Security Situation in Republic of Macedonia Using Semantic Algorithms for Open Data

1st Zivka Jovevska
FON University
Republic of Macedonia
zivkajaneva@gmail.com

2nd Daniel Jovevski
Ministry of information society and Administration
Republic of Macedonia
daniejovevski92@gmail.com

3st Leonid Djinevski
FON University
Republic of Macedonia
l.djinevski@fon.edu.mk

***Abstract*—Semantic algorithms are a web of information related to such a way that can easily be processed from the machines globally. It is an efficient display of information on the World Wide Web network or can be presented as a globally connected database. This paper aims to investigate the security situation in the Republic of Macedonia using Web Crawler and Semantic Algorithms for its research. The data was taken from the news page of the Ministry of Interior website. Some results-graphs of the research we conducted on the data obtained from the Ministry of Interior will be shown. The expected results of the paper are to get a clear picture of life in Macedonia, in individual cities and whether the security is improving or deteriorating.**

***Keywords:* Internet, Ministry of Interior, Security, Web Crawler**

I. Introduction

Many of the technologies and ideas developed during the creation of the semantic network are refined and live in various applications. What was not possible in 2001 is now possible: You can easily create applications that use data across the network. The difference is that you had to sign in for each API separately, which, beside the tedious manual integration, gives the one who hosts the API great control over how to access their data.

Because we hear and read about crimes daily over the news and on internet, we decided to make a research and find out what was the crime situation in Republic of Macedonia in 2019.

Using web crawling we will get data and based on the results we will get a clear picture about the security situation. Our motive for this research is to find out whether the security situation in our country is improving, or it is getting worse.

The research will be based on three parameters: Theft crime cases, Homicide cases and Drug trafficking cases.

The results will be shown in the graphs on a monthly basis. In the end, we will make a conclusion based on the results.

II. Semantic Web

In 2001, Tim Berners-Lee, inventor of the World Wide Web, published an article in Scientific American. Burners-Lee, along with two other researchers, Ora Lasila and James Chandler, who want to give the world an overview of the revolutionary new changes they've seen coming. Since its introduction just a decade ago, the network has become the world's fastest document sharing tool. Now, the authors promise that the network will evolve to include not only documents but any kind of data that could be imagined. They called this new website the Semantic Web.

The effort to build the Semantic Network consists of four phases. The first phase, which is from 2001 to 2005, is the golden age of semantic network activities. Between 2001 and 2005, the W3C issued new standards that represent the underlying technologies of the semantic future.

The most important of these is the Resource Description Framework (RDF).

Berner-Lee's article begins the second phase of the development of the Semantic Network, where the focus has shifted from standard settings and example building to the creation and popularization of large RDF databases. Perhaps the most successful of these databases was DBpedia, a huge repository of RDF triples extracted from Wikipedia articles.

The third phase of Semantic Network development involves adapting W3C[1] standards to suit real web developer practices and preferences. In 2008, JSON begins its rapid rise in popularity.

W3C is still working on the Semantic Network under the title "Data Activity", which can be called the fourth phase of the Semantic Network project. But that says the latest project "data activity" is a study of what the W3C must do to improve the standardization process. Even the W3C now seems to admit that few of its semantic web standards are widely accepted and that simpler standards would be more successful.

Semantic web technologies can be viewed as layers, each layer relying on and depending on the functionality of the layers beneath it. Although the semantic network is often presented as a separate entity, it is an extension and improvement of an existing website, not a replacement.

Semantic algorithms are a network of related information in such a way that it can be easily processed by machines globally. It is an effective representation of information on the WWW[2] network or can be presented as a globally connected database.

The semantic network is an abstract representation of WWW data, based on RDF standards and other standards to be defined. This is developed by the World Wide Web Consortium (W3C), with contributions from academic researchers and industry partners. Data can be defined and linked in such a way that there is more efficient detection, automation, integration, and reuse in different applications. The semantic network is a sequel to the Internet (WWW) that allows people to share content beyond the boundaries of applications and websites.

