Demo Paper: A Spatio-Temporal-Textual Crime Search Engine

Xutong Liu, Changshu Jian, Chang-Tien Lu
Department of Computer Science, Virginia Tech, USA
{xutongl, csjian, ctlu}@vt.edu

ABSTRACT

This paper proposes a STT(spatio-temporal-textual) search engine for extracting, indexing, querying and visualizing crime information. Until recently, it's a labor-intensive work to identify crime entities, cluster similar suspect activities, and discover patterns from massive online collections. It's a big challenge to reveal inherent ST(spatio-temporal) correlations among mass crime information. It's getting more difficult considering the subjectivity and vagueness of information retrieval from narratives of victims or witness and online documents of social networks. We have developed a crime search engine for Washington DC metropolitan area that includes geo-temporal-tagger, STT indexer, heuristic query and ranker and dynamical ST visualization. It assists crime detection for investigators, identification of crime trends and patterns for decision makers and researchers, and security of city life for residents and journalists.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Storage and Retrieval

Keywords

spatio-temporal search engine, crime analysis

1. INTRODUCTION

Crime activity reports available from victims, governmental organizations, news press, and social networks play a significant role in public safety, including crime prevention, suppression and investigation, uniformed patrol and response. These reports are organized either in well formed format, for example, public crime data from police, or in unstructured documents, for example, narratives from victims or witness. Crime information has inherent correlation in terms of spatial and temporal analysis and reasoning. Many spatial visualizer, for example CrimeStat[4], ST(spatio-temporal) visualizer [7] have been developed to alleviate the labor-intensive crime analysis work of law enforcement. However, this type of application only applies to data sets queried from traditional database. A spatio-textual search engine, for example, STEWARD[6, 3] can significantly contributes in this field although it lacks temporal features.

We developed a crime STT(spatio-temporal-textual) search engine for Washington DC metropolitan area to take advantages of both ST visualization and information retrieval methodologies. Both STT access methods and query operation are efficient to help a crime investigator retrieve all relevant information during hypothesis formulation and evidence collection. Dynamic ST visualization generated from real-time aggregation or data mining simplifies the process of identifying crime trends and patterns.

In this paper, we report the experience in how to integrate STT access methods and dynamic ST visualization in a search engine. The demonstration includes criminal detection, criminal trends and patterns identification utilizing this STT crime search engine.

2. ARCHITECTURE

This system is developed based on the general model of a textual search engine with the enhancement of integrating a series of additional modules to deal with spatial and temporal features of crime information. In addition to the general search engine components, it contains a geo-temporal-tagger, STT indexer, heuristic query and ranker, ST visualization.

Geo-temporal-tagger OpenNLP[2] is used to extract spatial, temporal and textual features from unstructured documents including narratives from victims, witness and online criminal articles. A crime-specific thesaurus and local gazetteer are developed to improve the precision of Name Entity Recognition. All the location names generated from geo-tagger are transformed into a geographic location and temporal information are transformed into a TimeML file.

STT indexer The indexer synthesizing inverted filed index and spatio-temporal index to efficiently index and search crime features. HR-tree [8] is chosen as ST access methods and implemented as a plug-in that can be seamlessly integrated into Solr[1]. Other optimization strategies including Document-At-A-Time[9], Buffer Management, are adopted to improve the query efficiency.

Heuristic query and ranker Query Expansion is performed for both textual query and ST query to recall crime information to the maximum extent. A ranker and pseudo relevance feedback based on user's clicks is used to improve the information quality for criminal detection. The ranker orders query results according to the combination of scores from textual (TF-IDF), spatial (Euclidean distance), and temporal (closeness). The calculation method includes weighted linear combination, step linear combination or product.

ST Visualization ST visualization includes two methods to present the query results: 1) crime-by-crime; 2) clustering and aggregation. This information is rendered on both Maps and Timelines to simplify the navigation through query results. Clustering methods include k-means, DBScan, hierarchical clustering, and kernel density clustering. Aggregation allows for browsing the hierarchical information over time or space.

3. FEATURES

The system is capable of answering various queries for crime information, including criminal detection, criminal trends and patterns identification. Figure 1 shows a screenshot of the system that consists of five parts.

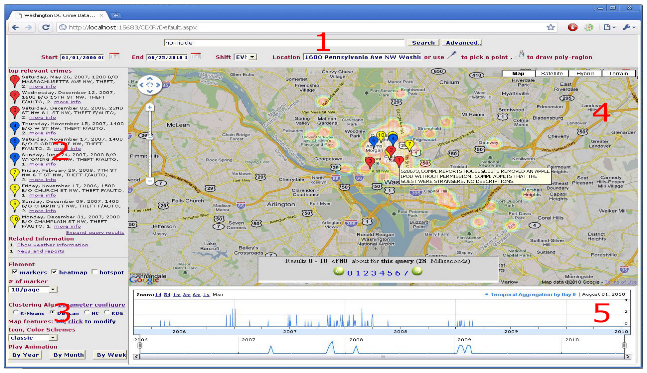
3.1 Criminal Detection

A crime investigator can explore and discovery relevant crimes due to the non-deterministic matching in the search process. The investigator constructs queries using arbitrary combinations of textual, spatial and temporal conditions (1 in Figure 1). The queries are expanded based on a crime domain thesaurus, Minimum Bounding Rectangle (MBR) based upward and downward spatial expansion, and temporal expansion. Retrieved documents are sorted based on their relevant scores and presented in a ST visualizer(2 4 5 in Figure 1). It recalls documents to the maximum extent to help the investigator build hypothesis. For example, when an investigator deals with a car break-in theft, he/she may start from ”car break-in, Capitol Hill, 6/1/2008”. ”car break-in” is expanded into a series of similar keywords with different weights, Capitol Hill is expanded into a MBR and ”6/1/2008” is expanded into a time window based on predefined rules. It helps the investigator fetch other crimes like ”driver’s window was smashed to remove a car navigator”, which contain no ”car break-in” keywords, or crimes that are very close to Capitol Hill but belongs to other residential neighborhoods.

