Using Map-Based Interactive Interface for Understanding and Characterizing Crime Data in Cities

Zhenxiang Chen[1](✉), Qiben Yan[2], Lei Zhang[1], Lizhi Peng[1],
and Hongbo Han[1]

[1] University of Jinan, Jinan 250022, People's Republic of China
{czx,Zhanglei,plz,nic_hanhb}@ujn.edu.cn
[2] Shape Security, Mountain View, CA 94040, USA
yanqiben@gmail.com

Abstract. Crime Data Analysis is vital to all cities and has become a major challenge. It is important to understand crime data to help law enforcement and the public in finding solutions and in making decisions. Data mining algorithms in conjunction with information system technologies and detailed public data about crimes in a given area have allowed the government and the public to better understand and characterize crimes. Furthermore, using visualization tools, the data can be represented in forms that are easy to interpret and use. This paper describes the design and implementation of a map-based interactive crime data web application that can help identify spatial temporal association rules around various types of facilities at different time, and extract important information that will be visualized on a map.

Keywords: Map-based interactive interface · Crime Data · Data Mining · Association rules

1 Introduction

A major challenge facing law enforcement agencies and the public is understanding crime patterns to predict future occurrences of offences. For example, in areas where there are repetitive offences, these offences can be spatially and/or temporally related. The banks in a local shopping center with low police coverage might be targets for robbery crimes. Traditional data mining techniques such as spatial-temporal association rules can be used to determine and predict these crime patterns.This information will help law enforcement and the public in making decisions and finding solutions, i.e. improving police coverage, avoiding certain locations at certain times, enhancing visibility, etc. A traditional data mining technique that has been used in crime data mining is the association analysis technique. It discovers frequently occurring items in a database and presents the association rules between events.

Y. Tan et al. (Eds.): ICSI-CCI 2015, Part III, LNCS 9142, pp. 479–490, 2015.
DOI: 10.1007/978-3-319-20469-7_51

In this study, the association rule technique will be used to determine crime patterns based on a spatial-temporal crime data. The association rules will be derived from patterns found in the dataset which will predict the occurrences of similar crimes pattern or other related crimes. The occurrences of some crimes can also be related to the occurrences of other crimes in an area.This research describes the design and implementation of an integrated data mining web application that can help identify spatial-temporal patterns and extract useful information from a crime database system. The data will be visualized on a map which can help researchers, law enforcement, and the public explore the dataset and retrieve valuable information. The components of such application consist of a database engine, a data analysis component, and a visualization Interface.

The remainder of this paper is organized as follows. Section 2 of this research discusses some related work in the field of crime data mining. Section 3 of this research discusses the proposed data analysis approach, and section 4 focuses on the detailed design of the web application. Section 5 presents the implementation of web interface. We show the association results in Section 6. Finally, we conclude this paper in section 7.

2 Related Work

There are several related work published in the field of crime data mining. Some of these publications are written on the subject of spatial and temporal pattern discovery and association rule mining techniques [1-6], while others are focusing on real-world applications [7-9]. The most related works to our study are [2, 8]. In [8], a general framework is suggested for crime data mining with some examples. The paper presents a general framework for crime data mining that draws from the experience gained with the Coplink project. The paper emphasizes using available data mining techniques to increase efficiency and reduce data analysis errors for real-world applications of law enforcement organizations. Our study is related to [8] in the sense that this general framework helps our study define different crime types and related security concerns. The suggested framework and the case study are informative and beneficial for this study.

In [2], the author discusses the methods for mining spatial association rules in a geographic information system (GIS) database. This paper demonstrates the effectiveness and efficiency of developing an association rule algorithm that can be used to provide spatial association rules based on users specific query. This paper provides methods that efficiently explore spatial association rules at multiple approximation and abstraction levels. Our study is related to [2] in the sense that a similar spatial association rule and technique can be adopted in exploiting the spatial association rules for crime data and facilities suggested in this study. Our method is different from [2] in three aspects: (1). we consider crime data applications; (2). we also consider many non-spatial objects in our rule discovery process; (3). our user query is more naïve than the complicated users query dealt with in their study.

