International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

Web Content Mining for Terrorism Analysis

Sonali R. Khelukar , Snehal A. Mane , Bhakti A. Sude
Mr. Narendra Gawai
Computer Science and Technology
USHA MITTAL INSTITUTE OF TECHNOLOGY
S. N. D. T. WOMEN'S UNIVERSITY, MUMBAI 400 049

Submitted: 05-11-2021 Revised: 12-11-2021 Accepted: 15-11-2021

ABSTRACT
In present scenario use of internet is in boom. But every coin has two sides, likewise the use of internet is beneficial as well as harmful to human being. Recent days many terror attacks were there on net. Such terror related activities are hazardous for peoples, organization and countries. Terrorist are using internet to spread terror and form terrorist groups. By using internet they easily do the same. To exchange information Internet infrastructure is used by different Terrorist cells and they recruit new members and supporters. This project aims to find out if a website is promoting terrorism or has content related to terrorism using Web Content Mining.

The World Wide Web has lot of information and continues to increase in size and complexity. It is very herculean task to search relevant information from huge amount of data. The data used for web content mining includes both text and graphical data. Content mining is divided into two parts, one is webpage content mining and other is search result mining. In webpage content mining web is search via content. The search result content mining searches from the previous search result. When you search any specific key word or any web page, number of links or result is displayed. But all the data which is displayed on the web is not relevant. So efficiently and effectively retrieve required data on the Web is becoming a challenge.
Internet infrastructure is used by Terrorist cells and they recruit new mem- bers and supporters. For example, high-speed Internet connections were used in- tensively by members of the infamous Hamburg Cell that was largely responsible for the preparation of the September 11 attacks against the United States. This is one reason for the major effort made by law enforcement agencies around the world in collecting the information from the Web about terror-related activities.
It is believed that the detection of terrorists activities on the Web might prevent further terrorist attacks. One way to detect terrorist activity on the Web is to eavesdrop on all traffic of Web sites associated with terrorist groups, organi- zations in order to detect the accessing users based on their IP address. Unfortu- nately it is difficult to detect terrorist sites since they do not use fixed.

CHAPTER 1

Introduction

Day to day the use of internet is going on increasing drastically, relatively the huge number of techniques are there emerging everyday on various dynamic platforms. Use of internet has increased but along with these destructive minded peoples are using internet for making harm to the society and peoples. To exchange information Internet infrastructure is used by Terrorist cells and they recruit new members and supporters. For example, high-speed Internet connections were used intensively by members of the infamous Hamburg Cell that was largely responsible for the preparation of the September 11 attacks against the United States. This is one reason for the major effort made by law enforcement agencies around the world in collecting the information from the Web about terror-related activities.

It is believed that the detection of terrorists activities on the Web might prevent further terrorist attacks. One way to detect terrorist activity on the Web is to eavesdrop on all traffic of Web sites associated with terrorist groups, organi- zations in order to detect the accessing users based on their IP address. Unfortu- nately it is difficult to detect terrorist sites since they do not use fixed.

Many terrorists or terrorist groups are used

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

to make use of such techniques, applications, and internet to spread the terror. They used to attract the youths to be involved in such activities. It is necessary to detect such attempts in order to prevent such hazardous things. Terrorists groups are using social sites. It is the big challenge to detect such hazardous terrorist attacks. It is similar to eavesdrop. Terrorists are using different IP addresses changing them frequently such a that it will become more hard to identify them. To detect them is so difficult.

Purpose

The purpose of this document is to give a detailed description of the re- quirements for the Web Content Mining for Terrorism Analysis project. It will illustrate the purpose and complete declaration for the development of system. It will also explain system constraints, interface and interactions with other external applications. This document is primarily intended to be proposed to a customer for its approval and a reference for developing the first version of the system for the development team.

Document Conventions

The format of this SRS is simple. Bold face and indentation is used on general topics and or specific points of interest. The title of document is Font size
16. Headings and Sub-Headings are written with Font 14 and 12 respectively. The remainder of the document will be written using the standard font size 12, New Times Roman.

Scope of the Project

The Web Content Mining for Terrorism Analysis is a Windows-based ap- plication which helps security agencies to find terrorism related content on website. The application will be more of finding out whether it has terrorism related con- tent or not. This can be just simple information or it can also be for destructive information present on website.

The system will analyze the content present on the website. This content will be refined and sanitized by removing unwanted stopping words, html tags. The final content will then be analyzed and reviewed for terrorism related words.

Definitions, Acronyms and Abbreviations

Term	Definition
HTML	Hyper Text Markup Language
SRS	System Requirement Specification
IP	Internet Protocol

Figure 1.1: Definitions, Acronyms and Abbreviations

Operating Environment

The system will be developed for Windows operating system. The mini- mum hardware and software requirements for the system are as follows:

Hardware

* Pentium Dual Core or above Processor
* 4 GB RAM
* 500GB HDD

Software

* Windows 7 / Windows 8 - 64 bits
* Visual Studio 2013
* .Net Framework 4.5 or above to the website.

CHAPTER 2

Overview of Stage I

Introduction

Problem Statement

The solution proposed is to perform web content mining to extract relevant information for wide range of web pages. The application will be more of finding out whether it has terrorism related content or not. This can be just simple information or it can also be for destructive information present on website. The system will analyze the content present on the website. This content will be refined and sanitized by removing unwanted stopping words, html tags. The final content will then be analyzed and reviewed for terrorism related words.

The scope of the project will be limited to tell users

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

whether a particular website has terrorism related content or not. It will not tell whether it promotes terrorism or not.

The system needs to be developed to efficiently identify Web sites hav- ing content related to terrorism. This will help government or security agencies to know which website needs to be blocked and how to prevent users from vis- iting these websites. The main users who will be using this system will be the Government and security agencies. The input to the system will be the web URL.

The entire system will be divided into four modules

1. Extracting website source content which will be in HTML format.
2. Cleansing the content to remove HTML tags and other stopping / un- wanted words.
3. Matching the word from the set of words stored in Database.
4. Calculating score to determine final output. The system flow is given below:

1. Get web page data from the website link
2. Clean the data by removing the HTML attributes and other formatting tags
3. Extract require content
4. Convert sentences into Words
5. Remove stop-words
6. With the remaining words perform matching test with keywords stored in database
7. Give the higher weight to the matched word
8. Compute the frequency of matched word
9. If frequency of matched words exceed the threshold value then the web page is related to terrorism

The output will be the whether the website has content related to terrorism or not. This application is used by Government. It is also used by Security Agencies and Cyber Security. Web Content Mining provides a path to screen more specific data. Techniques used are:

Classification :-
Given a collection of records (training set) Each record contains a set of attributes, one of the attributes is the class. Find a model for class attribute as a function of the values of other attributes.
Goal: previously unseen records should be assigned a class as accurately as possible. A test set is used to determine the accuracy of the model.
Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Clustering:-
Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters.
Help users understand the natural grouping or structure in a data set.
Clustering: unsupervised classi cation: no prede ned classes.
Used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms. Moreover, data compression, outliers detection, understands human concept formation.

