In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL'16)*
(M. Ali, S. Newsam, S. Ravada, M. Renz, and G. Trajcevski, Eds.), Burlingame, CA, November 2016.

CrimeStand: Spatial Tracking of Criminal Activity [*]

Faizan Wajid
Department of Electrical Engineering
University of Maryland, College Park
College Park, MD 20740
fwajid@umd.edu

Hanan Samet
Department of Computer Science
University of Maryland, College Park
College Park, MD 20740
hjs@umiacs.umd.edu

ABSTRACT

Pursuing criminal activity is tied with understanding illegal or unlawful actions taken on opportunity within a geographic location. Mapping such activities can aid significantly in determining the health of a region, and the vicissitudes of civilian life. Methods to track crime and criminal activity after the fact by mapping news reports of it to geographic locations using the NewsStand system are discussed. NewsStand provides a map-query interface to monitor over 10,000 RSS news sources and making them available within minutes after publication. NewsStand was designed to collect event data given keywords centered on locations specified textually and mapping these locations to their spatial representation, a procedure called geotagging. The goal is to demonstrate how to detect and classify criminal activity by geotagging keywords pertaining to crime, and, in effect, to enhance the capabilities of NewsStand to explicitly show this category of news. The resulting system is named “CrimeStand”.

CCS Concepts

•**Information systems** → **Geographic information systems; Content analysis and feature selection;** •**Human-centered computing** → *Geographic visualization;*

Keywords

NewsStand, GIS, geotagging, text mining

1. INTRODUCTION

The pervasiveness of criminal activity is an important criteria that determines the overall health of the region. Crime is defined as an action that is deemed injurious to the public welfare, and is legally prohibited [1]. With this in mind, capturing crime-related events will be given a broader scope as they will need to include events that proceed from criminal activity, such as the enactment of new laws and policies, human rights issues, and the like.

This paper discusses the techniques used to allow spatial querying of crime-related events with the use of NewsStand, a spatiotemporal news browser that enables querying news stories by the locations mentioned in them, achieved by using a map query interface [15, 24, 25, 26, 28]. NewsStand crawls the web seeking news articles and tags each article with an associated location along with other attributes [3, 13, 14, 16, 17, 19, 20, 21], NewsStand monitors the output of over 10,000 RSS news feeds which are made available within minutes of publication, and automatically clusters articles into categories, taking into account geographic references and presents articles on an interactive world map. NewsStand has a very intuitive user interface that can be used to present a variety of information related to the articles not limited to pictures and videos. It even includes dedicated *layers* whereby users can choose to view filtered news related to business brands, diseases, and people [2, 11, 12], rooted in our prior development of spatial browsers [4, 6, 22, 23].

By leveraging this system, we can explore the range of criminal activity by capturing news articles and associating them with our definition of crime. This yields a collection of crime-related news articles with varying degrees of differentiation — that is, events that are purely unlawful or events that result from crime. The utility of NewsStand is indispensable here as each news article will be *geotagged*, and it also allows us to see related (or similar) events in other parts of the world by virtue of the interactive world map.

Since our work follows a similar trajectory to previously implemented layers, namely disease and brand tracking [2, 11, 12], we omit describing the NewsStand Web interface and its modes for news querying. The presented demonstration allows users to query the system and browse the results using the map query interface.

The remainder of this paper is organized as follows. We first discuss the techniques that we used to obtain and format our data (Section 1), Second, we dive into our choice and design of classifiers (Section 2). Next, we report our results along with the shortcomings of each classifier (Section 3). We then briefly describe our demonstration (Section 4), and conclude with plans for improvements and directions for future research and extensions. Throughout the paper readers are provided pointers to the literature where more details about various aspects of NewsStand can be found. Of course, most of these papers are authored or co-authored by members of the NewsStand team.

2. METHODOLOGY

The NewsStand pipeline is structured to acquire news articles and transmit them to various other modules where they are ultimately tagged and stored. Each article can be independently re-

[*]This work was supported in part by the AWS research grant and the NSF under Grants IIS-12-19023 and IIS-13-20791.