3 Proposed Approach

Our general data analysis engine consists of three major parts: data preprocessing, ranking service and spatial association service. The data preprocessing will import the data sets, delete incomplete (or invalid) data and list the data in a way facilitating efficient data retrieval. In our project, we have two data sets to deal with: the crime data sets around the CITY1 area and spatial data sets for different facilities around the same area. The importation of data sets is explained in section 4.3. After the data importation, the crime data are listed in the database server, while the spatial data for facilities are also listed in the same database server. During the data importation, we search for the invalid crime data which are discovered to miss some important attributes for analysis, and delete them afterwards. Finally, both crime data and spatial data for facilities are ordered and organized in a way that efficient data retrieval can be performed.

3.1 Crime Ranking Service

Crime ranking service is the first service provided by the data analysis engine. Rank is one important outcome which is quite meaningful to the users, and can be visualized in various graphical manners. The users are prompted to input the addresses that they are interested in, which can be their travelling destinations, home addresses, workplaces, etc. Also, the users need to specify an area range around this specific location. The users can also input the time period that they are concerned with, and by default which is that of the whole crime data sets.

According to the users query, the crime data inside the interesting area within the time period are retrieved. We provide two ranking results for one specific query: (1). rank by crime types; (2). rank by TPD (time per day). To obtain the first ranking result, we can compute the numbers of crimes for each crime type in the retrieved data. Then, it becomes straightforward to rank different crime types according to their calculated occurrence frequencies. This rank provides the users a sense of which types of crimes are most likely to happen in the specified area. For the second ranking result, we first specify different time notations in one day, listed as:(1)Day:8am to 4pm (2)Evening:4pm to 6pm; (3)Mid-Night:12pm to 8am.Without losing generality, other notations can also be used in this service.

3.2 Multiple-Level Spatial Association Rules Discovery

The spatial association service is another major service for data analysis. The main technique follows the efficient algorithm for mining spatial association rules in [2]. Each spatial association rules will be evaluated according to their support and confidence:

1)The support of a conjunction of predicate P in a set S, denoted as $\sigma(P/S)$, is the number of objects in S which satisfy P versus the cardinality (i.e. the total number of objects) of S.

2)The confidence of a rule P→ Q is in S, ψ(P→ Q/S), is the ratio of σ(P∧ Q/S) versus σ(P/S), i.e. the possibility that Q is satisfied by a member of S when P is satisfied by the same member of S.

The users are mostly interested in the patterns which occur relatively frequently (or with large supports) and the rules which have strong implications (or with large confidence). The rules with the above patterns with large supports and large confidence are called strong rules. The main techniques of the association rule discovery service are illustrated as follows. Our data have the following database relations for representing spatial and nonspatial objects: (1)facility (name, type, geo);(2)crime (type, geo, time) and (3) TPD (name).

In the above relationship schemata, "geo" represents the spatial object. The facility "types" correspond to the types of facilities such as schools, hospitals, etc. "TPD" represents the time per day object, illustrated in the previous section. We consider multiple-level association rules. For example, high-level association rules are written as follows:

close_to(facility) ∧is_in(nigh time)→has(crime)

Basically, we consider simple two-level association rules, concept hierarchies are defined as follows:

-A concept hierarchy for facility:(facility(hospital, school, bar, shopping center))

-A concept hierarchy for crime:(crime(theft, robbery, homicide))

-A concept hierarchy for TPD: (TPD(daytime(morning, afternoon),nighttime (evening, late night)))

For the coarse concept level, Table 1 will be built according to [2]:

Table 1. First/coarse concept level

location	facility	crime	TPD
$around(x1, y1)$	$< close\ to, facility >$	$< has, crime >$	$< is\_in, daytime >$
$around(x2, y2)$	$< close\ to, facility >$	$< has, crime >$	$< is\_in, nighttime >$
...	...	...	...