Association rules:-
Given a set of transactions,nd rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

Many terrorists or terrorist groups are used to make use of such techniques, applications, and internet to spread the terror. They used to attract the youths to be involved in such activities. It is necessary to detect such attempts in order to prevent such hazardous things. Terrorists groups are using social sites. It is the big challenge to detect such hazardous terrorist attacks. It is similar to eavesdrop.

Terrorists are using different IP addresses changing them frequently such that it will become harder to identify them. To detect them is so difficult.

The World Wide Web has lot of information and continues to increase in size and complexity. It is very herculean task to search relevant information from huge amount of data. The data used for web content mining includes both text and graphical data. Content mining is divided into two parts, one is webpage content mining and other is search result mining. In webpage content mining web is search via content. The search result content mining searches from the previous search result. When you search any specific key word or any web page, number of links or result is displayed. But all the data which is displayed on the web is not relevant. So efficiently and effectively retrieve required data on the Web is becoming a challenge. The user issues the query terms (keywords) to a search engine and the search engine returns a set of pages that may be related to the query topics or terms. For a page, if the user wants to search the relevant pages further, he/she would prefer those relevant pages to be at hand. Here, a relevant Web page is the one that addresses the same topic as the original page, but is not necessarily semantically identical. On web data is updated at every second

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

so it is not necessary that a data or the web page that is retrieved by the user will be retrieved another time in the same structure or order.

Implementation Details
Algorithm

1) Web Page Content Extraction

This step extracts the contents of the web page on the basis of URL given by user. This step extracts the raw data of the page and gives it to data cleansing and standardizing module.

The web page contains so many contents in it and it is designed byusing a HTML language. To
nd out the terrorist related contents from that HTML pages is not so easy and it is not understandable to the common user. So, we are using content extraction by using which we can get only the original contents which are understandable to user.

2) Data Cleansing and Standardization

paragraph This step cleans the raw data that is extracted from a web page. It removes all the HTML tags, formatting tags and stopping words. To analyze such tremendous contents becomes difficult and the complexity of getting the required output becomes more.
So we need to clean all this HTML tags, formatting tags and stopping words to and make it more reliable for terrorist related content detection.

3) Keyword Matching

This step matches the word from the cleaned and standardized content with the inbuilt words stored in dictionary.

4) Score Calculation

This step finally calculates the density of the terrorism related words from the extracted content. On the basis of score, the application determines whether the webpage is related to terrorism or not.

Literature Survey

Web based Text data is the most common content type on the net when it comes to author's opinion. Recently, following the progress of wireless internet and smart phone devices, iPhones the amount of data on the web is dramatically increasing with no constrain to time or location. In this paper we suggested the method for extracting the words from document names as WordNet Hierarchy.

This method was tested with the sampled New York Times articles by querying four distinct words from four different areas. Experimental results show our proposed method effectively extracts context words from the text and identi- fies terrorism-related documents. Text analysis is used to discover unknown, valid patterns and relationships in large data sets. Even text analysis has a great po- tential to identifying unknown text documents, there is a limitation that human written language is still complicated for machine to understand semantic meanings of it.

The learning TypicalTerrorist-Behaviour part of the methodology defines and represents the typical behaviour of terrorist users based on the content of their Web activities. It is assumed that it is possible to collect Web pages from terror-related sites. The content of the collected pages is the input to the Vector Generator module that converts the pages into vectors of weighted terms (each page is converted to one vector).

One major issue of today is the representation of textual content of Web pages. More specifically, there is a need to represent the content of terror-related pages as against the content of a currently accessed page in order to efficiently compute the similarity between them. This study will use the vector-space model commonly used in Information Retrieval applications for representing terrorists interests and each accessed Web page.

Web mining is a rapidly growing research area. It consists of Web usage mining, Web structure mining, and Web content mining. Web usage mining refers to the discovery of user access patterns fromWeb usage logs. Web structure mining tries to discover useful knowledge from the structure of hyperlinks.

A. Web Content Mining

Web content mining is related but different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in Web content mining. It is related to text mining because much of the web contents are texts. However, it is also quite different from data mining because Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data. Web content mining is also different from text mining because of the semi-structure nature of the Web, while text mining focuses on unstructured texts. Web content mining thus requires creative applications of data mining and/or text mining techniques and also its own unique approaches. In the past few years, there was a rapid expansion of activities in the Web content mining area.

B. Web Structure Mining

World Wide Web can reveal more

information than just the information contained in documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. This can be compared to bibliographical citations. When a paper is cited often, it ought to be important. By means of counters, higher levels cumulate the number of arte- facts subsumed by the concepts they hold. Counters of hyperlinks, in and out documents, retrace the structure of the web artefacts summarized.

The main purpose for structure mining is to extract previously unknown relationships between Web pages. This structure data mining provides use for a business to link the information of its own Web site to enable navigation and cluster information into site maps. This allows its users the ability to access the desired information through keyword association and content mining. Hyperlink hierarchy is also determined to path the related information within the sites to the relationship of competitor links and connection through search engines and third party co-links.

C. Web Usage Mining System Structure

Web Usage Mining-Web Usage Mining is a part of Web Mining, which, in turn, is a part of Data Mining. As Data Mining involves the concept of extraction meaningful and valuable information from large volume of data, Web Usage mining involves mining the usage characteristics of the users of Web Applications. It i used for deciding business strategies through the efficient use of Web Applications. Web usage mining is the third category in web mining. This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. This web mining also enables Web based businesses to provide the best access routes to services or other advertisements. When a company advertises for services provided by other companies, the usage mining data allows for the most effective access paths to these portals. In addition, there are typically three main uses for mining in this fashion.

The first is usage processing, used to complete pattern discovery. This first use is also the most difficult because only bits of information like IP addresses, user information, and site clicks are available. With this minimal amount of information available, it is harder to track the user through a site, being that it does not follow the user throughout the pages of the site.

The second use is content processing, consisting of the conversion of Web information like text, images, scripts and others into useful forms. This helps with the clustering and categorization of Web page information based on the titles, specific content and images available.

Finally, the third use is structure processing. This consists of analysis of the structure of each page contained in a Web site. This structure process can prove to be difficult if resulting in a new structure having to be performed for each page.

