Government Information Quarterly 31 (2014) 534–544

Contents lists available at ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

A decision support system: Automated crime report analysis and classification for e-government

Chih-Hao Ku [a], Gondy Leroy [b,c,*]

[a] College of Management, Lawrence Technological University, 21000 W 10 Mile Rd, Southfield, MI 48075, United States
[b] Department of Management Information Systems, University of Arizona, Tucson, AZ 85721, United States
[c] School of Information Systems & Technology, Claremont Graduate University, 150 E 10th St, Claremont, CA 91711, United States

ARTICLE INFO

Available online 6 October 2014

Keywords:
Natural language processing
Similarity measures
Classification
Algorithms
Measurement
E-government

ABSTRACT

This paper investigates how text analysis and classification techniques can be used to enhance e-government, typically law enforcement agencies' efficiency and effectiveness by analyzing text reports automatically and provide timely supporting information to decision makers. With an increasing number of anonymous crime reports being filed and digitized, it is generally difficult for crime analysts to process and analyze crime reports efficiently. Complicating the problem is that the information has not been filtered or guided in a detective-led interview resulting in much irrelevant information. We are developing a decision support system (DSS), combining natural language processing (NLP) techniques, similarity measures, and machine learning, i.e., a Naïve Bayes' classifier, to support crime analysis and classify which crime reports discuss the same and different crime. We report on an algorithm essential to the DSS and its evaluations. Two studies with small and big datasets were conducted to compare the system with a human expert's performance. The first study includes 10 sets of crime reports discussing 2 to 5 crimes. The highest algorithm accuracy was found by using binary logistic regression (89%) while Naive Bayes' classifier was only slightly lower (87%). The expert achieved still better performance (96%) when given sufficient time. The second study includes two datasets with 40 and 60 crime reports discussing 16 different types of crimes for each dataset. The results show that our system achieved the highest classification accuracy (94.82%), while the crime analyst's classification accuracy (93.74%) is slightly lower.

1. Introduction

Protecting citizens from the harm and violence is one of e-government's priorities. Linders (2012) exams the evolving citizen–government relationship and believes that an internet-based reporting platform is an efficient and convenient method to report crimes and enhances the interaction between community members and law enforcement agencies. Crime reports are critical information to investigators and have led to a number of criminals arrested, cases cleared, and property recovered. For example, Los Angeles Regional Crime Stoppers[1] alone has received more than 31,000 tips, which have led to 1404 arrests, 121 weapons recovered, and $584,129 property recovered, since its inception on March 22, 2012. Crime Stoppers, a program that encourages citizens to report crimes anonymously, has proven the importance of anonymous tips with approximately 470,000 arrests and 800,000 cases cleared over 32 years (Kanable, 2008).

Today, a variety of crime reporting channels have been offered by law enforcement agencies and non-profit organizations. For example, Short Message Service (SMS) messages (Song, Kim, Schulzrinne, Boni, & Armstrong, 2009), iPhone, iPad, and Android applications[2,3] and even online tips reporting systems such as FBI's online tips submission system[4] and Newark Police's Crime Stoppers Twitter[5] have been used to report anonymous crime tips. Anonymous reporting channels allow citizens to submit crime tips without revealing their identities. Further, law enforcement agencies can save time and resources spent on collecting citizen reports (Cartwright, 2008). Unfortunately, such anonymous tips may also result in more false and duplicate reports being filed (Eric, 2005), for example, adversaries who accuse each other falsely of crimes or neighbors who do not get along and report on each other's fictitious crimes or transgressions. In addition, the crime reports filed online and stored in databases are written in natural language. Such unstructured free text, when available in large quantities, requires processing and analysis before it can be made useful. To manually filter, compare, and contrast a large set of crime reports is time- and labor-intensive. More efficient solutions are needed.

* Corresponding author at: School of Information Systems & Technology, Claremont Graduate University, 150 E 10th St, Claremont, CA 91711, United States.
E-mail addresses: cku@ltu.edu (C.-H. Ku), gondyleroy@email.arizona.edu, Gondy.Leroy@cgu.edu (G. Leroy).

[1] Los Angeles Regional Crime Stoppers, http://lacrimestoppers.org/.

[2] iWatch Harris County, http://iwatchharriscounty.com/.
[3] Tomball Police Department, http://www.ci.tomball.tx.us/police/tip-android.html.
[4] FBI's Tips Submission, https://tips.fbi.gov/.
[5] Newark Police's Crime Stoppers Twitter, https://twitter.com/#!/1877NWKTIPS.

http://dx.doi.org/10.1016/j.giq.2014.08.003
0740-624X/

Government agencies are responsible for a well-time response to analyze the increasing digitized text information and databases. A DSS integrated with text mining and classification techniques, for instance, could help crime analysts investigate crimes and enable citizens to use e-government programs to check neighborhood crimes in a timely manner. In this study, we investigate the use of NLP techniques combined with similarity measures, and classification approaches to automate and facilitate crime analysis. Especially filtering reports and identifying those that report on the same or similar crime is a necessary task. Finding reports on the same crime can increase the information available to catch the suspects or improve prevention. Finding similar crimes is important for analyzing crime trends and gang activities and for allocation law enforcement resources.

Our approach uses similarity measures and classification approaches to find similar or same crimes in reports. We compare the algorithm's efficiency with a trained analyst. To verify our DSS in a realistic setting, we conducted a completely new experiment with small and big datasets that compared the impact of dealing with more crime reports and different types of crimes. We evaluated our algorithm and compared our system with the crime analyst's classification performance.

2. Literature review

2.1. E-government and crime-related applications

E-government refers to the effective use of information and communication technologies (ICT) to enhance government agencies' performance and accordingly improve government services and operations in the public sector (Kushchu & Kuscu, 2003). The communication between citizens and government agencies is mostly through telephone, face-to-face-meetings and even internet-based activities, e.g., email, digital form, and online chatting. Most of these communications are saved or transformed into written text and then archived in a digital format, which has led to opportunities for automatic text analysis using NLP techniques to improve e-government agencies' workflow (Knutsson, Sneiders, & Alfalahi, 2012).

Several applications for crime data analysis have been studied. Most efforts focus on crime pattern discovery, spatiotemporal crime analysis (Roth, Ross, Finch, Luo, & MacEachren, 2013), geospatial visualization (Chen et al., 2003; Elnahrawy, 2002; Wu, Cao, Wang, & Wang, 2010), and criminal link analysis (Li, Wang, & Leung, 2009). To discover crime patterns, Buczak and Gifford (2010) applied fuzzy association rule mining for community crime pattern discovery and found that, e.g., dense-housing communities (e.g., apartment complexes) with large number of non-English speakers and heavy use of public transit are likely to experience higher volumes of robberies. Using a co-occurrence analysis and heuristic approach, Schroeder, Xu, Chen, and Chau (2007) also conducted link analysis to associated crime relevant entities, e.g., addresses, telephone numbers, and type of crimes from structured crime incident reports from the Tucson Police Department. In contrast, Kovachev, Reichert, and Speck (2008) used geospatial visualization to visualize patterns from incident data published by a Berlin police department allowing users to identify crime hot spots and trends visually.

While others make use of structured information in databases, information from unstructured sources is often ignored. However, it provides additional, complementary, and useful information for crime analysts. The proposed study focuses on extracting and comparing information in unstructured records and developing a DSS which integrates information extraction, similarity, and classification algorithms to assist crime analysts to analyze crime reports and identify similarity between the reports.

2.2. Natural language processing

NLP is a field that intersects with artificial intelligence and linguistics. NLP techniques are frequently used to explore how computers can process and understand natural language text or speech (Chowdhury, 2003). Major NLP tasks in any system that includes processed text include tokenization, sentence splitting, part-of-speech (POS) tagging, phrase segmentation, information extraction, and named entity recognition (Soon, Ng, & Lim, 2001; Wang, Zhang, Xie, Anvik, & Sun, 2008). The first four tasks are low-level tasks used to identify words, phrases, and sentences and their structures and boundaries (described in Section 3.2), while the last two tasks, built upon low-level tasks, are high-level tasks used to extract relevant information in a domain (Nadkarni, Ohno-Machado, & Chapman, 2011).