The above location is the location of the corresponding facility. The size of this table is equal to the number of facilities considered. According to the above coarse concept level table, we can compute the counts for large k-predicate set. During the computation, we compare the number of counts to the thresholds for minimal support and confidence. The entries with lower support or confidence than the thresholds are filtered out.Then, refined computations of detailed spatial relationships at the second concept level are done based on the following Table 2:

After that, the large k-predicate set for the second concept level can also be derived. Finally, the corresponding mining rules can be discovered. The whole association rules discovery method proceeds as follows:

Step1: Prepare the spatial relationships for two different concept levels;

Table 2. Second/fine concept level

location	facility	crime	TPD
$around(x1, y1)$	$< close\ to, hospital >$	$< has\_crime, theft >$	$< is\_in, morning >$
$around(x2, y2)$	$< close\ to, school >$	$< has\_crime, robbery >$	$< is\_in, afternoon >$
...	...	...	...

Step2: Compute the counts of k-predicates for the coarse concept level;

Step3: Filter out the predicates with supports or confidence lower than the minimal thresholds at the first level;

Step4: Compute the counts for k-predicates of refined concept level;

Step5: Find the large predicate at the refined concept level and find association rules.

3.3 Overall Structure for Data Analysis

The overall structure starts from users query, as illustrated in Fig. 1. Because we only have fixed crime and facility data sets, we are not able to update the association rules. Therefore, we only show the pre-computed association rules for the users. How to efficiently generate association rules for newly input data with an online manner will be a promising future direction.

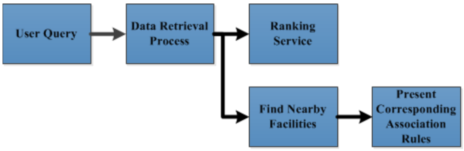


Fig. 1. Data Analysis Procedure

4 System Design

4.1 System Architecture

The system development architecture is drawn in Fig.2. The data analysis engine will deal with data preprocessing and data association,as mentioned in section 3.

4.2 Facility and Crime Data

Facility and crime data in the system architecture of Fig.2 is obtained from the CITY1 GIS Catalog website. The crime data was in XML format that was imported into the database engine. The facility data was in a Shape-file format (ESRI format). After importing the facility data into ArcMap, the facilities

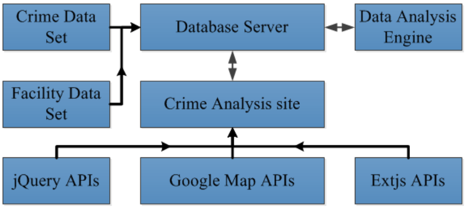


Fig. 2. Data Analysis Procedure

attributes tables were exported into an excel file. The most important attributes in the facility table are the type of a facility, its geographical location (x and y coordinates,) its address, its description, etc. Since the location data (x and y) are the projected values (in meters,) it must be converted to latitude and longitude (in degrees) using geographical coordinate system conversion methods (R tools.) Another way is to geocode an address of a facility using geocoding functions part of the Google Map API V3. We are planning to use the techniques of DIV and CSS to make the web-based application look more organized and fancy.

4.3 Data Set Preprocessing

The data set to be analyzed during this project has been provided to us. And all of the data we are planning to utilize are crime data sets which are associated with the areas around CITY1. With a relational database, we are able to analyze the crime data from various dimensions. Facility data in the system architecture of Fig.2 is obtained from the CITY1 GIS Catalog website. The facility data was in a Shape-file format (ESRI format file). The facilities attributes tables were exported into an excel file.Since we have the extracted text files for each facility where we separated each attribute with tabs, we are able to move forward to process every text file with a small program written in C♯. After we obtained all the text files we are willing to analyze, we can utilize the information inside to get the actual address for each facility. With obtaining the actual address to each facility, we can plug in the address into the function of geocoding, and get the precise latitude and the longitude as the result, in addition, we will generate a facility number to every facility, and then we can insert all the data we have into the database for the facilities. This is the basic idea of our geocoding part.

