Using Conditional Random Field in Named Entity Recognition for Crime Location Identification

Quintin Jackson Goraseb and Nathar Shah
Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia
Email: nathar.packier@mmu.edu.my

***Abstract*— Electronic data or information comes in different forms, some are structured data and others unstructured data. The act of collecting such data is known as data mining. This paper will discuss the mining of crime data from electronic news sources in Malaysia, and how this data is further transformed to extract meaningful information from it. Furthermore, the paper will demonstrate how crime locations can be identified within the various news articles. This is significant because there are cases where a location name is mentioned in the news article but that is not the true crime location. To help achieve this, the system makes use of Named Entity Recognition (NER) algorithms. They are task with identifying locations in various sentences. To bring more accuracy to the work, the system will employ machine learning technique known as Conditional Random Field (CRF) to recognize if a sentence is referring to a crime location.**

***Index Terms*—data mining, machine learning, text extraction**

I. Introduction

Social media, advances in smart-phone technology together with Internet of Things are all major contributors to the huge volumes of information available today. Online newspapers have increased in popularity in recent years as the information is timely and readily available. The articles contain a wide variety of information, ranging from sports, crime, politics, and popular culture.

With valuable information available from a variety of sources including electronic newspapers and archives, there is still a lack of software that can extract specific information and present it to the reader. The search engines (such as Google) only provide matching sources to the specific search done and still one needs to go through all the results one at a time (Cowie Lehnert 1996). The chosen domain in this paper is crime since this is one of the most important aspects of life which individuals are most interested in. Crime yields a lot of influence in society, as individuals make decisions like where to live and visit to name a few based on the level of crime in that place. Since these type of information is not readily available in a single application, it makes this project so much important as it can be used as a stepping stone for future advances in the study of information extraction.

The objective of this research project is to extract crime information available on online news articles and present it in a graphical form. To achieve this goal, the paper discusses a methodology that consists of six steps. We discuss the extraction techniques used to extract crime specific information from newspaper articles. We will then show the various techniques used in data transformation to achieve a more structured data format. To identify locations, we used Named Entity Recognition (NER) technique and then employed Conditional Random Field (CRF) to classify whether a sentence containing a location is a crime location sentence. Articles that focus on crime are extracted from three Malaysian online newspapers. The end product will be stored and presented in a graphical format.

II. Methodology

A. Data Acquisition and Transformation

Web crawlers or spiders as they are also known, are pieces of code which are capable of infiltrating targeted websites in order to extract information which they are assigned to retrieve. Web Scraper [1] is a company that focuses on extracting information from various web pages. They provide a free web scraper Google extension, that allows us to create sitemaps that are used to plan out how the website should be traversed. The scraped data is then exported out in a CSV format. Web Scraper was chosen for this project to be the tool that is used to collect data from the online news sources. The motivation behind the choice was the web scrapers ability to schedule and process request asynchronously. What this entail is that web scraper has no need to wait for request to finish processing, it can simply send another request and perform some other task in the meantime. Sending multiple concurrent request at the same time will enable us to perform fast crawls. Web scraper is also concerned with politeness and offers a variety setting so we can have more control of it.

Once the crawler has successfully collected all the data that has been requested, the next step is to process the data. The data is put through a preprocessing phase, and since data was collected using different sources, data integration

Manuscript received July 3, 2018; revised December 4, 2019.

doi: 10.18178/ijmerr.9.2.252-257

is performed. It is a process whereby data from different sources are combined to allow users to have a more uniform and unified view of the data. The collected datasets vary in terms of the noise of the data and its level of redundancy and consistency. Most of the data collected will turn out to be meaningless data and simply storing such data is pointless and expensive in regard to storage space. This is the primary reason why data collected should be put through a thorough preprocessing phase. Not only will this save on storage space but it will improve analysis accuracy [2].

Data redundancy elimination is a procedure used to eradicate data that is repeating or in surplus. Redundant data can bring about certain issues including an increase in unnecessary data transmission expense, it can waste storage space and lead to data inconsistency, reduction of data reliability and data damage. Some negative effects caused by data redundancy reduction include things like when data is compressed and decompressed it can cause additional computational burden. It is important to weight the costs and benefits of redundancy reduction. There are some methods that are used in redundancy reduction, such as redundancy detection, data filtering and data compression [3]. To avoid redundant data, we have created a python based program that eliminates unwanted data. Once the crawler delivers the data payload, its not just crime related articles but every article published on that particular day. The program then uses a keyword search technique that utilizes a list of crime terms to remove those articles not related to crime and keep those which are. The accuracy of this approach is only about 90 percent, as those articles which refer to crime in meaning don't get recognized.