Information extraction is a task used to automatically extract structured information such as vehicle, weapon, and type of crime from semi-structured and unstructured sources. For example, Pinheiro, Furtado, Pequeno, and Nogueira (2010) presented an IE framework and used NLP techniques to extract crime scenes and type of crimes from online texts and obtained 72%–87% precision and 68%–71% recall as a result. Both rule-based (Ananthanarayanan, Chenthamarakshan, Deshpande, & Krishnapuram, 2008; Jayram, Krishnamurthy, Raghavan, Vaithyanathan, & Zhu, 2006; Kozawa, Tohyama, Uchimoto, & Matsubara, 2008; Shen, Doan, Naughton, & Ramakrishnan, 2007) and statistical methods (Boiy & Moens, 2009; Guangpu, Xu, & Zhiyong, 2011; Haque, Dey, & Mahajan, 2009; Tatar & Cicekli, 2009) are widely used for IE; however, there is no clear winner (Sarawagi, 2007). When a large set of training data is available, statistical learning is preferred. When this is not available, a rule-based approach can be used.

Named entity recognition (Santos & Milidiú, 2012) is a subtask of IE used to recognize proper nouns such as location, organization, and personal name in text and to classify them into given categories. For example, Ananthanarayanan et al. (2008) used rule-based algorithms to extract named entities such as product and organization names from noisy text and achieved 61–85% precision and 42–73% recall.

A natural language text processing system is often comprised of several tasks described above and used to process a large amount of text. These tasks are usually modularized and combined in a *pipelined* system design (Nadkarni et al., 2011), so several different tasks can be performed in a single application. To achieve this, an NLP framework such as General Architecture for Text Engineering (GATE)[6] and Unstructured Information Management Architecture (UIMA)[7] can be used.

2.3. Similarity measures

Similarities are shared features between objects and concepts. Similarity measures are mathematical calculations to represent the degree of similarity between entities, and the sentences and documents they appear in. Similarity measures have been successfully applied in a number of domains such as text summarization (Aliguliyev, 2009; Wang, Li, Zhu, & Ding, 2008), plagiarism detection (Brixtel, Fontaine, Lesner, Bazin, & Robbes, 2010; Micol, Ferrández, & Muñoz, 2011), document clustering (Hatzivassiloglou, Gravano, & Maganti, 2000; Liu, Wang, & Liu, 2010), and even text classification (Lee & Chen, 2006; Liao & Jiang, 2005).

To identify similar documents, similarity measures such as the vector space model (Lakkaraju, Gauch, & Speretta, 2008), *Cosine*, *Dice*, and *Jaccard* (Runeson, Alexandersson, & Nyholm, 2007) are frequently used. In a vector space model, a document is represented as vectors of entities (also called Bag-of-Words) extracted from the document. The *Cosine* similarity measure is then used to calculate an angle between document vectors. To assign high weights to high frequency terms that appear in a small number of documents, term frequency–inverse document frequency (*tf–idf*) is a widely used weighting scheme (Lee, Chuang, & Seamons, 1997). To measure overlapping tokens and entities between sentences and document, *Jaccard* and *Dice* coefficient are commonly used. The difference between them is that *Dice* coefficient assigns a higher weight to overlapping items.

[6] GATE, http://gate.ac.uk/.
[7] UIMA, http://uima.apache.org/.

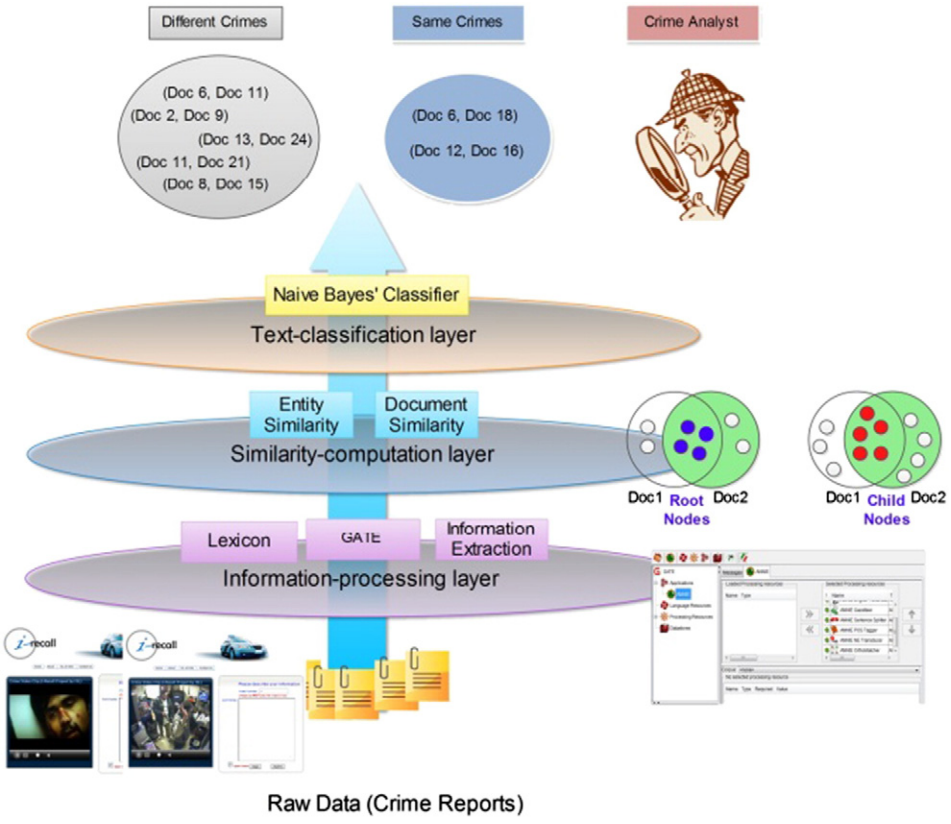

Fig. 1. Framework of decision support system.

To overcome a lack of semantic relationships between entities for vector-based approaches, lexical resources such as WordNet (Budanitsky & Hirst, 2006) and Wikipedia (Wang, Hu, Zeng, & Chen, 2009) can be included. Lexical resources are frequently included as a part of NLP component to provide semantic and syntactic information, typically for in-depth report analysis.

2.4. Text classification

Text classification is a process of labeling text documents into one or more pre-specified categories. Due to the vast majority of web pages, emails, social networking information, and even corporate information that is available in digital form, automatic text classification has gained attention in many research domains. For example, it has been used to filter spam (Delany, Buckley, & Greene, 2012; Lai, Chen, Laih, & Chen, 2009), categorize newspaper articles into topics (Mamakis, Malamos, & Ware, 2011; Suzuki & Hirasawa, 2007), and classify network intrusion attacks as Positive and Negative (David, 2008).

Current classification methods such as decision trees (Diao, Lu, & Wu, 2000; Vens, Struyf, Schietgat, Džeroski, & Blockeel, 2008), *k*-nearest neighbors (Li, Miao, & Wang, 2011; Tan, 2006; Wan, Lee, Rajkumar, & Isa, 2012), neural networks (Ghiassi, Olschimke, Moon, & Arnaudo, 2012; Rajan, Ramalingam, Ganesan, Palanivel, & Palaniappan, 2009), support vector machines (Li et al., 2011; Rajan et al., 2009; Wan et al., 2012), and Naïve Bayes (Bermejo, Gámez, & Puerta, 2011; Isa, Kallimani, & Lee, 2009; Chen, Huang, Tian, & Qu, 2009) have been successfully used in automated text classification. Among them, Naïve Bayes is a popular approach in text classification because of its simplicity, computational efficiency, and good performance (Chen et al., 2009). To evaluate the performance of different classification approaches, a domain expert's judgment is usually required to compare against the performance of classifiers (Ghiassi et al., 2012).

2.5. Decision support systems

A DSS is a computerized system that enhances decision-making processes to a given domain for decision makers. The DSS can be

Table 1
Components of information-processing layer.

Components	Description
Tokenizer	Segments the text into individual tokens such as words and punctuations.
Sentence splitter	Identifies boundaries of sentences in text reports.
POS tagger	Assigns parts of speech to each word such as noun, verb, and adjective.
Stemmer	Identifies the main part of tokens. For example, the stem of 'attacks' and 'attacked' is 'attack'.
Gazetteer	Each gazetteer is a list of words used to locate entities such as type of weapons and a suspect's age. The gazetteer lists have been collected and organized into a domain-specific, hierarchical lexicon. Our new lexicon in this study contains 20 semantic trees including 38,000+ words and phrases. Each tree has one root node and several levels of child nodes. Root and child nodes are the main classes and subclasses of the classification.
Ortho-matcher	Recognizes uppercase letters such as a brand of automobile 'Toyota'.
Noun phrase chunker	Extracts noun phrases such as 'a nine-inch knife' in text.
JAPE rules	JAPE rules have been developed to extract entities such as addresses, locations, and people's ages and names.
Information filtering	A process to remove stop words such as 'a', 'an', and 'the', remove duplicate entities, and keep relevant entities.

