Thamar University Journal of Natural & Applied Sciences ISSN: 2073-0764
2016 A(**6**)65-81 © 2016 Thamar University

A Proposed Model for Focused Crawling and Automatic Text Classification of Online Crime Web Pages

Muneer A. S. Hazaa[1], Fadl M. Ba-Alwi[2], Mohammed Albared[2] and Helmi Al-Salehi[1]

1 Faculty of Computer Science and Information Technology, Thamar University, Yemen
2 Faculty of Computer and Information Technology, Sana'a University, Yemen

ABSTRACT

With the exponential growth of textual information available from the Internet, there has been an emergent need to find relevant, in-time and in-depth knowledge about crime topic. The huge size of such data makes the process of retrieving and analyzing and use of the valuable information in such texts manually a very difficult task. In this paper, we attempt to address a challenging task i.e. a crawling and classification of crime-specific knowledge on the Web. To do that, a model for online crime text crawling and classification is introduced. First, a crime-specific web crawler is designed to collect web pages of crime topic from the news websites. In this crawler, a binary Naive Bayes classifier is used for filtering crime web pages from others. Second, a multi-classes classification model is applied to categorize the crime pages into their appropriate crime types. In both steps, several feature selection methods are applied to select the most important features. Finally, the model has been evaluated on manually labeled corpus and also on online real world data. The experimental results on manually labeled corpus indicate that Naive Bayes with mutual information and odd ratio feature selection methods can accurately distinguish crime web pages from others with an F1 measure of 0.99. In addition, the experimental results also show that the Naive Bayes classification models can accurately classify crime documents to their appropriate crime types with Macro-F1 measure of 0.87. Our results also on online real word data show that the focused crawler with two-level classification is very effective for gathering high-quality collections of crime Web documents and also for classifying them.

Keywords: Crime Data Mining, Web Mining, Focused Crawling, Classification

1. INTRODUCTION

As of today, the indexed web contains more than 50 billion web pages[1] and continues to grow at a rapid pace. With the rapid growth of the World Wide Web, the volume of the crime information that is available on the web is growing exponentially. Since there has been an explosion of media reports for different kind of crime news, this makes the process of analyzing and processing them manually a very difficult task. It is also widely known that general purpose search engines are not tailored at providing topic specific information (Samarawickrama and Jayaratne 2011). Finding relevant and in time information from these crime documents are crucial for many applications and can play a central role in improving crime-fighting capabilities, helping to enhance public safety and reducing future crime. Therefore, the huge amounts of crime news need to be organized in an effective way. One way of organizing this overwhelming amount of data is to gather these crime web pages from the Internet and classify them into their appropriate categories. This organized and classified data is essential to many information retrieval tasks such as constructing or expanding web directories (web hierarchies), improving search results, helping question answering systems and building domain-specific search engines. To gather such domain-specific web pages, domain- specific web crawler has to be developed to collect web pages from the Internet by choosing to gather only pages related to this domain. This type of web crawler does not need to gather every web page from the Internet. In fact, during the focused crawling process of a search engine, the crawler uses an automatic classification mechanism to determine whether the Web page being considered is ‘‘on the specific topic” or not(Özel 2011; Qi and Davison 2009).

Text classification or Web page classification is the task of classifying natural language documents or Web pages into a pre-defined set of categories from a predefined classes or topics. It has become one of the key methods for organizing online information. Many machine learning methods have been proposed for text classification in the previous years such as N-gram (Farhoodi et al. 2011; Suzuki et al. 2010), Naïve Bayes (Chen et al. 2009; Fan et al. 2001; Metsis et al. 2006), Nearest Neighbors (Sun and Lim 2001), decision tree (Li et al. 2011),and support vector machine (Joachims 2001).

Our contribution: In this paper, we describe a new model for online crime text crawling and classification which seeks, acquires, maintains and classifies pages on crime topic. This model consists mainly from two main modules: a crime-specific crawling system and a text classification system. The crime-specific crawler is used to collect as many crime web pages as possible from the news websites and avoid irrelevant ones. This focused crawler is guided by a binary supervised classifier (crime filter) which learns to recognize the relevance of a web page with respect to the crime topic and it is also utilized a set of domain specific keywords. The text classification system is a multi-class text classification model which has been applied to categorize the crime Web pages into their appropriate crime sub-classes. In addition, we also compare between the two-level classification approach and the flat classification approach in the context of crime Web pages categorization.

This paper is organized as follows: In Section 2, we give a summary of related works in focused crawling, Web page classification and Crime data mining. Section 3 describes our

[1] http://www.worldwidewebsize.com/

model for crime Web pages classification. Section 4 describes the evaluation methods. The data sets used in this study, the experimental results and discussion on the results are presented in Section 5. Finally, Section 6 concludes the study and gives some future work.

2. RELATED WORK

Focused crawling is a promising approach to improving the recall of expert search on the Web. A variety of methods for focused crawling have been proposed (Batsakis et al. 2009; Can and Baykal 2007; Ehrig and Maedche 2003; Hsu and Wu 2006; Liu et al. 2006; Samarawickrama and Jayaratne 2011; Wang et al. 2010; Yang 2010a, b; Zheng et al. 2008). The term focused crawler was first coined by Chakrabarti in 1999. Chakrabarti (1998) uses a canonical topic taxonomy and seed documents to build a model for classification of retrieved pages into categories. Earliest work on focused crawling dealt with simple key-word matching or regular expression matching. Related research in focused web crawling algorithms is presented in (Liu and Lu 2007; Novak 2004). Topic specific crawlers attempt to focus the crawling process on pages relevant to the topic. They try to keep the overall number of downloaded web pages for processing as minimum as possible and maximizing the percentage of relevant pages (Batsakis et al. 2009). Angkawattanawit (2002) deal with improving recrawling performance by utilizing several databases (seed URLs, topic keywords and URL relevance predictors) that are built from previous crawl process and used to improve harvest rate. Several works (Martin and Khelif 2011; Yang 2010b) utilize search engines as a source of seed URLs and back–references.

