Proceso de Descubrimiento de Patrones de Co-Localización alrededor de Tipos de Eventos de Referencia

Giovanni Daián Rottoli [1,2,3], Hernán Merlino[3], Ramón García-Martinez [3]

1 Programa de Doctorado en Ciencias Informáticas. Universidad Nacional de La Plata. Argentina.
2 Programa de Becas "Formación de Doctores para Fortalecer áreas de I+D+i". Universidad Tecnológica Nacional. Argentina
3 Grupo de Investigación en Sistemas de Información. Universidad Nacional de Lanús. Argentina.
gd.rottoli@gmail.com, hmerlino@gmail.com, rgm1960@yahoo.com

Resumen. El descubrimiento de patrones de co-localización revela subconjuntos de tipos de eventos espaciales cuyas instancias ocurren frecuentemente vecinas entre sí. Muchos algoritmos y métodos han sido desarrollados a través de los años, sin embargo, cuando se requiere encontrar estos patrones alrededor de tipos de eventos espaciales determinados, la alternativa existente resulta incompleta e incorrecta. En el presente trabajo, en consecuencia, se desarrolla un proceso de explotación de información para el descubrimiento de patrones de co-localización alrededor de tipos de eventos espaciales de referencia que utiliza cliques máximos y algoritmos TDIDT para brindar una solución a este problema. Se presenta una prueba de concepto del proceso propuesto.

Keywords: Patrones de Co-Localización, TDIDT, Cliques Máximos, Explotación de Información.

1 Introducción

Dado un conjunto de tipos de eventos espaciales booleanos y una relación de vecindad, el descubrimiento de patrones de co-localización permite encontrar subconjuntos de dichos tipos de eventos cuyas instancias se encuentran ubicadas frecuentemente vecinas entre sí [Shekhar & Huang, 2001]. En este contexto, se entiende evento espacial como un suceso que ocurre en un lugar del espacio determinado. En consecuencia, un tipo de evento espacial hace referencia a la clase de suceso que ocurre.

Para conseguir este objetivo, se han propuesto distintos algoritmos y métodos basados en análisis de asociaciones, los cuales se pueden dividir en dos grupos. El primero comprende a los Algoritmos No Transaccionales (*Transaction-Free Algorithms*), los cuales utilizan internamente algoritmos de minería de datos para el descubrimiento de reglas de asociación sobre la información correspondiente a instancias de eventos espaciales. Por otro lado, el segundo grupo abarca los algoritmos Transaccionales *(Transaction-Based Algorithms*), que generan información transaccional a partir de las instancias de eventos espaciales, para ser

utilizada como entrada de algoritmos de descubrimiento de reglas de asociación de manera explícita, siendo un enfoque más eficiente que la alternativa no transaccional. [Shekhar & Huang, 2001; Shekhar et al., 2011, Kim et al., 2014].

Por otro lado, existen tres diferentes modelos de generación de transacciones para resolver el problema de descubrimiento de patrones de co-localización [Shekhar & Huang, 2001; Xiong et al., 2004]. La primera forma, denominada Modelo Centrado en Eventos (*Event Centric Model*), se utiliza cuando existen muchos tipos de eventos espaciales y se desea encontrar subconjuntos de los mismos que sucedan frecuentemente juntos y ha sido ampliamente utilizada en trabajos como [Agrawal & Srikant, 1994; Shekhar & Hung, 2001; Huang et al., 2003, 2006; Yoo et al., 2004; Xiong et al., 2004; Yoo & Shekhar, 2006; Celik et al., 2007; Eick, 2008; Adilmagambetov et al., 2013; Kim et al., 2011, 2014].

El segundo modelo de generación de transacciones, denominado Modelo Centrado en Ventanas (*Window Centric Model*), permite descubrir patrones dentro de subdivisiones del espacio de datos, llamadas ventanas. Este modelo es utilizado en áreas como la Minería, donde cada ventana se correspondería con parcelas de terreno.

Por último, el tercer modelo es llamado Centrado en Tipos de Eventos de Referencia (*Reference Feature Centric Model*) o Basado en Referencias (*Reference Based*), consiste en encontrar patrones de co-localización generando transacciones alrededor de determinados tipos de eventos espaciales, utilizándose por ejemplo para la determinación de factores ambientales condicionantes de casos de Cáncer.