SIGSPATIAL'16 October 31 - November 03, 2016, Burlingame, CA, USA
© 2016 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-4589-7/16/10.
DOI: http://dx.doi.org/10.1145/2996913.2997006

trieved and if so configured, it can also be associated with one (or many) layers. These layers, as the name implies, are superimposed upon NewsStand and allow the user to view filtered news respective to the categories under which they were *tagged*. Following this, we extend NewsStand to include a *crime layer* as a platform to only view crime-related articles. However, two non-trivial challenges must first be addressed in order for the correct articles to be displayed: *context* and *relevance*.

A typical problem within NLP is providing *context* around a word to reduce misclassification, i.e. battery — legal jargon relating to beating — could yield Duracell or Energizer batteries, or theft to yield Grand Theft Auto — not a mention of an actual act but rather the famous video game. The latter example contains three words which only make sense if they appear together, but also note that there are four (sensible) permutations that this sequence can take. Of course, these examples can also be applied to the problem of *relevance*, which plays a larger role since not all crime-related articles will explicitly mention hit-friendly keywords, or at least with the frequency and granularity with which we wish to find them. We now discuss two methods employed to discriminate between articles.

2.1 Data Processing

We started by creating a rudimentary list of crime-related keywords to serve as the initial dictionary. Primarily, this list consisted of common nouns such as murder, homicide, burglary, extortion, etc. including certain drugs and mental disorders. It totaled about 100 entries.

We then obtained over 5,000 miscellaneous news articles from NewsStand and formatted this collection to only return the article headline, body, and a unique identifier. A preliminary scan cross-referenced each article body for occurrence of keywords present in the dictionary, and binned the article with the matching keyword as the label. Each bin contained unique entries as we were not interested in the context of the event at this point, therefore repeats were not necessary.

With the articles successfully organized, we went through each bin and manually classified the articles within it as being crime-related or not, and modified the dictionary accordingly. Albeit a tedious task, it was done in order to ensure that all crime-related news was accounted for, that is to say, not just an incident or action representative of common words (theft, murder, burglary), but also actions or activity that took place after the onset of criminal activity (government talks and legislation, police force training, arms reduction, influence and awareness through entertainment to name a few). Dictionary keywords with low hits were removed, and new words (and some important word-combinations) were added to strengthen the efficacy of the dictionary.

2.2 Data Formatting

Refinement of the dictionary made obvious the extent to which our list could potentially grow given the breadth of data being processed. More so the issue of language in its presentation can change among media. In attempts to remedy this, we used the Porter stemming algorithm [18] for suffix stripping. As the name implies, the algorithm was created to find the stem in a word and simplify it, or to give it a more phonetic wording to enlarge the matching body. By modifying prefixes and suffixes, the overall word count can be reduced and simplified to root words allowing us to omit gerunds ("-ing"), plurals and contractions ("'s"), and other language-based word manipulations (for example, *explosion* can be represented as *explos*, *obscenity* can be represented as *obscen*). We can now capture larger words with more flexibility. We complement this technique with additional text pre-processing to remove non-essentials:

1. Alphabetize emergency digits — i.e. replacing "911" with *nineoneone*
2. Remove numbers: metadata is not collected.
3. Remove symbols: all non-alphanumerics are eliminated.
4. Lower-case the dictionary and articles: Normalize text and simplify pattern matching.
5. Simplifying the text: Applying Porter stemmer on text body to obtain *simple format*.

2.3 Classification

With a sizable collection of articles, we turned our attention to designing a system to solve a binary classification problem to automate the classification of incoming articles as crime-related.

Our first classification approach leveraged StanfordNER, an open-source library built for name entity recognition [7]. StanfordNER is also an integral part of the NewsStand pipeline so it's also a natural way-point for consideration. It includes several predefined feature extractors and classifiers with a high degree of customization making it a valuable tool for text classification. Specifically, StanfordNER performs entity recognition on the following three classes: **Person**, **Organization**, **Location**. To integrate with StanfordNER, the most intuitive method was to transpose our dictionary to fit under **Organization** as this category is broad enough to contain different parts of speech. With the exception of acronyms, all keywords will be lower-cased.