Web classification is considered to be an important and challenging task(Özel 2011). It has attracted more and more research work in recent years(HUSSAIN and ASGHAR 2012). In the detailed survey of Qi and Davison (2009) , the Web page classification problem has been divided into many problems such as subject classification, functional classification, genre classification, binary/multi-class classification, single/multi-label classification, and hard/soft classification. Due to domain diversity and complexity, there remain many problems not solved.

In the recent decade, several studies have been performed on crime data mining. The results are usually used in developing new software applications for detecting and analyzing crime data. Oatley et al. (2005) introduce a general overview on applying intelligent crime analysis methods including neural networks, Bayesian networks, and genetic algorithms in predicting and matching crime incidents. Adderley (2007) have applied neural networks for crime data clustering and crime data classification through using both supervised and unsupervised learning methods. Keyvanpour et al. (2011) applied a SOM clustering method in the scope of crime analysis and then they use the clustering results to identify crime matching patterns. The COPLINK project (Atabakhsh et al. 2001; Chung et al. 2005; Hauck et al. 2002; Hauk and Chen 1999) represents a prominent framework for text mining, classification and clustering of crime data aiming. Nath (2006) and Phillips and Lee (2009) used clustering algorithms to detect the crimes patterns and speed up the process of solving crime. (Ghosh et al. 2016) used a Machine Learning approach to automate and help crime analysts identify the connected entities and events by collecting, integrating and analyzing diverse data sources to generate alerts and predictions for new knowledge and insights that lead to better decision making and optimized actions.. (Sharef and Martin 2015) introduces the evolving fuzzy grammar (EFG) method for crime texts categorization. The learning model is built based on a set of selected text fragments which are then transformed into their underlying structure called fuzzy grammars.

3. ONLINE CRIME TEXT CLASSIFICATION

The objective of our proposed task is to classify crime-specific web pages to help the user to find relevant, in time and in-depth knowledge about crime topic. In this study, we use a two stage classification approach to online crime text classification. In this scenario, web pages are first crawled and classified to crime or irrelevant class. Then, web pages which are classified as crime, as shown in (Figure 1), are passed to a second level of classification (multiclass classification system) which then classifies this crime information to their classes. Before execution of any of the classification level in the crime web classification, the features should be extracted and expressed at first.

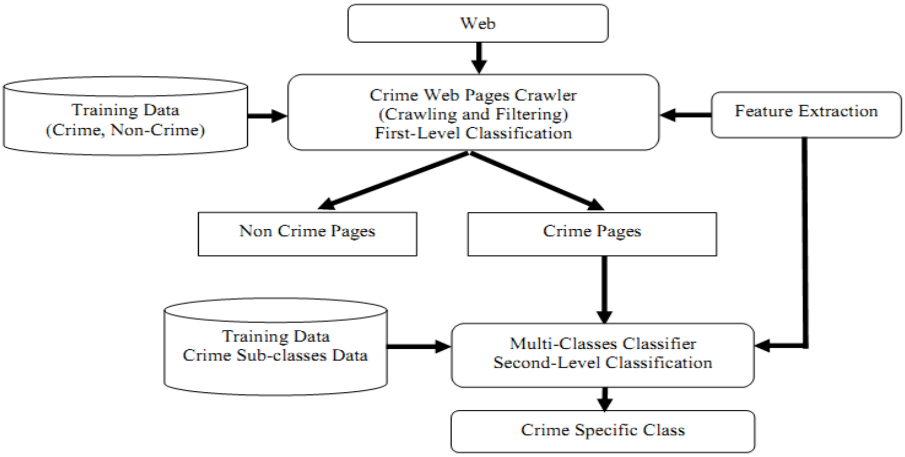


Figure 1: The Two Level Crime Web Pages Classification System.

3.1 Architecture of the Focused Crawler

As the size of the Web grows, topic-specific Web Crawlers are becoming more important due to their ability to acquire, and maintain a collection of Web pages relevant to a certain topic .The success of such topic-specific search tools depends on their ability to locate topic-specific pages on the Web while using limited resources.

In this section, we describe the architecture of our crime-specific crawler which represent the first level. (Figure 2) shows the architecture of our crime-topic crawler. The proposed crawler has the following components: 1) a URL Ranking algorithm which determines the relevancy of each URL in the starting URLs to the crime topic. 2) A URL Queue which contains ordered URLs obtained from search engine's starting URLs and from crawled pages. 3) A preprocessor module which is used for parsing the web page to remove stop words and the entire HTML tags. 4) A Page Downloader which download pages from WWW. 5) A binary classifier which makes relevance judgments on pages crawled and filters crime pages from non crime pages.