B. Sentence Tokenization

Once we have extracted all the data and transformed it into a more structured form, the following step is tokenization. Here the individual articles need to be split into individual sentences. This action allows for the algorithm to easily identify location names in a given sentence. The best tool for this is the PunktTokenizer from the NLTK toolkit. The tokenizer will divide an article into a list of sentences [4].

C. Location Identification

The next step is to identify the locations in the individual sentences. To achieve this, we used a Named Entity Recognition algorithm [5].

D. Feature Identification

One of the main problems we identified earlier was the fact that an article can have more than one location, so how do we tell which is the crime location. One way of going about this is to assign features and labels to the individual sentences. If a sentence contains crime term and a location, it will be labeled as a crime location. The ones with only a location or crime will not be considered.

E. Training CRF

The CRF algorithm will learn the difference between the features from a training dataset and create a model, which it will use to automatically dedicate labels to each sentence in the article.

F. Store and Display

Once we have all the locations and crime instances extracted and organized, those articles have to be stored in a database, from where it can be queried and viewed by the public. In this case the best option for a database system was a MySQL database. Since the program is only concentrating on three news sources, the amount of crime data extracted is not much and is easily handled by the database mentioned.

III. Experiment Design

Web crawlers or web spiders as they are also known are pieces of code which are capable of entering different websites and extracting certain information which they are assigned to from those sites. First, the Scrapy Framework will be examined. Scrapy is an application framework for online source crawling and the extracting of structured data. This is a Python based application designed specifically for web scraping. The one thing that sets Scrapy apart is its ability to schedule and process request asynchronously. What this entail is that Scrapy has no need to wait for request to finish processing, it can simply send another request and perform some other task in the meantime. Sending multiple concurrent request at the same time will enable you to perform fast crawls. Scrapy is also concerned with politeness and offers a variety setting so you can have more control of it.

The Beautiful-Soup is a also a Python based library that is used to extract data from HTML and XML mark-up languages. Beautiful-Soup allows to extract specific content from web pages, remove the HTML mark-up and save the information for later analysis. This is a tool that helps cleanse and parse information that has been scraped from the web. It is very good at accessing information held deep within HTML structure. Beautiful-Soup makes use of tags to select content and then uses the "prettify" module to create a better view of the information within the tag.

Crawl-Jax is Java based open source tool for crawling web pages. It uses an event-driven dynamic crawling engine to access JavaScript-based Ajax applications. Crawl-Jax is able to create a state-flow graph of the dynamic Document Object Model (DOM) states and the transitions between them. This state flow-flow graph allows for many types of web analysis and testing techniques: detecting broken links, detecting unused code, invariant-based testing, test generation, non functional testing.

Once the crawlers have successfully collected all the data that has been requested, the next step is to process the data. The data will be put through a preprocessing phase. Data is collected using many different sources and therefore the collected datasets vary in terms of the noise of the data and its level of redundancy and consistency. Most of the data collected will turn out to be meaningless data and simply storing such data is pointless and expensive in regard to storage space. This is the primary reason why data collected should be put through a

thorough preprocessing process. Not only will this save on storage space but it will improve analysis accuracy. We will take a look into some relational data preprocessing techniques in use.

Data integration is the process whereby data from different sources are combined. This allows users to have a more uniform and unified view of the data. One of the key approaches used in integration is known as data Warehousing. This system is typically used for reporting and data analysis. Data warehouses are considered to be central repositories of integrated data from one or more disparate sources. This approach involves three techniques, Extract, Transform, and Load (ETL)data from heterogeneous sources into a single view schema, which allows data from different sources to become more compatible. A set back of data warehouse is that it is less feasible for data that is frequently updated. This means that the ETL process must be continuously re-executed for synchronization. The ETL phases are typically executed in parallel, while data is being extracted, another transformation process executes. It processes the already received data and prepares it for loading.