Our second approach was inspired by the Spam Classification problem using Support Vector Machines. This method consists of a *hit-on-occurrence* generation of feature vectors before SVM is employed. We modified this method to include word locality mainly to provide significance to adjectives and other word-modifiers (i.e. illegal, brutal) to better draw conclusions about context and relevance to reinforce the competency of the classifier. We also repositioned these word-modifiers before primary keywords to give us control on where to demarcate hits. Linguistically, these primary keywords only appear together if they are provided in a comma-separated list. In order to use SVM, we must first generate feature vectors from the articles, thereby ensuring that only relevant keywords are highlighted, and all articles can be represented by the same length such that matrix operations can be performed. The number of entries in the dictionary will serve as the length of the feature (row) vector because the dictionary will remain static (of length 318). To do this, we simply iterated over every article and if it contained a word from the dictionary, that element in the feature vector would be set to **1**, and to **0** otherwise. If a word-modifier is found within two units to the left and right of a keyword, we place a **1** adjacent (one unit to the left and right) to the keyword essentially simulating a tiny cluster which pushes SVM into widening the margin between support vectors. The same concept applies if the feature vector contains many **1**'s in sequence before word locality is measured — it implies that the text body contained a comma-separated list of crimes that were committed in the event. Therefore, word-locality alignment will not be important and the keyword with the most hit frequency will be used to label the crime. Finally, we are left with a 5000x318 binary-valued matrix, and the corresponding label vector is 5000x1 (where again, 5,000 represents the number of articles and 318 is the dictionary length). We used LIBSVM [5] for the classification task by first converting the feature matrix into a LIBSVM-friendly file format and then training with different kernels: linear, polynomial, and radial basis function (RBF).

3. RESULTS

Following another common heuristic, we randomized and split the hand-classified news articles (4,000 entries for training, and 1,000 for testing) to measure the accuracy of each classifier. As such, Table 1 outlines our results, and for simplicity, only final accuracy results are shown (after three revision iterations per test). We also provide dictionary length as a metric of expansion; can more keywords be added without loss of accuracy and/or loss of *context* and *relevance*?

Classifier	Accuracy	Dict Length
Baseline	45.5%	153 entries
StanfordNER	68.3%	181 entries
LIBSVM (RBF kernel)	87.7%	318 entries

Table 1: Accuracy and Dictionary Lengths

The baseline test functions by simply returning all articles that contain the keywords present in the dictionary, non-uniquely. Essentially, it is akin to *grepping* the keywords within the articles. An elementary technique appropriately yielded a very low accuracy of 45.5% on a stringent and limited dictionary, the same unmodified list we began our experiment with. It stands to reason that additions or further modifications to the dictionary would not improve the accuracy enough to balance the cost of corpus bloat as the technique is fundamentally a greedy search. The StanfordNER classifier performed significantly better than our baseline test at 68.3% while also allowing for an 18% increase in dictionary size. Lack of further improved performance ostensibly comes from missing word locality application. Usage of adjectives and other word-modifiers serve a specific purpose, that is, to strengthen the surrounding words in a given sentence. Such words by themselves would unnecessarily inflate the dictionary and flag false-positives. Both the baseline and StanfordNER limited us internally from implementing methodologies to contextualize word-modifiers, however there are open-source toolkits available that could presumably help remedy this. Finally the SVM method proved to be the most outstanding and robust, with an accuracy of 87.7% operating on an RBF kernel (linear and polynomial yielded 87.1% and 82.1% respectively), not to mention a 75.7% overall increase in dictionary size with respect to StanfordNER's. The nature of SVM's performance on two-label classification problems along with our text locality augmentation helped in arriving at this number, with room for improvement.

4. DEMONSTRATION

We now describe how users can access CrimeStand and its features. Alongside this step-by-step tutorial, we will also display the results of both the NER and SVM classifiers.