The detailed process of our crime-topic crawler as suggested in Figure 3. In Algorithm 1, the first step is to determine the starting URLs or the starting point of a crawling process. The crawler is unable to traverse the Internet without starting URLs. Moreover, the crawler cannot discover more relevant web pages if starting URLs are not good enough to lead to target web pages. In this step, the crawler sends crime keywords to a search engine 2 in order to build an initial set of seed URLs. The search engine returns a set of URLs with their titles and description texts, the crawler will use both titles and description texts to compute topic similarity scores to order the URLs in that set. The topic similarity score (rank) of each URL is easily obtained by computing the similarity between the profile of the URL title and description text and the profile of the crime class that were calculated from a manually crawled crime corpus. In this work, we use a vector matching operation, based on the cosine similarity (in Eq. (1)), to compute URLs similarity scores and to rank them.

$$score(URL_i) = sim(d_i, q) = \frac{\vec{d}_i \cdot \vec{q}}{|\vec{d}_i| \cdot |\vec{q}|} = \frac{\sum_{j=1}^{n} w_{j,i} \times w_{j,q}}{\sqrt{\sum_{j=1}^{n} w_{j,i}^2} \times \sqrt{\sum_{j=1}^{n} w_{j,q}^2}} \qquad (1)$$

where q is a crime corpus, d_i is the title and the description text of the ith URL in the starting URLs, q is the profile of the joined crime documents in the manually crawled corpus, and $w_{j,i}$ stands for the weight of the term w_i in document j.

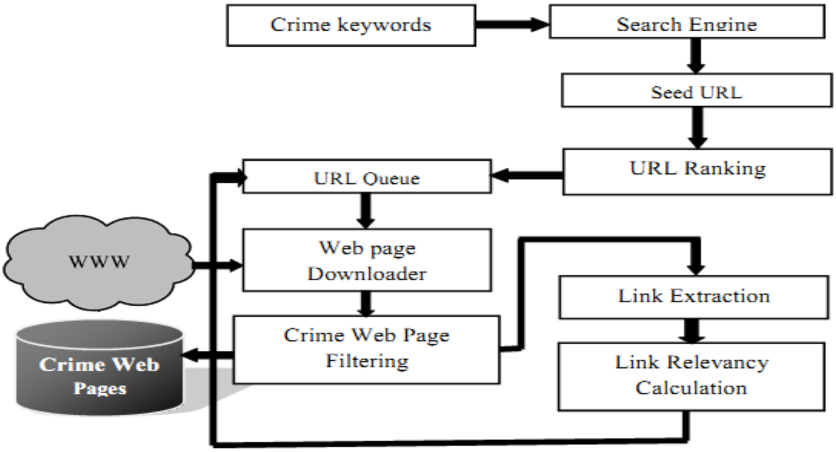


Figure 2: The Architecture of the Crime-Topic Crawler.

[2] www.Bing.com

Starting URLs will then be put into a URL queue and sorted by their scores. Then, the crime crawler selects the highest score URL from the URL Queue. A page downloader downloads the page associated with this URL. Then, a preprocessing module is used for parsing the web page, performing stop word removal, and removing all the HTML tags. After that, a crime filtering (binary classifier) is used determine whether the downloaded web page is crime page or not. If the web page is classified as crime web page, the algorithm extracts all the new URLs from this page. For each new URL, its relevancy (priority) to crime is deduced by merging its page score and the similarity score of the crime corpus with the following three values together(the link information); the anchor text associated with the link, the heading of the section where the link is found and the text of the paragraph containing the link. The priority of link i in the crawled page p is depend on two values; the similarity of the page to the crime corpus and the similarity of the link information to crime corpus:

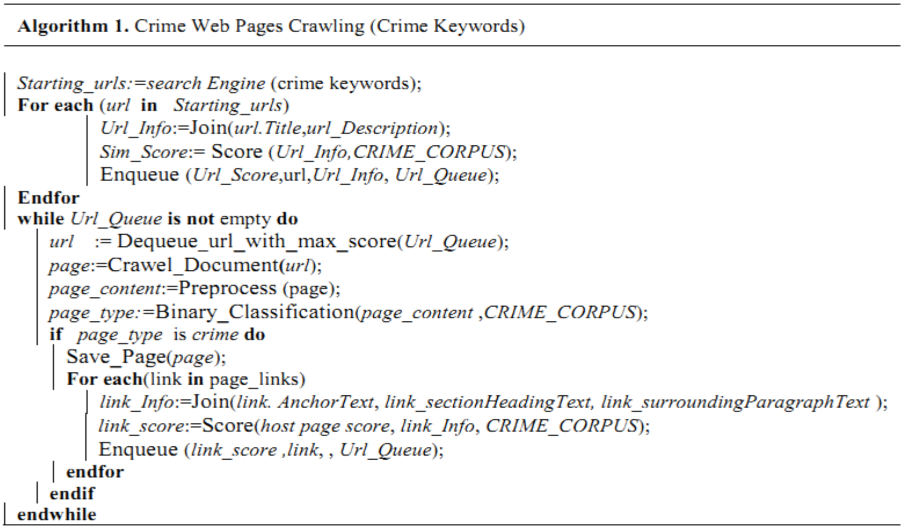
$$Priority\,(link_i) = \lambda\ score(p) + (1-\lambda)score(link\ \_\inf ormation) \qquad (2)$$

New extracted URLs will be ordered by their scores and then be put into a URL queue.

3.2 Crime Classification Phase

The objective of phase is to classify crime-specific web pages to their crime topics: Arson, Drugs, Fraud, Kidnap, Money Laundering, Murder, Sexual Crime, Theft, and Traffic Violation. To do this, we use a two level classification model to online crime text classification. As shown in (Figure 2), the major function of the crime web pages filtering module (first-level classification) is to define whether a web page has crime information or not. For those web pages, which have been classified as crime web page in the first level, there is a further classification (second level classification) which will be conducted for the crime subtype classification. The second level of classification is a multi-classes classification which classifies each crime web page to their specific crime topic.