Analysis of this usage data will provide the companies with the informa- tion needed to provide an effective presence to their customers. This collection of information may include user registration, access logs and information leading to better Web site structure, proving to be most valuable to company online market- ing. These present some of the benefits for external marketing of the companies products, services and overall management.

Internally, usage mining effectively provides information to improvement of communication through intranet communications. Developing strategies through this type of mining will allow for intranet based company databases to be more effective through the provision of easier access paths. The projection of these paths helps to log the user registration information giving commonly used paths the forefront to its access.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

System Features
Block Diagram/Architecture Diagram

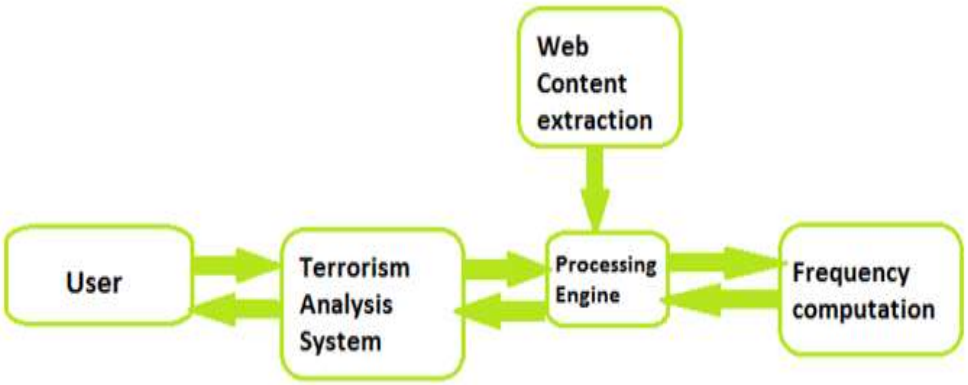


Figure 2.1: Block Diagram

Functional Description
Object Model

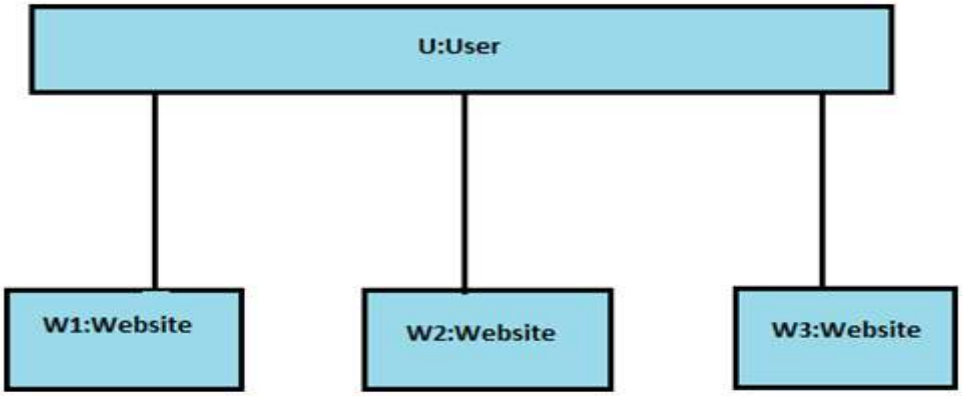


Figure 2.2: Obeject Diagram

Dynamic Model

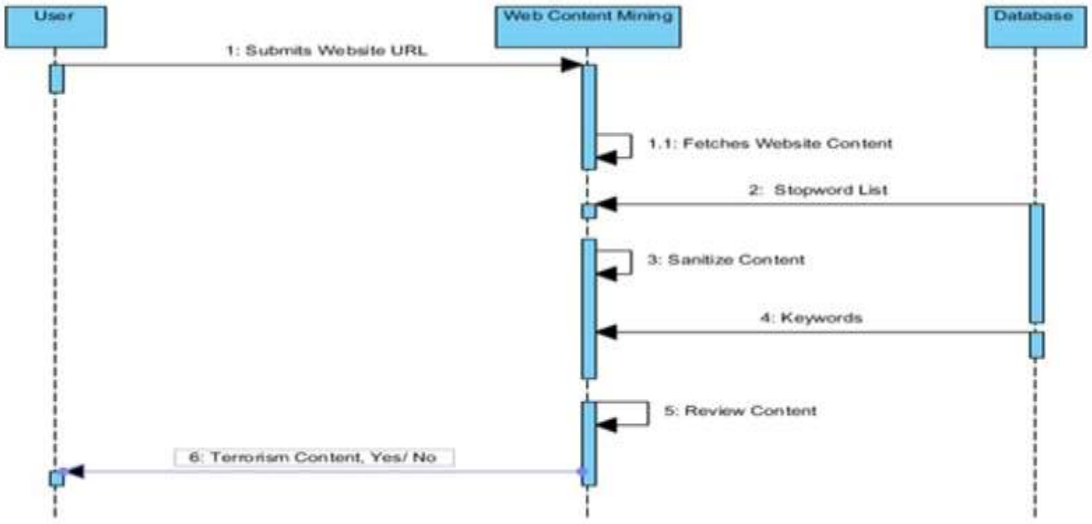


Figure 2.3: Dynamic Model

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

Data Flow Diagram with Data Dictionary
Level 0

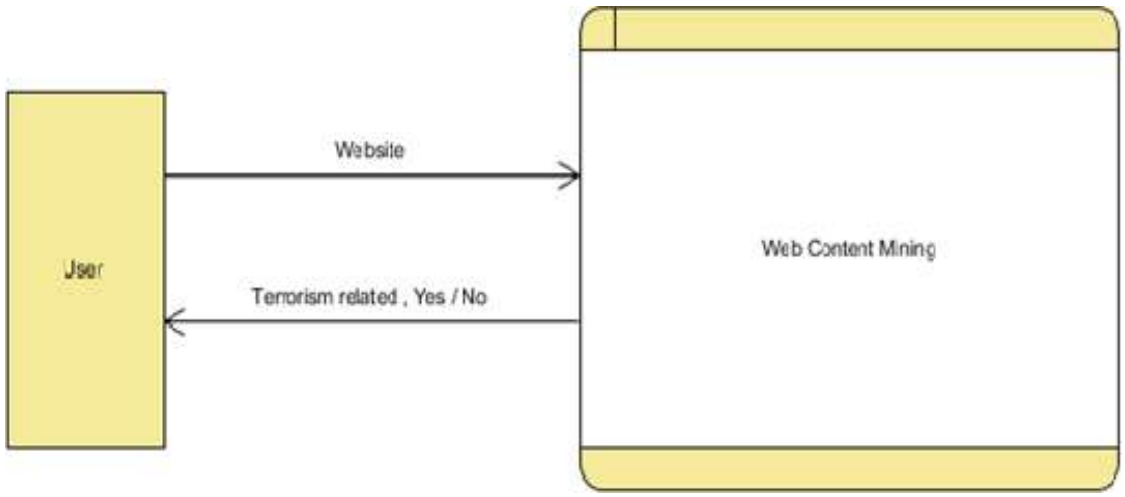


Figure 2.4: Data Flow: Level 0

Level 1

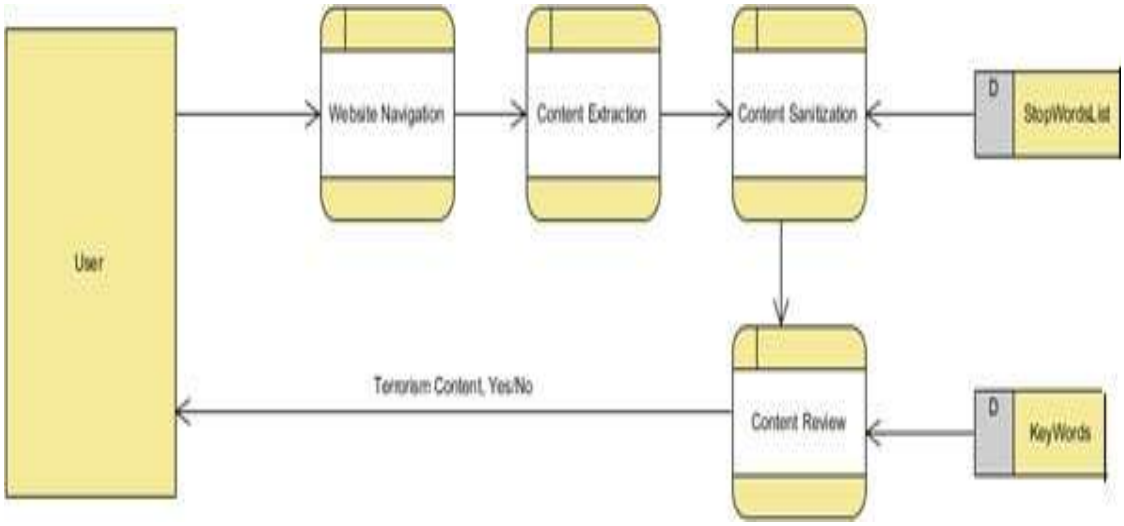


Figure 2.5: Data Flow: Level 1

CHAPTER 3

Overview of Stage II
System Architecture Description
Overview of Modules / Subsystem
Algorithm used Web content mining for terrorism analysis

The word extraction algorithm is given below.
Input: Web URLs i.e. log file
Output: Website is either terrorism related or not.
Begin
Step 1: Get web page data from the website link.
Step 2: Clean the data by removing the HTML attributes and other for- matting tags.
Step 3. Extract require content.
Step 4. Convert sentences into Words. Step 5. Remove stop-words.
Step 6. With the remaining words perform matching test with keywords stored in database.
Step 7. Give the higher weight to the matched word. Step 8. Compute the frequency of matched word.
Step 9. If frequency of matched words exceed the threshold value then the web page is related to terrorism.
Exit.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