[1] World Wide Web Consortium

[2] World Wide Web

17th International Conference on Informatics and Information Technologies

If HTML and Web pages together make online documentation look like one big book, the semantic network makes all the data in the world look like a global and huge database.

The great promise of the Semantic Network is that it can be understood not only by humans but also by machines. Web pages will be important for software programs - they would have semantics - allowing programs to communicate with the network in the same way as people. Programs can exchange data over the Semantic Network without being explicitly designed to talk to each other.

XML bits are a way of expressing metadata for a website. We are all familiar with metadata in the context of a data system: When we look at a file on our computers, we can see when it was created, when it was last updated, and by whom it was originally created. Similarly, Semantic Web sites will be able to tell our browser who is the author of the site and perhaps even where he went to school, or where that person is currently employed. Theoretically, this information will enable Semantic Web browsers to answer questions through a large collection of web pages. In his article for Scientific American, Berner-Lee and his co-authors explained that you can, for example, use the Semantic Network to look for a person you met at a conference whose name you only partially remember.

Indeed, in many cases, the <meta> HTML tag is abused in an attempt to improve the visibility of their websites in search results. Search engines have once experimented with using keywords, delivered via the <meta> tag, to index results, but soon discovered that unscrupulous website authors include tags that are not related to the actual content of their website. As a result, search engines are starting to ignore the <meta> tag in favor of using complex algorithms to analyze the actual content of a website.

III. Web Crawling

Web Crawler also known as a web spider or web robot is a program or automated script that can browse any site in a methodological, automated way. This process is called web crawling or networking. Many legitimate sites, especially in search engines, use crawling as a tool of providing up-to-date analytics data.

There are several libraries for downloading and parsing in C#. Some of them are:

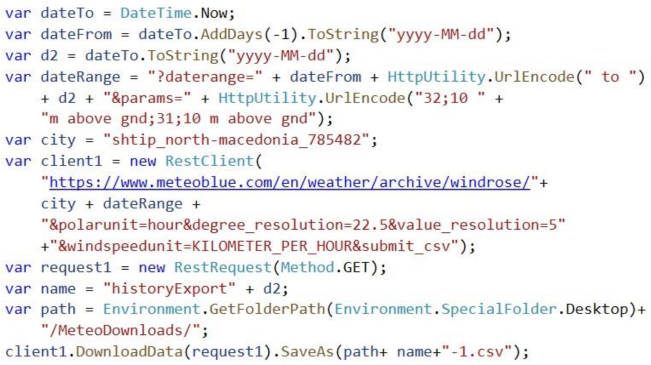
- HtmlAgilityPack- HTML parser that builds read/write DOM and supports plain XPATH or XSLT
- Abot- is a free C # web crawler that is fast and flexible. Refers to the low level of HTTP requests, deployment, link parsing. Only event registration is required to process data from websites.
- Wangkanai.Detection.Crawler – is a reference to ASP.NET Core, which is used to crawl data from web pages.
- AbotX Web Crawler - a C# web crawler that simplifies advanced crawling features. It is an upgrade to the Abot crawler, which offers many extensions.
- Spidey – is a library designed for crawling and parse the web content.
- InfinityCrawler – a simple but powerful library for web content crawling.
- Aspose.HTML for .NET is a cross-platform library that allows you to perform a wide range of HTML tasks directly within your .NET applications. Aspose.HTML supports parsing HTML5, CSS3, SVG and HTML to construct the DOM.
- DotnetSpider.Core is a standard .NET crawling library similar to Web Magic and Scrapy.
- RestSharp is a large library, free HTTP client that works with all kinds of .NET technologies. It can be used to build robust applications that will facilitate the interface to public APIs and fast access.

In my data download research, we use the RestSharp library with C# language. It is one of the many ways you can create a web service or web application in .NET. RestSharp is a comprehensive open-source library that works with all types of .NET technologies. It can be used to build robust applications by simplifying the interface to public APIs and allowing quick and easy access to data without the hassle of sending a large number of HTTP requests. RestSharp offers enormous advantages and saves time with a simple, clean interface, making it one of the best and most used tools today.

