An Integrated Framework for Spatio-Temporal-Textual Search and Mining

Bingsheng Wang[1], Haili Dong[1], Arnold P. Boedihardjo[2], Chang-Tien Lu[1], Harland Yu[2], Ing-Ray Chen[1], Jing Dai[3]
[1]Virginia Tech, Falls Church, VA 22043
[2]U. S. Army Corps of Engineers, Alexandria, VA 22315
[3]Google Inc., New York, NY 10011
{claren89, hailid}@vt.edu, arnold.p.boedihardjo@usace.army.mil, ctlu@vt.edu,
harland.yu@usace.army.mil, irchen@cs.vt.edu, jddai@google.com

ABSTRACT

This paper presents an integrated framework for Spatio-Temporal-Textual (STT) information retrieval and knowledge discovery system. The proposed ensemble framework contains an efficient STT search engine with multiple indexing, ranking and scoring schemes, an effective STT pattern miner with Spatio-Temporal (ST) analytics, and novel STT topic modeling. Specifically, we design an effective prediction prototype with a third-order linear regression model, and present an innovative STT topic modeling relevance ranker to score documents based on inherent STT features under topical space. We demonstrate the framework with a crime dataset from the Washington, DC area from 2006 to 2010 and a global terrorism dataset from 2004 to 2010.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search Process.

General Terms

Algorithms, Management.

Keywords

Spatio-temporal-textual search engine, Spatio-temporal analytics, topic modeling.

1. INTRODUCTION

In the last two decades, crime incidents and terrorism threats have caused increased global anxiety and unrest. For example, the 9/11 attacks resulted in approximately 3,000 deaths and more than 6,000 injuries [1]. To learn, understand, and plan mitigating strategies for crime incidents and terrorism attacks, it is critical to design a framework that enables users to efficiently locate the most relevant events and effectively discover important patterns from data. Furthermore, with the prevalence of mobile devices, it is important to design and implement a search and mining system that adapts to the needs of both desktop and mobile users.

Although there exist search engines [2] and pattern mining techniques [3] for crime incidents and terrorism threats, they all lack the critical integration of Information Retrieval (IR) and Knowledge Discovery (KD). Specifically, the pattern mining module is unable to utilize fast indexing provided by the search engine and the learned patterns have no contribution to the performance of the search engine. With a sophisticated integration, fast query processing could increase the effectiveness of pattern mining and mined patterns could be leveraged to retrieve relevant information. Hence, it is crucial to integrate the search engine and the pattern mining module into an ensemble framework and synergize the mutual interaction between the IR and KD processes. This is the motivation behind the STT IR and KD System. The IR module allows users to search based on space, time, and keywords. The KD module provides ST analytics and STT topic modeling to discern novel patterns from the search results.

The STT search engine encapsulates STT access methods, query operations, and STT relevance rankings to provide analysts with efficient retrieval of relevant information during hypothesis formulation and evidence collection. The ST analytics are designed to alleviate the labor-intensive tasks of spatiotemporal crime and terrorism analysis, such as prediction and visualization. Particularly, we employ STT access methods to accelerate ST analytics/pattern mining. Given a spatial region and time period, crime incidents and terrorism attacks can be found with inherent correlations in terms of unstructured content analysis [3]. In order to support efficient discovery of the ST based textual patterns, we develop STT topic modeling to automatically visualize the dominant or prevailing concepts for abstracting elemental relations. Additionally, we improve the STT ranker with STT topic modeling through integrating latent semantics for grading documents under topical space.

The major contributions of this paper are as follows:

- **Ensemble framework:** The proposed integration framework of STT search engine and STT pattern mining greatly enhances the efficiency and effectiveness of the knowledge discovery process through their mutual utilization.
- **ST analytics & STT topic modeling:** The designed analytical algorithms automatically and efficiently discover relationships within complex ST data. STT topic modeling is proposed to uncover the evolution and distributions of thematic concepts in unstructured data within an ST framework.
- **STT search engine:** STT search engine is designed to rapidly process users' queries and provide a ranked list of results based on multiple relevancy measures. Specialized indexes are developed for STT data to provide efficient query processing. Several variants of ranking methods are proposed that consider different topological configurations of STT.