Use Case Diagram

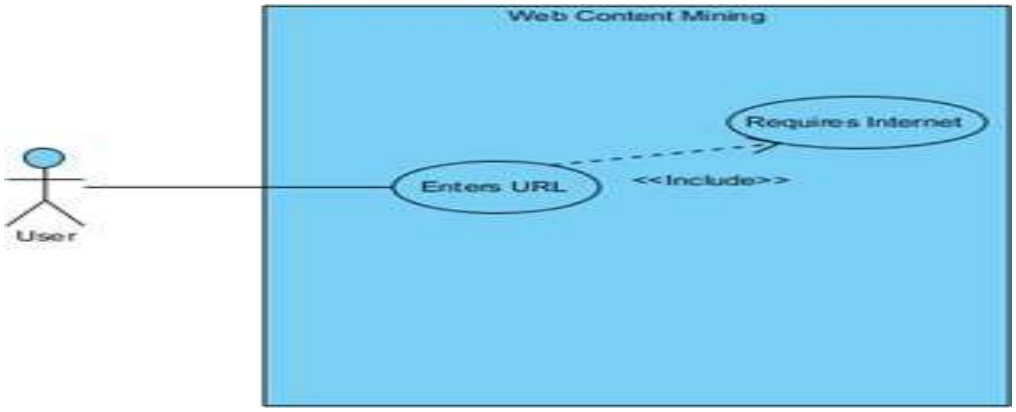


Figure 3.1: Use Case Diagram

Sequence Diagram

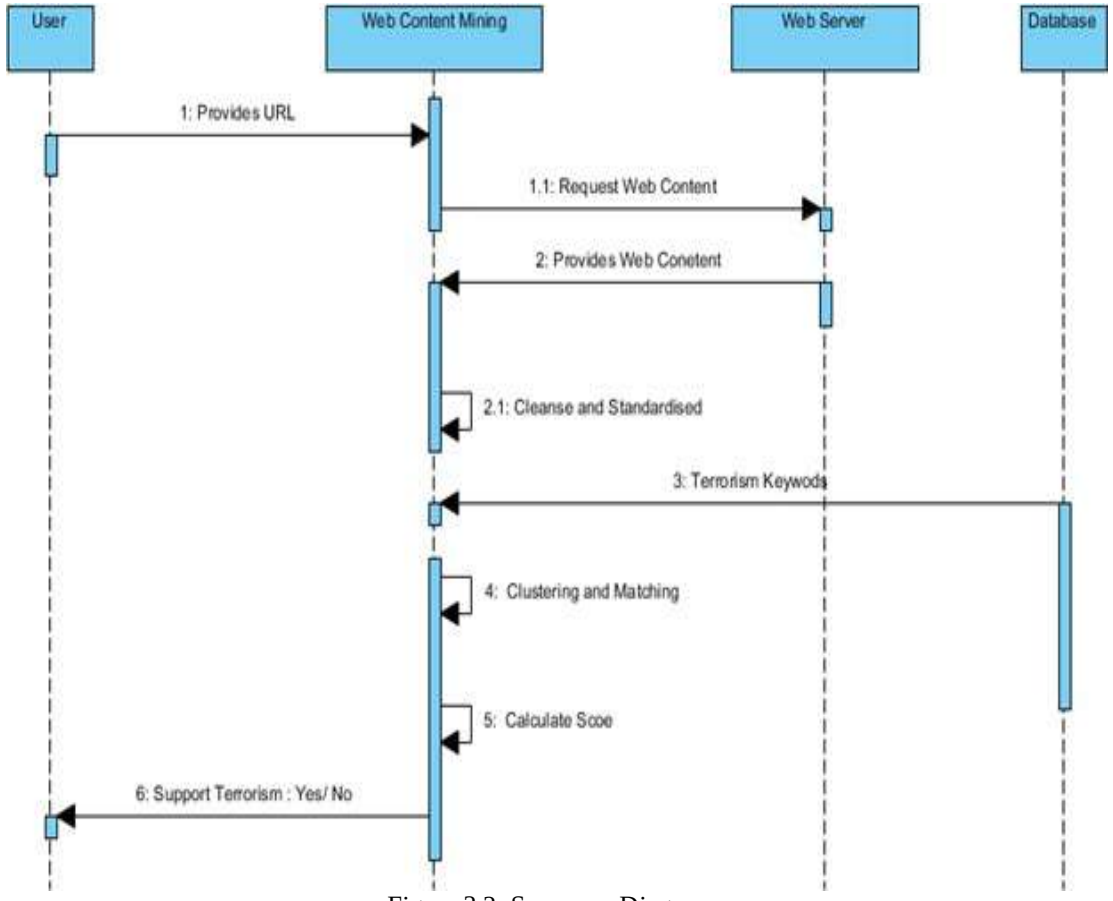


Figure 3.2: Sequence Diagram

Structure and Relationships

Introduction to .NET Framework

The Microsoft .NET Framework is a software framework available with several MicrosoftWindows operating systems. It includes a large library of coded solutions to prevent common programming problems and a virtual machine that manages the execution of programs written specifically for the framework. The .NET Framework is a key Microsoft offering and is

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

intended to be used by most new applications created for the Windows platform.

The framework's Base Class Library provides a large range of features including user interface, data and data access, database connectivity, cryptography, web application development, numeric algorithms, and network communications. The class library is used by programmers, who combine it with their own code to produce applications.

Programs written for the .NET Framework execute in a software environ- ment that manages the program's runtime requirements. Also part of the .NET Framework, this runtime environment is known as the Common Language Run- time (CLR). The CLR provides the appearance of an application virtual machine so that programmers need not consider the capabilities of the specific CPU that will execute the program. The CLR also provides other important services such as security, memory management, and exception handling. The class library and the CLR together compose the .NET Framework.

Principal design features:

Interoperability

Because interaction between new and older applications is commonly re- quired, the .NET Framework provides means to access functionality that is im- plemented in programs that execute outside the .NET environment. Access to COM components is provided in the System.Runtime.InteropServices and Sys- tem.EnterpriseServices namespaces of the framework; access to other functionality is provided using the P/Invoke feature.

Common Runtime Engine

The Common Language Runtime (CLR) is the virtual machine component of the .NET framework. All .NET programs execute under the supervision of the CLR, guaranteeing certain properties and behaviors in the areas of memory management, security, and exception handling.

Language Independence

The .NET Framework introduces a Common Type System, or CTS. The CTS specification defines all possible datatypes and programming constructs sup- ported by the CLR and how they may or may not interact with each other. Because of this feature, the .NET Framework supports the exchange of instances of types between programs written in any of the .NET languages.

Base Class Library

The Base Class Library (BCL), part of the Framework Class Library (FCL), is a library of functionality available to all languages using the .NET Framework. The BCL provides classes which encapsulate a number of common functions, including file reading and writing, graphic rendering, database interac- tion and XML document manipulation.

Simplified Deployment

The .NET framework includes design features and tools that help manage the installation of computer software to ensure that it does not interfere with previously installed software, and that it conforms to securityrequirements.

Security

The design is meant to address some of the vulnerabilities, such as buffer overflows, that have been exploited by malicious software. Additionally, .NET provides a common security model for all applications.

Portability