Table 2
Components of entity similarity algorithm.

Components	Description
Entity similarity algorithm	*Entity Similarity* $(c_1, c2) = \frac{sem(c1,c2)+Jaccard(c1,c2)}{2} \times \frac{(depth(c2)+depth(c2))/2}{\max depth(c)}$
Semantic similarity	Calculates the shortest distance between nodes in a tree. A shorter distance between nodes represents higher similarity. *Sem* (c_1, c_2) represents a semantic score between two concepts c_1 and c_2.
Jaccard coefficient	*Jaccard* coefficient is used to measure overlapping tokens between entities. *Jaccard* (c_1, c_2) represents a *Jaccard* coefficient between concepts c_1 and c_2.
Depth-of-node weighting	More specific information is weighted more heavily. *Max depth* (c) represents the maximum depth of a tree. *depth* (c_1) *and depth* (c_2) represent the depth of c_1 and c_2 in a tree.

classified into six categories: text-oriented DSS, database-oriented DSS, knowledge-oriented DSS, model-oriented DSS, communications-oriented DSS, and compound DSS as explained in Power (Power, 2004; Power, 2004) and Holsapple and Whinston (1996). A text-oriented DSS uses a computer-based system to process and analyze documents, e.g., product specifications, business documents, and crime reports for decision making, while a database-oriented DSS stresses how to access, manipulate, and analyze large databases of structured data, e.g., Online Analytical Processing (OLAP), Access, and Excel associated with Business Intelligence. A knowledge-oriented DSS focuses on the development of knowledge, heuristic rules, and problem-solving strategies associated with a specific domain, while a model-oriented DSS emphasizes access and manipulation of statistical, financial, and simulation models to help business managers find cost-effective solutions. A communications-oriented DSS uses communications and network techniques, e.g., groupware, video conferencing, and social networking tools, to support decision making, while a compound DSS uses two or more techniques described above to support decision making.

DSS have been popular applications in many different fields. For example, there are clinical DSS for medical and disease diagnosis (Graber & Mathew, 2007; Sanchez, Toro, Carrasco et al., 2011), financial analysis DSS for the assessment of a company's financial condition (Swiderski, Kurek, & Osowski, 2012), bankruptcy prediction (Olson, Delen, & Meng, 2012), and financial sequence prediction (Chan & Franklin, 2011), and the crime-associated DSS for fighting identity theft for online business and consumers (Lai, Li, & Hsieh, 2012) or for identifying matching of customers, vendors, and employees to support crime investigation.

2.6. Research question

The number of published, crime-related studies has focused on structured text mining (Piskorski et al., 2010), crime classification (Borg, Boldt, Lavesson, Melander, & Boeva, 2014), crime analysis and visualization (Kovachev et al., 2008), and even similarity between criminal investigations (Cocx & Kosters, 2006). However, many of these methods are not designed to explore large amounts of unstructured crime reports (Helbich, Hagenauer, Leitner, & Edwards, 2013) and not able to extract more information from text reports because of the lack of a domain-specific lexicon (Pinheiro et al., 2010). In addition, most similarity algorithms are not specially developed to compare and contrast crime reports, but general text reports. The aforementioned gaps had led the study to investigate the *general question*:

> Can a DSS that integrates the proposed domain-specific lexicon, algorithms with classification approaches and a human expert achieve high performance in classifying small and big datasets of crime reports?

Our interests lie in the development of a compound DSS (a text- and knowledge-oriented DSS) that extracts and uses information from unstructured text. This requires our domain-specific lexicon and information extraction and similarity algorithms. First, we learn how to integrate semantic weighting techniques and heuristics into our similarity algorithms. Our weighted algorithm combines a depth-of-nodes weighting approach and leverages a crime analyst's heuristic knowledge coded as rules. Then, we aim to learn how best to use this measure for identifying reports on the same crime using classification approaches. We compare Naïve Bayes' classifier, a machine learning approach, logistic regression, a statistical approach, and a human expert for crime report classification.

3. System development and design

3.1. Overview

To compare and contrast a large set of text reports, crime relevant information must be extracted and similarities between crime reports must be measured. To this end, we have developed algorithms that extract crime-associated entities such as crime scenes, people, weapons, vehicles, and types of crimes based on our hierarchical lexicon, measure document similarity between crime reports, and classify the reports into the same and different crime. We report here on our next generation document similarity algorithms. They differ from our previous version since they combine our original rule-based approach with a new depth-of-nodes and expert weighing approaches. In addition, a new, complete experiment that uses small and bigger, realistic datasets with more crime reports and several different crimes was conducted with a logistic regression and Naïve Bayes' approach for the crime report classification.

Table 3
Components of document similarity algorithm.

Components	Description
Document similarity algorithm	*Document Similarity* = (*average weighted entity similarity* + *root-node Dice Coefficient* + *Weighted child-node Dice Coefficient*) / 3.
Entity similarity	See Table 2.
Dice coefficient	Measures overlapping information for higher level concepts between documents. *Dice* coefficient between documents D1 and D2 is defined as: $\frac{2\lvert D1 \cap D2\rvert}{\lvert D1\rvert+\lvert D2\rvert}$ The numerator, $\lvert D1 \cap D2\rvert$, represents the nodes common to both documents while the denominator represents the combined number of nodes in both documents.
Weighted *Dice* coefficient	*Weighted Dice Coefficient* $(D1, D2) = \frac{2\lvert D1 \cap D2\rvert}{\lvert D1\rvert+\lvert D2\rvert} \times$ *Expert Weighting*
Expert weighting	Emphasizes the importance of specific information such as specific race and people's name for crime analysis. Table 4 shows an overview of the adjusted weights for such nodes in our hierarchical lexicon with higher importance. Expert weighting = Sum of weightings for all overlapping child nodes / Max weighting.

Table 4
Expert weighting — nodes with increased importance.

Weightings	Nodes
Triple weighting	• Specific race and ethnicity • Specific weight, height, and age • Specific people's names
Double weighting	• Specific brands and types of vehicles • Specific types of weapons • Type of crimes • Specific locations and stores • Specific electronic devices • Jewelry • Specific date and time
Equal weighting	• All other general information, e.g., general people, clothing, and personal belongings.

3.2. DSS design and development

We develop a three-layer DSS which is comprised of information-processing layer, similarity-computation layer, and text-classification layer (see Fig. 1).

3.2.1. Information-processing layer

We used and adapted several components from the GATE to process text reports and extract relevant information. The information-processing layer is comprised of eight components: tokenizer, sentence splitter, POS tagger, stemmer, gazetteer, ortho-matcher, noun phrase chunker, Java Annotations Pattern Engine (JAPE), and information filtering. The first four components are used to process words and sentences in text reports, while the next three components are used to extract relevant entities such as noun and verb phrases out of reports. The last component is used to remove stop words and duplicate entities. We have evaluated the extraction algorithm earlier with good results: 93–96% precision and 77–83% recall (Ku, Iriberri, & Leroy, 2008a, 2008b) so we limit this section to a short overview of the components as shown in Table 1.

3.2.2. Similarity-computation layer

The similarity-computation layer is comprised of two main components: the entity and document similarity algorithm. Our algorithms combine two most common similarity measures *Jaccard* (a component of the *entity similarity algorithm*) and *Dice* coefficient (a component of the *document similarity algorithm*) and the semantic similarity measure. We have evaluated both algorithms earlier with good results and found 87–92% classification accuracy (Ku & Leroy, 2011) for the document similarity algorithm to identify reports describing the same crime. Therefore, we limit this section to a short overview shown in Tables 2 and 3.

Our new generation similarity algorithms include two weighting approaches: the depth-of-nodes (a component of the *entity similarity algorithm*) and expert weighting (a component of the *document similarity algorithm*) shown in Table 4. The usefulness of a depth-of-nodes weighting to compute similarity scores has been demonstrated by others, but to our knowledge has never been tested for crime-related information. For example, Leacock and Chodorow (1998) used the depth in WordNet to compute semantic similarity. Sussna (October 2010) used a weighting approach, depth-relative scaling, for word sense disambiguation. Wu and Palmer (1994) used the depths of two concepts in a tree to compute conceptual similarity for verb translations between English and Chinese. In our earlier study, we compared our entity similarity algorithm with two WordNet-based similarity algorithms: Wu and Palmer and Leacock and Chodorow because both similarity measures also are based on a hierarchical lexicon to compute similarity scores and use depth-of-nodes in a tree to refine the similarity scores.