ACM SIGSPATIAL GIS '12, November 6-9, 2012. Redondo Beach, CA, USA.
Copyright (c) 2012 ACM ISBN 978-1-4503-1691-0/12/11...$15.00.

- **Desktop & mobile visualizations:** Desktop & mobile visualizations are developed to demonstrate the adaptability and capability of the system.

2. SYSTEM OVERVIEW

In this section, we discuss the system architecture (Figure 1).

Data Processing: We first remove data items which do not exhibit significant attributes (significant attributes are defined as features that correspond to spatiotemporal properties). Particularly, geo-parsing and geo-coding techniques are applied to the terrorism dataset to extract and disambiguate locations, respectively.

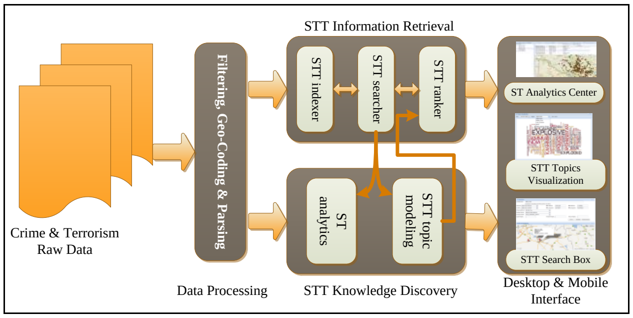
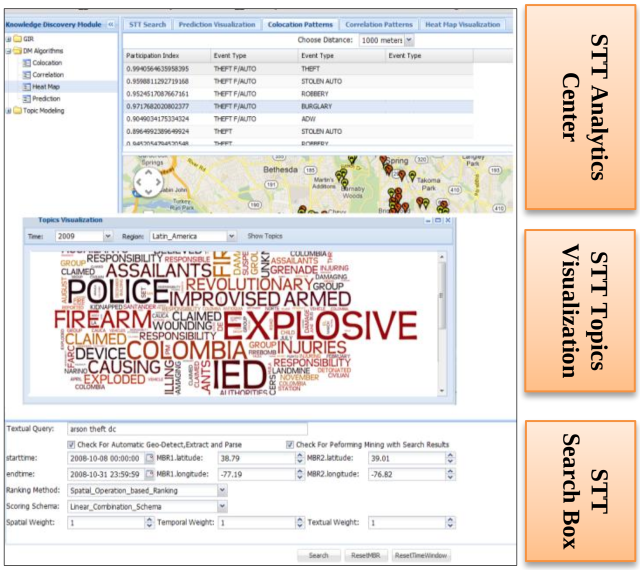


Figure 1. Overview of system architecture

STT Information Retrieval: This module is developed through the integration of the ST indexer, searcher, and ranker into Solr [4]. ST index is implemented via HR+-tree [5] and B+-tree with the Hilbert curve [6]. In addition to implementing a set of traditional rankers [7], we have developed an STT topic based ranker, which is an extension to traditional ranking schemes aimed at grading documents in the topical space.