3.2 Criminal Trends and Patterns Identification

Decision makers or researchers can use this system to identify criminal trends and patterns. Both heatmap and timeline are generated and associated for each query, which makes it possible to play heatmap animations by year, month or weekday. It simplifies the identification of criminal hot spots that are critical parameters of police patrol deployment and security device installation. Observations from aggregation over timeline can also find many interesting trends. For example, the timeline generated from the query on ”GPS iPOD, Washington DC, 2006-2010” indicates the crime wave introduced in [5]. Map features are also customizable, which can be used to identify the correlation between criminal patterns and map features, such as the relationship between ”theft f/auto” crime and business park lots.

This feature can also help Washington DC’s citizens and journalists have a better understanding about city security. For example, journalists can directly pick the interested region from the map and get the recent hotspot of the area instead of navigating crime-by-crime, and accordingly make a safer trip plan.



Washington DC Crime Data...

http://localhost:15683/CDIR/Default.aspx

homicide

Start 01/01/2006 0(End 06/25/2010 1 Shift EVI Location 1600 Pennsylvania Ave NW Washi or use to pick a point , to draw poly-ragion

top relevant crimes

1 Saturday, May 26, 2007, 1200 B/O MASSACHUSETTS AVE NW, THEFT, 2. more info

2 Wednesday, December 12, 2007, 1600 B/O 15TH ST NW, THEFT F/AUTO, 2. more info

3 Saturday, December 02, 2006, 22ND ST NW & L ST NW, THEFT F/AUTO, 2. more info

4 Thursday, November 15, 2007, 1400 B/O W ST NW, THEFT F/AUTO, 2. more info

5 Saturday, November 17, 2007, 1400 B/O FLORIDA AVE NW, THEFT F/AUTO, 2. more info

6 Sunday, June 24, 2007, 2000 B/O WYOMING AVE NW, THEFT F/AUTO, 1. more info

7 Friday, February 29, 2008, 7TH ST NW & T ST NW, THEFT F/AUTO, 2. more info

8 Friday, November 17, 2006, 1500 B/O CHURCH ST NW, THEFT F/AUTO, 1. more info

9 Sunday, December 09, 2007, 1400 B/O CHAPIN ST NW, THEFT F/AUTO, 2. more info

10 Monday, December 31, 2007, 2300 B/O CHAMPLAIN ST NW, THEFT F/AUTO, 1. more info

Expand query results

Related Information

1 Show weather information

1 News and reports

Element

☑ markers ☑ heatmap □ hotspot

# of marker

10/page

Clustering Alg parameter configure

○ K-Means ● DBScan ○ HC ○ KDE

Map features: all, click to modify

Icon, Color Schemes

classic

Play Animation

528673,COMPL REPORTS HOUSEGUESTS REMOVED AN APPLE IPOD WITHOUT PERMISSION. COMPL ADMITS THAT THE GUEST WERE STRANGERS. NO DESCRIPTIONS.

Results 0 - 10 of 80 about for this query.(28 Milliseconds)

Temporal Aggregation by Day 0 | August 01, 2010

Figure 1: a screenshot of the user interface: 1)query input panel, 2) relevant ranking results, 3) parameters setting, 4) geographical rending, 5)timeline

4. CONCLUSION

We attempt to integrate spatio-temporal-textual techniques and interactive spatio-temporal visualization into search engine in this project. The application in crime domain reveals its power of answering users’ information need of real-world scenarios. As exemplified in the features, it can dig in-depth information in crime investigation using query expansion and relevance feedback. Furthermore, it can generate hotspots, aggregations, or animations on query basis to simplify the identification of crime trends and patterns. It can significantly alleviate much of the labor-intensive and manual work of knowledge discovery of crime information.

5. REFERENCES

[1] http://lucene.apache.org/solr/.

[2] http://opennlp.sourceforge.net/.

[3] Y.-Y. Chen, T. Suel, and A. Markowetz. Efficient query processing in geographic web search engines. In *SIGMOD ’06: Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 277–288, New York, NY, USA, 2006. ACM.

[4] C. G. Heraux. Software review: Spatial data analysis of crime. *Soc. Sci. Comput. Rev.*, 25(2):259–264, 2007.

[5] A. C. John Roman. Is there an icrime wave? In *Urban Institute. Justice Policy Center (Washington, DC)*, pages 1–10, Urban Institute. Justice Policy Center (Washington, DC), 2007. Urban Institute.

[6] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. Steward: architecture of a spatio-textual search engine. In *GIS ’07: Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, pages 1–8, New York, NY, USA, 2007. ACM.

[7] S. K. Lodha and A. K. Verma. Spatio-temporal visualization of urban crimes on a gis grid. In *GIS ’00: Proceedings of the 8th ACM international symposium on Advances in geographic information systems*, pages 174–179, New York, NY, USA, 2000. ACM.

[8] Y. Tao and D. Papadias. Efficient historical r-trees. In *SSDBM ’01: Proceedings of the 13th International Conference on Scientific and Statistical Database Management*, page 223, Washington, DC, USA, 2001. IEEE Computer Society.

[9] H. Turtle and J. Flood. Query evaluation: strategies and optimizations. *Inf. Process. Manage.*, 31(6):831–850, 1995.