Crime Analytics: Analysis of Crimes Through Newspaper Articles

Isuru Jayaweera, Chamath Sajeewa, Sampath Liyanage, Tharindu Wijewardane, Indika Perera
Department of Computer Science and Engineering, Faculty of Engineering
University of Moratuwa
Sri Lanka
{jayaweera.10, chamaths.10, sampath.10, tharinduwije.10, indika}@cse.mrt.ac.lk

Adeesha Wijayasiri
Department of Computer and Information Science and Engineering
University of Florida
Gainesville, FL 32611, United States
adeeshaw@ufl.edu

***Abstract*—Crime analysis is one of the most important activities of the majority of the intelligent and law enforcement organizations all over the world. Generally they collect domestic and foreign crime related data (intelligence) to prevent future attacks and utilize a limited number of law enforcement resources in an optimum manner. A major challenge faced by most of the law enforcement and intelligence organizations is efficiently and accurately analyzing the growing volumes of crime related data. The vast geographical diversity and the complexity of crime patterns have made the analyzing and recording of crime data more difficult. Data mining is a powerful tool that can be used effectively for analyzing large databases and deriving important analytical results. This paper presents an intelligent crime analysis system which is designed to overcome the above mentioned problems. The proposed system is a web-based system which comprises of crime analysis techniques such as hotspot detection, crime comparison and crime pattern visualization. The proposed system consists of a rich and simplified environment that can be used effectively for processes of crime analysis.**

Keywords—crime analysis, web crawling, document classification, entity extraction, near duplicate detection

I. INTRODUCTION

Crime analysis has become one of the most vital activities of the modern world due to the high magnitude of crimes which is a result of technological advancements and the population growth. Law enforcement organizations and the intelligence gathering organizations all around the world usually collect large amounts of domestic and foreign crime data (intelligence) to prevent future attacks. As this involves a large amount of data, manual techniques of analyzing such data with a vast variation have resulted in lower productivity and ineffective utilization of manpower. This is one of the most dominant problems in many law enforcement and intelligence organizations.

Unlike developed countries such as United Kingdom and United States of America, Sri Lanka is still using a manual crime recording and analysis system which gives little support for decision making. In [1], this manual process is described in detail. First the complaint is recorded as described by the complainant. Then the crime is investigated physically and depending on its results, complaint is included into the GCR (grave crime record) or MOR (minor offences record). In addition to that the crime scene is spatially mapped onto a physical map by using a pin or a dot with a specific color assigned to the type of crime. During further investigation, discovered data is maintained in GCR or MOR in a categorical manner. Court data is recorded in another book and GCR or MOR reference numbers are used for cross referencing. The Sri Lankan police department has practiced this particular manual method for a long period of time to plan regional security arrangements and placements of police patrols.

There are several significant reasons for crime analysis such as to identify general and specific crime trends, patterns, and series in an ongoing, timely manner, to maximize the usage of limited law enforcement resources, to access crime problems locally, regionally, nationally within and between law enforcement agencies, to be proactive in detecting and preventing crimes and to meet the law enforcement needs of the changing society. There are various crime data mining techniques available [2] such as clustering techniques, association rule mining, sequential pattern mining, and classification and string comparison.

Several web based crime mapping systems are available on the Internet such as narcotics network in Tucson police department, but majority of them have been custom made for legislative authorities in different countries and those systems are not accessible to parties outside that particular law enforcement or legislative authorities [3] [4].

This paper presents a web based crime analysis system. Sri Lankan English newspapers (Daily Mirror, The Island, and Ceylon Today) are used as the source for details of crime incidents. Newspaper articles are crawled using a focused crawler and they are classified using a SVM based classifier. Required entities are extracted from classified crime articles and duplicate detection is performed. By using preprocessed data, crime analysis operations are performed and results are displayed using web based GUI. Unlike most systems, this system is open to anyone who is interested in crime analysis.