De estos modelos existentes es notable mencionar que el modelo centrado en tipos de eventos de referencia no ha sido implementado en demasiados algoritmos y métodos para el descubrimiento de patrones de co-localización, resultando incorrectos e incompletos aquellos que sí lo implementan [Adilmagambetov et al., 2013; Kim et al., 2014]. Por estas razones, en la sección 2 del presente trabajo se introduce a la problemática derivada del análisis del estado del arte, se presenta en la sección 3 una solución al problema planteado, se realizan en la sección 4 pruebas de concepto mediante herramientas estadísticas, y por último se presentan conclusiones en la sección 5.

2 Definición del problema

Ante una gran cantidad de tipos de eventos espaciales las búsquedas de patrones de co-localización con generación de transacciones basada en eventos puede resultar extenuante, demandando entonces una gran cantidad de recursos.

Ante la presencia de cierto tipo de evento de interés para un dominio del problema dado, un enfoque de generación de transacciones alrededor de tipos de eventos de referencia, resulta una alternativa más adecuada, sin embargo, la solución que utiliza este enfoque utiliza verificaciones de la relación de vecindad de cada evento espacial, con las instancias del tipo de evento espacial seleccionado como referencia, junto con consideraciones especiales en los cálculos de prevalencia para la creación de patrones [Shekhar & Huang, 2001; Xiong et al., 2004].

Este acercamiento propuesto, sin embargo, no permite generar transacciones correctas, debido a que no se asegura que todos los elementos de la misma sean vecinos entre sí, o completas, pudiéndose perder ciertos vecindarios en el proceso

[Adilmagambetov et al., 2013; Kim et al., 2014]. Este motivo hace necesaria el desarrollo de una solución que permita el descubrimiento de patrones de co-localización alrededor de tipos de eventos espaciales determinados de una forma correcta y completa.

En el presente trabajo se desarrolla un proceso de explotación de información [Britos, 2008] para dar solución a este problema.

3 Solución propuesta

Como se mencionó anteriormente, se propone en el presente trabajo un proceso de explotación de información que permita la obtención de patrones de co-localización correctos y completos alrededor de tipos de eventos espaciales determinados.

A tal fin, se toma como base el trabajo de Kim et al., (2014), en el cual se desarrolla un *framework* transaccional para el descubrimiento de patrones de co-localización, y se evalúa la conveniencia de utilizar cliques máximos sobre las relaciones de vecindad como forma de generar transacciones que aseguren completitud y correctitud.

Dado un grafo de vecindad un clique es un subgrafo completo, lo cual significa que en sí mismo todos los nodos son vecinos entre sí. Un clique máximo, entonces, es un clique de dicho grafo que no está incluido en ningún otro clique. Cada clique máximo correspondería entonces a una transacción donde todos los elementos son vecinos entre sí asegurando la correctitud del método [Kim et al., 2014, Lemma 1]. A su vez, la utilización de cliques máximos como transacciones asegura la completitud del método, al estar todas las relaciones de vecindad consideradas por lo menos en un cliqué máximo [Kim et al., 2014, Lemma 2].

A fin de resolver el problema presentado, se propone un proceso de explotación de información para el descubrimiento de patrones de co-localización alrededor de tipos de eventos espaciales de interés que utiliza cliques máximos bajo un modelo centrado en eventos, sometiendo a las transacciones creadas a un proceso de explotación de información para el descubrimiento de reglas de comportamiento utilizando algoritmos de la familia *Top-Down Induction of Decision Trees* – TDIDT – [Britos, 2008].

Como puede observarse en la Figura 1, el proceso parte de un conjunto de información espacial representada en distintos formatos (texto plano, bases de datos, mapas referenciados geográficamente, entre otras), la cual es integrada para formar un único repositorio donde conste, por cada evento espacial, un identificador de dicha instancia, el tipo de evento espacial al cual pertenece, y su ubicación en el espacio.

Posteriormente, la información integrada es utilizada para generar información transaccional. Este subproceso, como se observa en la Figura 2, calcula primero todas las relaciones de vecindad entre las instancias de eventos espaciales, evaluando la distancia que existe entre ellos, para posteriormente encontrar los cliques máximos en el grafo de vecindad construido, generando una transacción por cada uno de ellos, en la cual conste los tipos de eventos espaciales de cada evento que forma parte de dicho clique.

Una vez obtenida la información transaccional, es necesario especificar el tipo de evento espacial alrededor de las cuales se desea hallar los patrones de co-localización

para ser utilizado como atributo objetivo o *target* de un algoritmo TDIDT, utilizando los demás tipos de evento espacial como atributos de entrada del mismo.