To compare different similarity measures, a gold standard for the comparison with human ratings was developed. The gold standard is based on similarity evaluations by fifteen human evaluators of word pairs (Ku & Leroy, 2011). The original gold standard contained 179 entities pairs. However, since pronouns are ignored by the new algorithm, we eliminated them from the gold standard resulting in 143 entity pairs. The reliability and agreement of ratings among fifteen raters were measured by Interclass Correlation Coefficient (ICC) analysis. The result shows that all Cronbach's Alpha values are higher than .9, which indicate that the rater scores are reliable. Our weighted entity similarity measure showed the strongest correlation ($r = .783$, $p < .001$) with 15 human raters (a gold standard), followed by the Leacock and Chodorow similarity measure ($r = .679$, $p < .001$), and the Wu & Palmer similarity measure ($r = .573$, $p < .001$).

3.2.3. Text-classification layer

To identify whether crime reports discuss the same crime, there are multiple approaches possible to classify crime reports ranging from completely manual to automated approaches. We opted for a semi-automated and automated approach. The text-classification layer includes two excellent candidates for making this classification: binary logistic regression (a semi-automated approach) and Naïve Bayes' classifier (an automated approach).

Logistic regression is a statistical technique that uses a logistic function to classify cases into categories. Binary logistic regression is commonly used to predict dichotomous results from predictor variables. In our system, the predicted variable is a function of the probability that a similarity score will be in one of two categories — the same crime or different crime. We use the SPSS binary logistic regression tool to conduct such classification and identify cutoff values by the author. A cutoff value can be used to distinguish between high and low similarity. If two reports describing the same crime, their content will be very similar which will result in a high similarity score. If two reports describe completely different crimes, their content will be different and similarity scores will be low. By training datasets, we can locate the best probability cutoff value that leads to the highest percentage of correct classification results.

Table 5
Small dataset — video set selection.

Datasets (N reports)	Number of different crimes	Video labels (N reports)	Average number of words in reports
Dataset 1A (10)	2	CV11 (5), CV01 (5)	111
Dataset 1B (10)	2	CV02 (5), CV12 (5)	77
Dataset 1C (10)	2	CV13 (5), CV03 (5)	67
Dataset 1D (10)	3	CV04 (3), CV14 (3), CV08 (4)	89
Dataset 1E (10)	3	CV09 (4), CV15 (3), CV05 (3)	104
Dataset 1F (10)	4	CV06 (2), CV04 (3), CV15 (3), CV16 (2)	86
Dataset 1G (10)	4	CV05 (3), CV07 (2), CV10 (3), CV17 (2)	71
Dataset 1H (10)	4	CV06 (3), CV08 (2), CV01 (2), CV16 (3)	107
Dataset 1I (10)	5	CV07 (2), CV09 (2), CV14 (2), CV11 (2), CV02 (2)	96
Dataset 1J (10)	5	5: CV10 (2), CV03 (2), CV12 (2), CV13 (2), CV017 (2)	68
Total (100)	34		87

Table 6
Big dataset — video set selection.

Datasets (N reports)	# of different crimes	Video labels (N reports)	Average number of words in reports
Dataset 2A (40)	16	CV01 (2), CV02 (2), CV03 (2), CV04 (3), CV05 (3), CV06 (2), CV07 (3), CV08 (3), CV09 (3), CV10 (3), CV12 (3), CV13 (2), CV14 (2), CV15 (2), CV16 (3), CV17 (2)	96
Dataset 2B (60)	16	CV01 (3), CV02 (3), CV03 (3), CV04 (4), CV05 (4), CV06 (4), CV07 (4), CV08 (4), CV09 (4), CV10 (4), CV12 (3), CV13 (4), CV14 (4), CV15 (4), CV16 (4), CV17 (3)	113
Total (100)			106

When a significant amount of training dataset is available, automated approaches also can be used to classify crime reports to minimize human involvement and training processes. Naïve Bayes' classifier is based on a probabilistic learning method. The WEKA framework, an open-source data mining toolbox, was used to train and apply the Naïve Bayes' classifier. We split crime reports into training and test datasets and the unused report pairs, 13,915 pairs, were used as the training dataset for Naïve Bayes' classifier.

4. Evaluation

The evaluation focuses on our weighted document similarity with different sizes of datasets and classification methods.

4.1. Weighted document similarity study design

Two studies were conducted with the small and big datasets where both crime analyst and system performance were compared. The first study used small datasets containing 10 crime reports and up to 5 different crimes per set, while the second study used two big datasets containing 40 and 60 crime reports respectively and 16 different crimes per set. The second study differs in the much larger number of crime reports and the number of different crimes in each dataset. This allows us to evaluate our approach in a realistic setting when the number of reports and crimes increases because a crime analyst's effort and time needed will increase, and accuracy may decrease. We expect that with an increasingly large dataset, the usefulness of our system will also increase.

4.2. Crime report collection

4.2.1. Small datasets

To build a collection of crime reports, we first recruited 40 volunteers (18 years of age or older) participated. None of the participants had a law enforcement background. Seventeen different video clips obtained from police training videos, surveillance systems shared on the internet, and commercial movies were used. The average length of a clip was 2 min, with the shortest 1 min and the longest 5 min. Each participant was shown 4 video clips and each video was viewed, in total, 10 times. Each participant was asked to write down what they had "witnessed." We used 100 reports out of 170 crime reports for the first study.

4.2.2. Big datasets

The majority of crime reports have been used in our earlier studies either training or testing our system. We then recruited 30 more participants from different schools for our second study. We used the same video clips described earlier for the crime report collection. Each video was viewed 7–8 times resulting in 120 crime reports for this study test bed. We used 100 reports out of 120 crime reports for the second study. Examples of crime reports are shown in Appendix A.

4.3. Methodology

4.3.1. Crime video labeling

Seventeen video clips (see Appendix B) were reviewed by two researchers. The video clips were classified into five types of crimes: burglary, robbery, assault, theft, and assault with three degrees of violence: low, moderate, and high based on the definition found on National Criminal Justice Reference Service,[8] Bureau of Justice Statistics,[9] and Federal Bureau of Investigation.[10] The video clips were collected from the police training video, commercial movie, surveillance video and even TV program.

4.3.2. Small datasets

We compiled 10 datasets, referred to from Dataset 1A to Dataset 1J. From our test bed with crime reports, we randomly selected 2–5 reports from each crime video clip for dataset (17 different crimes). Each dataset has 10 crime reports discussing 2 to 5 different crimes (each video clip shows a different crime). Table 5 shows an overview. The average number of words in the reports is 87 for the 10 datasets.

4.3.3. Big datasets

We compiled 2 datasets, referred to as Dataset 2A and Dataset 2B shown in Table 6. From our test bed with crime reports, we randomly selected 2–3 reports for Dataset 2A and 3–4 reports for Dataset 2B from each video clip (16 different crimes). Dataset 2A contained 40 and Dataset 2B had 60 crimes, each on the same 16 different crimes. The 16 crimes in each dataset are based on different videos. The average number of words in the reports is 96 and 113 for Datasets 2A and 2B respectively.

4.3.4. Independent variable

We compare binary logistic regression, Naive Bayes' classifier, and a human approach (crime analyst) to identify reports as discussion the same crime.

4.3.5. Measurement (dependent variables)

Receiver operating characteristic (ROC) curve and grouping accuracy were used for our first study, while grouping accuracy and the time spent were used for the second study. We included a ROC curve analysis to show the strengths and limitations of our algorithm.

We define accuracy as follows:

$$\text{Accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}.$$

In addition, additional qualitative measures for the crime analysis condition are subjective opinion measured with a short follow-up questionnaire, shown in Table 7 for our second study.

[8] Justice Reference Service, http://www.ci.farmington-hills.mi.us/.
[9] Bureau of Justice Statistics, http://www.bjs.gov/.
[10] Federal Bureau of Investigation, http://www.fbi.gov/.

Table 7
The user experience in grouping crime reports.

Experience

Q1: Which dataset was the most difficult for you to group similar crime reports together?

☐ Dataset A (40 documents) ☐ Dataset B (60 documents)

Q2: As the number of crime reports increased, how would you rate the ease of grouping crime reports?