In this study, we used the Naive Bayes (NB) classifier which is used in both level of crime text categorization due to their simplicity and they are very effective in text categorization (Feng et al. 2015; Metsis et al. 2006; Tang et al. 2016). The main advantages of Naive Bayes classifiers is that they are easy to implement, they have a linear computational complexity, and their accuracy especially in filtering, which is comparable to that of more elaborate learning algorithms (Metsis et al. 2006; Nath 2006). For more details about NB classifier, see (Chen et al. 2009; Schneider 2003).



Algorithm 1. Crime Web Pages Crawling (Crime Keywords)

```
Starting_urls:=search Engine (crime keywords);
For each (url in Starting_urls)
        Url_Info:=Join(url.Title,url_Description);
        Sim_Score:= Score (Url_Info,CRIME_CORPUS);
        Enqueue (Url_Score,url,Url_Info, Url_Queue);
Endfor
while Url_Queue is not empty do
    url  := Dequeue_url_with_max_score(Url_Queue);
    page:=Crawel_Document(url);
    page_content:=Preprocess (page);
    page_type:=Binary_Classification(page_content ,CRIME_CORPUS);
    if page_type is crime do
        Save_Page(page);
        For each(link in page_links)
            link_Info:=Join(link. AnchorText, link_sectionHeadingText, link_surroundingParagraphText );
            link_score:=Score(host page score, link_Info, CRIME_CORPUS);
            Enqueue (link_score ,link, , Url_Queue);
        endfor
    endif
endwhile
```

Figure 3: The Crime-Topic Crawler. Algorithm

3.3 Feature Selection

Feature selection is a process which selects a subset of features that is considered as important. Such selection can help in building faster, cost effective and accurate models for data processing. The process typically involves certain metrics that are used for finding utilities or importance level of features. However, in this work, in the first level of classification, the feature selection methods extract the global feature vector to optimize the work of crime filter. The global feature vector is extracted from all the crime documents. In the second level of classification, these feature selection methods extract and construct feature vectors of each crime type. However, we have used the following feature selection methods in both classification levels to calculate the score of term (t) belonging to category c_i (Chen et al. 2009; Feng et al. 2015; Li et al. 2015) :

$$\chi^2(c,t) = \frac{N \times (\text{AD-BC})}{(A+C)(B+C)(A+B)(C+D)} \tag{3}$$

$$MI(c,t) = \log_2\left(\frac{A \times N}{(A+C)\times(A+B)}\right) \tag{4}$$

$$OddR(c,t) = \frac{AD}{CB} \tag{5}$$

$$\text{GSS}(c,t) = \frac{(\text{AD-BC})}{N^2} \tag{6}$$

where ***A*** is the number of times *t* and *c* co-occur, ***B*** is the number of times *t* occurs without *c*, ***C*** is the number of times *c* occurs without *t*, ***D*** is the number of times neither *c* nor *t* occurs, and ***N*** is the total number of documents

4. EVALUATION METHODS

In order to evaluate the components of our model, several experiments have been conducted. Fist, we have evaluated the performance of the classification modules in both levels i.e. the binary NB classification algorithms and the multi-label classification modules. We have measured the performance of these classification algorithms on manually labeled crime data sets. For binary classification task, we use the following metrics:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \qquad (7)$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \qquad (8)$$

$$\text{F}_1 = \frac{2 * \text{Recall} * \text{Precision}}{(\text{Recall} + \text{Precision})} \qquad (9)$$

In the multi-label classification, the Macro-averaged (Macro-F1) (Forman 2003) is used. The macro-averaged F-measure is the traditional arithmetic mean of the F-measure computed for each problem.

$$F_1^{\text{macro}} = \frac{1}{m}\sum_{i=1}^{m} F_1(i) \qquad (9)$$

In addition, we also evaluate and compare both the two-level classification approach and the flat classification approach in the context of crime text categorization. In this context, we use the accuracy measure (Sun and Lim 2001) denoted by Ac_i for category C_i :

$$Ac_i = \frac{\text{TP}_i + TN_i}{\text{TP}_i + TN_i + FP_i + FN_i} \qquad (10)$$

Secondly, we empirically evaluate the effectiveness of our model through evaluating the components of the model on online real world data. The crawler performance is typically measured by the harvest rate i.e. the percentage of downloaded pages that are relevant to the crime topic.

$$Harvest\_rate = \frac{relevant\_pages}{pages\_downloaded} \qquad (10)$$

5. EXPERIMENTAL RESULT

5.1 First-Level Classification Experiment

In this experiment, our corpus consists of 2179 crime documents and 2257 non-crime documents, collected from Malaysian news web pages, is used to train the classifier. A test set consists of 472 files are used to test the classifier. The feature selection methods, that have been discussed, have been implemented to reduce the size of each feature set under each category.

For testing the effectiveness of the binary classification models phase, we have investigated the performance of the NB classifier according to each feature selection method (MI, CHI, GSS and OddR) by varying the size of the top rated features. These features are selected from feature space at different size 500, 750, 1000, 1500, 1750 and 2000.

However, we first examined the impact of the number of features on the effectiveness of the NB classifier with the feature selection methods (MI, CHI, GSS and OddR). As shown in (Figure 4), the NB-MI and NB-GSS achieve important improvement, when the size of features is 2000, of approximately 1.3% over their results, when the size of features is 500. The best performances in term of Macro-F1 achieved by NB-MI and NB-GSS are 0.99 for both when the number of terms is 2000. From these results, we can conclude that Naïve Byes classifiers are highly accurate crime filtering models.