Extract is the first part of the ETL process. It involves extracting data from different sources. It is important that be extracted correctly since it sets the stage for the success of subsequent processes.

Transform involves cleaning the data, applying a series of business rules or functions, the checking of data integrity, the creation of aggregates or disaggregate to the extracted data to prepare it for the loading stage. Not all data require transformation, those that don't are known as direct move or pass through data. Data is transformed in order to get it into a standard format since they all are derived from separate sources with different formats. All data cleaning is typically done in a separate data stage area before loading the transformed data into the ware-house.

Load stage is where transformed data is loaded into an end target like a data warehouse. Some data warehouses may override existing information with cumulative information, updating extracted data is done on a daily, weekly or monthly basis. They might also add new data in a historical form at regular intervals. It is a good idea to utilize an established ETL framework, it may lead to better connectivity and scalability. A good ETL tool should be able to communicate with relational databases and read the all kinds of file formats available.

The second approach used in integration is known as data Federation. It is a software type that standardizes the integration of data from different sources. It has a wrapper per data source for extraction and a mediator for integration. These systems mainly focus on data transformations for schema translation and schema integration, while providing limited support for data cleaning. Unlike with data warehouse, data is not per-integrated but needs to be extracted from multiple sources, transformed and combined during query run-times. There are no acceptable response times as the corresponding communications and processing delays can be significant.

Data entry and acquisition is inherently prone to errors both simple and complex. These front-end processes should be taken into account with respect to reduction in entry error, even though errors are a common fixture in large data sets. Data cleaning is a process that is used to improve data quality by identifying inaccurate, incomplete, or unreasonable data and then to modify or delete such data. Serious data cleansing practices involve decomposing and reassembling the data. Data warehouses require and provide extensive support for data cleaning. Data is loaded and continuously refreshed in large amounts from a variety of sources, so the probability that some of the data sources contain dirty data is high. Some complimentary procedures in data cleaning are defining and determining error types, searching and identifying errors, correcting errors. Manual processes of data cleansing is common but not advised. This is because it is itself error prone and takes up a lot of resources. The need to use powerful tools that automate or greatly assist in the data cleansing process are recommended. They are a more practical and cost effective way to achieve a reasonable quality level in an existing data set. New and improved computers allow performing the data cleansing process in acceptable time on large amounts of data. The following are some issues in data cleansing.

- Missing data
- Determining record usability
- Erroneous data

The definition of data cleansing depends on the particular in which it is applied. These major areas that include data cleansing as part of their defining processes are, data warehousing, data mining, and data/information quality management. Data cleansing should be viewed as a process, and the following three phases define a data cleansing process.

- Defining and determine error types
- Search and identify error instances
- Correct the uncovered errors

Knowing what data is supposed to look like will allow errors to be detected. This is not the case though in the real world as data often is very diverse and rarely conforms to any standard. Therefore, more than one method for outlier detection is necessary to capture most of the outliers. The following methods are utilized for error detection.

- Statistical-The values of mean, standard deviation, and range are used to identify outliers fields and records.
- Clustering-Outlier records are identified using clustering on Euclidean distance.
- Pattern-based-Fields and records who do not fit with existing patterns in data are identified.

Data redundancy elimination is a procedure used to eradicate data that is repeating or in surplus. Redundant data can bring about certain issues including an increase in unnecessary data transmission expense, it can waste storage space and lead to data inconsistency, reduction of data reliability and data damage. Some negative effects caused by data redundancy reduction include things like when data is compressed and decompressed it can cause

additional computational burden. It is important to weight the costs and benefits of redundancy reduction. There are some methods that are used in redundancy reduction, such as redundancy detection, data filtering and data compression.

Data preprocessing plays an important role in the big data arena, it is a time consuming but important step towards achieving the final product. In this project our main concern is with when preprocessing should take place efficiently without slowing down the entire process. We will be acquiring our data through the web crawling process, which will yield massive amounts of unstructured data. The next step will be to put this data through a preprocessing phase. There are two options when it comes to executing this, the first is to directly store the data into our data store as unstructured data and then remove it to perform preprocessing or to implement the preprocessing phase before the storage. If we are to store the data directly and then cleanse it, it will take more time and processing compared to preprocessing first. The diagram below illustrates the two options mentioned above.