Como salida del paso anterior obtendremos un conjunto de reglas de comportamiento, en función al árbol de decisión generado. Debido a que las transacciones poseen valores booleanos que evidencian la presencia o ausencia de los tipos de evento espacial presentes en los vecindarios, es necesario filtrar solo aquellas que muestren como consecuente la presencia del tipo de evento seleccionado como objetivo.

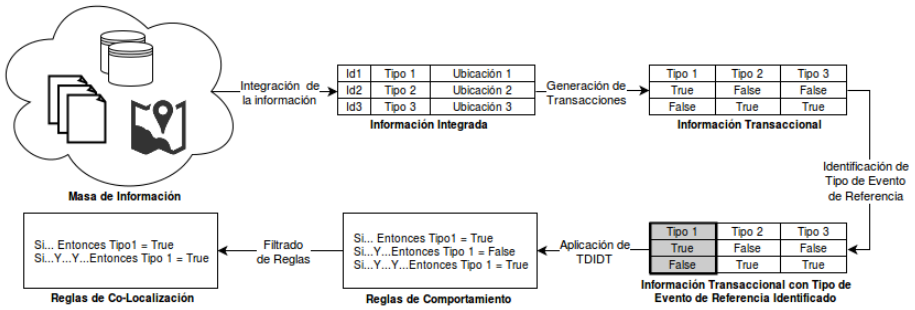

Fig. 1. Proceso para el descubrimiento de patrones de co-localización alrededor de tipos de eventos de referencia

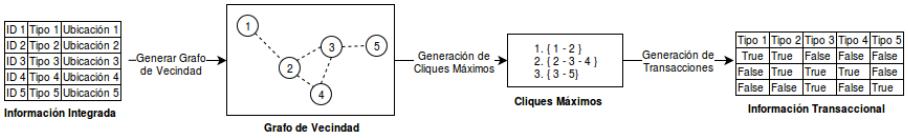

Fig. 2. Generación de transacciones

Esta secuencia de pasos hace posible la reutilización de la información transaccional para descubrir patrones de co-localización alrededor de distintas tipos de evento espacial de interés, sin necesidad de realizar el cálculo de vecindarios en cada oportunidad.

Las reglas obtenidas, por otro lado, no solo describen los tipos de eventos espaciales vecinos, sino también las condiciones que deben reunir los vecindarios, esto es, si es necesaria en algunos casos la ausencia de ciertos tipos de eventos, agregándose de esta forma información a los resultados.

4 Prueba de Concepto

Se realiza una prueba de concepto comparando el proceso propuesto con un algoritmo basado en tipos de eventos de referencia para determinar si el primero puede detectar mayor cantidad de patrones correctos que los métodos existentes hasta el momento.

Para ello, se cuenta con 10 conjuntos sintéticos de 500 puntos generados y clasificados en 7 tipos de manera aleatoria, con coordenadas en el plano en el intervalo [0; 40] tanto del eje de las ordenadas como del de las abscisas. A modo de ejemplo, la Figura 3 ilustra la distribución de los puntos del primer conjunto de datos, donde cada símbolo representa un tipo de evento diferente.

Los conjuntos en cuestión son usados como entrada tanto del proceso propuesto, cómo del algoritmo seleccionado para comparar: Co-Location Miner bajo un modelo centrado en tipos de evento de referencia [Shekhar & Huang, 2001].