Choose Distance: 1000 meters

Participation Index	Event Type	Event Type	Event Type
0.9940564635958395	THEFT F/AUTO	THEFT	
0.9598811292719168	THEFT F/AUTO	STOLEN AUTO	
0.9524517087667161	THEFT F/AUTO	ROBBERY	
0.9717682020802377	THEFT F/AUTO	BURGLARY	
0.9049034175334324	THEFT F/AUTO	ADW	
0.8964992389649924	THEFT	STOLEN AUTO	
0.9452054794520548	THEFT	ROBBERY	

Time: 2009 Region: Latin_America

Textual Query: arson theft dc

☑ Check For Automatic Geo-Detect,Extract and Parse ☑ Check For Peforming Mining with Search Results

starttime: 2008-10-08 00:00:00 MBR1.latitude: 38.79 MBR2.latitude: 39.01

endtime: 2008-10-31 23:59:59 MBR1.longitude: -77.19 MBR2.longitude: -76.82

Ranking Method: Spatial_Operation_based_Ranking

Scoring Schema: Linear_Combination_Schema

Spatial Weight: 1 Temporal Weight: 1 Textual Weight: 1

Figure 2. System interface

STT Knowledge Discovery: This module contains two components: ST analytics and STT topic modeling. ST analytics provides prediction, heat map visualization, and colocation & correlation analysis. STT topic modeling is designed to learn the underlying semantics based on Latent Dirichlet Allocation (LDA) [8].

Mobile & PC Interface: Figure 2 shows the major components of the user interface for both desktop and mobile platforms. The "ST Analysis Center" is the entrance to inspecting patterns mined from the dataset. "STT Topics Visualization" provides an interactive interface for users to view the words that contribute to the topics found in both crime and terrorism datasets. "STT Search Box" presents an interactive window for users to conveniently customize a comprehensive set of query parameters.

2.1 ST Analytics

The ST analytics module is used to reveal patterns from the dataset. In this section, we look specifically into the task of mining the crime dataset.

2.1.1 *Prediction Model*

We first split the DC area into equal-sized grids. Then, we apply a third-order autoregressive model to predict crime frequency for a specific crime type and grid cell. Formally, the prediction model is,

$$C_t^{(i)} = w_1^{(i)} * C_{t-1}^{(i)} + w_2^{(i)} * C_{t-2}^{(i)} + w_3^{(i)} * C_{t-3}^{(i)} + \varepsilon_t$$

where parameter $w_j^{(i)}$ is the influence weight from $C_{t-j}^{(i)}$ to $C_t^{(i)}$, j represents the time interval, and ε_t denotes white noise. $C_t^{(i)}$ refers to the number of i^{th} crime type at time t. $i = 1, \dots, N$ (N is the number of crime types). In addition, $w_j^{(i)}$ is subject to exponential distribution, $w_j^{(i)} \sim \exp(-a_i * j)$ where a_i is a deterministic parameter.

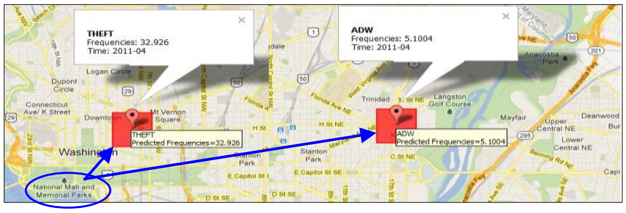


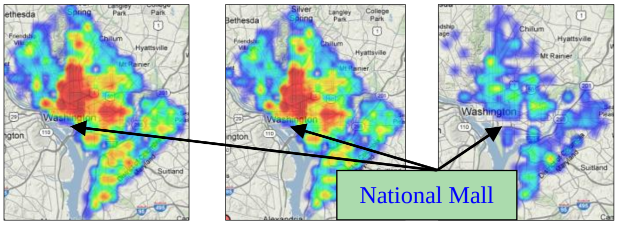
Figure 3. Predicted crime frequencies

We exemplify the results of the prediction model in Figure 3. Figure 3 shows areas (red rectangles) with the highest predicted frequencies of "THEFT" and "ADW (Assault with a Deadly Weapon)". We can observe that the spatial location of "ADW" is much farther away from the National Mall than that of "THEFT". This suggests that city officials should institute policies and schedule patrols to prevent "THEFT" around the National Mall, and employ tactics to reduce "ADW" near Kingman Park.