☐ Very easy ☐ Quite easy ☐ Slightly easy ☐ Neither ☐ Slightly difficult ☐ Quite difficult ☐ Very difficult

Q3: As the number of crime reports increased, how would you rate your overall confidence in correctly grouping crime reports discussing the same crime?

☐ Very confident ☐ Quite confident ☐ Slightly confident ☐ Neither ☐ Slightly uncertain ☐ Quite uncertain ☐ Very uncertain

4.3.6. Procedure

Two studies were conducted on different dates with the same crime analyst. The study with smaller datasets had been first conducted. The second study with the big datasets was then conducted after 6 months. The crime analyst met with the researcher at his local Police Department in California and was shown the crime reports in a paper format only. Each dataset of crime reports was given each at a time.

First, a set of paper-based crime reports were given. The crime analyst was asked to group the reports that, according to him, discuss the same crime and label each group with type-of-crime information (see Fig. 2). The number of different crimes and the label information were not provided to the crime analyst. Upon completion of the first task, the crime analyst was given the next dataset. The same grouping process was required to complete the task for both studies. Each set of crime reports was also processed by the system and grouping accuracy was measured. The weighted document similarity shown in Table 3 was used.

5. Study results and discussion

5.1. ROC analysis

The ROC curve and the area under the ROC curve (AUC) are commonly used to visualize trade-offs between sensitivity (true positive rates) and 1 − specificity (false positive rates) for possible cutoff values. The closer the ROC curve is to the upper left corner, the better the performance. The AUC is used to measure accuracy. The higher the AUC value, the better the overall performance.

Fig. 3 shows the performance of the weighted document similarity algorithm. The ROC curve is far above the reference line (the worst case scenario) and close to the upper left corner (most accurate). The AUC of the weighted document similarity algorithm shows the approach to be significantly better than chance, with the AUC (AUC = .912, $p < 0.001$) significantly above 0.5.

The lower bound of AUC is .879 while the upper bound is .945 with asymptotic 95% confidence interval.

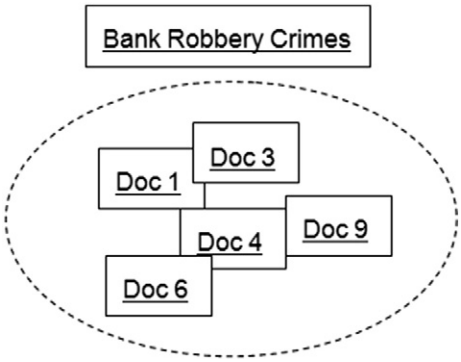

Fig. 2. Grouping crime reports with a type-of-crime label.

5.2. Accuracy analysis

5.2.1. Small datasets

Table 8 presents the detailed results for the 10 datasets for the binary logistic regression, Naïve Bayes' classifier, and the manual approach. The crime analyst achieved the highest accuracy 96%. The second best score, 88.89% accuracy, is obtained with the semi-automated classification approach (binary logistic regression) and a slightly lower accuracy 86.67% was obtained for the automated classification approach (Naïve Bayes' classifier). A closer look at individual datasets reveals that the accuracy was in some cases very high, e.g., above 95% in Dataset 1B with binary logistic regression and in Dataset 1D and 1H with Naïve Bayes' classifier, while sometimes as low as 73.33% accuracy in Dataset 1A (Naïve Bayes' classifier).

Table 9 shows the average accuracy based on the number of different crimes. When three different crimes were tested, the highest average accuracy (>91%) was found for all classification approach. Surprisingly, when only two different crimes were tested, the lower average accuracy (80%–94%) was found for three classification approaches.

5.2.2. Big datasets

To simplify the classification process and reduce the human involvement, Naïve Bayes' classifier was selected to process the big datasets (batch processing evaluation). Table 10 presents the results of the batch processing evaluation, including the time spent by the crime analyst and the accuracy of both Naïve Bayes' classifier. Surprisingly, the system scores with the Naïve Bayes' classifier achieved higher accuracy (94.82%) than the crime analyst's grouping accuracy (93.76%). For Dataset 2A with 40 crime reports, the system achieved the highest accuracy (95.26%), while the crime analyst's grouping accuracy

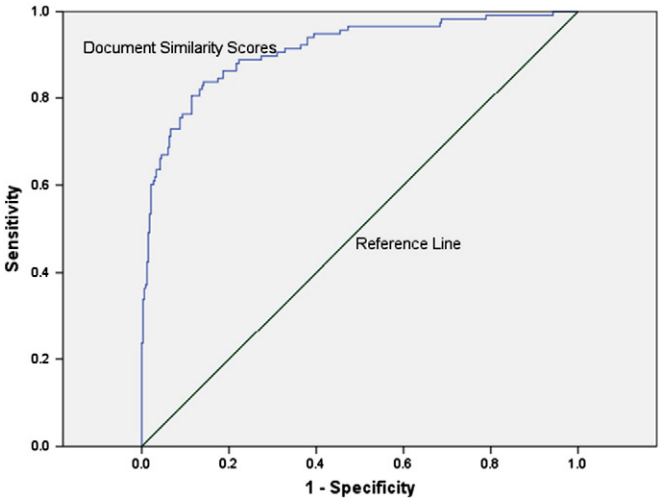

Fig. 3. The ROC curve for all 450 crime report pairs.

Table 8
Small datasets — accuracy for each dataset.

Datasets (N)	Nr. of different crimes	Binary logistic regression cutoff	Naïve Bayes' classifier accuracy	Crime analyst accuracy	Accuracy
Dataset 1A (10)	2	0.255	80.00%	73.33%	100%
Dataset 1B (10)	2	0.243	95.56%	91.11%	100%
Dataset 1C (10)	2	0.256	84.44%	77.78%	84%
Dataset 1D (10)	3	0.276	93.33%	95.56%	100%
Dataset 1E (10)	3	0.247	93.33%	86.67%	100%
Dataset 1F (10)	4	0.241	91.11%	86.67%	100%
Dataset 1G (10)	4	0.253	80.00%	82.22%	100%
Dataset 1H (10)	4	0.257	91.11%	95.56%	87%
Dataset 1I (10)	5	0.251	84.44%	88.89%	100%
Dataset 1J (10)	5	0.243	93.33%	88.89%	91%
Total (100)	34		88.89%	86.67%	96%

(94.74%) is slightly lower. For Dataset 2B with 60 crime reports, the system achieved the highest accuracy (94.63%), while the crime analyst's grouping accuracy (93.33%) is slightly lower. The crime analyst spent more time on Dataset 2A (87 min) than on Dataset 2B (62 min).

For the user experience, the crime analyst rated Dataset 2A as the most difficult to group crime reports describing the same crimes together, rated "slightly easy" for grouping the increasing crime reports, and rated "slightly confident" for grouping crime reports correctly.

6. Conclusion

Numerous law enforcement agencies have started using online crime reporting systems, providing more information but also leading to an ever-expanding digital crime reports and more time-consuming work in analyzing crime reports by crime analysts. The challenge of crime report analysis and classification lies no longer in similarity measures, but in a need of domain-specific lexicon and an appropriate weighting approach that is particularly tailored for automated crime analysis. In response, we developed a DSS that can identify crime reports on similar or the same crimes. The *document similarity algorithm* generated document similarity scores which were used in conjunction with an automated approach, a Naive Bayes' classifier, to identify reports discussing the same crime. Our algorithms were fine-tuned with the depth of the hierarchy and expert knowledge coded as rules to scale the similarity scores. This is to maximize the expertise of crime analysts without increasing the financial expenditure and labors. The evaluation results of high accuracy justify this approach.

6.1. Contributions

The DSS system contributes to both the current body of research, e-government and law enforcement. The contributions of this study are twofold.

6.1.1. Contributions to e-government and law enforcement

This paper presents how NLP techniques can be used to increase e-government agencies' efficiency (Knutsson et al., 2012) without sacrifice the quality of text report analysis. Automated text reports analysis is attractive because it can enhance and accelerate the discovery and analysis process and consequently shorten government agencies' response time to the public sector. A DSS can be useful in processing and analyzing crime information efficiently and generate decision support information necessary to solve crimes. Such DSS can alleviate information overload, reduce the time to search, process, and analyze crime information, and thus reduces crime analysts' workload and save the precious police resources to more important tasks, typically when budget cuts confronted by law enforcement agencies.