While the data preprocessing didnt just stop here, there are still something to be taken care of. For instance, when we are applying the geocoding to the addresses that we got from the text files, we cannot send the request too fast, since if we send the request too fast to the Google server, the Google server will send one error result back to us: "OVER_QUERY_LIMIT", which means the server is too busy to respond all the requests we are sending it. And if such situation happened the overall database will be ruined, since when the server is sending the "OVER_QUERY_LIMIT" back, the result latitude will not be updated, which means, several facility will share one same pair of latitude and longitude as a terrible result.

4.4 Visualization Component Architecture

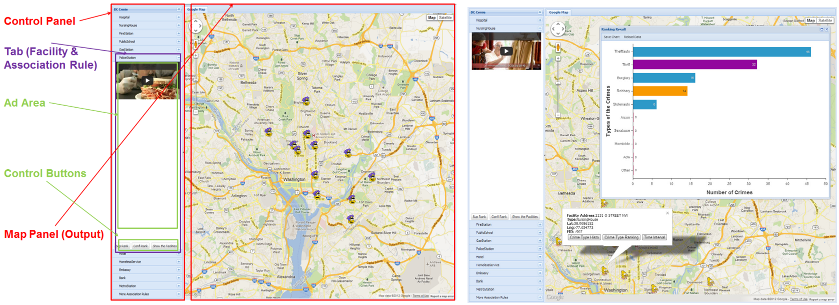
This is the visualization component architecture, it consists of several parts. The layout of the web-based application will be formed by the Ext Js 4.1 which is a Javascript application framework that works on all modern browsers from IE6 to the latest version of Chrome. It enables us to create the best cross-platform applications using nothing but a browser, and has a phenomenal API. With the utilization from the Ext Js 4.1 we can make our application more fancy and user friendly, for instance, with applying the APIs from libraries which are provided by the Ext Js, we can arrange our layout with windows, tabs, forms and charts that can make our web-based application visualized with a better understanding of the boring raw data output from the analysis.

The most important part of visualization is to show every facility on the Google map and show their ranking status and distribution to the users. Here we applied the functions of addMarkers (IMMarkerConfigs) and removeMarkers (IMCurMarkers) to control the markers on the Google map. The parameter of IMCurMarkers is an array of the result from the json data from upper steps. It contains the information about the latitude and the longitude, with these parameters, we are able to pinpoint the icons on the Google map. The layout of the web page is designed under the framework named Extjs, this is a javascript API library, we utilized some of its APIs to fulfill some of our web applications function and the layout format of our web page on the left hand side is named "accordion", under this layout each window can be expanded or collapsed, with this layout we can make our web page more compact.

5 System Implementation

We can firstly start from the association rules data table generation. The way we generate the association rules table is that, we utilize the two datasets that we have generated in the previous months, namely, the crime data and the facilities data. Since we have more than 120,000 crime cases, and 988 facilities contained in the facility dataset. It is much easier for us to generate an association rules table with the association rules starts from the facilities to the crime counts, for example, the order of our association rules table is like this: FID, OTYPE, LAT, LON, STOLENAUTO, ROBBERY, ARSON, THEFTFAUTO, SEXABUSE, HOMICIDE, BURGLARY, ADW, THEFT, OTHER, EVE, MID, DAY, UNK, ADD, which are the facility id number, facility type , the latitude of the facility, the longitude of the facility, number of stolen auto cases, number of robbery cases, number of arson cases, number of theft from auto cases, number of sex abuse cases, number of homicide cases, number of burglary cases, number of assault with deadly weapon cases, number of theft cases, other cases, number of case happened in the evening, midnight and daytime, the last one is the case with no specific time period is noted. This table is the heart for our research, since all the data generated here will be later visualized and displayed on our web-based application.