2.1.2 *Heat Map Visualization*

Crime heat map gives a density measure of a given spatial region and time period for a set of crime types. Hence, users are able to customize the heat maps across a dynamic range of spatiotemporal parameters and crime types.

The heat map of DC area is shown in Figure 4 (a). The deeper red color indicates higher crime occurrence. Figure 4 (b) shows the heat map of most frequent crimes ("BURGLARY", "ROBBERY", "STOLEN AUTO", "THEFT", and "THEFT F/AUTO"). Most hot spots in Figure 4 (a) and (b) are located at the north or northeast of National Mall. A potential reason for this pattern is that the crowd of parking lots near this area leads to a large amount of "THEFT", "THEFT FROM AUTO" or "STOLEN AUTO". Figure 4 (c) shows the heat map of serious crimes ("ADW", "ARSON", "HOMICIDE", and "SEX ABUSE") indicating that the hot spots of these serious crimes are far from the National Mall. The described visualizations allow important contextual information (e.g., National Mall landmarks) to be referenced so as to facilitate the analytical process.



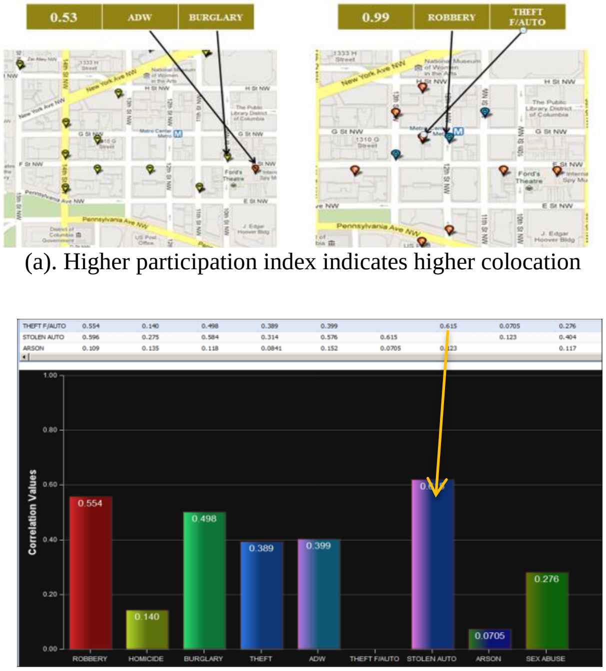
(a). All crimes (b). Frequent crimes (c). Serious crimes

Figure 4. Crime heat map visualization

2.1.3 Colocation & Correlation Pattern Mining

Colocation analysis identifies crime types that co-occur spatially based on participation index [9]. The analysis is performed on all of the extracted events of the retrieved document set. The mining operations are hence executed in real-time. In order to reduce processing time, we cache the analysis results which can be used to answer future queries.

For correlation analysis, we first aggregate crime data with respect to crime type and grid cell. Then, we compute the correlation between crime types through cosine similarity under the grid space. The advantages of using cosine similarity are that it is faster to compute (compared to Pearson's correlation) and can effectively reveal the relationships of crimes at a larger (global) level.



THEFT F/AUTO	0.554	0.140	0.498	0.389	0.399		0.615	0.0705	0.276
STOLEN AUTO	0.596	0.275	0.584	0.314	0.576	0.615		0.123	0.404
ARSON	0.109	0.135	0.118	0.0841	0.152	0.0705	0.123		0.117

(a). Higher participation index indicates higher colocation

(b). The most correlated crimes

Figure 5. Crime colocation & correlation patterns

As shown in Figure 5 (a), the participation index of "ROBBERY" and "THEFT FROM AUTO" (0.99) is higher than that of "ADW" and "BURGLARY" (0.53). This pattern is demonstrated in the map visualization of the two groups of collocated crime types. Figure 5 (b) shows the global correlation color matrix between "THEFT F/AUTO" and nine other crime types where big value denotes high correlation and little value indicates low correlation. From this figure, we can observe that "THEFT F/AUTO" and "STOLEN AUTO" have the highest correlation (0.615) which indicates "THEFT F/AUTO" and "STOLEN AUTO" have similar patterns under the grid space, resulting in a high cosine similarity score.