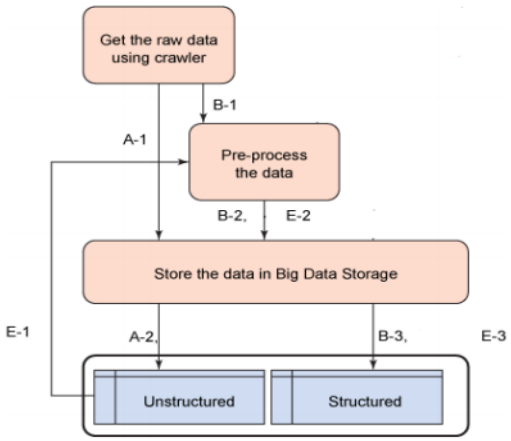

Figure 1. Big data accessing steps

Description for Fig. 1:

• A-1, The crawler reads raw data
• A-2, Data is stored in unstructured storage
• B-1, The crawler reads raw data
• B-2, Data gets preprocessed
• B-3, Data is stored in structured storage
• E-1, Data which is stored without preprocessing is unstructured and cannot be useful unless it is in structured format.
• E-2, Data is preprocessed
• E-3, Data that is preprocessed and in structured format is stored in structured storage.

When we choose to perform preprocessing before storing the data, we have to find a platform on which to perform this operation. Hadoop ecosystem offers the best solution to this problem when you want to manipulate and apply transformations to data before its loaded into a data warehouse. The transformations and data flows are coded in MapReduce and the approach has been used with the extract, transform, load processes. So when we use a Hadoop based landing zone for the data, all that we will require to use Hadoop as a transformation engine is in place. After the data extraction process is done, we will implement our transformation logic into MapReduce and after the data is transformed it will be loaded into the data warehouse using Sqoop. Warehouse data is able to be moved into storage using some standard connectors, which are provided by various vendors. Moving preprocessed data is much convenient because it is in structured form. This data movement into storage can take place using simple extract, transform, and load processes. The advantage of NoSQL data stores is that the architectures can lend themselves to cloud implementations because they are structured consistently with cloud architecture. NoSQL databases are high performance, non-relational databases, which make use of a variety of data models. These include document, graph, key-value, and columnar. They are known for their ease of development, high availability, and scalable performance. Key-Value databases are the most simple databases around, there is a key and there is the rest of the data (the values) and that is it. Searching the key field will allow us to find the value data. Key-Value databases are much faster than relational databases and they have the ability to scale allowing them to grow by hundreds and thousand times without significant redesign. Amazon DynamoDB is a NoSQL database service that has good performance with seamless scalability. This database can be used to create tables which are able to store and retrieve any volume of data, and service any level of traffic. The database is capable of spreading the data and traffic for the table over a number of servers to handle the request capacity, which the customer has determined and the amount of data stored. This is all done while it maintains a consistent and fast performance.

Big data analysis involves the collecting, organizing and analysing of large data sets to try and identify patterns and other useful information. Big data analytics is a powerful tool in the sense that it can help organizations and individuals gain a more clear understanding of the information making up the data. They will be able to identify the data that is most important. The following are a list of things which are possible and not possible with data analysis.

Things it can do:

- Diagnostic analysis
- Predictive analysis
- Prescriptive analysis
- Monitoring an event as it happens

Things it cannot do:

- Predict a definitive future
- Imputation of new data sources
- Find a creative solution to a business problem
- Find a solution to a not so well defined problem

Issues facing data analysis

- Dealing with large amounts of data can create bias.
- False positives are created when individuals rush to make decisions based on a subset of data.
- When dealing with huge volumes of data, it can be hard to find true value from the data.

In terms of analytic methods, the best option is to utilize parallel processing as the main method for the

extraction of information. MapReduce is the perfect tool to use in this case as we will conduct an off-line analysis on the data because we don't have high requirements on response time. The aim throughout was to extract and import the data onto a platform to conduct preprocessing and data analytics. The Hadoop ecosystem provides the best environment to achieve this. Tools like Microsoft Excel which provide powerful data processing and statistical capabilities can also be implemented at a later stage when the need arises.