The design of the .NET Framework allows it to theoretically be platform agnostic, and thus cross-platform compatible. That is, a program written to use the framework should run without change on any type of system for which the framework is implemented. Microsoft's commercial implementations of the frame- work cover Windows, Windows CE, and the Xbox 360.In addition, Microsoft sub- mits the specifications for the Common Language Infrastructure (which includes the core class libraries, Common Type System, and the Common Intermediate Language), the C sharp language, and the C++/CLI language to both ECMA and the ISO, making them available as open standards. This makes it possible for third parties to create compatible implementations of the framework and its languages on other platforms.

Visual overview of the Common Language Infrastructure (CLI) Common Language Infrastructure (CLI)

The core aspects of the .NET Framework lie within the Common Language Infrastructure, or CLI. The purpose of the CLI is to provide a language-neutral platform for application development and execution, including functions for ex- ception handling, garbage collection, security, and interoperability. Microsoft's implementation of the CLI is called the Common Language Runtime

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

orCLR.

Assemblies

The intermediate CIL code is housed in .NET assemblies. As mandated by specification, assemblies are stored in the Portable Executable (PE) format, common on the Windows platform for all DLL and EXE files. The assembly consists of one or more files, one of which must contain the manifest, which has the metadata for the assembly. The complete name of an assembly (not to be confused with the filename on disk) contains its simple text name, version number, culture, and public key token. The public key token is a unique hash generated when the assembly is compiled, thus two assemblies with the same public key token are guaranteed to be identical from the point of view of the framework. A private key can also be specified known only to the creator of the assembly and can be used for strong naming and to guarantee that the assembly is from the same author when a new version of the assembly is compiled (required to add an assembly to the Global Assembly Cache).

Metadata

All CLI is self-describing through .NET metadata. The CLR checks the metadata to ensure that the correct method is called. Metadata is usually gener- ated by language compilers but developers can create their own metadata through custom attributes. Metadata contains information about the assembly, and is also used to implement the reflective programming capabilities of .NET Framework.

Security

.NET has its own security mechanism with two general features: Code Access Security (CAS), and validation and verification. Code Access Security is based on evidence that is associated with a specific assembly. Typically the evidence is the source of the assembly (whether it is installed on the local machine or has been downloaded from the intranet or Internet). Code Access Security uses evidence to determine the permissions granted to the code. Other code can demand that calling code is granted a specified permission. The demand causes the CLR to perform a call stack walk: every assembly of each method in the call stack is checked for the required permission; if any assembly is not granted the permission a security exception is thrown.

When an assembly is loaded the CLR performs various tests. Two such tests are validation and verification. During validation the CLR checks that the assembly contains valid metadata and CIL, and whether the internal tables are correct. Verification is not so exact. The verification mechanism checks to see if the code does anything that is 'unsafe'. The algorithm used is quite conservative; hence occasionally code that is 'safe' does not pass. Unsafe code will only be executed if the assembly has the 'skip verification' permission, which generally means code that is installed on the local machine.