2.2 STT Topic Modeling

Topics present the thematic structure of documents, indicating the underlying latent semantics in a corpus. A topic can be regarded as a Dirichlet distribution over words, and the words with high probability in the topic represent features of the topic. For example, topic "Police" consists of words: "Police" with probability 0.3674, "Officer" with probability 0.1267, and "Station" with probability 0.0966. These top words (*dominant words*) present the primary profile for shaping the "Police" topic.

Learning topics based on spatial location and time period can elucidate important thematic patterns of crime incidents and terrorism attacks. There are three steps to learn and visualize topics. The first step is to filter documents based on a spatial MBR and temporal period. The second step is to apply LDA to learn topics. The final step is to visualize *dominant words* with tag clouds.

For our analysis, we define five spatial areas: Afghanistan, Middle East, Latin America, Africa, and North America and Europe, and define the time period to be one year. Given a set of documents filtered with ST condition, LDA is applied to extract topics and capture the inherent core themes in the documents. Topics are ordered with probabilities, representing the correlation between topics and documents. Topics of documents are selected from a subset of words collected from documents with probability subject to the Dirichlet distribution. The *dominant words* in the topic set are visualized using tag clouds where the word size is proportional to their probability.

Figure 6 shows the pattern that topics vary greatly with spatial regions for a fixed time period, but vary very little (i.e., slow evolution) with time periods for a fixed spatial region. The *dominant words* in Middle East ("POLICE" and "VEHICLE") are much different from those in Latin America ("COLOMBIA" and "ARMED"). From the *dominant words* in Latin America, we might be able to conclude that terrorism attacks have high probability of occurring in "COLOMBIA".

2.3 STT Information Retrieval

STT IR consists of STT indexer, STT searcher and STT ranker. As shown in Figure 7, users are able to customize a rich set of parameters for their STT query. Users can also view the spatial location and details of selected items. Moreover, we apply paging mechanism to accelerate the response times of users' requests. The sub components of STT IR are stated in details in the following subsections.

2.3.1 STT Indexer

For the textual index, we employed Solr's inverted file index. For the ST index, we employed the HR+-tree index [2] and B+-tree with the Hilbert curve. Because the HR+-tree lacks robust implementation for concurrency and transaction controls, we focused on extending the B+-tree based indexes. There are two B+-tree alternatives: (1) B+-tree with 2D Hilbert curve for spatial index and B+-tree for temporal index (B2SBT), and (2) B+-tree with 3D Hilbert Curve for ST index (B3ST). B2SBT occupies less storage, but consumes higher search cost; while B3ST occupies more storage, but requires less search time.