6.1.2. Technical contributions

The information extraction algorithm and domain-specific lexicon were developed to extract crime-related entities from unstructured text reports. The lexicon we constructed focuses on a single domain so we can ignore the words in other domains to reduce errors. The lexicon and extracted entities are reusable, since high precision and recall were achieved for the proposed algorithm. They can be used to measure entity, string, sentence, and even document similarity and to highlight words and phrases.

We provided an integrated solution including a domain-specific lexicon, NLP techniques, similarity measures specially tailored for crime analysis, and an automatic classification approach to automate the process of analyzing crime reports. For example, the experimental results show how the similarity scores can be combined with an automated classification approach, i.e., Naïve Bayes' classifier and achieved high accuracy when the number of crime reports and of different crimes significantly increase. Furthermore, the similarity scores generated by the algorithms we developed are reusable, since high classification accuracy was obtained for automated and semi-automated classification approaches. They can be used for crime report classification, clustering, and even information visualization.

6.2. Limitations

There are two major limitations to this work. First, the system might occasionally classify two reports describing different crimes as the same crime due to a high similarity score obtained. This is because there is always some overlap in reports, e.g., suspects of different crimes may steal similar cell phones or have similar hair styles. More advanced NLP, text mining, and classification techniques may be helpful in making this distinction. Second, several studies have been conducted, but only one expert has participated in our studies. To make the system more practical and useful to law enforcement agencies, the document similarity and classification algorithm can be augmented and fine-tuned by working

Table 9
Small datasets — average accuracy based on the number of different crimes.

Nr. of different crimes	Binary logistic regression average accuracy	Naïve Bayes' classifier average accuracy	Crime analyst average accuracy
2	86.87%	80.62%	94.81%
3	93.33%	91.11%	100.00%
4	86.96%	88.46%	95.56%
5	89.47%	88.89%	95.56%

Table 10
Big datasets — accuracy for each dataset and time spent by the crime analyst.

Datasets (N)	Time spent by the crime analyst (min)	Naïve Bayes' classifier accuracy	Crime analyst accuracy
Dataset 2A (40)	87	95.26%	94.74%
Dataset 2B (60)	62	94.63%	93.33%
Total (100)	149	94.82%	93.76%

with several more different people and experts and conducting more user studies.

6.3. Future directions

In the future, we plan to fine-tune the IE, similarity, and classification algorithms. For the IE algorithm, future work includes adding lexical resources such as Wikipedia and Urban Dictionary to enhance precision and recall in entity extraction. For similarity algorithms, future work includes testing more weighting schemes and datasets to enhance the performance of the system.

Acknowledgment

The authors would like to thank Matt Hopkins at the Fontana Police Department in California for serving as our expert.

Appendix*A. Examples of crime reports*

Doc 23 (Video 1)

A tanned skin Caucasian male, perhaps Hispanic, with dark brown messy hair, possibly dreadlocks, about 6 in. long, a thin mustache and 6 O'clock shadow, possibly a thin goatee, weighing about 150 lbs wearing a light blue/gray sleeveless shirt with white texture, and blue jeans attacked another Caucasian male, blond hair, 160 lbs, yellow shirt and blue jeans with a household appliance perhaps a lamp pole. The attack lasted for approximately 15 s with repeated hitting with the device on the victim.

Doc 33 (Video 1)

There are two guys (1st guy and 2nd guy) working on some sort of painting projects.

Their relationship seems to be friendly with each other. However one of the two guys (the 1st guy) seems to be bored of the work and he wanted to have some fun by turning on the music very loudly.

The 1st guy also happens to know another guy (the 3rd guy), who appears to have a negative relationship with the 2nd guy.

The 1st guy also knows that the 3rd guy is in the "house" and he actually tried to prevent the 2nd guy going around the house to meet the 3rd guy. The video didn't tell me what are the conflicts between the 2nd guy and the 3rd guy, but when they saw each other, they just happened to have a very hostile attitude toward each other. The 2nd guy acted in a very unpleasant way when he saw the 3rd guy and he questioned the 1st guy "why are you doing this?" The 3rd guy was also excited in a negative way when he saw the 2nd guy and for some reason, he decided to use some heavy metal to attack him. The 3rd guy attacked the 2nd guy so hard to cause him bleeding on the stomach and unable to get up from the floor. The 1st guy (here, I'm not sure if there was another person, because it seems that there was one or more person also present in the house) tried to stop the 3rd guy from hitting the 2nd guy, but it seems like he was too late for that. After the 1st guy stopped the 3rd guy, they escaped out of the house. Then, "somebody" called the police and they arrived at the scene.

Doc 18 (Video 3)

There are three people involved. They look like Indian. One person with gun in his hand was hitting a man by the gun.

The man was trying to fight back but he couldn't. Then a woman on the bed was crying.

She couldn't help the man. The criminal hit the man so hard and the mirror was broken.

The man fainted. Then the criminal went to the woman and wanted to rape her.

The woman was crying. The man over the mirror saw what was happening but he couldn't move himself.

The woman couldn't do anything. Finally the criminal raped the woman. The man over the mirror lost consciousness.

Doc 16 (Video 5)

A robber covered by a mask hands a stick or a long gun (?) trying to rob the staff in the counter.

Instead of giving out all the cash, the staff defends himself with an iron stick (no clue why he hides an iron stick at the counter)

Then, when he reports to the policemen, two more robbers appear from the parking lot described by the narrator, so they have surveillance system, but they don't have automatic system to report to 911, like a button underneath the counter? Or at least, he should close the door temporarily, not keeping on running business because the latter two robbers have guns, robbed all the cash, and ran away….

Doc 8 (Video 9)

A white (almost bald) guy (I think he's American) with dark blue sweater and jeans pants carrying pliers walked on the road. He just used pliers to cut a chain of a red bicycle and rode it away like nothing happened. There was a woman standing there but she didn't notice anything. The next case, he used the handsaw to cut the chain which took about 6 min. No one said anything. Then, he finished cutting the chain and rode it away. The next one, he used an electric cutter to cut the chain and it took about 20 s to finish cutting. The next one, he used a hammer and a nail to destroy the chain. A messenger seemed to worry about him so he suggested him use another tool to make it faster. All of the cases were in the public locations but no one tried to stop him because he acted so calmly and smoothly as if the bicycles he stole were his.

Doc 14 (Video 10)

The scene at the beginning has jewelry exchange store, security from the bank (maybe), and then the screen shows title — car thief and XXX. Then, a guy walks on the street and when he has eye contact with a lady randomly, the lady tries to avoid him and walks away quickly. I thought he is a car thief at the beginning, but after he sneaks into a BMW, which looks more like a clunker, not a BMW though. He doesn't start the engine.

Bust, he starts looking for something specific things in the car.

When he finds a used boarding pass and reads the information on it.

Soon, a guy shows up suddenly, grabs his collars and asks the first guy what did you do here and what did you find, so he doesn't think he is a car thief, but he knows something else going on there….

Doc 8 (Video 13)

There is a man in white Muslim outfit, sitting in the book store or the library. The room has many book shelves and plenty of books. There is another man sitting in the desk, wearing purple shirt, bold. When the bold man walk out of the room, the other man opens the drawer, takes out a laptop, put it

underneath his clothes, and stands up. He tries to walk away from the room.

Doc 16 (Video 14)

The couple got in an argument and then the guy smacked the girl. However, the girl got bloodied and fainted.

The guy thought that he killed the girl, because the girl would not wake up when he shouts at the her.

Then he felt guilty and started saying sorry and was not sure if he should touch or move the girl.

After that, he got frustrated and found his pistol on the top of the fridge.

He started to yell at the girl again, and there was no response at all.

The man sat down on the couch and decided to commit-suicide. After that the next scene shows that the girl actually woke up.

At the end the message of this video is to stop the domestic violence in family, especially beating women. 10 women die from this everyday.

Doc 5 (Video 17)

It's a jewelry store again. Two men walked in, seems like they came in to shop. There is another guy who walks in, he straight-up stabs the security guard who was sitting right inside next to the entrance. He stabs the guard numerous times but the guard wouldn't go down. The owner smashed a wooden chair in to the robber's head while the guard and the robber fight on the ground. The owner smashed a metal mop handle in to head this time after no effect on the first attack. He smashes until the mop handle bended. The robber tries to escape but too late, the cops are waiting for him outside. I don't know that this guy was going after the guard only when he was getting attacked by the owner.