One of the major problems facing big data scientist today is storage. The rate at which data is accumulating is at a massive scale and adequate data storing methods are still being explored. Big data storage refers to the storage and management of huge datasets while achieving reliability and availability of data accessing. There are a few storage systems that have been developed to meet the demand of massive data. Choosing the right database is very important, simply because it has to be able to handle all the data flowing in and store it in a manner that is easy to access. The following is a list of databases which are considered for this project.

Key-Value database: These databases are used for quick storing of information both basic and complex. They are known to be very performant, efficient and easily scalable. The data is stored in correspondence to key-values. Amazon Dynamo is a distributed key-value data storage system. DynamoDB eliminates the need to perform database administration or maintenance. This data store is great for ad tech, big data, gaming, mobile apps and other applications which need fast response time and easily scale with demand. DynamoDB backs up your data in three separate facilities.

Document database: Document databases are able to handle any of the short comings experienced by key-value and column databases. Any structure that is able to form a document can be stored using this data storage system. Some of the few setbacks of this database system includes; when retrieving a value, means that the entire lot has to be brought out. The same applies for updates, causing the system to have low performance. SimpleDB is also a distributed data store and a web service of Amazon. Data is distributed into various domains where it can be stored, acquired, and queried.

Column-Oriented database: This type of database can be utilized when key-value pairs are not sufficient enough, and storing of large volumes of data needs to be done. Database management systems which implement the use of column-based, schema-less models can scale well. Cassandra is a distributed data store that can handle large volumes of structured data. This system incorporates the concepts of both DynamoDB and Googles Big Table.

IV. Experiment Results

For this experiment, we collected 5 articles from a local newspaper and tokenized these articles into sentences. Then we investigated two different NER algorithms to find the one that accurately identify a location correctly. The details of the two NER algorithms compared are given below.

A. Stanford NER

Java based Stanford Named Entity Recognizer is used to identify four types of entities: location, organization, person and miscellaneous [6]. We used the location information in this work. The algorithm uses the CRF classifier to train the model for identifying named entities.

B. LBJ Tagger

LBJ NER Tagger is one of the models of the Named Entity Recognition system developed at the University of Illinois [7]. This model is based on regularized average perception. It uses gazetteers extracted from Wikipedia, where the models for word class are derived from unlabeled text and expressive non-local features. We used the classic 4-label type model to identify the locations and organizations

C. Results

Algorithm	Precision	Accuracy
Stanford NER	0.93	0.76
LBJ Tagger	0.98	0.94

The results of these two algorithms to identify correct locations on the set of 5 local newspaper articles is given above. As can be seen, LBJTagger algorithm has the highest accuracy (94 percent) followed by the Stanford NER (76 percent).

V. Conclusion

The key problem address in this article is associating semantic to an entity through features and labeling. Identifying the exact crime location from a crime report in an article can be achieved to a degree of success with this approach. A straight thru identification of location without any features and labeling would give a poor accuracy as an article could reference many other locations not specific to a crime location. The key difference with other works is the CRF algorithm will learn the difference between the features from a training dataset and create a model, which it will use to automatically dedicate labels to each sentence in the article.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Web Scraper. [Online]. Available: http://webscraper.io/

[2] E. Rahm, H. H. Do. Data Cleaning: Problems and Current Approaches. Accessed August 23, 2016, [Online]. Available: http://betterevaluation.org/sites/

[3] L. Maurizio, "Data integration: A theoretical perspective," in *Proc. the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ACM, 2002.

[4] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python, Oreilly, 2009.

[5] D. Nadeau and S. Sekine. (2007). A survey of named entity recognition and classification. [Online]. Available: http://nlp.cs.nyu.edu/sekine/papers/li07.pdf

[6] J. R. Finkel, T. Grenager, and C. Manning, "In-corporating non-local information into information extraction systems by gibbs

sampling," in *Proc.of the 43rd Annual Meeting on Association for Computational Linguistics*, *Association for Computational Linguistics*, pp. 363370, 2005.

[7] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proc. the Thirteenth Conference on Computational Natural Language Learning*, *Association for Computational Linguistics*, pp. 147155, 2009.

Quintin Jackson Goraseb is a student at Multimedia University, Cyberjaya, Malaysia pursuing Bachelor of Computer Science majoring in Software Engineering.

Nathar Shah is an academic at Multimedia University, Cyberjaya, Malaysia with years of experiences teaching computer science subjects. He is a graduate of the University of York, United Kingdom in Software Engineering.