When newspapers are considered, they contain articles only for a subset of total crime population. However they contain most of the major crimes that take place in Sri Lanka. Often major crimes gain more interest than minor crimes. Most of the

978-1-4799-1740-2/15/$31.00 ©2015 IEEE

time the police and other interested parties are more concerned about major crime incidents rather than minor crime incidents when taking decisions. Therefore crime analysis results based on newspaper articles will be useful to interested parties (police, researchers, investors and tourists) as means of assistance for their respective tasks even though newspapers cannot reveal the exact number of crimes.

The proposed system cannot be directly validated using records of the police department because police records include both major and minor crime incidents. The proposed system is based on newspaper articles so it includes only a subset of total crime incidents. So individual components of the proposed system are evaluated and results of that evaluation are used to measure the effectiveness of proposed system.

The remainder of this paper is structured as follows. An overview of related work and preliminaries are provided in Section 2. A detailed description of the proposed system is given in Section 3. A discussion on experiments and an analysis on the proposed system is provided in Section 4. The conclusion and future improvements are discussed in Section 5.

II. Related Work & Preliminaries

A. Related Work

Crime data mining is the application of data mining techniques for crime analysis [5]. Various researches have been carried out in this domain and few of them are given in [1], [2], [6] and [7]. Crimes can be divided into subcategories based on different criteria. In [2] eight crime categories are given. They are traffic violations, sex crimes, theft, fraud, arson, drug offenses, cybercrimes and violent crimes. They have given definitions for each category in local law enforcement level and national law enforcement level. IPTC [8] (international press telecommunication council) too has given a different categorization where crimes are divided into war crime, corporate crime, organized crime etc.

There are various crime data mining techniques available [2]. The most commonly used methods are, entity extraction, clustering techniques, association rule mining, sequential pattern mining and classification.

There were many efforts to analyze different types of crimes using automated techniques but there is no unified framework describing how to apply those techniques to different crime types. In [2], they have proposed a framework which includes a relationship between the crime data mining technique and crime type characteristics.

There are several existing systems which use crime data mining techniques for crime analysis such as, regional crime analysis program [9], data mining framework for crime pattern identification [7] and narcotics network in Tucson police department [2].

In [1] a collection of criminal analysis steps are given. Among them, steps such as hotspot detection, crime comparison, crime pattern visualization are significant. In crime pattern visualization, a time series can be drawn between the crime frequency and the time and using it interesting crime trends can be identified. In addition to these steps, [1] has given some other analysis steps such as crime clock, outbreaks detection and nearest police station detection.

Using the above techniques, crime data can be analyzed more effectively and efficiently and law enforcement organization and other interested parties will be able to get more accurate decisions based on them.

An intelligent crime identification system is described in [6] which can be used to predict possible suspects for given crime. They have used five types of agents namely, message space agent, gateway agent, prisoner agent, criminal agent and evidence agent.

B. Preliminaries

1) Web Crawling: Web crawler is an Internet bot. It systematically browses the World Wide Web typically for the purpose of web indexing [10]. Crawlers can be selective about the pages they fetch (Ex- crawl only the pages of selected newspaper sites) and are then referred to as preferential or heuristic-based crawlers [11]. Preferential crawlers built to retrieve pages within a certain topic are called topical or focused crawlers.

2) Document Classification: Document classification is defined as assigning predefined categories to text documents, where documents can be news stories, technical reports, web pages, etc., and categories are most often subjects or topics [12]. To perform an effective document classification, syntactic (arrangement of words) and semantic (meaning of words) aspects of the natural language have to be addressed [13]. There are a lot of algorithms available for text classification and among them SVM (support vector machine) is the best [14]. The most commonly used approach for document classification is the utilization of SVM in combination with other linguistic related resources such as ontology or word net [15] [16]. Classifiers which are built based on imbalanced datasets usually perform well on the majority class data but they perform poorly on the minority class data [17] [18]. Various approaches such as different error costs and sampling techniques [17] [18] have been proposed in order to handle this problem.