B. Type of crimes

Video #	Type of crime	Degree of violence	Type of video
CV01	Burglary	Moderate	Commercial movie
CV02	Robbery	Moderate	Police training video
CV03	Assault	High	Commercial movie
CV04	Robbery	Moderate	Surveillance video
CV05	Robbery	Moderate	Surveillance video
CV06	Robbery	Moderate	Surveillance video
CV07	Robbery	Moderate	Surveillance video
CV08	Robbery	Low	Surveillance video
CV09	Theft	Low	TV program
CV10	Theft	Moderate	TV program
CV11	Theft	Moderate	TV Program
CV12	Theft	Low	Surveillance video
CV13	Theft	Low	Surveillance video
CV14	Assault	High	TV program
CV15	Assault	Moderate	TV program
CV16	Assault	Moderate	TV program
CV17	Assault	High	Surveillance video

References

Aliguliyev, R. M. (2009). A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Systems with Applications*, *36*(4), 7764–7772.

Ananthanarayanan, R., Chenthamarakshan, V., Deshpande, P.M., & Krishnapuram, R. (2008). Rule based synonyms for entity extraction from noisy text. *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data (AND '08)* (pp. 31–38). Singapore: ACM.

Bermejo, P., Gámez, J. A., & Puerta, J. M. (2011). Improving the performance of naive Bayes multinomial in E-mail foldering by introducing distribution-based balance of datasets. *Expert Systems with Applications*, *38*(3), 2072–2080.

Boiy, E., & Moens, M. -F. (2009). A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, *12*(5), 526–558.

Borg, A., Boldt, M., Lavesson, N., Melander, U., & Boeva, V. (2014). Detecting serial residential burglaries using clustering. *Expert Systems with Applications*, *41*(11), 5252–5266.

Brixtel, R., Fontaine, M., Lesner, B., Bazin, C., & Robbes, R. (2010). Language-independent clone detection applied to plagiarism detection. *SCAM '10 Proceedings of the 2010 10th IEEE Working Conference on Source Code Analysis and Manipulation* (pp. 77–86). Washington, DC, USA: IEEE Computer Society.

Buczak, A. L., & Gifford, C. M. (2010). *Fuzzy association rule mining for community crime pattern discovery*. Washington, D.C.: ACM, 1–10.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computer Linguistic*, *32*(1), 13–47.

Cartwright, A. (2008). Beyond the paper chase. *Law Enforcement Technology*, *35*(11), 58.

Chan, S. W. K., & Franklin, J. (2011). A text-based decision support system for financial sequence prediction. *Decision Support Systems*, *52*(1), 189–198.

Chen, H., Atabakhsh, H., Petersen, T., Schroeder, J., Buetow, T., Chaboya, L., et al. (2003). Coplink: Visualization for crime analysis. *Proceedings of the 2003 Annual National Conference on Digital Government Research* (pp. 1–6). Boston, MA: Digital Government Society of North America.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with naïve Bayes. *Expert Systems with Applications*, *36*(3), 5432–5435.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89.

Cocx, T., & Kosters, W. (2006). A distance measure for determining similarity between criminal investigations. *Advances in Data Mining* (pp. 511–525).

David, C. W. (2008). Rubee: Applying low-frequency technology for retail and medical uses. *Management Research News*, *31*, 549–554.

Delany, S. J., Buckley, M., & Greene, D. (2012). SMS spam filtering: Methods and data. *Expert Systems with Applications*, *39*(10), 9899–9908.

Diao, Y., Lu, H., & Wu, D. (2000). A comparative study of classification based personal e-mail filtering. In T. Terano, H. Liu, & A. Chen (Eds.), *PADKK '00 Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications, Kyoto, Japan* (pp. 408–419). Berlin, Heidelberg: Springer.

Elnahrawy, E. M. (2002). Log-based chat room monitoring using text categorization: A comparative study. *The International Conference on Information and Knowledge Sharing (IKS 2002)* (pp. 111–115). St. Thomas, Virgin Islands: ACTA Press.

Eric, S. (2005). Online crime reporting: Should law enforcement turn to the internet for savings? *Public Management*, *87*(6), 26–30.

Ghiassi, M., Olschimke, M., Moon, B., & Arnaudo, P. (2012). Automated text classification using a dynamic artificial neural network model. *Expert Systems with Applications*, *39*(12), 10967–10976.

Graber, M. L., & Mathew, A. (2007). Performance of a web-based clinical diagnosis support system for internists. *Journal of General Internal Medicine*, *23*(S1), 37–40.

Guangpu, F., Xu, C., & Zhiyong, P. (2011). *A rules and statistical learning based method for Chinese patent information extraction, Web Information Systems and Applications Conference (WISA) (pp. 114–118)*. Washington, DC, USA: IEEE Computer Society.

Haque, S., Dey, L., & Mahajan, A. (2009). A news analysis and tracking system. *PReMI '09 Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence* (pp. 231–236). New Delhi, India: Springer-Verlag.

Hatzivassiloglou, V., Gravano, L., & Maganti, A. (2000). *An Investigation of linguistic features and clustering algorithms for topical document clustering*. New York, NY, USA: ACM, 224–231.

Helbich, M., Hagenauer, J., Leitner, M., & Edwards, R. (2013). Exploration of unstructured narrative crime reports: An unsupervised neural network and point pattern analysis approach. *Cartography and Geographic Information Science*, *40*(4), 326–336.

Holsapple, C. W., & Whinston, A. B. (1996). *Decision support systems: A knowledge based approach*. West Group.

Isa, D., Kallimani, V. P., & Lee, L. H. (2009). Using the self organizing map for clustering of text documents. *Expert Systems with Applications*, *36*(5), 9584–9591.

Jayram, Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., & Zhu, H. (2006). Avatar information extraction system. *IEEE Data Engineering Bulletin*, *29*(1), 40–48.

Kanable, R. (2008). Talking to tipsters. *Law Enforcement Technology*, *35*(11), 10–18.

Knutsson, O., Sneiders, E., & Alfalahi, A. (2012). Opportunities for improving egovernment: Using language technology in workflow management. *ICEGOV '12 Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance* (pp. 495–496). Albany, New York: ACM.

Kovachev, S., Reichert, P., & Speck, H. (2008). Crimeblips: Web based framework for crime incident analysis and visualization. *10th International Conference on Information Integration and Web-based Applications and Services (iiWAS2008)* (pp. 694–697). Linz, Austria: ACM.

Kozawa, S., Tohyama, H., Uchimoto, K., & Matsubara, S. (2008). Automatic acquisition of usage information for language resources. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, & D. Tapias (Eds.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)* (pp. 1227–1232). Marrakech, Morocco: European Language Resources Association (ELRA).

Ku, C. H., Iriberri, A., & Leroy, G. (2008a). Crime information extraction from police and witness narrative reports. *IEEE International Conference on Technologies for Homeland Security* (pp. 193–198). Boston, USA: IEEE.

Ku, C. H., Iriberri, A., & Leroy, G. (2008b). Natural language processing and e-government: Crime information extraction from heterogeneous data sources. *International Conference on Digital Government Research* (pp. 162–170). Montreal, Canada: Digital Government Research Center.

Ku, C. H., & Leroy, G. (2011). A crime reports analysis system to identify related crimes. *Journal of the American Society for Information Science and Technology*, *62*(8), 1533–1547.

Kushchu, I., & Kuscu, H. (2003). From e-government to m-government: Facing the inevitable. *The 3rd European Conference on e-Government* (pp. 253–260). Ireland: MCIL Trinity College Dublin.