First, users will need to visit NewsStand by pointing their browsers to: `http://newsstand.umiacs.umd.edu/`. Following this, the user must enable the crime layer by clicking the "Settings" button on the bottom right corner of the screen. A menu pane will appear listing the "Crime Layer" under the "Layers" column. The map-view will update showing various crime-related tags relating to recent news. Users can hover over these tags, which will display the article directly associated with the tag. It will also provide a list of a representative article around the same location (blue dots), as well as other locations where the tag was mentioned (orange dots). Users may zoom in to inspect certain locations with more attention.

Figure 1 provides a snapshot of the CrimeStand map interface. This view is also the default mode, known simply as Map Mode. This mode allows users to search for articles based on Location and/or Keyword, with a slider to control how many results should be visible.

5. CONCLUSIONS AND FUTURE WORK

This paper detailed our work to extend the NewsStand system by enabling a dedicated layer to view news and events related to crime. To this end, our findings have been large and broad, and have left us somewhat unhinged having found the sheer amount of crime that's being reported around the world.

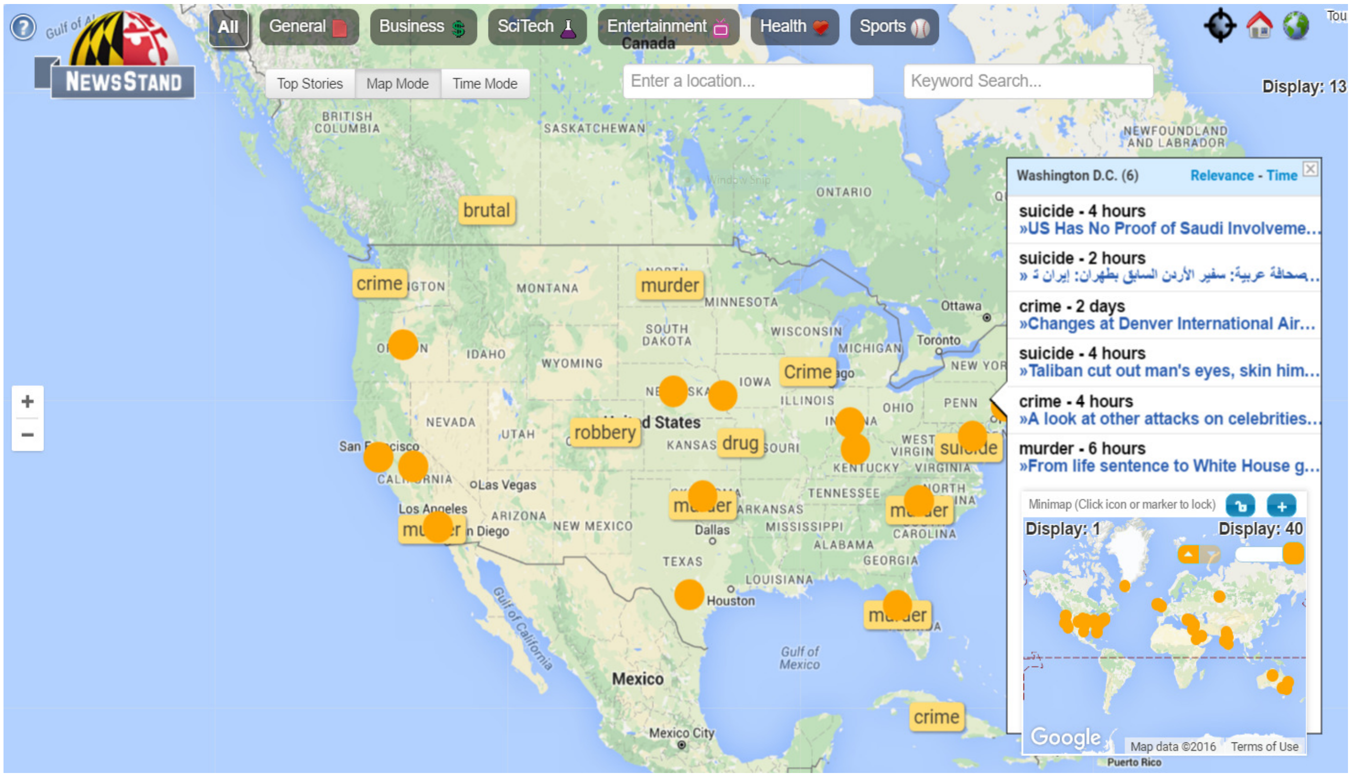
The diversity of crime-related news alone made the case to focus on manual classification of crime-related events. We still feel our data fell short, but increasing training data by the same laborious method might not be the best way, so we are investigating alternatives. We chose to ignore context during our article binning phase despite its importance, especially for emergency services. An event that consists of multiple crimes, e.g. "arson" and "murder", might not receive adequate priority if incorrectly categorized. We found out from the Prince Georges County Police Department that events are ordered in terms of whichever is capable of being remedied first. Classifying based on article title alone is a viable shortcut, however it will miss more nuanced events as the headline will certainly omit key details (our results showed this method fell short). Another consideration is time bias — our data was obtained for the month of October 2015. Our initial pass of classification returned a high number of misclassified events because the crime-related keywords would appear in Entertainment news. We are certain this was due to preparations to celebrate Halloween in Western-nations. Minor offenses also appear and dictionary bloat would exacerbate this issue even more, such as the injuring of a bald eagle, which is an offense in the United States of America, and was flagged. It is peculiar that we saw only three mentions of *cybersecurity* and *cybercrime*, which will certainly be pronounced in years to come.