3) Entity Extraction: This is the main step of transforming the unstructured data into the structured format. A named entity is typically a name of a person, a place, an organization or a date. Extraction of named entities involves identification of small chunks of appropriate texts and classification of them into one of such predefined categories of interest [19]. For an example if "Jim bought 300 shares of ABC Corp. in 2006" is used for entity extraction, and then the output will be [Jim] $_{\text{Person}}$ bought 300 shares of [ABC Corp.] $_{\text{Organization}}$ in [2006] $_{\text{Time}}$ where Person, Organization and Time are predefined categories. In order to do a successful entity extraction there is some amount of preprocessing needs to be done on the unstructured data (e.g. sentence splitting, tokenizing, POS tagging, etc.) [20].

4) Duplicate Detection: Duplicates (i.e. documents that are exact duplicates of each other due to mirroring and plagiarism) are easy to identify by standard check summing techniques. A more difficult problem is the identification of near-duplicate documents. Two such documents are identical in terms of content but differ in a small portion of the document [21]. Document similarities are measured usually in high dimensional vector space [22], but it is really computationally expensive. Fingerprints such as simhash and shingles are used to produce sketches of the documents to process them efficiently [21].

III. The Proposed System

The proposed system consists of seven major components. They are crawler, classifier, entity extractor, duplicate detector, data base handler, analyzer and graphical user interface. High-level architecture of the system including the above components is given in Fig. 1.

A. Crawler

The main responsibility of the crawler is to crawl news articles of a given newspaper. Required content of the crawled articles is stored in the database for further processing. A focused crawler has been implemented by extending a generic crawler known as Crawler4j [23]. It is an open source web crawler which is used widely for simple applications and is easily modifiable to fit into the application. It supports multi-threaded and resumable crawling. It is implemented in java so that it can be easily integrated into java projects. Jsoup [24] is a java library that can be used for extracting and manipulating HTML data. It has been used to extract the required data (page content etc.) from the crawled HTML pages. Crawler4j cannot handle cookies therefore it cannot be used for web sites which use cookies to identify users. For that, a cookie handler has been implemented which acts as a mediator between the crawler and the web site.

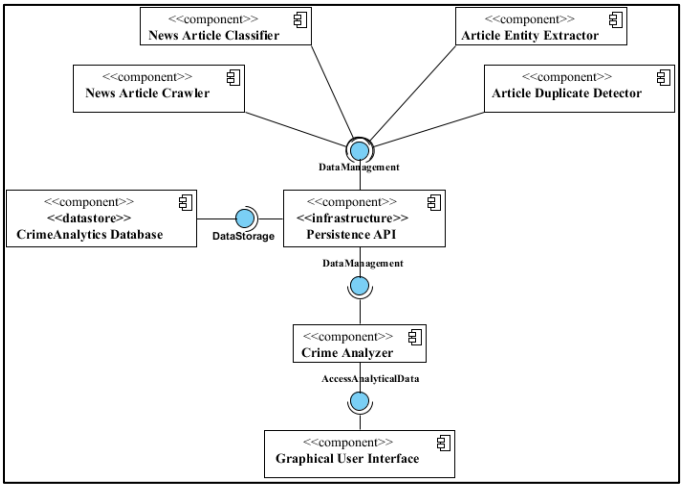

Fig. 1. High level architecture of the proposed system.

B. Document Classifier

Responsibility of document classifier is to classify crawled newspaper articles as crime and non-crime articles. Classified articles will be stored in data base for entity extraction. Documents have to be processed before using them for the classification process. Weka library [25] has been used for this purpose. Documents have been transformed to feature vectors while removing stop words from them using tf-idf transformation. Stemming/ lemmatization has been performed in order to reduce words into their base forms. LibSVM [26] library has been used to implement the classifier. It is an open source implementation of support vector machine algorithm. C-SVC SVM type and RBF kernel has been used for classification [27]. Grid search has been performed in order to find appropriate values for cost and gamma values [27]. Training article collection was highly unbalanced because there are large numbers of non-crime articles compare to crime articles when we consider a particular sample from an article population. Initially hybrid sampling technique was used in order to handle data unbalanced problem. Minority class has been up sampled using SMOTE [28] and majority class has been down sampled using support vectors. However it was not very effective so that different error cost (DEC) approach has been used. In that approach a higher error cost has been assigned to the minority class in order to reduce the effect of the majority class. Summary of the classification results is given in TABLE I.

