measure results. KNN has the lowest classification accuracy also it performs well in other measures.

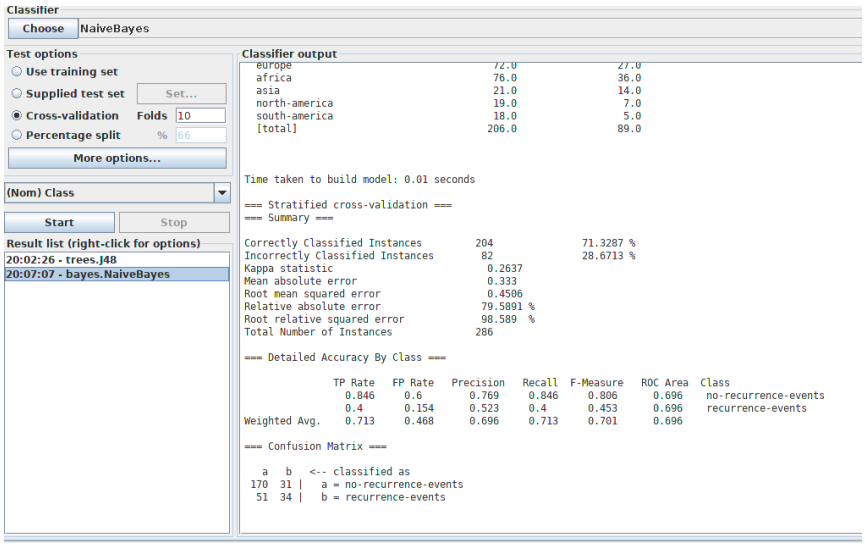

Figure III.2: Performance measures results of used classifiers Naive bayes algorithm.

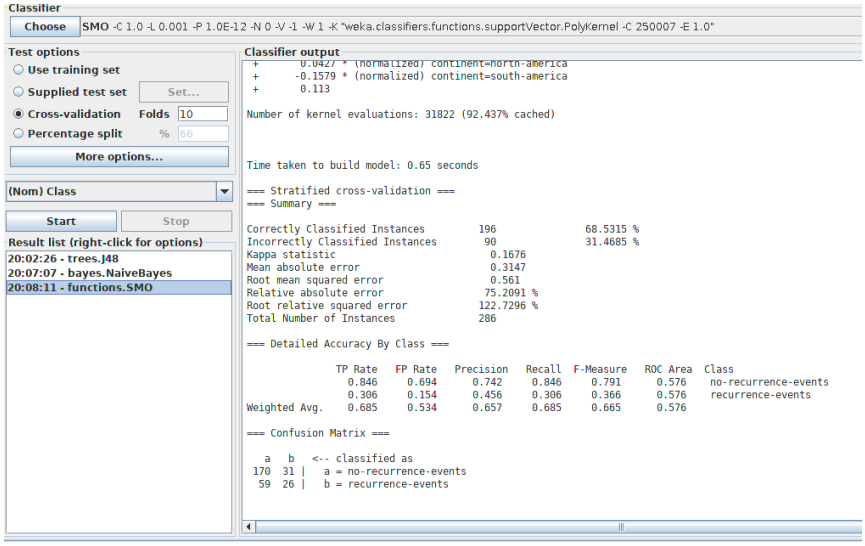

Figure III.3: Performance measures results of used classifiers svm algorithm.

CONCLUSION

Weka software is very powerful software for study or research needs in the field of prediction and automatic data learning. He has a plurality of classification filtering and learning algorithms that vary from one others according to their use and performance. This architecture available and accessible open source allows to add to this powerful tool other algorithms. Indeed, some classifying and learning algorithms written in Python could be introduced in the software by python communications interfaces. The translation of many mathematical functions in the Python language especially for series geometric and arithmetic is very advanced thanks to the development of numerous libraries and its proximity to computing units. This project allows us to have a broad initiation to data analysis with artificial intelligence algorithms by increasing knowledge about the different classifiers and their implementation in the language python.

During this thesis, we presented several models of prediction, some of regressive type and other classification. We have been able to detect and predict the hints of terrorism from many explanatory variables. So we responded to our objective which was to detect the hints of terrorism.

In the case of test split of the input data with divisions, 75% for training data and 25% for testing data showed that J48 is more accurate than another classifier especially Naive Bayes, SVM and Bayes Net, the overall performance of NB and SVM is very near.

KNN has the lowest accuracy, but it performs well in other measures. In 10-fold cross validation case, NB classifier is very near to the Bayes Net accuracy, recall, and f-measure.

NB classifier performs as Bayes Net in most measures, and J48 performs worse than other classifiers in precision and F- measure.

Finally, some researchers could perform a modification of this research by using different methods. Others could use different test options to test the performance of the classification algorithms.