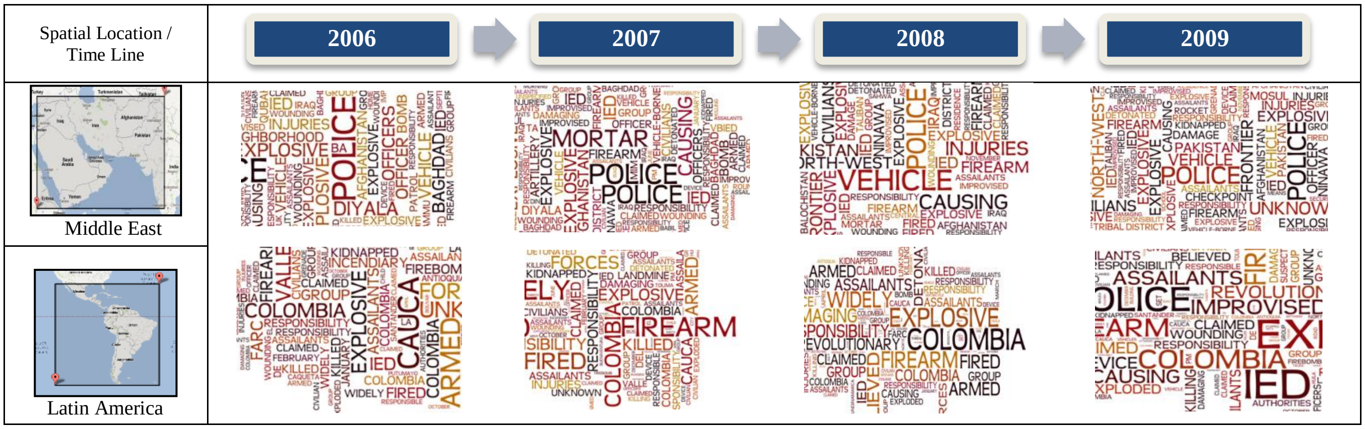
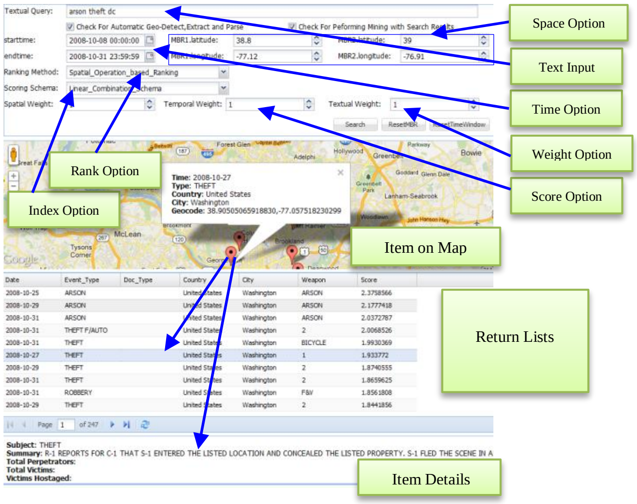


Figure 6. Topics in Middle East and Latin America from 2006 to 2009



Textual Query: arson theft dc
☑ Check For Automatic Geo-Detect,Extract and Parse ☑ Check For Peforming Mining with Search Results
starttime: 2008-10-08 00:00:00 MBR1.latitude: 38.8 MBR2.latitude: 39
endtime: 2008-10-31 23:59:59 MBR1.longitude: -77.12 MBR2.longitude: -76.91
Ranking Method: Spatial_Operation_based_Ranking
Scoring Schema: Linear_Combination_Schema
Spatial Weight: 1 Temporal Weight: 1 Textual Weight: 1

Time: 2008-10-27
Type: THEFT
Country: United States
City: Washington
Geocode: 38.90505065918830,-77.057518230299

Date	Event_Type	Doc_Type	Country	City	Weapon	Score
2008-10-25	ARSON		United States	Washington	ARSON	2.3758566
2008-10-29	ARSON		United States	Washington	ARSON	2.1777418
2008-10-31	ARSON		United States	Washington	ARSON	2.0372787
2008-10-31	THEFT F/AUTO		United States	Washington	2	2.0068526
2008-10-31	THEFT		United States	Washington	BICYCLE	1.9930369
2008-10-27	THEFT		United States	Washington	1	1.933772
2008-10-29	THEFT		United States	Washington	2	1.8740555
2008-10-31	THEFT		United States	Washington	2	1.8659625
2008-10-31	ROBBERY		United States	Washington	F&W	1.8561808
2008-10-29	THEFT		United States	Washington	2	1.8441856

Page 1 of 247

Subject: THEFT
Summary: R-1 REPORTS FOR C-1 THAT S-1 ENTERED THE LISTED LOCATION AND CONCEALED THE LISTED PROPERTY. S-1 FLED THE SCENE IN A
Total Perpetrators:
Total Victims:
Victims Hostaged:

Figure 7. STT search engine

2.3.2 *STT Searcher*

STT searcher is based on the STT indexer. We first perform textual search with inverted files. Then we process the ST search with one of our ST indexes (HR+-tree, B2SBT or B3ST). HR+-tree and B3ST simultaneously perform queries on both spatial and temporal attributes. B2SBT performs two searches (space and time) and join the results. Finally, the ST results are joined with the textual results.

