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/.