As a result to this approach, we obtained a table with expanded columns, from original 5 columns to finally 19 columns. Since some of the area near one specific facility was not containing any crime cases, we regarded these facilities as “safe” area, in order to make the further analysis more convenient, we also transformed this table into binary format that if the area contains no crime case, we assign binary number ”0” to the crime type, otherwise, we assign binary number ”1” to the crime type indicating that the type of crime was observed within the defined area around the facility. We applied the approach to obtain a better insight of the correlation between one specific facility and one type of crime in the perspective of association rule. We also take the time interval into consideration. According to the flaws in the raw data provided by the CITY1 catalog web site, all the crime data are lack of the information about what time the crime case was reported, for every case, the information in the column of reported time was all 00:00:00. The layout of our web-based application was written in Javascript mostly, and there are several API were utilized to fulfill the function of the interaction for the data and control requests. The Extjs Javascript APIs library was utilized mostly for the generation of our layout to the application. There are mainly two parts of components embedded in the layou. the control panel and the web-based interface was displayed in the Fig.3.



Facility Address:2131 O STREET NW
Type:NursingHouse
Lat:38.9086152
Lng:-77.054773
FID: :907

Fig. 3. Application Layout and Web-based Application Interface

In order to fulfill the function of data and command interaction, we applied two libraries to complete the functionality, the jQuery and the Extjs both applied Ajax technology. The basic idea to the data interaction is this: we formed a web service in the back stage of our website, and all the direct control commands such as the SQL queries are functioned in the web service class. And then, we applied the APIs provide by the Javascript libraries to send some web request back to the web service class formed in the back stage of the site via the Ajax technology. Later, as long as the web service get the web request it will send the direct SQL command to the database where store all the data we have generated earlier through our data analysis process. Then, the database will send all the requested data to the web service, after a brief data format transformation, the data will be sent to the front stage web application similarly via Ajax.

6 Results

We have obtained a lot of interesting association rules from crime to facility indicating which facility has high frequency of a specific crime, from crime to time per day indicating which time period has high frequency of a specific crime,in Fig.4. We have association rule from another direction, i.e. from facility to crime indicating which crime is most likely to happen in this specific facility. Also we have association rule from facility to time per day indicating which time-per-day is most likely to have crimes around this specific facility.

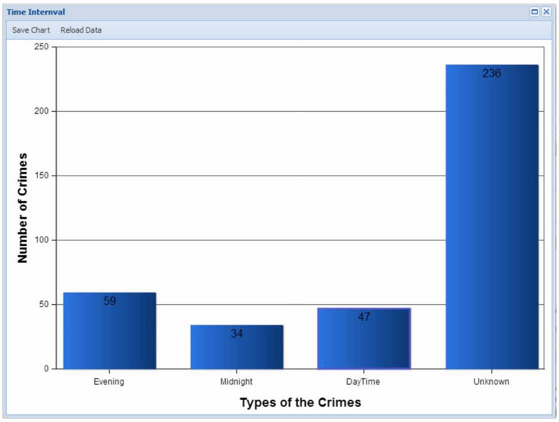
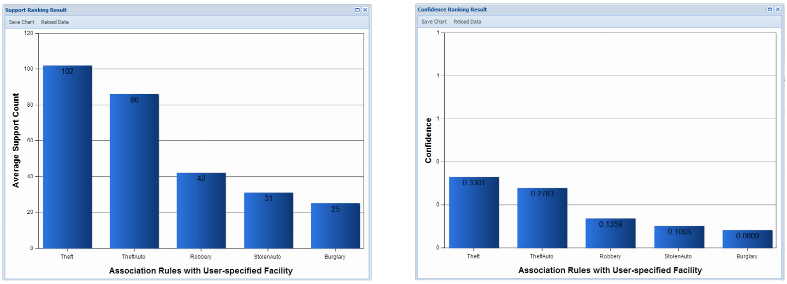