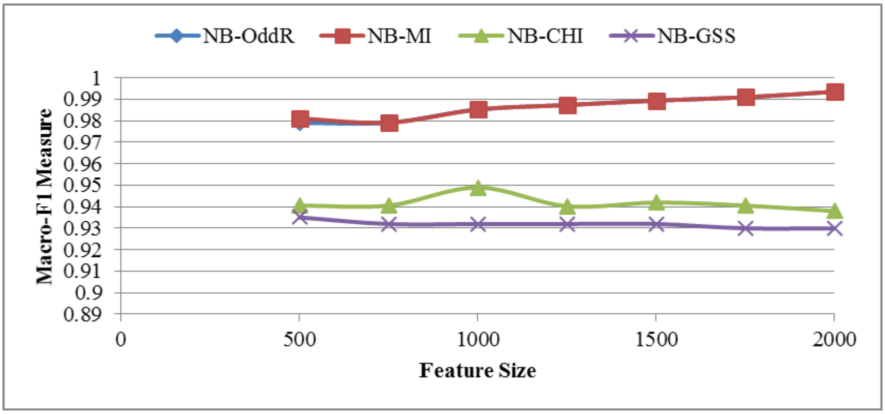


Figure 4: Macro-F1 Values of the Feature Selection Methods with NB Classifier (Crime Text Filtering)

5.2 Second-Level Classification Experiment

In the second experiment, our training corpus which consists of 2179 crime documents is used. The data in the corpus is divided into nine crime categories i.e. Traffic Violation, Theft, Sex Crime, Money laundering, Murder, Kidnap, Fraud, Drugs, and Arson (Chen et al., 2004). A test set which consists of 532 files is used to test the classifiers. The feature selection methods that have been discussed have been implemented to reduce the size of the feature set under each category. For testing the effectiveness of these classification models, we investigated the effectiveness of the four NB models namely, NB-MI, NB-CHI, NB-OddR, and NB-GSS. From (Figure 5), we can see that in NB-MI gets the highest categorization performances. We can also see that NB-CHI model leads to the worst performance overall.

In addition, we have investigated the performance of the NB classifier with each feature selection method (MI, CHI, GSS and OddR) by varying the size of the top rated features. These features are selected from feature space at different size 500, 750, 1000, 1500, 1750 and 2000. Fig. 4 shows the macro-averaged F-measure for each of the feature selection metrics as we vary the number of features to select. The results of these experiments indicate that overall the best results are achieved with 2000 features, as one might have expected. As shown in (Figure 5), the performances of the classification models, NB-MI, NB-CHI, NB-OddR, and NB-GSS, are generally improved as the size of features used is increased. The best performance is obtained by NB-MI which is according to Macro-F1 is 0.84 when the number of features is 2000.

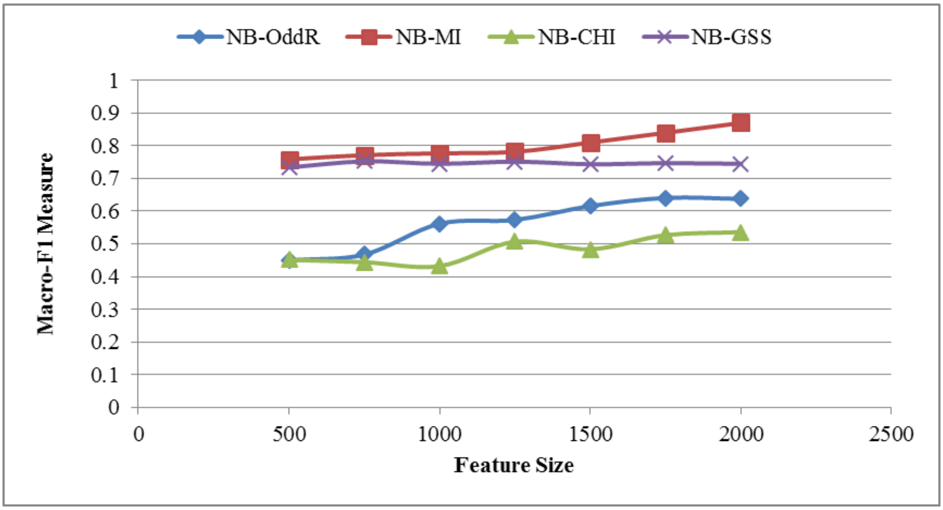


Figure 5: Macro-F1 values of the feature selection methods with NB Classifier (Crime Text Classification

5.3 Flat Classification VS Two Level Classification

In this experiment, we also evaluate both the two level (hierarchical) classification approach and the flat classification approach in the context of crime text categorization. The data set used in this experiment is classified to ten categories: traffic violation, theft, sex crime, money laundering, murder, kidnap, fraud, drugs, and arson and the others.

In the flat classification experiment, 2000 features have been extracted according to the type of the crime web pages, and each of them was assigned a weight based on their emerge frequency. The performance of classifying the crime dataset with a single NB-MI classifier is evaluated. The detailed results are shown in Table 1.

In the second experiment (two levels), first the same data set is separated into two sets: crime data set and non-crime data set. Then, we utilize the global feature vector. After that, the performance of classifying the dataset with first NB-MI classifier is evaluated. After the results has been produced. The documents which have been classified as crime documents by the first classifier have been used as the test set for the second level classifier.

The results of web page classification using the two-level classification model are list in Table 1. Comparing the two-level approach with one-level (flat) classification, the classification accuracy achieved by the two-level approach is 88% while the classification accuracy achieved by one-level approach is 83%. A 5% increase is gained. So it is obviously that the two-level classification is better than the one- level classification.