C. Entity Extractor

This module is used to extract important entities from the classified newspaper articles. From each crime article, entities

TABLE I. Overall Classification Results for Different Approaches.

Method	Accuracy	Sensitivity (true positive rate	Specificity (true negative rate)	G-mean	ROC
Generic	92.7547	63.38	97.96	78.79	81
Grid Search	92.8165	71.35	96.62	83.02	84
SMOTE Up Sampling	95.8876	97.76	94.02	95.87	96
Ensemble Boosting	88.6499	88.18	88.73	88.45	92
Ensemble Bagging	86.8571	92.45	85.87	89.09	91
DEC	95.7163	96.22	95.21	95.71	95

TABLE II. COMPARISON OF PERFORMANCE BETWEEN HYBRID, MACHINE LEARNING AND CONTEXT PATTERN BASE APPROACHES

Approach	Precision%	Recall%	F1 Measure%
Hybrid	82.93	80.95	81.93
Machine Learning	53.84	43.75	48.28
Context Pattern	91.1	10	18.02

such as crime date, location, police, court, victim count etc. are extracted if possible. Several ancillary processes have been carried out to do the required preprocessing on the document corpus to prepare documents for named entity extraction. GATE (General Architecture for Text Engineering) [29] has been used for text processing as well as for entity extraction. It consists of several ready-made applications such as ANNIE, LingPipe and plugins such as date normalizer, OpenNLP and others. Text content of each article has been tokenized using ANNIE English tokenizer and each single word has been considered as a separate token. Sentences have been segmented using ANNIE sentence splitter. The weaknesses that it possessed have been minimized using custom JAPE rules. ANNIE POS tagger has been used to perform POS tagging and identify phrases, referential proper nouns etc. ANNIE POS tagger has been combined with Stanford POS tagger in order to enhance the POS tagging process. Dates given in article have been normalized with respect to the published date using GATE date normalizer plugin. After identifying the location of a crime, the corresponding district has been identified using Google Maps API. Extracted addresses are sent to Map API to obtain the details of the district. Proposed hybrid approach is compared with conventional machine learning and context based named entity extraction approaches. Results are given in TABLE II.

These measurements were taken based on the crime location entity extraction. The results clearly display that the proposed hybrid approach is better than the two conventional methods.

D. Duplicate Detector

The main purpose of this module is to identify exact/near duplicates of newspaper articles and remove them from the database. Newspaper articles have been represented using 64 bit simhash values. The entire contents of newspaper article have not been used to generate simhash values as noise may distort the simhash value. Therefore in order to generate the corresponding simhash value, extracted entities of each article have been used. Crime type, crime date and crime location is used to generate the representation of the document. The steps to calculate Simhash value for an article is given in [30]. To hash each feature, murmer hash implementation [31] has been used. It is a non-crypto graphic hash function. Hamming distance has been used to calculate the difference between calculated simhash values.

E. Database Handler

All database transactions are handled using this module. This has been implemented using Hibernate framework [32].

F. Analyzer

Analyzer module will perform crime analysis operations on processed crime articles. It will perform following crime analysis steps,

- Hot spot detection – identifying a number of crimes in each district and assigning a color depending on the frequency of crime.
- Crime comparison – comparing different types of crimes to get an idea about the growth of a particular crime type over the others.
- Crime pattern visualization - A time series plot is generated to represent the changes in frequency of crime types.

G. Web based GUI

This module is used to visualize crime statistical details of the previous years. The client side of the GUI uses a JavaScript library called high charts [35] for visualizing data with maps, graphs and pie charts. The server side of the web application has been written in JavaScript in order to interact with client side JavaScript library efficiently. The server side analyses data fetched from the database and converts them to a format for visualization. The web application is implemented as a single page web application (SPA) and the communication between the server and client happens through AJAX in order to improve responsiveness.

