A Framework for Mining Co-evolving Spatial Events

Jin Soung Yoo and Shashi Shekhar
Department of Computer Science and Engineering
University of Minnesota, Minneapolis, MN
jyoo,shekhar@cs.umn.edu

Abstract

A spatial co-located event set represents a subset of spatial events whose instances are located in a spatial neighborhood. The discovery of co-evolving spatial event sets involves finding co-located event sets whose spatial prevalence variations over time are similar to a specific query sequence. For example, the frequency of drought and wild fire events in Australia over the last 50 years shows similarity with an El Niño index sequence. Mining co-evolving spatial event sets is computationally challenging due to the high computational cost of finding co-located event instances on continuous geographic space, large temporal space and a composite interest measure, i.e., the spatial prevalence time sequence of a co-located event set. We propose a framework for mining co-evolving spatial event sets from a spatio-temporal dataset.

I. Introduction

A spatial co-located event set represents a subset of spatial events whose instances are located in a spatial neighborhood area. Examples of spatial events include outbreaks of disease, climate observations, distributions of plant species, crime hot-spots, mobile service request types, etc. Frequent spatial co-located event sets, i.e., co-location patterns [9], [11], [12], give important insights for many application domains such as earth science, ecology, public health, business, etc. However, the spatio-temporal nature of datasets generated in the various application domains raises intriguing questions regarding co-location pattern analysis. Scientists in these domains are often interested in understanding the evolution of co-location patterns among events. In this paper, we tackle the problem of temporal aspects of co-location pattern analysis, i.e., how the co-location patterns change over time. Specifically, we focus on identifying *co-evolving spatial event sets*, i.e., co-located event sets whose temporal occurrences are correlated with a special time series. If we consider a specification to be the similarity with El Niño index values, one example is the frequent co-occurrences of climate events with the El Niño phenomenon over the last 50 years [15]. El Niño, an abnormal warming in the eastern tropical Pacific Ocean[7], has been linked to climate phenomena such as droughts and wild fires in Australia [10]. NASA Earth scientists are very interested in understanding the tele-connection between special events such as El Niño and various related events, such as an increase of rain in the Midwest or drought or other seasonal changes in particular locations around the world [1].

Mining co-evolving spatial event sets presents challenges due to the following reasons: First, identifying spatial co-located event sets is computationally expensive by itself since the instances of spatial events are embedded in a continuous space and share neighbor relationships. Second, we have to consider a composite interest measure, e.g., a spatial prevalence time sequence, rather than a scalar numeric interest measure such as a spatial prevalence value. Exponentially increasing computational costs of generating the spatial prevalence time sequences of all combinatorial candidate event sets become prohibitively expensive. Third, the similarity functions for measuring the degree of consistency with a query time sequence are also computationally expensive with increases of time space. In this paper, we propose a framework to efficiently mine co-evolving spatial event sets from a spatio-temporal dataset. The detailed description of this work is presented in [13].

II. Problem Statement and Related Work

We provide the formal problem statement for the discovery of co-evolving spatial event sets, and then discuss related work.

A. Problem Statement

Given:

1) A spatial framework SF
2) A time framework TF which can be divided into a set of disjoint time slots, $TF = t_0 \cup \ldots \cup t_{n-1}$.
3) A set of spatio-temporal events $E = \{e_1, \ldots, e_m\}$ and a set of their instance objects ST where each instance object $\in ST$ is a vector $<$ event type, instance id, location, time $>$, where location $\in SF$ and time $\in TF$.
4) A spatial neighbor relationship R over locations
5) A query time sequence $\vec{Q} =< q_0, \ldots, q_{n-1} >$ over TF
6) A time sequence similarity function: $f_{similarity}(\vec{P}, \vec{Q})$
7) A similarity threshold θ.

Develop:

An algorithm to find spatial co-located event sets whose prevalence variations over time are similar to a given query time sequence.

Objective:

Find a complete and correct set of co-located events $C \subseteq E$ which satisfies $f_{similarity}(\vec{P_C}, \vec{Q}) \leq \theta$, where $\vec{P_C} =< p_0, \ldots, p_{n-1} >$ is the time sequence of spatial prevalence values of a co-located event set C over time slots $t_0, \ldots, t_{n-1}$.

B. Related Work

To our knowledge, researchers have yet to tackle the problem of mining co-evolving spatial event sets. In the spatial association mining literature, [6], [9], [11], [12] proposed different approaches for mining spatial co-location patterns. [6] adopted space partitioning for identifying neighboring objects for frequent neighboring feature sets, and used support count as the interest measure. [9] defined a statistically meaningful interest measure for spatial co-location patterns and proposed an instance join-based co-location mining algorithm. [11], [12] proposed to materialize spatial neighbor relationships for efficient co-location pattern mining. However, none of these works considers the temporal domain of the co-location pattern. Otherwise, in the temporal association mining literature, recent efforts have attempted to capture special temporal profiles of association patterns in market basket transaction datasets. [8] identified cyclic association rules, which discover periodically repetitive frequent patterns. [5] explored the problem of finding frequent itemsets along with calendar-based patterns which are defined with a calendar schema, e.g, year, month, and day. [14] proposed a similarity-based time-profiled association mining in a time-stamped transaction dataset. These methods are not directly applicable for mining co-evolving spatial event sets since there is no explicit transaction concept in a spatio-temporal dataset.