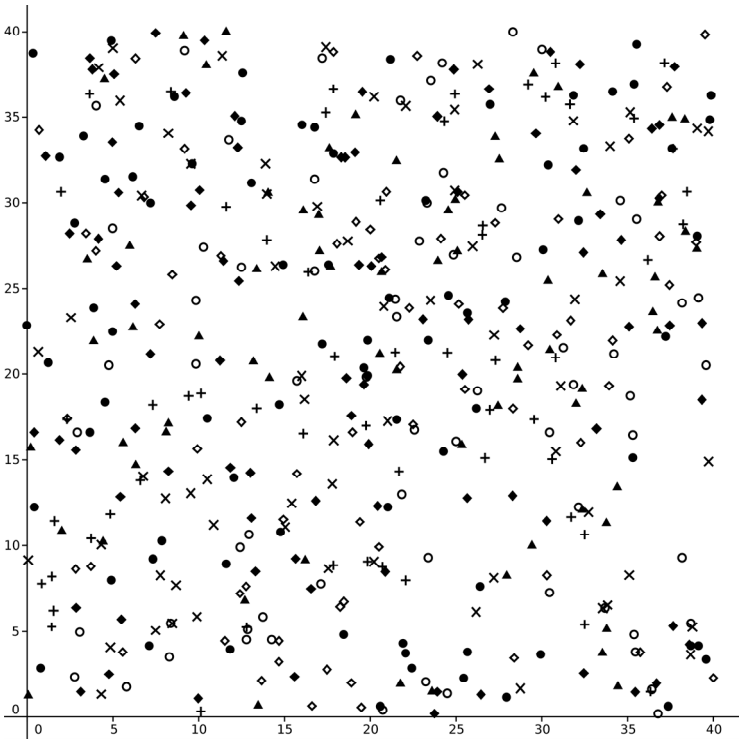

Fig. 3. Distribución de puntos del primer conjunto sintético de datos, donde cada símbolo corresponde a un tipo de evento diferente

Por otro lado, ejecución del proceso de explotación de información se ha realizado calculando en una primera instancia las relaciones de vecindad entre los puntos, verificando si la distancia entre cada par de puntos es menor al cierto umbral de vecindad especificado. En esta etapa nos valemos de la propiedad de simetría de la función de distancia a fin de reducir las comparaciones y aumentar la velocidad de ejecución del proceso. Se utilizó además el algoritmo CLIQUES para la generación de cliques máximos sobre las relaciones de vecindad por demostrar eficiencia superior a otros métodos [Uno, 2005; Tomita et al., 2006], y el software Tanagra [Rakotomalala, 2005] para la ejecución del algoritmo TDIDT seleccionado, C4.5 [Quinlan, 1993].

Luego de la ejecución de tanto el proceso presentado como del algoritmo Co-Location Miner, los patrones de co-localización obtenidos fueron evaluados para corroborar la correctitud de los mismos, a fin de determinar la cantidad de patrones correctos detectados en cada oportunidad. Para demostrar que el proceso propuesto posee un mejor comportamiento, se utilizó el test estadístico no paramétrico de rangos con signo de Wilcoxon [Wilcoxon, 1945], buscando rechazar la hipótesis nula H_0 y aceptar la hipótesis alternativa H_A, siendo estas las que se muestran en la tabla 1.

La ejecución del test de rangos con signos de Wilcoxon puede observarse en la Tabla 2, junto con los valores obtenidos durante las pruebas, utilizándose posteriormente el método del P-Valor sobre la suma de los rangos positivos con un nivel de significación del 1%, obteniéndose un valor igual a 0,0038, por lo cual se rechaza la hipótesis nula, aceptándose la alternativa en consecuencia, confirmando de esta forma que el proceso de explotación de información permite encontrar mayor cantidad de reglas correctas que el método existente hasta el momento con un 99% de confianza.

Tabla 1. Hipótesis nula e hipótesis alternativa consideradas en el test de rangos con signos de Wilcoxon

H_0:	La cantidad de Patrones Correctos detectados por el algoritmo Co-Location Miner es Mayor o Igual a la cantidad detectada por el proceso propuesto.
H_A:	La cantidad de Patrones Correctos detectados por el proceso propuesto es Mayor a la cantidad detectada por el algoritmo Co-Location Miner.

Tabla 2. Ejecución del test de rangos con signo de Wilcoxon sobre los conjuntos de datos sintéticos, ordenados según las diferencias positivas

Conjunto	Patrones correctos con el proceso propuesto	Patrones correctos con Co-Location Miner	Diferencias positivas	Rangos
Conjunto 7	3	3	0	-
Conjunto 8	2	1	1	2
Conjunto 9	2	1	1	2
Conjunto 10	5	4	1	2
Conjunto 1	3	1	2	5
Conjunto 2	3	1	2	5
Conjunto 5	4	2	2	5
Conjunto 4	4	1	3	7
Conjunto 3	6	2	4	8.5
Conjunto 6	7	3	4	8.5
Suma de rangos:				45

5 Conclusiones

Se ha presentado un proceso de explotación de información para el descubrimiento de patrones de co-localización de manera correcta y completa. El proceso propuesto utiliza cliques máximos para la generación de información transaccional, y la

aplicación a esta de algoritmos TDIDT para la obtención de reglas de comportamiento sobre los tipos de eventos espaciales asociados, resultando un método original a tal fin.

Se ha presentado una prueba de concepto que muestra con un método estadístico no paramétrico un mejor desempeño del proceso propuesto frente al que actualmente está en uso.

El proceso propuesto permite además (i) la búsqueda de patrones alrededor de distintos tipos de eventos espaciales sin necesidad de realizar el cálculo del grafo de vecindad en cada oportunidad, y (ii) agregar información a los resultados determinando bajo qué condiciones ocurren los patrones hallados.