Table 1: Result of Crime Text Classification with One-Level and Two-Level Classifier

Quantity	Class	NB-MI			
		Hierarchal Classification Two Levels		Flat Classification One Level	
		Accuracy	Correct Classification	Accuracy	Correct Classification
Arson	41	1.00	41	0.83	34
Drugs	65	0.69	45	0.91	59
Fraud	56	0.88	49	0.82	46
Kidnap	71	0.94	67	0.94	67
Money Laundering	49	0.80	39	1.00	49
Murder	55	0.69	38	0.93	51
Sexual Crime	56	0.68	38	0.66	37
Theft	57	1.00	57	0.89	51
Traffic Violation	27	0.89	24	0.85	23
Non-crime	192	0.69	132	0.98	188
Average		**0.83**		**0.88**	

5.4 Crawler Evaluation

In this section, we present the evaluation of our crime specific crawler and classifier. At the initial stage, the crawler is initialized using a set of seed URLs for each keyword. These URLs are selected automatically from the fetched results of Bing search engine. All crawling and searching processes are limited to Malaysian news web pages. Then, the URLs are put into a priority queue according to the cosine similarity of their description and the crime corpus. After that, the crawler begins to download these web pages according to their URLs priority. Hyperlinks are extracted from the downloaded web pages and put into the priority queue.

The results for each method are represented by a plot showing the number of relevant pages returned by the method as a function of the total number of downloaded pages. The results of the crawler using both two levels classification and one-level classification approaches are shown in (Figure 6). The x-axis shows the total number of pages crawled. The y-axis shows the number of crawled crime web pages. As shown in (Figure 6), the total number of relevant crawled web pages of the crawler which uses two-level classification is higher than the total number of relevant crawled web pages of the crawler which uses flat classification. However, the harvest rate of both crawlers are high, all crawlers download large number of crime web pages. In the crawler with the flat classification model, the crawled web page is classified directly to irrelevant class or one of the crime classes. This means the multi-classes classification module is a part of the crawler. While in hierarchal crawling and classification, two-level classifier, the crawled web pages are first classified as crime page or irrelevant. Then, it is classified to its appropriate crime class using the second level classifier.

To analyze the data further, we studied the nature of these downloaded crime web pages and the percentage of each crime type during different stages of the crawling process. For each crime type, we calculated the number of relevant pages that belong to this type along the crawling process. (Figure 7) shows the size of each crime type from the crawled web pages. As shown in the (Figure 7), sexual crimes constitute the highest percentage of crimes followed by theft, murder….etc. Although these experiments are carried at certain time, the results may indicate the most frequent types of crimes in the community or it may indicate the types of crimes topics that attract the press more to write about them.

In general, it can be observed that all NB models are highly accurate crime filters. In addition, their performances are increased when the number of features is increased. The two-level NB classification approach has better performance than flat NB classification approach based on classification accuracy. In the crime web pages crawling and classifying task, the crime-focused crawling approach with two-level classification can crawl relevant crime web pages more effectively than its counterparts with flat classification throughout the whole crawling process. However, both approaches have high harvest rates.

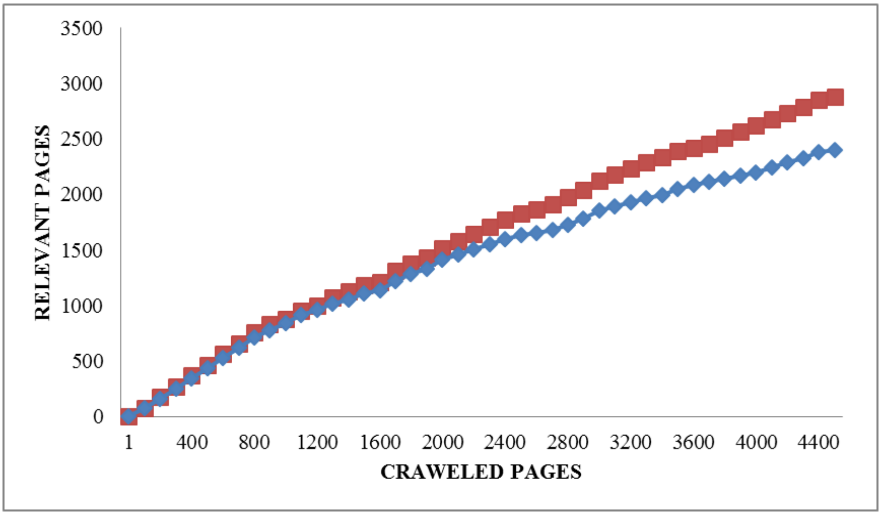


Figure 6: Harvest Rate of the Crime-Topic Crawler with Two-Level and Flat Classification methods

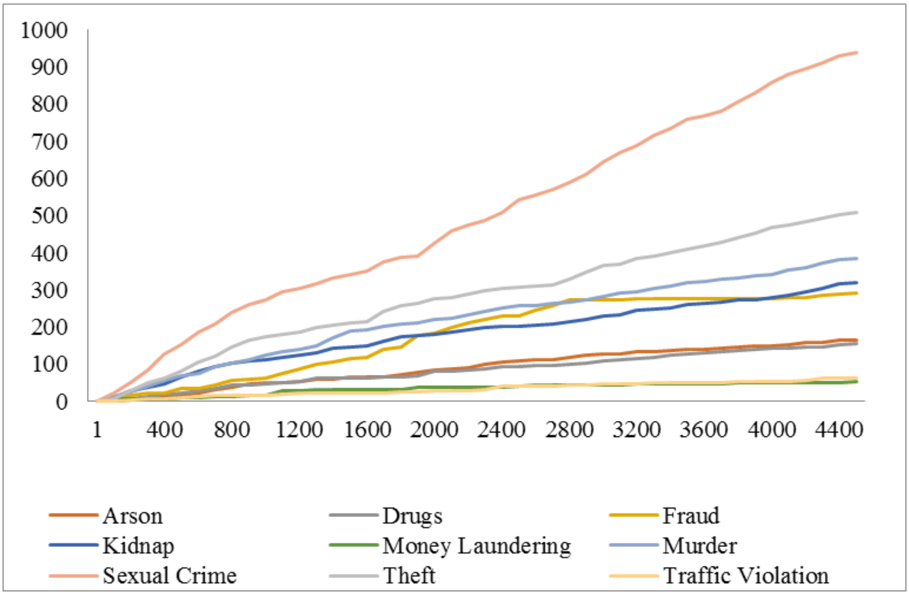


Figure 7: Harvest rates of the crawled pages for different crime classes (Percentage of Each Crime Type from the Crawled Pages)