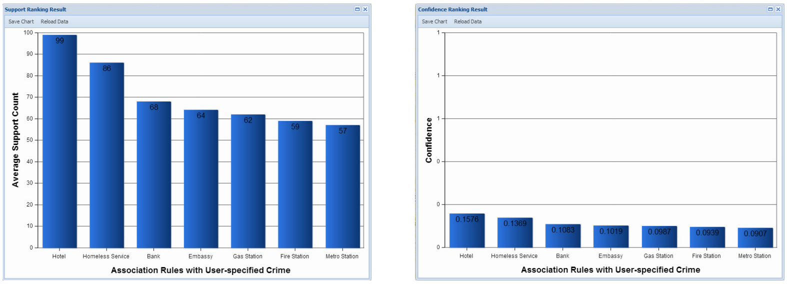
Fig. 4. Time Interval result

On the MAP interface, we only show the high frequency association results, and delete all the infrequent association results. We determine the high frequency by manually setting a threshold for support and confidence. In general, each case will generate five or six rules. We show the most useful results for our users. Homeless service area is the most dangerous place to stay. Homeless service area has highest density of StolenAuto crimes, Robbery crimes, SexAbuse crimes, Burglary, A sault with Deadly Weapon. Hotel has the highest density of Theft from Auto crime, because a lot of cars parked around hotel. Bank has highest density of Theft crimes, which is very straightforward. Criminals like to wait for the persons coming out of the banks to steal their money. One interesting point we find out is that Police station has very high density of ADW crimes, i.e. people are likely to use deadly weapons to deal with police officers. Many police stations are located in low income residential areas where crime is high, top five facilities Most crimes happen are Homeless Service,Public School,Metro Station,Hotel and Gas Station (Fig.5).



(a) The support count result (b) The confidence result

Fig. 5. Association Rules Results from Homeless Service Facility to Crimes

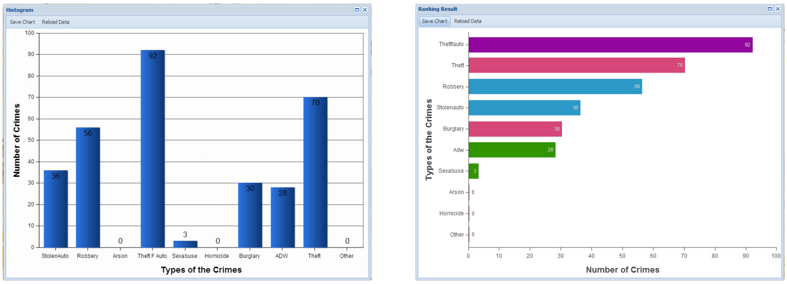


(a) The support count result (b) The confidence result

Fig. 6. Association Rules Results from Theft-From-Auto Crimes to Facilities

Regarding time, Bank, metro station and hotels have high density of crimes in day and evening. In midnight, Bank, Homeless service, and hotel have high density of crimes. Look at that, homeless service again is very dangerous even in mid night. For example, during the night areas around homeless shelters are usually crowded with homeless persons seeking shelters.From another association direction, we find that, around most facilities, theft and theft from auto are highest density crimes. Except in public school and police station, ADW is very high. Criminals are more likely to use deadly weapons around police station and public schools. So, stay safe if you are around these areas. Overall,around most facilities, the top five crimes are:Theft,Theft from Auto,Robbery,Stolen Auto,Burglary.(Fig.6)

Regarding time, most crimes happen during unknown time. If we do not consider unknown time. Around almost all facilities, evening has the highest crime density, midnight has the lowest , except around police station and embassy. Both of them have very high crime density during Midnight. Which probably means around police station and embassy, many officers are patrolling during midnight. So they will more likely to find crimes happening during midnight.Based on the findings of the study, it is vital for police and the public to consider such findings



(a) The histogram for the number of crimes

(b) Ranking for the number of the crimes

Fig. 7. The Crime Count Ranking Results

when making decisions. Police may consider where and how to allocate resources, what measures are needed to prevent crimes, and what other options are available. The public may also consider these finding to determine which areas are safe and which areas they should avoid.