IV. EXPERIMENTS AND ANALYSIS

As mentioned in the Introduction sections crime analysis functionality of the proposed system cannot be evaluated directly using police crime records. Instead individual components of the proposed system are evaluated for their comprehensiveness and accuracy. Evaluation is focused more on data preprocessing components since all other subsequent crime analysis operations are based on preprocessed data.

Article crawler is evaluated to measure its comprehensiveness. A news article sample during a specific period from CeylonToday, DailyMirror, NewsFirst and The Island is used for the experiment. Results are given in TABLE IV.

The article classifier is evaluated to measure its classification accuracy. Different classification accuracy measurements such as accuracy, sensitivity, specificity and F-score are used. Results are obtained using cross validation method. They are given in TABLE III.

Article entity extractor is evaluated for measuring its accuracy and comprehensiveness of identifying entities. Entities such as crime location, crime date and crime type are used to obtain values for accuracy measurements. Results of the evaluation are given in TABLE V.

Article duplicate detector is evaluated for its ability to detect exact and near duplicates. Several documents have been selected as query documents and precision and recall values for each document was calculated. Average precision and recall values are given in TABLE VI.

From above given results, it is clear that each component of data preprocessing process has acceptable values for accuracy/ comprehensiveness.

Hotspot detection tool which is shown in Fig. 2 aids in identifying the areas with higher crime density than others so that the law enforcement resources can be utilized appropriately. For example the police department can utilize the police patrols in an efficient manner so that the patrol services in areas with high crime density can be increased and the patrol services in other areas can be decreased accordingly. Also travelers and tourists can identify regions which are more suitable for travel and visiting. In addition to this investors can get assistance to find out suitable areas for investments.

Crime comparison tool helps an analyzer to compare the crime frequencies within a particular time period (year) which is shown in Fig. 3. This uses a pie chart to represent frequencies of crimes for each year. By using this tool law enforcement officers can understand clearly what type of crimes needs more attention, so that necessary actions can be taken.

Crime pattern visualizer tool is shown in Fig. 4 and it supports the police officers and crime analyzers to identify gradual changes of crime frequencies with respect to time so that they can change security arrangements accordingly. This has an option to generate a time series plot for each crime type separately. In addition it also has an option to generate time series plots for all the crime types in the same graph making the comparison between graphs easier. Law enforcement officers can use this tool to easily identify the trends involved in different types of crimes. Also if they have followed a security plan, they can evaluate the progress of it.

TABLE III. ARTICLE CLASSIFIER ACCURACY EVALUATION RESULTS FOR DIFFERENT CLASSIFICATION ACCURACY MEASUREMENTS

Accuracy Measurement	Value %
Accuracy	95.7163
Sensitivity	96.22
Specificity	95.21
G-mean	95.71
ROC	95

TABLE IV. AN ANALYSIS OF THE COMPREHENSIVENESS OF THE CRAWLER FOR EACH NEWSPAPER

Paper	No. of articles in the online Paper	No. of articles in our database	Comprehen-siveness %
Ceylon Today	151	129	85.4
Daily Mirror	230	182	79.1
News First	314	287	91.4
The Island	402	330	82.0

TABLE V. EVALUATION OF CRIME TYPE AND CRIME DATE ENTITIES

Entity Category	Accuracy of Extractions %
Crime Location	82
Crime Type	82
Crime Date	74

TABLE VI. EVALUATION OF ARTICLE DUPLICATE DETECTOR IN DETECTING EXACT AND NEAR DUPLICATES

Measurement	Average Value %
Precision	95
Recall	96

V. CONCLUSION

This paper proposed a web based crime analysis system. The proposed system performs crime analysis operations such as hotspot detection, crime comparison and crime pattern visualization. Graphical user interface of the system uses graphs and diagrams to display the results which make crime analysis a very simple task. Then law enforcement officers and other interested users will be able to use this system effectively and efficiently for crime analysis processes. Also this is a public accessible system so that anyone who is interested in this area will be able to use this system.