III. A Framework for Mining Co-evolving Spatial Event Sets

A. Modeling Co-evolution Patterns

The participation index measure [9] has been successfully used in spatial co-location mining since it represents the spatial statistical significance of a pattern. We use a time sequence of participation index values as an interest measure for the variation of prevalence of a co-located event set over time.

Definition 1: Given a spatio-temporal dataset $ST = ST_0 \cup \ldots \cup ST_{n-1}$ where ST_i is a set of spatio-temporal event objects occurring in time slot i, i=0, $\ldots, n-1$, the **spatial prevalence time sequence** of a co-located event set C, $\vec{P}_C$ is the sequence of the participation index values of C over time slots, i.e., $\vec{P}_C =< PI_{ST_0}(C), \ldots, PI_{ST_{n-1}}(C) >$, where $PI_{ST_j}(C)$, $0 \leq j < n-1$, is the participation index value of a co-located event set C in a dataset ST_j at time slot j.

Next, we propose using *normalized Euclidean distance* [3] as a similarity function between a query time sequence and a prevalence time sequence.

Definition 2: For two time sequences $\vec{P} = < p_0, \ldots, p_{n-1} >$ and $\vec{Q} = < q_0, \ldots, q_{n-1} >$, the normalized Euclidean distance between $\vec{P}$ and $\vec{Q}$, $D(\vec{P}, \vec{Q})$, is defined as $D(\vec{P}, \vec{Q}) = \sqrt{\frac{\sum_{i=0}^{n-1}(p_i - q_i)^2}{n}}$, where n is the number of time slots.

B. Algorithmic Design Concepts

A naive method for finding co-evolving spatial event sets can follow a two-step procedure. First, it finds spatial instances of all possible co-located event sets, calculates their participation index values, and generates their prevalence time sequences over all time slots. Second, it searches the generated prevalence time sequences similar to a query sequence. In this step, advanced time series search methods using spatial indexing schemes [2], [4] can be used. However, as the number of both the event types and the time points increases, the computation cost to calculate the spatial prevalence values of all combinations of event sets becomes prohibitively expensive. We present our algorithmic design concepts to combine the generation of prevalence time sequences with the sequence search.

B.1. Co-located Event Instance Filtering

Identifying the instances of co-located event sets having a clique neighbor relationship is computationally expensive since the instances of spatial events are embedded in a continuous space. [12] proposed first to find relatively inexpensive star neighbor relationships from an input dataset instead of finding all maximal clique relationships directly. We adopt this join-less co-location mining approach [12] to find co-located clique instances from co-located star instances at each time slot.

B.2. Similar Co-located Event Set Filtering

We propose several filtering steps for efficiently finding co-evolving co-located event sets. First, we present an important property related to our spatial prevalence time sequence and similarity function.

Definition 3: For a prevalence time sequence $\vec{P} = < p_0, \dots, p_{n-1} >$ and a query time sequence $\vec{Q} = < q_0, \dots, q_{n-1} >$, the **lower bounding distance** between $\vec{P}$ and $\vec{Q}$ is defined as $D_{lb}(\vec{P}, \vec{Q}) = \frac{\sum_{i=0}^{n-1} f(p_i, q_i)}{n}, where\, f(p,q) = 0,\, if\, p \geq q,\, and\, f(p,q) = (q-p)^2,\, if\, p < q.$

The lower bounding distance considers the subsequences of the prevalence time sequence and the query time sequence at time slots where the element participation index is less than the corresponding query sequence value.

Lemma 1: The lower bounding distance between the prevalence time sequence of a co-located event set and a query time sequence is **monotonically non-decreasing** with size of the co-located event set. Lemma 1 ensures that the lower bounding distance can be used to effectively reduce the co-evolving co-located event set search space. Next, we propose filtering schemes using the monotonicity property.

1) Event-level filtering : We have two event-level filtering procedures that reduce examining co-located event instances and calculating true prevalence time sequences. The first event-level filtering prunes a candidate set if the lower bounding distance of a subset of the candidate event set does not satisfy a given similarity threshold. The second event-level filtering is done by the estimated upper bound of the prevalence time sequence of a candidate event set and its lower bounding distance. We define the event-level upper bound of the prevalence time sequence of a co-located event set using the prevalence time sequences of its subsets.