6. CONCLUSION

This paper describes our work in crawling and classification the crime Web pages. In particular, we have proposed a new model for online crime text crawling and classification. The model consists of two levels. In the first level, crime-topic web pages is proposed to crawl web pages and to filter them to crime and non-crime documents using Naïve Bayes binary classification. In the second level, ,a multi-classes Naïve Bayes classification models are proposed to assign each crime documents to its appropriate crime type. The empirical study conducted using offline crime dataset has verified that the two-level proposed two level Naïve Bayes classification models are reliable classifiers for crime web pages. The empirical study also conducted using online Malaysian news web pages has verified that the proposed crawling and classification models creates high-quality web pages collections for each crime type.

In the future, we have identified several important directions for future research. We plan to will expand the multi-class classification methods into multi-label and multi-classes classification models in which a crime document can be assigned to more than one class. We also plan to integrate utilizes natural language processing technology and machine learning algorithms to analyze content, extracting useful information and to provide clear insight into the content of crime Web pages.

7. REFRENSE

Adderley, R. (2007). The use of data mining techniques in crime trend analysis and offender profiling

Atabakhsh, H., Schroeder, J., Chen, H., Chau, M., Xu, J.J., Zhang, J., & Bi, H. (2001). COPLINK knowledge management for law enforcement: text analysis, visualization and collaboration

Batsakis, S., Petrakis, E.G., & Milios, E. (2009). Improving the performance of focused web crawlers. *Data & Knowledge Engineering, 68,* 1001-1013

Can, A.B., & Baykal, N. (2007). MedicoPort: A medical search engine for all. *Computer methods and programs in biomedicine, 86*, 73-86

Chakrabarti, S., Dom, B., & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. In, *ACM SIGMOD Record* (pp. 307-318): ACM

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications, 36,* 5432-5435

Chung, W., Chen, H., Chaboya, L.G., O'Toole, C.D., & Atabakhsh, H. (2005). Evaluating event visualization: a usability study of COPLINK spatio-temporal visualizer. *International Journal of Human-Computer Studies, 62,* 127-157

Ehrig, M., & Maedche, A. (2003). Ontology-focused crawling of Web documents. In, *Proceedings of the 2003 ACM symposium on Applied computing* (pp. 1174-1178): ACM

Fan, Y., Zheng, C., Wang, Q., Cai, Q., & Liu, J. (2001). Using naive bayes to coordinate the classification of web pages. *Journal of software, 12,* 1386-1392

Farhoodi, M., Yari, A., & Sayah, A. (2011). N-gram based text classification for Persian newspaper corpus. In, *Digital Content, Multimedia Technology and its Applications (IDCTA), 2011 7th International Conference on* (pp. 55-59): IEEE

Feng, G., Guo, J., Jing, B.-Y., & Sun, T. (2015). Feature subset selection using naive Bayes for text classification. *Pattern Recognition Letters, 65,* 109-115

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research, 3,* 1289-1305

Ghosh, D., Ae Chun, S., Shafiq, B., & Adam, N.R. (2016). Big Data-based Smart City Platform: Real-Time Crime Analysis. In, *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research* (pp. 58-66): ACM

Hauck, R.V., Atabakhsb, H., Ongvasith, P., Gupta, H., & Chen, H. (2002). Using Coplink to analyze criminal-justice data. *Computer, 35,* 30-37

Hauk, R.V., & Chen, H. (1999). COPLINK: A case of intelligent analysis and knowledge management. In, *Proceedings of the 20th international conference on Information Systems* (pp. 15-28): Association for Information Systems

Hsu, C.-C., & Wu, F. (2006). Topic-specific crawling on the web with the measurements of the relevancy context graph. *Information Systems, 31*, 232-246

HUSSAIN, T., & ASGHAR, S. (2012). Web Mining: Approaches, Applications and Business Intelligence. *Ann Arbor MI, 25*

Joachims, T. (2001). A statistical learning learning model of text classification for support vector machines. In, *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 128-136): ACM

Keyvanpour, M.R., Javideh, M., & Ebrahimi, M.R. (2011). Detecting and investigating crime by means of data mining: A general crime matching framework. *Procedia Computer Science, 3*, 872-880

Li, B., Yan, Q., Xu, Z., & Wang, G. (2015). Weighted Document Frequency for feature selection in text classification. In, *2015 International Conference on Asian Language Processing (IALP)* (pp. 132-135): IEEE

Li, Y., Hung, E., & Chung, K. (2011). A subspace decision cluster classifier for text classification. *Expert Systems with Applications, 38*, 12475-12482

Liu, H., Janssen, J., & Milios, E. (2006). Using HMM to learn user browsing patterns for focused web crawling. *Data & Knowledge Engineering, 59*, 270-291

Liu, J.-H., & Lu, Y.-L. (2007). Survey on topic-focused Web crawler. *Application Research of Computers, 10*, 006