With its simple API and powerful library, Rest Architecture is a tool for developers who want to build detailed programs and applications. The RESTful architecture provides resource-oriented information access for creating web applications. It also offers common tasks such as generating URIs, load parsing, and authentication as configuration options, ensuring that developers no longer have to worry about low-level tasks such as networking.

We made 2 crawlers. The code is written in a C# console application where the RestSharp and dotNetRDF libraries are linked.

The first crawler download 3 .csv files from www.meteoblue.com. From this site, each file is downloaded by clicking on a radio button and then “Download as CSV” button. Each radio button displays data about a particular wind direction. My crawler downloads files by entering the city take today's date for “date to” and takes 1 day to get the “date from”, which we will use for variables to form a link from where the data for the given day is downloaded. However, if we want data for an extended period, it can only be inserted into the code at a specific date and subtracted from the appropriate number of days to retrieve from when we want it and by activating the console application the data is downloaded to the CSV file locally. For downloading and writing files locally, we use the following code:



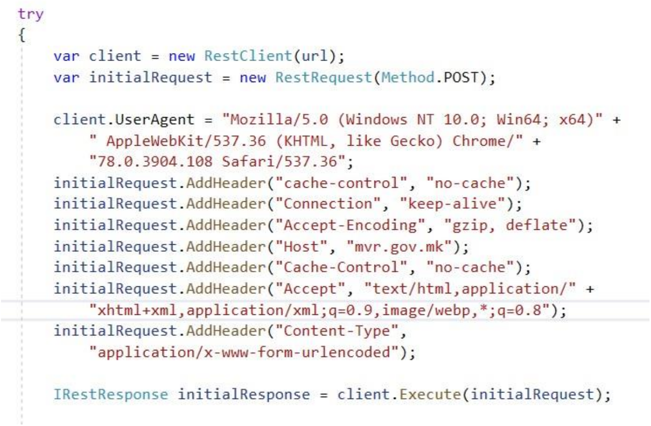
```
var dateTo = DateTime.Now;
var dateFrom = dateTo.AddDays(-1).ToString("yyyy-MM-dd");
var d2 = dateTo.ToString("yyyy-MM-dd");
var dateRange = "?daterange=" + dateFrom + HttpUtility.UrlEncode(" to ")
    + d2 + "&params=" + HttpUtility.UrlEncode("32;10 " +
    "m above gnd;31;10 m above gnd");
var city = "shtip_north-macedonia_785482";
var client1 = new RestClient(
    "https://www.meteoblue.com/en/weather/archive/windrose/"+
    city + dateRange +
    "&polarunit=hour&degree_resolution=22.5&value_resolution=5"
    +"&windspeedunit=KILOMETER_PER_HOUR&submit_csv");
var request1 = new RestRequest(Method.GET);
var name = "historyExport" + d2;
var path = Environment.GetFolderPath(Environment.SpecialFolder.Desktop)+
    "/MeteoDownloads/";
client1.DownloadData(request1).SaveAs(path+ name+"-1.csv");
```

Fig. 1. Download first file from meteoblue about Wind(10m above ground)

The code about the other 2 downloads is similar, the only parameter for wind is different. The downloaded results is for Shtip, if we want the other city we need to change value for variable "city".

The second courier is downloading news from the Ministry of Interior website on which we will conduct my security research in the Republic of Macedonia.