Bibliography

[1] Vivek Kumar, Brojo Kishore Mishra, Manuel Mazzara, Dang N. H. Thanh and Abhishek Verma. Feb 2019, *«Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications»*, p 8. Google Scholar

[2] Delen, D., G. Walker and A. Kadam. 2005, *«Predicting breast cancer survivability : a comparison of three data mining methods»*,Artificial intelligence in medicine, vol. 34,no2, p. 113–127. Google Scholar

[3] Dangare, C. S. and S. S. Apte. 2012, *«Improved study of heart disease prediction system using data mining classification techniques»*, International JoGoogle Scholarurnal of Computer Applications, vol. 47, no10, p. 44–48.

[4] Lee, S.-M., J.-O. Kang and Y.-M. Suh. 2004, *«Comparison of hospital charge predictionmodels for colorectal cancer patients : neural network vs. decision tree models»*, Journal of Korean medical science, vol. 19, no5, p. 677–681. Google Scholar

[5] Bing Bu, Zhenyang Pi and Lei Wang. 2019, *«Support Vector Machine for Classification of Terrorist Attacks Based on Intelligent Tuned Harmony Search»*, article, p. 12. Google Scholar

[6] Allemar Jhone P. Delima. August 2019, *«Applying Data Mining techniques in Predicting Index and non-Index Crimes»*,International Journal of Machine Learning and Computing, Vol. 9, No. 4, p. 6.

[7] Jeff Gruenewald, Brent R. Klein, Joshua D. Freilich and Steven Chermak. October 2016, *«American jihadi terrorism: A comparison of homicides and unsuccessful plots»*, Terrorism and Political Violence, p. 22. Google Scholar

[8] Bart Schuurman. 2019, *«Topics in terrorism research: reviewing trends and gaps, 2007-2016»*, article, p. 12. Google Scholar

[9] Bhavani Thuraisingham, *«Data Mining for Counter-Terrorism»*, The MITRE Corporation Burlington Road, Bedford, MA, On leave at the National Science Foundation, Arlington, VA, p. 28. Google Scholar

[10] Bellavance, F. 2017, *«Préparation de données pour le data mining»*, Cours universitaire HEC Montréal.

[11] Cristóbal Romero,Sebastián Ventura, Pedro G. Espejo and César Hervás. 2017, *«Data Mining Algorithms to Classify Students»*, p 273. Google Scholar

[12] Byung-Hoon Park and Hillol Kargupta. 2002, *«Distributed Data Mining: Algorithms, Systems, and Applications»*, article, p 22. Google Scholar

[13] Kalpana Rangra and Dr. K. L. Bansal. 2014, *«Comparative Study of Data Mining Tools»*, International Journal of Advanced Research in Computer Science and Software Engineering, p 8. Google Scholar

[14] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti and A.S.K.Ratnam. July 2012, *«A Study of Data Mining Tools in Knowledge Discovery Process»*, International Journal of Soft Computing and Engineering (IJSCE)ISSN: 2231-2307, Volume-2, Issue-3, p 22. Google Scholar

[15] Thuraisingham, B. M. and M. G. Ceruti. 2000, *«Understanding data mining and applying it to command, control, communications and intelligence environments»*, Computer Software and Applications Conference, 2000. COMPSAC 2000. The 24th Annual International, IEEE, p. 171–175. Google Scholar

[16] Verma, T., Renu, R., Gaur, *Tokenization and Filtering Process in Rapid Miner.* Int. J. Appl. Inf. Syst. 7(2), 16–18 (2014). Google Scholar

[17] Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer et R. Wirth. 2000, *«Crisp-dm 1.0 step-by-step data mining guide»*. Google Scholar

[18] Bellavance, F. 2017, *«Préparation de données pour le data mining»*, Cours universitaire HEC Montréal.

[19] Berry, M. J. and G. Linoff. 1997, *Data mining techniques : for marketing, sales, and customer support*, John Wiley and Sons, Inc. Google Scholar

[20] Saadoun, M. *«I–du data warehouse au data mart»*,.

[21] ROMAN URBAN and JIŘÍ F. URBÁNEK, *«Computer-Aided Expert System for a Ranking of New Terrorist Threats»*,Proceedings of the European Computing Conference,p.6. Google Scholar

[22] Bhavani Thuraisingham, *Data Mining for Counter-Terrorism*, The MITRE Corporation Burlington Road, Bedford, MA On leave at the National Science Foundation, Arlington, VA,p.28. Google Scholar

[23] Paul HOFFMAN, 2004, *Human Right and Terrorism*,26 Hum. Rts. Q. 932 Google Scholar

[24] Basuchoudhary A, Shughart WF, 2010 *On ethnic conflict and the origins of transnational terrorism.* Defence and Peace Economics, (21): 65-87. Google Scholar

[25] Nabie Y. Conteh and Paul J. Schmick, *Cyber-security:risks, vulnerabilities and counter measures to prevent social engineering attacks*, Google Scholar

[26] Luo X, Brody R, Seazzu A,and Burd S, *Social engineering: the neglected human factor for information security management*, Information Resources Management Journal, 2011,p1-8. Google Scholar

[27] Bisson D,*Social engineering attacks to watch out for. The state of security*,Accessed 13 November 2019. Read more