CHAPTER 4

Implementation Details

Pseudocode for Components

1. Web Content Extraction

a. Provide URL.
b. Initialize Components.
c. Initialize VariableScore.
d. Get website contents using.

WebBrowser.DocumentCompleted= WebBrowserDocumentCompleted;

f. If website is loaded successfully. Process extraction method . webBrowser.DocumentText.ToString();
g. If variable score is greater than 5.

MessageBox.show(web page contains terrorist related information.):

h. Else

MessageBox.show(web page does NOT contain terrorist related informa-tion.);

2. Cleaning HTML Tags

a. Replace HTML tags /r , /t , nbsp with blank space.

3. Extract Required Content

a. While check variable index count is less than arrayContent.length.
b. If variable line contains ¡p , ¡h , and ¡a.
c. Calculate variable startIndex and char/count using ¿ , ¡/ .
d. If variable startIndex and charCount is greater than zero.
e. Find Substring between startIndex to endIndex. endPosition startPosi- tion 1 used to remove tags like ¡ and ¿.

4. Convert Sentences into Words

a. If variable webData is not equal to null

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

Insert range between 0 to variableData and split with blank space.And initialize function RemoveStoppingWords.

b. If stopWord count is equal to zero . Execute SQLquery Select stopwords from StopWordList ;

c. For each string in variable Word in array WebContent. If array stop- WordContains variable word Remove variableWord and clean content.

5. Perform Matching Test with Keywords

a. For each string variableWord in array cleaned content. If variableWord contains arrayKeyword Count the word .

Calculate frequency using .

variableScore = Math. ceiling (((double) varCount / (double) arrCleaned- Content.count * 100)) ;

b. If the frequency matches or exceeds threshold. MessageBox.show(web page contains terrorist related information.);

c. Else MessageBox.show(web page does not contain terrorist related information.).

Source Code Details

```
using System;
using System.Collections.Generic; using System.Linq;
using System.Threading.Tasks; using System.Windows.Forms;
namespace WebContentMiningTerrorismAnalysis
static class Program
/// ¡summary¿
/// The main entry point for the application.
/// ¡/summary¿
STAThread
static void Main()
Application.EnableVisualStyles();
Application.SetCompatibleTextRenderingDefault(false); Application.Run(new FrmMain());
end function

Function funcExtractContent( input varText)
varIndexCount = 0;
arrContent = varText.Replace('r', 'ˆ', ”nbsp;”,””).Split('n');
while varIndexCount ¡ arrContent.Length
do
varTemp = ””;
varLine = arrContent[varIndexCount];
if varLine.Contains(”¡p”) —— varLine.Contains(”¡h”) —— varLine.Contains(”¡a”) then
RecursiveLabel:
varStartIndex = varLine.Contains('¿') then varLine.IndexOf('¿') else -1; varCharCount = varLine.Contains(”¡/”) then varLine.IndexOf(”¡/”) else 0;
if varStartIndex ¿ 0 varCharCount ¿ 0 then
varTemp = varLine.Substring(varStartIndex + 1, varCharCount - varStartIn- dex - 1);
if varTemp.Contains('¿') then
varLine = varTemp.Substring(varTemp.LastIndexOf('¿') + 1);
goto RecursiveLabel;
else if varTemp.Trim() != ”” then
varWebData = varWebData +” ”+ varTemp.Trim();
arlCleanedContent.Add(varTemp.Trim());
Console.WriteLine(varTemp);
end if end if
if varTemp.Contains(”¡”) then varLine = varTemp;
goto RecursiveLabel; end if
else if varStartIndex ¿ 0
varLine = varTemp.Substring(varTemp.LastIndexOf('¿')+1);
goto RecursiveLabel;
else if varLine.Trim() != ””
varWebData = varWebData + ” ” + varLine.Trim();
arlCleanedContent.Add(varLine.Trim());
Console.WriteLine(varLine);
end if end if
varIndexCount++; end while
if varWebData != null then
arlCleanedContent.Clear();
arlCleanedContent.InsertRange(0, varWebData.Split(' '));
call funcRemoveStoppingWord(varWebData.Split(' '));
end if
function end
function funcRemoveStoppingWord(input array arrWebContent ) if arlStopWords.Count ==0 then

excute query ”select StopWords from StopWordsList”; while results end of file
do
arlStopWords.Add(varSqlReader[0].ToString().ToLower().Trim()); end while
end if
for each varWord in arrWebContent
do
if
arlStopWords.Contains(varWord.ToLower().Trim()) then arlCleanedContent.Remove(varWord);
end if end for
for i = 0; i ¡ arlCleanedContent.Count; i++ then
```

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

```
do
arlCleanedContent[i] = arlCleanedContent[i].ToString().Replace(']', '[', '','','!','?','(',')','.',',',' ').Trim().ToLower();
end for
call funcMatchKeywords(); end function
function funcMatchKeywords()
if arlKeyWords.Count == 0 then
execute query "select Word from KeyWords"; while results end of file
do
arlKeyWords.Add(varSqlReader[0].ToString().ToLowe
r().Trim()); end while
end if varCount = 0;
for each varWord in arlCleanedContent
do
if
arlKeyWords.Contains(varWord.ToLower().Trim())
then arlCandidateWords.Add(varWord);
varCount++; end if
end for
varScore = varCount / arlCleanedContent.Count * 100; end function
```

Testing Details

Test Plan

Test Case Id	Test Case	Expected Result	Actual Result
1	Main form: User provides URL	Application should display corresponding Webpage under "Site View" and corresponding page code under "Site Code" tab	Application should displayed corresponding Webpage under "Site View" and corresponding page code under "Site Code" tab
2	Web Page Code cleaning	The web page code should be free from HTML tags	The web page code was free from HTML tags
3	Web Page Code Standardization	The web page code should be free from stopping and unwanted words	The web page code was free from stopping and unwanted words
4	Non Terrorism URL	The application should respond with "Website does not contain terrorist related information"	The application responded with "Website does not contain terrorist related information"
5	Terrorism URL	The application should respond with "Website contains terrorist related information"	The application responded with "Website contains terrorist related information"

Figure 4.1: Test Plan Representation

Kinds of Testing

Black box Testing

Black-box testing is a method of software testing that examines the func- tionality of an application without peering into its internal structures or workings. Specific knowledge of the application's code/internal structure and programming knowledge in general is not required. The tester is aware of what the software is supposed to do but is not aware of how it does it. For instance, the tester is aware that a particular input returns a certain, invariable output but is not aware of how the

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

software produces the output in the first place.