We first call the Ministry of Interior website, the news section, and in response, we get their homepage. We use this answer to send another call where we send the login information called cookies, and the generated data we already received from the first call for __VIEWSTATE and __EVENTVALIDATION.



```
try
{
    var client = new RestClient(url);
    var initialRequest = new RestRequest(Method.POST);

    client.UserAgent = "Mozilla/5.0 (Windows NT 10.0; Win64; x64)" +
        " AppleWebKit/537.36 (KHTML, like Gecko) Chrome/" +
        "78.0.3904.108 Safari/537.36";
    initialRequest.AddHeader("cache-control", "no-cache");
    initialRequest.AddHeader("Connection", "keep-alive");
    initialRequest.AddHeader("Accept-Encoding", "gzip, deflate");
    initialRequest.AddHeader("Host", "mvr.gov.mk");
    initialRequest.AddHeader("Cache-Control", "no-cache");
    initialRequest.AddHeader("Accept", "text/html,application/" +
        "xhtml+xml,application/xml;q=0.9,image/webp,*;q=0.8");
    initialRequest.AddHeader("Content-Type",
        "application/x-www-form-urlencoded");

    IRestResponse initialResponse = client.Execute(initialRequest);
```

Fig. 2. Part of the code about the crawling

The web site form is made in ASP.NET which works with the doPostBack() function and an application model that enables a particular page to validate and process its own data.

When we have already received the data from their site, it is then parsed in RDF format.

We use the dotNetRDF library to parse in RDF format. This library is written in C# and is designed to get a simple yet powerful RDF data API. As such it has many classes for executing different tasks for reading and writing RDF data and more. Creating triples of data in the form of subject, predicate, and object.

An RDF document can be considered to form a graph, so we represent it as a set of RDF triples as graphs. The assignment of values to locations is done using the appropriate triple in which the object represents the location which has a proper identifier, the predicate or link represents the name of the location and the object represents the value of the location data. Other triplets for other locations are appropriately defined. All library graphs are IGraph interface implementations and generally derive from the abstract BaseGraph class that implements some of the basic interface methods, allowing specific implementations to focus on specifications as storage/security thread persistence. Implementation of IGraph is the representation of memory in the RDF document. The most commonly used application of IGraph is the Graph class. The library operates first on the level of triplets, graphs and Triple Stores and provides very limited interface support and no direct OWL support. The triples can be added to the graph using the Assert (..) method. The method takes a triple or list of triples.

IV. Results from the Research

Having already written the data in the FDF format, we did some research on them with SPARQL queries.

If we enter a word in the text field and search for it, then we check to see if the data contains the appropriate term. Since the word may be singular or plural so that we have no problems in the search we cut the last letter and extract the relevant data. If we enter a sentence or several words, then we remove the words consisting of 1 or 2 letters and search the content for the remaining words.

We first did research into which part of Macedonia had the most thefts during 2019.

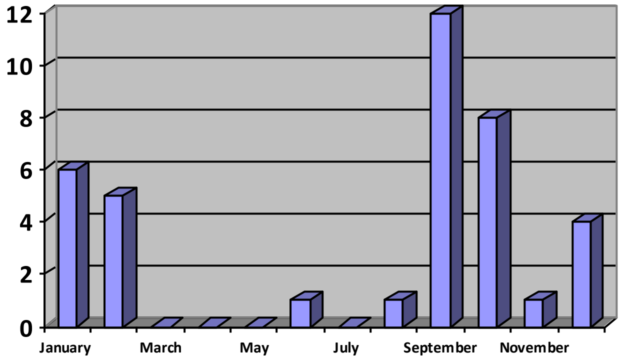


Fig. 3. Results from the first search about "theft"

From the results, we can conclude that the most thefts has been reported in September.

The second survey is in which part of Macedonia has the highest number of homicides in 2019.

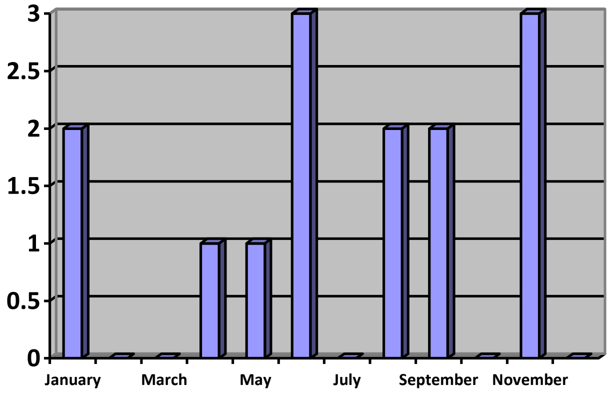


Fig. 4. Results from the second search about “killings”

We can see that the most homicides were reported in June and November.