[28] Bart Schuurman, 2019, *Topics in terrorism research: reviewing trends and gaps, 2007-2016*, Institute of Security and Global Affairs (ISGA), Leiden University, The Hague, The Netherlands. Google Scholar

[29] Mikhail Petrovskiy, 30 April 2003,*A Hybrid Method for Patterns Mining and Outliers Detection in the Web Usage Log*, International Atlantic Web Intelligence Conference AWIC 2003, Advances in Web Intelligence pp 318-328. Google Scholar

[30] Yacine Slimani, Abdelouahab Moussaoui, Yves Lechevallier and Ahlem Drif, 2012 *A community detection algorithm for Web Usage Mining systems*, International Symposium on Innovations in Information and Communications Technology. Google Scholar

[31] Bhupendra Kumar Malviya and Jitendra Agrawal, 2015, *A Study on Web Usage Mining Theory and Applications*, sFifth International Conference on Communication Systems and Network Technologies. Google Scholar

[32] Shyam Varan Nath, 2007, *Crime Pattern Detection Using Data Mining*, International Conference on Web Intelligence and Intelligent Agent Technology Workshops. Google Scholar

[33] Bhavani Thuraisingham, 2002, *Data Mining, National Security, Privacy and Civil Liberties*, article, ACM SIGKDD Explorations Newsletter,p.5. Google Scholar

[34] Bhavani Thuraisingham, 2009,*Data Mining for Malicious Code Detection and Security Applications*, p6-7. Google Scholar

[35] Martin Krallinger, Alfonso Valencia and Lynette Hirschman, 2008, *Linking genes to literature: text mining, information extraction, and retrieval applications for biology.*Google Scholar

[36] Un Yong Nahm and Raymond J. Mooney, 2002, *Text Mining with Information Extraction*, p.6. Google Scholar

[37] Ah-Hwee Tan, *Text Mining:The state of the art and the challenges* Google Scholar

[38] Fred Morstatter, Jürgen Pfeffer, Huan Liu, Kathleen and M. Carley,ICWSM (2013),*Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehouse*. Google Scholar

[39] Jiang Long, Yu Mo , Zhou Ming, Liu Xiaohua and Zhao Tiejun,2011, *Target-dependent twitter sentiment classification. In: Proceedings of the 49th annual meeting of the association for computational linguistics, human language technologies*, vol. 1, p. 151–60. Google Scholar

[40] Alexander Pak and Patrick Paroubek, *Twitter as a corpus for sentiment analysis and opinion mining*,LREC (2010), pp. 1320-1326. Google Scholar

[41] Johan Bollen, Huina Mao and Xiaojun Zeng, *Twitter mood predicts the stock market J Comput Sci*, (2011), pp. 1-8. Google Scholar

[42] Michael J. Paul and Mark. Dredze, *You are what you tweet: analyzing twitter for public health*, Proceedings of the fifth international AAAI conference on weblogs and social media (2011), pp. 265-272. Google Scholar

[43] Johan Bollen, Huina Mao and Alberto Pepe, *Modeling public mood and emotion: twitter sentiment and socio-economic phenomena*,Proceedings of the fifth international AAAI conference on weblogs and social media (2011), pp. 450-453. Google Scholar

[44] Conover Michael D, Gonçalves Bruno, Ratkiewicz Jacob, Flammini Alessandro and Menczer Filippo, *Predicting the political alignment of twitter users*, In: 3rd IEEE international conference on privacy, security, risk and trust (Passat); 2011. p. 192–9. Google Scholar

[45] Saptarsi Goswami, Sanjay Chakraborty, Sanhita Ghosh, Amlan Chakrabarti and Basabi Chakraborty, 2018, *A review on application of data mining techniques to combat natural disasters*, Ain Shams Engineering Journal. Google Scholar

[46] Dilkhaz Yaseen Mohammed and Murat KARABATAK, 2018, *Terrorist Attacks In TURKEY*,6th International Symposium on Digital Forensic and Security (ISDFS),p.3.

[47] Dilrukshi, Inoshika, Kasun De Zoysa, and Amitha Caldera, 2013, *"Twitternews classification using SVM."*, In Computer Science and Education (ICCSE), 2013 8th International Conference on, pp. 287-291. Google Scholar

[48] Patil, Pritam H., Suvarna Thube, Bhakti Ratnaparkhi, and K. Rajeswari, 2014, *"Analysis of Different Data Mining Tools using Classification,Clustering and Association Rule Mining."*, International Journal of Computer Applications 93, no. 8.

[49] Nijhawan, Vani Kapoor, Mamta Madan, and Meenu Dave, 2017, *"The Analytical Comparison of ID3 and C4. 5 using WEKA."* International Journal of Computer Applications 167, no. 11 .

[50] Thankachan, Tulips Angel, and Kumudha Raimond, 2017, *"A Survey on Classification and Rule Extraction Techniques for Data mining."*, IOSR Journal of Computer Engineering 8, no. 5.

[51] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, October 2006, *"Introduction to Data Mining"*, International Edition,770 p

[52] *"Techopedia explains Web Mining"*, Web mining definition