Crime prediction is expected to be implemented in future to enhance the functionality of the system. Comprehensiveness of the news article collection can be further improved by extending the news article crawler to crawl more news websites. Linguistic knowledge (WordNet, Ontology, etc.) can

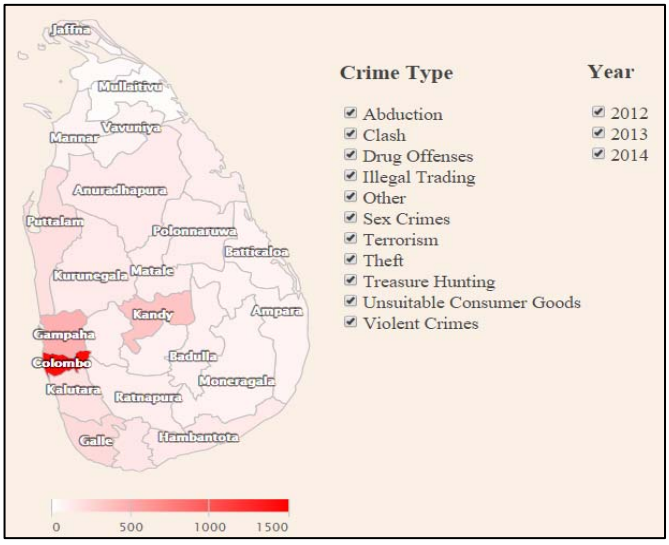

Fig. 2. Crime hotspot analysis.

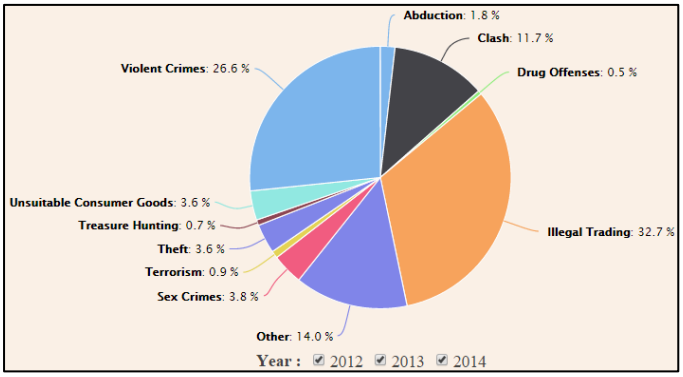

Fig. 3. Crime comparison.

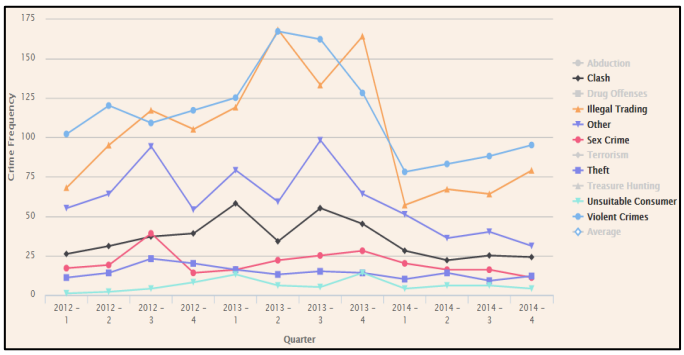

Fig. 4. Crime pattern visualization.

be incorporated with the document classifier module in order to improve the accuracy of the classification process. In addition the entity extraction module can be improved by incorporating more rules which will improve the accuracy and the comprehensiveness of the entity extraction process.

REFERENCES

[1] P. Chamikara, D. Yapa, R. Kodituwakku and J. Gunathilake, "SL-SecureNet : intelligent policing using data mining techniques," International Journal of Soft Computing and Engineering, vol. 2, no. 1, pp. 175-180, 2012.

[2] H. Chen, W. Chung, J. Xu, G. Wang , Y. Qin and M. Chau, "Crime data mining: a general framework and some examples," IEEE Explore-Computer, vol. 37, no. 4, pp. 50-56, 2004.

[3] *Crime Mapping and Reporting System* [Online]. Available: https://www.crimereports.com/

[4] *Intelligent Mapping System* [Online]. Available: http://maps.met.police.uk/

[5] R. Krishnamurthy and S. Kumar, "Survey of data mining techniques on crime data analysis," International Journal of Data Mining Techniques and Applications, vol. 1, no. 2, pp. 117-120, December 2012.