And the third research concerns where there was drug trafficking in 2019.

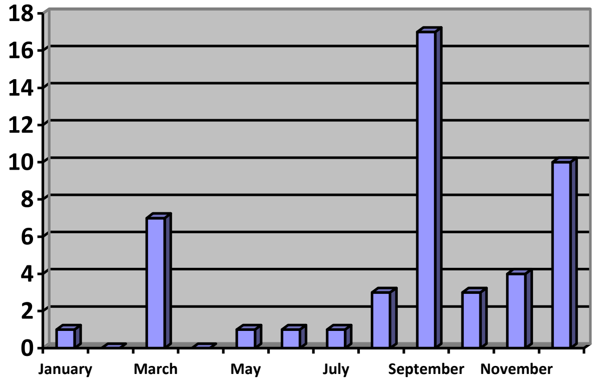


Fig. 5. Results from the third search about “drug trafficking”

Figure 5 has shown us that most reported cases about “drug trafficking” were reported in September.

V. Conclusion

Using semantic web and web crawling we can get all the data from any public news that are available and based on the collected data we can make research about any specific field. With this method we can get all the data for longer period which then is parsed in RDF format, from where using queries we filter the news with the parameters we need. We used this method to gather the needed data for our research and based of the data we conduct our investigation.

We can conclude that there were 38 thefts in all of Macedonia. Most thefts were reported in Skopje, a total of 10. No thefts were reported in Kriva Palanka.

By the time of execution, most were in September, a total of 12, while none in April and May.

There was a total of 14 homicides in Macedonia, mostly in Skopje, 5. Most of the cities do not mention the word homicide.

3 murders were committed on June, 2 in Skopje and 1 in Strumica.

In February and March, there was no mention of murder.

"Drug trafficking" is mentioned 49 times in total. It is often mentioned about the city of Skopje, 14 times in total, and never in Kriva Palanka.

It is mentioned 17 times in September, while in February and April it is never mentioned.

According to these data, we can conclude that Kriva Palanka is the safest city to live in, and Skopje is the riskiest city to live in. We can conclude that security in Macedonia is not improving at all, as drug trafficking is often mentioned in December, and there have been several thefts. The investigation is not 100% valid because, in the context of the extract, it has published a story which states, for example, "murder", but refers to finding a murderer for an old crime, or for "drug trafficking" news that there was raid in certain cities but no drug-trafficking was found and committed in that month.

References

[1] Introduction to Semantic Algorithms https://twobithistory.org/2018/05/27/semantic-web.html May, 2018

[2] Berners-Lee, Tim, James Hendler, and Ora Lassila. “The Semantic Web.” Scientific American, May 2001

[3] Introduction to Semantic Algorithms. https://en.wikipedia.org/wiki/Semantic_Web#Applications

[4] Considered nuggets for data mining. https://nugetmusthaves.com/Tag/crawler

[5] Downloaded data on the nuggets https://www.nuget.org/.

[6] Crawler for downloading data from open pages. https://nugetmusthaves.com/Package/HtmlAgilityPack

[7] Ministry of Interior of the Republic of Macedonia. Published news. https://mvr.gov.mk/vesti 2019

[8] Data Writing Tool in .RDF format. https://www.nuget.org/packages/dotNetRDF

[9] Description of the dotNetRdf frame. https://github.com/dotnetrdf/dotnetrdf/wiki/UserGuide-Library-Overview June, 2017

[10] RDF Schema Syntax. https://www.w3.org/2001/sw/RDFCore/Schema/200212/ November, 2002

[11] Meteoblue wind directions. https://www.meteoblue.com/en/weather/archive/windrose/shtip_north-macedonia_785482