NewsStand allows provisioning news in both the spatial and temporal domains, However, our module currently does not provide functionality for temporal querying. Integrating the *time-mode* feature can open many avenues for further exploration by giving us the ability to track events of interest around the globe through time and view the concentrations they exist in. We are also looking to integrate with TwitterStand [9, 10, 27] and WeiboStand [8]. Although the architectures of NewsStand and TwitterStand being similar, our classifier will need to be modified to handle tweets. Tweet messages are of a different nature than published news due to their word-limit, as well as the prevalence of colloquialisms, and as such, need to be accounted for.

Making criminal events visible on a map interface can be a valuable tool for social scientists seeking to study the effects of prolonged exposure to crime, it's affects on human behavior and on mental health. We wish to further enhance the system by introducing more nuances and linguistic-specific components to enable these interactions and provide further discrimination between events, i.e. crimes caused by mental disorders, etc. Could our classifier begin to distinguish within-criteria of reports, such as identify murder from extortion, or help to identify if refugees and minorities within sovereign nations do in fact have the propensity to be criminals, or perform terrorist acts? We hope to continue to enhance our crime layer to shed more light upon an oft-missed issue.

6. ACKNOWLEDGMENTS

A big thanks to both Shangfu Peng and Hong Wei for helping us integrate the crime layer into NewsStand!



Washington D.C. (6) Relevance - Time

suicide - 4 hours
»US Has No Proof of Saudi Involveme...

suicide - 2 hours
»صحافة عربية: سفير الأردن السابق بطهران: إيران تـ..

crime - 2 days
»Changes at Denver International Air...

suicide - 4 hours
»Taliban cut out man's eyes, skin him...

crime - 4 hours
»A look at other attacks on celebrities...

murder - 6 hours
»From life sentence to White House g...

Figure 1: NewsStand's interactive map with Crime Layer enabled.

7. REFERENCES

[1] Dictionary.com definition of crime; http://dictionary.reference.com/browse/crime?s=t.

[2] A. Abdelkader, E. Hand, and H. Samet. Brands in newsstand: Spatio-temporal browsing of business news. In *GIS*, pages 97:1–97:4, Bellevue, WA, Nov 2015.

[3] M. D. Adelfio and H. Samet. Structured toponym resolution using combined hierarchical place categories. In *Proceedings of 7th ACM SIGSPATIAL Workshop on Geographic Information Retrieval (GIR'13)*, pages 49–56, Orlando, FL, Nov. 2013.

[4] F. Brabec and H. Samet. Client-based spatial browsing on the world wide web. 11(1):52–59, Jan 2007.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for Support Vector Machines. pages 27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] C. Esperança and H. Samet. Experience with sand-tcl: A scripting tool for spatial databases. 13:220–255, Apr 2002.