2.3.3 *STT Ranker*

STT scores for ranking are computed based on different scoring schemes. Textual score is calculated based on term frequency-inverse document frequency (TF-IDF) model [4, 10]. Spatial score is calculated with spatial Euclidean distance, geographical metrics (contain, inside, overlap, nearby) based similarity, and geographic coverage between document and query [7]. Temporal score is calculated with time distance, which is inversely proportional to the length between document time and query time. We implemented multiple combination schemes to provide several ranking types, including linearly combining STT scores, the product of STT scores, the maximization of STT scores, and step linear combination [7]. Although the combination of STT scores can model relations between documents and query under the STT constraints, they are unable to account for the inherent STT concepts or context that reside in the documents. This motivates us to propose an STT topic modeling ranker to score documents based on STT topic/theme relevance.

2.3.3.1 *STT Topic Modeling Ranker*

Topic modeling improves the traditional ranking model by embodying textual thematic knowledge and inherent ST features into the ranking scores.

We first filter the dataset with spatial MBR and temporal period. Then we learn STT topics from the filtered results and textual inputs based on LDA and infer STT topic proportions for both documents and query, where we approximate exact solution with Gibbs Sampling [11]. Finally, the relevance scores between documents and query are computed with normalized STT topic proportions using cosine similarity under the topical space. With the extracted topics, the STT ranker incorporates critical thematic features (and thus contextual information) into its relevancy measure.

3. CONCLUSION

The ensemble framework which integrates STT IR and STT KD is developed to enhance the performance of the entire search and analysis work flow. The STT IR employs indexing approaches that accelerate both the processing of search queries and pattern mining. A variety of relevancy ranking schemes is proposed to model different STT relations that consider contextual information through topic modeling. The ST analytics provides predictions, hotspots, correlations analysis, and topic exploration to facilitate the identification of spatial trends and patterns. These analytical algorithms leverage the STT access methods in STT IR to rapidly process large and complex datasets. The case studies demonstrate that non-trivial and crucial ST patterns can be learned from crime and terrorism datasets.

4. REFERENCES

[1] "September 11 attacks," in http://en.wikipedia.org/wiki/September_11_attacks.

[2] X. Liu, C. Jian, and C.-T. Lu, "A spatio-temporal-textual crime search engine," in Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2010, pp. 528-529.

[3] L. K. Siebenecka, R. M. Medinaa, I. Yamadaa *et al.*, "Spatial and Temporal Analyses of Terrorist Incidents in Iraq, 2004-2006," *Studies in Conflict & Terrorism*, 2009.

[4] "Solr," in http://lucene.apache.org/solr/.

[5] Y. Tao, and D. Papadias, "Efficient Historical R-trees," *Proceedings of the 13th International Conference on Scientic and Statistical Database Management*, 2001.

[6] H. Sagan, "Hilbert's Space-Filling Curve," *Space-Filling Curves*, pp. 9-29: Springer-Verlag, 1994.

[7] C. Jian, "GIR and Ensemble Relevance Ranking Independent Study Report," 2011.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[9] Y. Huang, S. Shekhar, and H. Xiong, "Discovering colocation patterns from spatial data sets: a general approach," *IEEE Transactions on Knowledge and Data Engineering*, 2004.

[10] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," in Department of Computer Science, Rutgers University, Piscataway, NJ, 2002.

[11] D. Blei, L. Carin, and D. Dunson, "Probabilistic Topic Models," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 77-84, 2010.