Como próximo paso se prevé realizar casos de validación en los dominios de accidentología, incidentes de defensa civil, y factores eco-ambientales condicionantes de enfermedades.

Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por el Programa para la Formación de Doctores para Fortalecer áreas de I+D+i (2016-2020) de la Universidad Tecnológica Nacional (Argentina) y por los Proyectos de Investigación 33B133 y 33A205 de la Secretaria de Ciencia y Técnica de la Universidad Nacional de Lanús (Argentina).

Referencias

Adilmagambetov, A., Zaiane, O. R., & Osornio-Vargas, A. (2013). Discovering co-location patterns in datasets with extended spatial objects. In Data Warehousing and Knowledge Discovery (pp. 84-96). Springer Berlin Heidelberg.

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

Britos, P. V. (2008). Procesos de explotación de información basados en sistemas inteligentes. Tesis de Doctorado en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata.

Celik, M., Kang, J. M., & Shekhar, S. (2007, October). Zonal co-location pattern discovery with dynamic parameters. In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on (pp. 433-438). IEEE.

Chicago Police Department (2016). Reported Incidents occurred in the City of Chicago from 2001 to present [On-Line]. Chicago, USA. [Consultado el 3 de Febrero de 2016]. Disponible en: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Eick, C. F., Parmar, R., Ding, W., Stepinski, T. F., & Nicot, J. P. (2008, November). Finding regional co-location patterns for sets of continuous variables in spatial datasets. In Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems (p. 30). ACM.

Huang, Y., Xiong, H., Shekhar, S., & Pei, J. (2003, March). Mining confident co-location rules without a support threshold. In Proceedings of the 2003 ACM symposium on Applied computing (pp. 497-501). ACM.

Huang, Y., Pei, J., & Xiong, H. (2006). Mining co-location patterns with rare events from spatial data sets. Geoinformatica, 10(3), 239-260.

Kim, S. K., Kim, Y., & Kim, U. (2011). Maximal cliques generating algorithm for spatial co-location pattern mining. In Secure and Trust Computing, Data Management and Applications (pp. 241-250). Springer Berlin Heidelberg.

Kim, S. K., Lee, J. H., Ryu, K. H., & Kim, U. (2014). A framework of spatial co-location pattern mining for ubiquitous GIS. Multimedia tools and applications, 71(1), 199-218.

Quinlan, J. R. (1993). C4. 5: programs for machine learning.

Rakotomalala, R. (2005). TANAGRA: a free software for research and academic purposes. In Proceedings of EGC (Vol. 2, pp. 697-702).

Shekhar, S., & Huang, Y. (2001). Discovering spatial co-location patterns: A summary of results. In Advances in Spatial and Temporal Databases (pp. 236-256). Springer Berlin Heidelberg.

Shekhar, S., Evans, M. R., Kang, J. M., & Mohan, P. (2011). Identifying patterns in spatial information: A survey of methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3), 193-214.

Tomita, E., Tanaka, A., & Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. Theoretical Computer Science, 363(1), 28-42.

Uno, T. (2005). MACE_GO: MAximal Clique Enumerator (CLIQUES Implementation) [C Code]. Versión 2.0.. Disponible desde: http://research.nii.ac.jp/~uno/code/ macego10.zip

Venkatesan, M., Thangavelu, A., & Prabhavathy, P. (2011). Event Centric Modeling Approach in Colocation Pattern Snalysis from Spatial Data. arXiv preprint arXiv:1109.1144.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics bulletin, 1(6), 80-83.

Xiong, H., Shekhar, S., Huang, Y., Kumar, V., Ma, X., & Yoo, J. S. (2004, April). A Framework for Discovering Co-Location Patterns in Data Sets with Extended Spatial Objects. In SDM (pp. 78-89).

Yoo, J. S., Shekhar, S., Smith, J., & Kumquat, J. P. (2004, November). A partial join approach for mining co-location patterns. In Proceedings of the 12th annual ACM international workshop on Geographic information systems (pp. 241-249). ACM.

Yoo, J. S., Shekhar, S., & Celik, M. (2005, November). A join-less approach for co-location pattern mining: A summary of results. In Data Mining, Fifth IEEE International Conference on (pp. 4-pp). IEEE.

Yoo, J. S., & Shekhar, S. (2006). A joinless approach for mining spatial colocation patterns. Knowledge and Data Engineering, IEEE Transactions on, 18(10), 1323-1337.