Definition 4: Let C_k be a size k co-located event set and $A = \{C_{k-1}^1, \dots, C_{k-1}^k\}$ be a set of all size $k-1$ subsets of C_k, where $C_{k-1}^j \subset C_k$, $1 \leq j \leq k$. Let $\vec{P}_{C_{k-1}^j} = < p_0^{C_{k-1}^j}, \dots, p_{n-1}^{C_{k-1}^j} >$ be the spatial prevalence time sequence of $C_{k-1}^j \in A$. The **event-level upper bound** of the spatial prevalence time sequence of C_k, $\vec{EU}_{C_k} = < u_0^{C_k}, \dots, u_{n-1}^{C_k} >$ is $< min\{p_0^{C_{k-1}^1}, \dots, p_0^{C_{k-1}^k}\}, \dots, min\{p_{n-1}^{C_{k-1}^1}, \dots, p_{n-1}^{C_{k-1}^k}\} >$.

2) Coarse filtering : We have a scheme to filter candidate event sets before doing expensive clique check operations for finding the co-located event instances. We explore the instance-level upper bound of the prevalence time sequence using the star instances of a candidate co-located event set.

Definition 5: Let C_k be a size k co-located event set. The **instance-level upper bound** of the prevalence time sequence of C_k, $\vec{IU}_{C_k} = < u_0^{C_k}, \dots, u_{n-1}^{C_k} >$ is $< p'_0, \dots, p'_{n-1} >$, where p'_i is the participation index of star instances of a co-located C_k at time slot i, where $1 \leq i \leq n-1$.

If the lower bounding distance to the instance-level upper bound of a co-located event set does not satisfy a given threshold, the candidate co-located event set is pruned.

3) Refinement filtering : Finally, we generate prevalence time sequences using the true participation index values from exact clique co-located instances, and find similar co-located event sets satisfying a given threshold value.

IV. Conclusions and Future Work

We explored the problem of mining co-evolving spatial event sets from spatio-temporal data, and presented a framework to discover co-evolving co-located event sets. The proposed framework is expected to substantially reduce the search space of spatio-temporal event sets using several filtering schemes. In the future, we plan to examine the behavior of our mining algorithm with real-world datasets, e.g., NASA climate data. In addition, we plan to consider other similarity functions[4] rather than a Euclidean distance based measure for discovering co-evolving spatial event sets, and to explore the computational structures in spatio-temporal data.

REFERENCES

[1] NASA Workshop on Issues in the Application of Data Mining to Scientific Data. In *http://datamining.itsc.uah.edu/meeting06/*, October 19-21, 1999.

[2] C. Faloutsos, M. Ranganathan, and Y.Manolopoulos. Fast subsequence matching in time-series database. In *Proc. of the ACM SIGMOD Conference*, 1993.

[3] Dian Q. Goldin, Todd D. Millstein, and Ayferi Kutlu. Bounded similarity querying for time-series data. *Information and Computation*, 2004.

[4] D. Gunopulos and G. Das. Time Series Similarity Measures and Time Series Indexing. *SIGMOD Record*, 30(2), 2001.

[5] Y. Li, P. Ning, X. S. Wang, and S. Jajodia. Discovering Calendar-Based Temporal Assocation Rules. In *Proc. of the International Symposium Temporal Representation and Reasoning(TIME)*, 2001.

[6] Y. Morimoto. Mining Frequent Neighboring Class Sets in Spatial Databases. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001.

[7] NOAA. El Nino Page. http://www.elnino.noaa.gov/.

[8] B. Ozden, S. Ramaswamy, and A. Silberschatz. Cyclic Association Rules. In *Proc. of the IEEE International Conference on Data Engineering*, 1998.

[9] S. Shekhar and Y. Huang. Co-location Rules Mining: A Summary of Results. In *Proc. of the International Symposium on Spatio and Temporal Database*, 2001.

[10] G. H. Taylor. Impacts of el nino on southern oscillation on the pacific northwest. http://www.ocs.orst.edu/reports/enso_pnw.html.

[11] J.S. Yoo and S. Shekhar. A Partial Join Approach for Mining Co-location Patterns. In *Proc. of the ACM International Symposium on Advances in Geographic Information Systems(ACM-GIS)*, 2004.

[12] J.S. Yoo and S. Shekhar. A Join-less Approach for Co-location Pattern Mining: A Summary of Results. In *Proc. of the IEEE International Conference on Data Mining(ICDM)*, 2005.

[13] J.S. Yoo, S. Shekhar, S. Kim, and M. Celik. Discovery of Co-evolving Spatial Event Sets. In *Proc. of the SIAM International Conference on Data Mining(SDM), http://www-users.cs.umn.edu/ jyoo/source/publication.html*, 2006.

[14] J.S. Yoo, P. Zhang, and S. Shekhar. Mining Time-Profiled Associations: An Extended Abstract. In *Proc. of the Pacific-Asia Conference on Data Mining and Knowledge Discovery(PAKDD)*, 2005.

[15] P. Zhang, M. Steinbach, V. Kumar, S. Shekhar, P. Tan, S. Klooster, and C. Potter. Discovery of Patterns of Earth Science Data Using Data Mining. In Mehmed M. Kantardzic and Jozef Zurada, editors, *Next Generation of Data Mining Applications*. IEEE Press, 2004.