[6] S. Adhikari and K. Bogahawatte, "Intelligent criminal identification system," in The 8th International Conference on Computer Science & Education, Colombo, Sri Lanka, 2013, pp. 633-638.

[7] V. Nath, "Crime pattern detection using data mining," in Web Intelligence and Intelligent Agent Technology Workshops, Hong Kong, 2006, pp. 41-44.

[8] *International Press Telecommunications Council* [Online]. Available: http://www.iptc.org/site/Home/

[9] E. Brown, "The regional crime analysis program: a framework for mining data to catch criminals," in IEEE International Conference on Systems, Man, and Cybernetics, San Diego, 1998, pp. 2848-2853.

[10] *Web Crawler* [Online]. Available: http://www.sciencedaily.com /articles/w/web_crawler.htm

[11] G. Pant, I Srinivasan and F. Menczer, "Crawling the Web," in Web Dynamics.: Springer Science & Business Media, 2004, pp. 153-178.

[12] W. Zhang, T. Yoshida and X. Tang, "Text classification based on multi-word with support vector machine," Knowledge-Based Systems, vol. 21, no. 8, pp. 879-886, 2008.

[13] K. Cheng, S. Pan and F. Kurfess, "Ontology-based semantic classification of unstructured documents," Adaptive Multimedia Retrieval in Computer Science, vol. 3094, pp. 120-131, 2004.

[14] K. Mertsalov and M. McCreary, "Document classification with support vector machines," 2009.

[15] N. Guarino, D. Oberle and S. Staab, "What is an ontology?," in Handbook on ontologies, R. Studer S. Staab, Ed.: Springer Berlin Heidelberg, 2009, pp. 1-17.

[16] G. Millar, "WordNet: a lexical database," Communications of the ACM, vol. 38, no. 11, pp. 39-41, 1995.

[17] M. Choi, "A selective sampling method for imbalanced data learning on support vector machines,", 2010.

[18] R. Batuwita and V. Palade, "Classs imbalance learning methods for support vector machines," in Imbalanced Learning: Foundations, Algorithms, and Applications, Y. Ma H. He, Ed.: Wiley-IEEE Press, 2012, ch. 6.

[19] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Linguisticae Investigationes, 2007.

[20] P. Osenova and S. Kolkovska, "Combining the named-entity recognition task and NP chunking strategy for robust pre-processing," 2002.

[21] S. Charikar, "Similarity estimation techniques from rounding algorithms," in ACM symposium on Theory of computing, New York, 2002, pp. 380-388.

[22] P. Runeson, M. Alexandersson and O. Nyholm, "Detection of duplicate defect reports using natural language processing," in Intenational Conference in Software Engineering, Minneapolis, 2007, pp. 499-510.

[23] *Crawler4j* [Online]. Available: https://code.google.com/p/cra- wler4j/

[24] *JSoup* [Online]. Available: http://jsoup.org/

[25] *Weka* [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[26] *LIBSVM: A Library for Support Vector Machines* [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[27] C. Hsu, C. Chang and C. Lin, "A practical guide to support vector classification," 2010.

[28] W. Bowyer, O. Hall, P. Kegelmeyer and V. Chawla, "SMOTE: synthetic minority over-sampling technique," Artificial Intelligence Research, pp. 321-357, 2002.

[29] *GATE* [Online]. Available: https://gate.ac.uk/

[30] C. Sadowski and G. Levin, "SimHash: hash-based similarity detection," 2007.

[31] *Smhasher* [Online]. Available: https://code.google.com/p/smha-sher/wiki/MurmurHash

[32] *Hibernate ORM* [Online]. Available: http://hibernate.org/orm/

[33] *Commons Math: The Apache Commons Mathematics Library* [Online]. Available: http://commons.apache.org/proper/commons-math/

[34] *Machine Learning Tool Kit* [Online]. Available: https://githu-b.com/yinlou/mltk

[35] *HighCharts* [Online]. Available: http://www.highcharts.com/