Lai, G. -H., Chen, C. -M., Laih, C. -S., & Chen, T. (2009). A collaborative anti-spam system. *Expert Systems with Applications, 36*(3, Part 2), 6645–6653.
Lai, F., Li, D., & Hsieh, C. -T. (2012). Fighting identity theft: The coping perspective. *Decision Support Systems, 52*(2), 353–363.
Lakkaraju, P., Gauch, S., & Speretta, M. (2008). Document similarity based on concept tree distance. *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia* (pp. 127–132). Pittsburgh, PA, USA: ACM.
Leacock, C., & Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. In Christiane Fellbaum (Ed.), *Wordnet: An electronic lexical database* (pp. 265–283). Cambridge, MA: MIT Press.
Lee, L. W., & Chen, S. M. (2006). New methods for text categorization based on a new feature selection method and a new similarity measure between documents. *IEA/AIE'06 Proceedings of the 19th international conference on Advances in Applied Artificial Intelligence: industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 1280–1289). Berlin: Springer-Verlag.
Lee, D. L., Chuang, H., & Seamons, K. (1997). Document ranking and the vector-space model. *Software, IEEE, 14*(2), 67–75.
Li, W., Miao, D., & Wang, W. (2011). Two-level hierarchical combination method for text classification. *Expert Systems with Applications, 38*(3), 2030–2039.
Li, L., Wang, J., & Leung, H. (2009). A knowledge-based similarity classifier to stratify sample units to improve the estimation precision. *International Journal of Remote Sensing, 30*(5), 1207–1234.
Liao, S., & Jiang, M. (2005). An improved method of feature selection based on concept attributes in text classification. In L. Wang, K. Chen, & Y. Ong (Eds.), *ICNC'05 Proceedings of the First international conference on Advances in Natural Computation — Volume part I* (pp. 1140–1149). Heidelberg: Springer Berlin.
Linders, D. (2012). From e-government to we-government: Defining a typology for citizen coproduction in the age of social media. *Government Information Quarterly, 29*(4), 446–454.
Liu, M., Wang, X., & Liu, Y. (2010). Clcl — A clustering algorithm based on lexical chain for large-scale documents. *Computer and Information Science, 3*(1), 91–100.
Mamakis, G., Malamos, A. G., & Ware, J. A. (2011). An alternative approach for statistical single-label document classification of newspaper articles. *Journal of Information Science, 37*(2), 1–11.
Micol, D., Ferrández, O., & Muñoz, R. (2011). Information retrieval techniques for corpus filtering applied to external plagiarism detection. *Proceedings of the 16th International Conference on Natural Language Processing and Information Systems* (pp. 100–111). Berlin, Heidelberg: Springer-Verlag.
Nadkarni, P.M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association, 18*(5), 544–551.
Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems, 52*(2), 464–473.
Pinheiro, V., Furtado, V., Pequeno, T., & Nogueira, D. (2010). Natural language processing based on semantic inferentialism for extracting crime information from text. *2010 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 19–24). Vancouver: BC.
Piskorski, J., Atkinson, M., Belyaeva, J., Zavarella, V., Huttunen, S., & Yangarber, R. (2010). Real-time text mining in multilingual news for the creation of a pre-frontier intelligence picture. *ACM SIGKDD Workshop on Intelligence and Security Informatics* (pp. 1–9). Washington, D.C.: ACM.
Power, D. (2004a). Decision support systems: From the past to the future. *2004 Proceedings of the Americas Conference on Information Systems (pp. 2025–2031)*. New York: Association for Information Systems.
Power, D. J. (2004b). Specifying an expanded framework for classifying and describing decision support systems. *Communications of the Association for Information Systems, 13*(1), 158–166.
Rajan, K., Ramalingam, V., Ganesan, M., Palanivel, S., & Palaniappan, B. (2009). Automatic classification of Tamil documents using vector space model and artificial neural network. *Expert Systems with Applications, 36*(8), 10914–10918.
Roth, R. E., Ross, K. S., Finch, B. G., Luo, W., & MacEachren, A.M. (2013). Spatiotemporal crime analysis in U.S. law enforcement agencies: Current practices and unmet needs. *Government Information Quarterly, 30*(3), 226–240.
Runeson, P., Alexandersson, M., & Nyholm, O. (2007). Detection of duplicate defect reports using natural language processing. *ICSE '07 Proceedings of the 29th international conference on Software Engineering* (pp. 499–510). Washington, DC, USA: IEEE Computer Society.
Sanchez, E., Toro, C., Carrasco, E., Bonachela, P., Parra, C., Bueno, G., et al. (2011). *A knowledge-based clinical decision support system for the diagnosis of Alzheimer disease.* IEEE, 351–357.
Santos, C. N., & Milidiú, R. L. (2012). Named entity recognition. *Entropy guided transformation learning: Algorithms and applications* (pp. 51–58). London: Springer.
Sarawagi, S. (2007). Information extraction. *Foundations and Trends in Databases, 1*(3), 261–377.
Schroeder, J., Xu, J., Chen, H., & Chau, M. (2007). Automated criminal link analysis based on domain knowledge: Research articles. *Journal of the American Society for Information Science and Technology, 58*(6), 842–855.
Shen, W., Doan, A., Naughton, J. F., & Ramakrishnan, R. (2007). Declarative information extraction using datalog with embedded extraction predicates. *VLDB '07 Proceedings of the 33rd International Conference on Very Large Databases* (pp. 1033–1044). Vienna, Austria: VLDB Endowment.
Song, W., Kim, J. Y., Schulzrinne, H., Boni, P., & Armstrong, M. (2009). Using IM and SMS for emergency text communications. *Proceedings of the 3rd International Conference on Principles, Systems and Applications of IP Telecommunications* (pp. 4:1–4:7). New York, NY, USA: ACM.
Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics, 27*(4), 521–544.
Suzuki, M., & Hirasawa, S. (2007). Text categorization based on the ratio of word frequency in each categories. *Systems, MAN AND CYBERNETICS, 2007. ISIC. IEEE International Conference on (pp. 3535–3540)*. Montreal, Quebec, Canada: IEEE.
Swiderski, B., Kurek, J., & Osowski, S. (2012). Multistage classification by using logistic regression and neural networks for assessment of financial condition of company. *Decision Support Systems, 52*(2), 539–547.
Tan, S. (2006). An effective refinement strategy for KNN text classifier. *Expert Systems with Applications, 30*(2), 290–298.
Tatar, S., & Cicekli, I. (2009). Two learning approaches for protein name extraction. *Journal of Biomedical Informatics, 42*(6), 1046–1055.
Vens, C., Struyf, J., Schietgat, L., Džeroski, S., & Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning, 73*(2), 185–214.
Wan, C. H., Lee, L. H., Rajkumar, R., & Isa, D. (2012). A Hybrid Text Classification Approach with Low Dependency on Parameter by Integrating K-Nearest Neighbor and Support Vector Machine. *Expert Systems with Applications, 39*(15), 11880–11888.
Wang, P., Hu, J., Zeng, H. -J., & Chen, Z. (2009). Using Wikipedia knowledge to improve text classification. *Knowledge and Information Systems, 19*(3), 265–281.
Wang, D., Li, T., Zhu, S., & Ding, C. (2008). *Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization.* Singapore, Singapore: ACM, 307–314.
Wang, X., Zhang, L., Xie, T., Anvik, J., & Sun, J. (2008). An approach to detecting duplicate bug reports using natural language and execution information. *ICSE '08 Proceedings of the 30th International Conference on Software Engineering* (pp. 461–470). New York, NY, USA: ACM.
Wu, Y., Cao, C., Wang, S., & Wang, D. (2010). A Laplacian eigenmaps based semantic similarity measure between words. In Z. Shi, S. Vadera, A. Aamodt, & D. Leake (Eds.), *Intelligent information processing V* (pp. 291–296). Berlin, Heidelberg: Springer.
Wu, Z., & Palmer, M. (1994). Verb Semantics and Lexical Selection, Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94). *Las Cruces* (pp. 133–138). New Mexico: Morgan-Kaufman.

Chih-Hao Ku is an Assistant Professor in the College of Management (IT) at the Lawrence Technological University. He received his M.S. and Ph.D. in Information Systems and Technology at Claremont Graduate University. He received his B.S. and M.S. degree in Information Management in Taiwan. He is an active ACM, IEEE, ISSA, and AIS Member. His research currently focuses on natural language processing, decision support systems, and information visualization. He has published his work in the *Journal of the American Society for Information Science and Technology* (*JASIST*) and in conferences among others.

Gondy Leroy is an Associate Professor in the Department of Management Information Systems at the University of Arizona. She was educated at the Catholic University of Leuven, Belgium, where she earned a combined B.S. and M.S. in Experimental Psychology (1996) and the University of Arizona where she earned a M.S. and Ph.D. in Management Information Systems (2003). She is an IEEE Senior Member and serves on the editorial board of three journals and multiple conferences. Her research focuses on natural language processing in medical informatics and digital government. Her projects have been funded by the National Institutes of Health, the National Science Foundation, Microsoft Research and several foundations. She has published her work in *ACM computing Surveys*, *Journal of the American Medical Informatics Association* (*JAMIA*), *Journal of the American Society or Information Science and Technology* (*JASIST*), *International Journal of Medical Informatics*, *Empirical Software Engineering* among others. She authored the book "*Designing User Studies in Informatics*", published by Springer, and conducts tutorials on this topic in the United States, Europe, Canada and Asia. She is active in outreach and has organized several workshops and contributed to doctoral consortia, workshops and conferences aiming to encourage women to enter and remain in the field of computing.