Integration Testing

Integration testing is the phase in software testing in which individual software modules are combined and tested as a group. It occurs after unit testing and before validation testing. Integration testing takes as its input modules that have been unit tested, groups them in larger aggregates, applies tests defined in an integration test plan to those aggregates, and delivers as its output the integrated system ready for system testing.

Acceptance Testing

Acceptance testing is a test conducted to determine if the requirements of a specification or contract are met.

CHAPTER 5

Software Maintenance

Software maintenance is widely accepted part of SDLC now a days. It stands for all the modifications and updations done after the delivery of software product. There are number of reasons, why modifications are required, some of them are briefly mentioned below:

Market Conditions

Policies, which changes over the time, such as taxation and newly intro- duced constraints like, how to maintain bookkeeping, may trigger need for modi- fication.

Client Requirements

Over the time, customer may ask for new features or functions in the software.

Host Modifications

If any of the hardware and/or platform (such as operating system) of the target host changes, software changes are needed to keep adaptability.

Organization Changes

If there is any business level change at client end, such as reduction of organization strength, acquiring another company, organization venturing into new business, need to modify in the original software may arise.

In a software lifetime, type of maintenance may vary based on its nature. It may be just a routine maintenance tasks as some bug discovered by some user or it may be a large event in itself based on maintenance size or nature. Following are some types of maintenance based on their characteristics:

Corrective Maintenance

This includes modifications and updations done in order to correct or fix problems, which are either discovered by user or concluded by user error reports.

Adaptive Maintenance

This includes modifications and updations applied to keep the software product up-to date and tuned to the ever changing world of technology and business environment.

Perfective Maintenance

This includes modifications and updates done in order to keep the software usable over long period of time. It includes new features, new user requirements for refining the software and improve its reliability and performance.

Preventive Maintenance

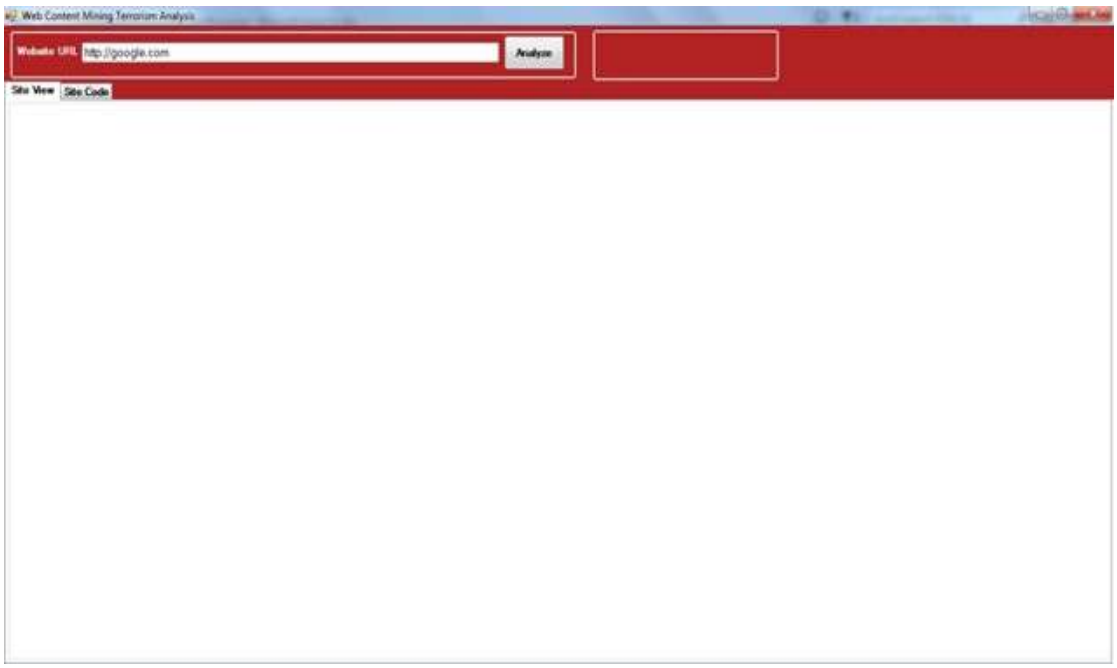
This includes modifications and updations to prevent future problems of the software. It aims to attend problems, which are not significant at this moment but may cause serious issues in future.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

CHAPTER 6

User Interface

Figure 6.1: Web Content Mining Main Form



Main Form
Terrorism Related URL

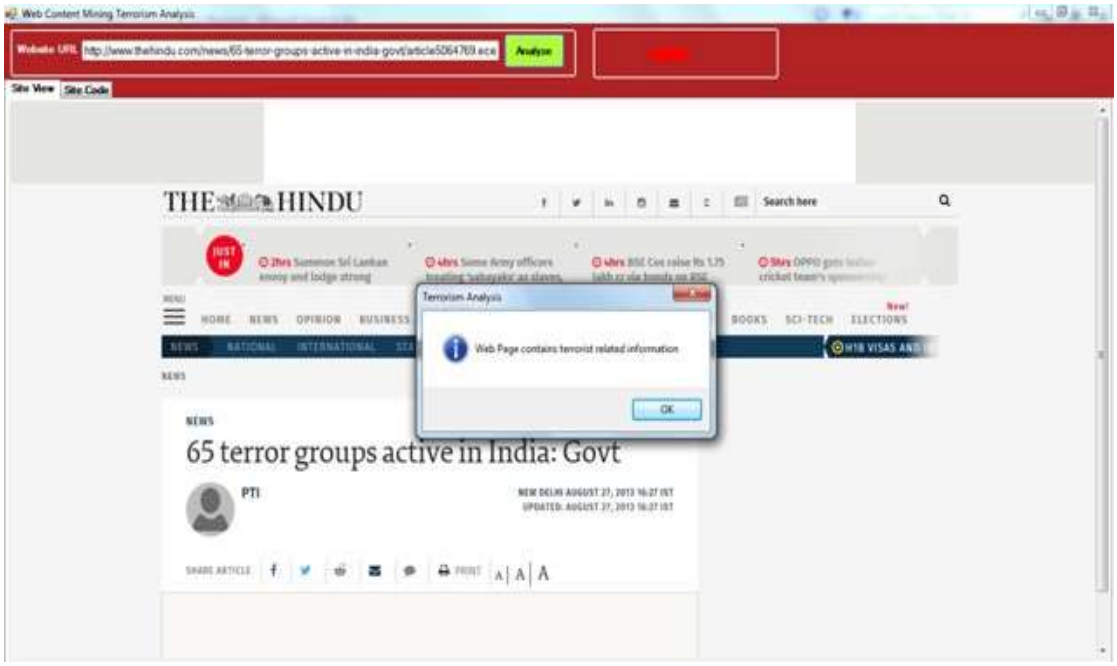


Figure 6.2: Terrorism Related URL

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 11 Nov 2021, pp: 320-335 www.ijaem.net ISSN: 2395-5252