[7] J. R. Finkel, T. Grenager, and C. Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *ACL*, pages 363–370, Ann Arbor, MI, Jun 2005.

[8] C. Fu, J. Sankaranarayanan, and H. Samet. Weibostand: Capturing Chinese breaking news using Weibo. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN'14)*.

[9] N. Gramsky and H. Samet. Seeder finder: Identifying additional needles in the twitter haystack. In *LSBN*, pages 44–53, Orlando, FL, Nov 2013.

[10] A. Jackoway, H. Samet, and J. Sankaranarayanan. Identification of Live News Events Using Twitter. In *LBSN*, pages 25–32, Chicago, IL, Nov 2011.

[11] R. Lan, M. D. Lieberman, and H. Samet. The Picture of Health: Map-based, Collaborative Spatio-temporal Disease Tracking. In *HealthGIS*, pages 27–35, Redondo Beach, CA, Nov 2012.

[12] R. Lan, M. D. Lieberman, and H. Samet. Spatio-temporal disease tracking using news articles. In *HealthGIS*, pages 31–38, Dallas, TX, Nov 2014.

[13] M. D. Lieberman and H. Samet. Multifaceted toponym recognition for streaming news. In *SIGIR*, pages 843–852, Beijing, China, Jul 2011.

[14] M. D. Lieberman and H. Samet. Adaptive context features for toponym resolution in streaming news. In *SIGIR*, pages 731–740, Portland, OR, Aug 2012.

[15] M. D. Lieberman and H. Samet. Supporting rapid processing and interactive map-based exploration of streaming news. In *SIGSPATIAL*, pages 179–188, Redondo Beach, CA, Nov 2012.

[16] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE*, pages 201–212, Long Beach, CA, Mar 2010.

[17] M. D. Lieberman, H. Samet, and J. Sankaranayananan. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups. In *GIR*, pages 6:1–6:8, Zurich, Switzerland, Feb 2010.

[18] M. F. Porter. An algorithm for suffix stripping. pages 130–137, 1980.

[19] G. Quercini and H. Samet. Uncovering the spatial relatedness in wikipedia. In *SIGSPATIAL*, pages 153–162, Dallas, TX, Nov 2014.

[20] G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. Determining the spatial reader scopes of news sources using local lexicons. In *SIGSPATIAL*, pages 43–52, San Jose, CA, Nov 2010.

[21] H. Samet. Using minimaps to enable toponym resolution with an effective 100% rate of recall. In *Proceedings of 8th ACM SIGSPATIAL Workshop on Geographic Information Retrieval (GIR'14)*, pages 9:1–9:8, Dallas, TX, Nov 2014.

[22] H. Samet, A. Rosenfeld, C. A. Shaffer, and R. E. Webber. A geographic information system using quadtrees. *Pattern Recognition*, 17(6):647–656, November/December 1984.

[23] H. Samet, H. Alborzi, F. Brabec, C. Esperança, G. R. Hjaltason, F. Morgan, and E. Tanin. Use of the SAND spatial browser for digital government applications. *Communications of the ACM*, 46(1):63–66, Jan. 2003.

[24] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and B. E. Teitler. Porting a web-based mapping application to a smartphone app. In *GIS*, pages 525–528, Chicago, IL, Nov 2011.

[25] H. Samet, B. E. Teitler, M. D. Adelfio, and M. D. Lieberman. Adapting a map query interface for a gesturing touch screen interface. In *Proceedings of the Twentieth International Word Wide Web Conference (Companion Volume)*, pages 257–260, Hyderabad, India, March-April 2011.

[26] H. Samet, J. Sankaranarayanan, M. D. Lieberman, M. D. Adelfio, B. C. Fruin, J. M. Lotkowski, D. Panozzo, J. Sperling, and B. E. Teitler. Reading news with maps by exploiting spatial synonyms. *Communications of the ACM*, 57(10): 64–77, Sep 2014.

[27] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: News in Tweets. In *GIS*, pages 42–51, Seattle, WA, Nov 2009.

[28] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: A new view on news. In *GIS '08*, pages 18:1–18:10, Irvine, CA, Nov 2008.