Martin, N., & Khelif, K. (2011). Focused crawling using name disambiguation on search engine results. In, *Intelligence and Security Informatics Conference (EISIC), 2011 European* (pp. 340-345): IEEE

Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). Spam filtering with naive bayes-which naive bayes? In, *CEAS* (pp. 27-28)

Nath, S.V. (2006). Crime pattern detection using data mining. In, *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on* (pp. 41-44): IEEE

Novak, B. (2004). A survey of focused web crawling algorithms

Oatley, G., Zeleznikow, J., & Ewart, B. (2005). Matching and predicting crimes. *Applications and Innovations in Intelligent Systems XII* (pp. 19-32): Springer

Özel, S.A. (2011). A Web page classification system based on a genetic algorithm using tagged-terms as features. *Expert Systems with Applications, 38*, 3407-3415

Phillips, P., & Lee, I. (2009). Mining top-k and bottom-k correlative crime patterns through graph representations. In, *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on* (pp. 25-30): IEEE

Qi, X., & Davison, B.D. (2009). Web page classification: Features and algorithms. *ACM Computing Surveys (CSUR), 41*, 12

Samarawickrama, S., & Jayaratne, L. (2011). Automatic text classification and focused crawling. In, *Digital Information Management (ICDIM), 2011 Sixth International Conference on* (pp. 143-148): IEEE

Schneider, K.-M. (2003). A comparison of event models for Naive Bayes anti-spam e-mail filtering. In, *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1* (pp. 307-314): Association for Computational Linguistics

Sharef, N.M., & Martin, T. (2015). Evolving fuzzy grammar for crime texts categorization. *Applied Soft Computing, 28*, 175-187

Sun, A., & Lim, E.-P. (2001). Hierarchical text classification and evaluation. In, *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on* (pp. 521-528): IEEE

Suzuki, M., Yamagishi, N., Tsai, Y.-C., Ishida, T., & Goto, M. (2010). English and taiwanese text categorization using n-gram based on vector space model. In, *Information Theory and its Applications (ISITA), 2010 International Symposium on* (pp. 106-111): IEEE

Tang, B., He, H., Baggenstoss, P.M., & Kay, S. (2016). A Bayesian classification approach using class-specific features for text categorization. *IEEE Transactions on Knowledge and Data Engineering, 28*, 1602-1606

Wang, W., Chen, X., Zou, Y., Wang, H., & Dai, Z. (2010). A focused crawler based on naive Bayes classifier. In, *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on* (pp. 517-521): IEEE

Yang, S.-Y. (2010a). A focused crawler with ontology-supported website models for information agents. *Advances in Grid and Pervasive Computing* (pp. 522-532): Springer

Yang, S.-Y. (2010b). OntoCrawler: A focused crawler with ontology-supported website models for information agents. *Expert Systems with Applications, 37*, 5381-5389

Zheng, H.-T., Kang, B.-Y., & Kim, H.-G. (2008). An ontology-based approach to learnable focused crawling. *Information Sciences, 178*, 4512-4522

نموذج مقترح لاسترجاع وفلترة
وتصنيف بيانات الجرائم اون لاين من صفحات الويب

منير عبد الله سعيد هزاع[1]، فضل مختار باعلوي[2]، محمد البارد[2]، حلمي الصالحي[1]

1 كلية علوم الحاسوب وتكنولوجيا المعلومات ،جامعه ذمار ، اليمن.
2 كلية الحاسبات وتكنولوجيا المعلومات،جامعه صنعاء،اليمن.

ملخص

مع النمو الهائل للمعلومات النصية المتاحة من خلال الإنترنت، تواجدت الحاجة الملحة لاستخلاص المعرفة في الوقت المناسب حول موضوع الجريمة. الحجم الضخم لهذه البيانات يجعل من عملية استرجاع وتحليل واستخدام المعلومات القيمة في هذه النصوص يدويا مهمة صعبة جدا. هذه الورقة البحثية تحاول معالجة مهمة صعبة ونقصد بها هنا استرجاع واستخلاص بيانات الجرائم على شبكة الانترنت. للقيام بذلك، قدمت هذه الورقة البحثية لاستنباط المعلومات المتعلقة بالجرائم وتصنيفها. أولا، تم تصميم زاحف عنكبوتي على شبكة الإنترنت متخصص في ايجاد واسترجاع بيانات الجرائم الموجودة على الشبكة من المواقع الإخبارية. في هذا الزاحف، يتم استخدام نموذج تصنيف ثنائي بتقنية بايز لفلترة صفحات الجرائم من الصفحات الاخرى. ثانيا، يتم تطبيق نموذج تصنيف الى فئات متعددة لتصنيف صفحات الجرائم وتحديد نوع الجريمة. في كل الخطوات، يتم تطبيق عدة طرق لاختيار أفضل الميزات الأكثر أهمية. وأخيرا، تم تقييم النموذج على بيانات معدة يدويا وأيضا على بيانات العالم الحقيقي على الانترنت. النتائج التجريبية تظهر أن نموذج تصنيف بايز يمكن أن تصنف بدقة بيانات الجرائم وتحدد نوع الجريمة المناسب بنسبة 87 في المائة. وتشير نتائجنا أيضا على بيانات العالم الحقيقي على الانترنت ان الزاحف الذي يحتوي على مستويين لفلترة وتصنيف صفحات الجرائم فعال جدا وله قدرة جيدة لاسترجاع وتصنيف بيانات الجرائم من صفحات الويب

كلمات البحث: تنقيب بيانات الجرائم، تنقيب صفحات الويب، تصنيف الجرائم.