Non-Terrorism Related URL

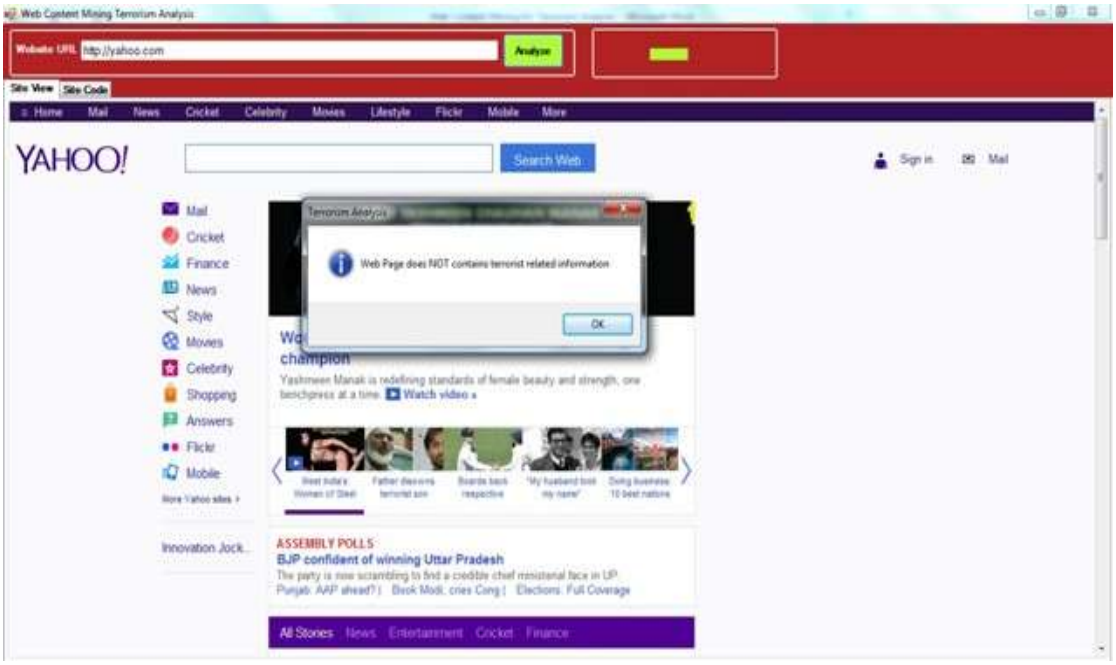


Figure 6.3: Non-Terrorism Related URL

CHAPTER 7

Conclusion

In this project, a concept of checking a web page related to terrorism is implemented. The application will just tell whether the given URL /Web page has terrorism related content or not. The content can be anything i.e. it can be informative as well as sensitive for spreading terrorism.

As per the goal of this project an innovative, knowledge-based methodology for terrorist activity detection on the Web is presented. The results of an initial case study suggest that the methodology can be useful for detecting terrorists and their supporters using a legitimate ways of Internet access to view terror related content at a series of evasive web sites. Present system just recognize the the words of terrorists language. By developing such system, relationship between human and computer becomes much closer and secure. Thus it helps in overcoming the problem of Terrorism on net.

The Web Content Mining System is developed to find whether a web page has terrorism related content or not. The focus is given on the content rather than web logs and structure. The content helps in better accuracy as compared to web logs and structure. The page content is checked for probable keywords which directly or indirectly relates to terrorism.

The application was tested with different sets of URL. The application detected the URLs correctly which were related to terrorism and which not. The response time of the system was also good. On an average it toll less than a minute to analyze and detect the URL.

The proposed approach is efficient one to detect terror related activities.

FUTURE WORK

The current implementation of the system just checks whether the page has terrorism related content or not. Future enhancement to this project would be to identify with what respect is the content i.e. informative or spreading terrorism.

There can be module where terrorist communicating in codewords can be tracked by the system itself.
In proposed system tracks susceptible IP address and provides the in- formation to the offices using the system. In later, system will track particular susceptible person who is sending massages relevant to terrorism.

REFERENCES

[1]. Theint Theint Aye ,Web Log Cleaning for Mining of Web Usage Patterns 2011 IEEE.
[2]. Shaily Langhnoja,Mehul Barot, Darshak Mehta, Pre-Processing: Procedure on Web Log File for Web Usage Mining International Journal of Emerging Technology and Advanced Engineering December 2012.
[3]. L.K.JoshilaGrace,V.Maheswari,Dhinaharan Nagamalai, Analysis of Web Logs and Web User In Web Mining International Journal of Network Se- curity and Its Applications (IJNSA),Vol.3,No.1,January 2011.
[4]. J.Vellingiri,S. Chenthur Pandian,A Survey on Web Usage Mining Volume 11 Issue 4 Version 1.0 xMarch 2011.
[5]. Abbasi, A., and Chen, H. (2005). Applying authorship analysis to extrem- istgroup Web forum messages. IEEE Intelligent Systems,
[6]. Special Issue on Artificial Intelligence for National and Homeland Security, 20(5),[6775].
[7]. Dongjin Choi , Byeongkyu Ko, Heesun Kim, Pankoo Kim, Text analysis for detecting terrorism-related articles on the web, Journal of Network and Computer Applications 38 (2014) 1621.
[8]. Mohammad Javad Hosseinpour,Mohammad Nabi Omidvar, Detecting Terror- Related Activities on the Web with Using Data Mining Techniques, 2009 Second International Conference on Computer and Electrical Engineering.
[9]. Ramesh Yevale, Mayuri Dhage, Tejali Nalawade,.Trupti Kaule, Unautho- rized Terror Attack Tracking Using Web Usage Mining, (IJCSIT) Inter- national Journal of Computer Science and Information Technologies,ISSN: 0975-9646, Vol. 5 (2) , 2014,1210-1212.
[10]. Y.Elovici, A.Kandel, M.Last, B.Shapira, O. Zaafrany, Using Data Mining Techniques for Detecting Terror-Related Activities on the Web University of South Florida,4202 E. Fowler Ave. ENB 118 Tampa, FL, 33620, USA.
[11]. Nisha Chaurasia1, Mradul Dhakar1, Akhilesh Tiwari2 and R. K. Gupta2, A Survey on Terrorist Network Mining: Current Trends and Opportunities, International Journal of Computer Science Engineering Survey (IJCSES) Vol.3, No.4, August 2012.