When it comes to crimes occurring at specific locations(Fig.7), the crime patterns inferred will motivate the police to create a list of potential measures that may prevent or reduce these crimes. For example, at locations where crimes such as theft, burglary, robbery, auto theft, etc., the police may allocate resources such as personals, cameras, signs, and other security measures to prevent or ease the crime rate at these locations.In the specific case of locations such as Homeless Services and shelters where many different types of crimes occur, the police may find ways to reduce the crimes in these areas using several methods and techniques such as allocating more resources in the area for monitoring or through community programs and services. The same can be said about other locations such as police station, banks, Hotels, etc. Decisions can also be made regarding high density crimes during the night where crimes are much higher than during the day.

In regard to the public, they become more educated and aware of the hot spot areas.Their decisions become based on facts not on myths. As an example, a person may not wonder in a hot spot area in the middle of the night, a person may become more aware and alert when visiting a bank area where crime density is high, and finally, a person may avoid hanging around location that serves the homeless population. Most importantly the public will collaborate with police to reduce crime. Crime rates will become much less and safety is much better.

7 Conclusion

In this paper, we have laid out the objective of the project, technologies used, proposed approach, and the system design. Our main objective is to analyze a given crime dataset and present the information to users in a meaningful and more understandable forms so that it can be utilized by law enforcement and the general public in solving problems and making decision concerning safety. To achieve such objective, we proposed to implement a web based application.The application consists of several components: Database engine, Analysis engine, and a visualization component. The result of the project showed that such technique is very beneficial in presenting data meaningfully.The findings of the study showed the vital rule that data mining techniques plays in helping the police and the public make decision regarding crimes and safety. An informative data will enable the police to finding solutions by creating measures against crime. The same data will also help the public understand crime data and act upon it.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grants No.61472164 and No.61203105,the Natural Science Foundation of Shandong Province under Grants No.ZR2014JL042 and No.ZR2012FM010.

References

1. Mohan, P., Shekhar, S., Shine, J.A., Rogers, J.P.: Cascading spatio-temporal pattern discovery: A summary of results. In: Proceedings of the 10th SIAM Data Mining (SDM 2013), pp. 327–338. SIAM (2013)
2. Koperski, K., Han, J.: Discovery of spatial association rules in geographic information system. In: Egenhofer, M., Herring, J.R. (eds.) SSD 1995. LNCS, vol. 951, pp. 47–66. Springer, Heidelberg (1995)
3. Huang, Y., Shekhar, S., Xiong, H.: Discovering Colocation Patterns from Spatial Data Sets: A General Approach. IEEE Transactions on Knowledge and Data Engineering **16**(12), 1472–1485 (2004)
4. Clementini, E., Felice, P.D., Koperski, K.: Mining Multiple-level Spatial Association Rules for Objects with A Broad Boundary. Journal of Data and Knowledge Engineering **34**(3), 251–270 (2000)
5. Bogorny, V., Kuijpers, B., Alvares, L.: Reducing Uninteresting Spatial Association Rules in Geographic Databases Using Background Knowledge: A Summary of Results. International Journal of Geographical Information Science **22**(4), 361–386 (2008)
6. Bogorny, V., Kuijpers, B., Alvares, L.: Semantic-based Pruning of Redundant and Uninteresting Frequent Geographic Patterns. GeoInformatica **14**(2), 201–220 (2011)
7. Liu, X., Jian, C., Lu, C.T.: Demo paper: A spatio-temporal crime search engine. In: Proceedings of the ACM GIS 2012, pp. 528–529. ACM (2012)
8. Chen, H., Jie, J., Qin, Y., Chau, M.: Crime Data Mining: A General Framework and Some Examples. IEEE Computer **37**(4), 50–56 (2004)
9. Shah, S., Bao, F., Lu, C.T., Chen, I.R. : CROWDSAFE : Crowd sourcing of crime incidents and safe routing on mobile devices (demo paper). In: Proceedings of the ACM GIS 2014, pp. 521–524. ACM (2014)