MINING ASSOCIATION PATTERNS IN SPATIO-TEMPORAL DATA

Martin Pešek
Master Degree Programme (2), FIT BUT
E-mail: xpesek07@stud.fit.vutbr.cz

Supervised by: Jaroslav Zendulka
E-mail: zendulka@fit.vutbr.cz

ABSTRACT

Spatio-temporal data mining is currently a rapidly evolving area of research. This paper deals with the spatio-temporal association patterns and the possibilities of their mining. It describes types of association patterns and the main ideas of algorithm GenSTMiner for mining one of them – generalized spatio-temporal patterns.

1 ÚVOD

V poslední době, zejména s rozvojem lokalizačních a dohledových systémů, dochází k velkému nárůstu objemu ukládaných časoprostorových dat. Vzhledem k tomu se výrazně zvyšuje význam těchto dat a časoprostorové databáze se postupně stávají velmi aktivní oblastí výzkumu. Kromě rozvoje technologií pro modelování, dotazování a indexování časoprostorových dat je velký důraz kladen i na oblast získávání znalostí z těchto dat.

Časoprostorová data jsou data, která kromě vlastních rysů obsahují jak časovou tak prostorovou informaci. Může jít například o popis objektů pohybujících se v prostoru v určitém časovém období. Během prvních výzkumů v oblasti dolování z časoprostorových dat se nabízela možnost použít existující techniky dolování v časových a prostorových datech pro časoprostorová data. Časoprostorová data však obsahují složité vztahy, které nelze odhalit pohledem na časovou a prostorovou dimenzi dat odděleně. Druhým problémem, který se při tomto přístupu často projevuje, je příliš rozsáhlý prohledávací prostor časové i prostorové dimenze. Z těchto důvodů jsou zapotřebí techniky, které sjednotí časovou a prostorovou informaci dohromady za účelem nalezení zajímavých a užitečných časoprostorových vzorů.

V současnosti lze dolování v časoprostorových datech klasifikovat podle vzorů, které vyhledávají, do těchto základních skupin: evoluční vzory přirozených jevů, frekventované pohyby objektů, časoprostorová klasifikace nebo predikce, časoprostorové shlukování a časoprostorové asociační vzory. Tento příspěvek se zabývá popisem časoprostorových asociačních vzorů se zaměřením na zobecněné časoprostorové vzory a implementací algoritmu GenSTMiner pro jejich dolování.

2 ASOCIAČNÍ VZORY V ČASOPROSTOROVÝCH DATECH

Mezi typy časoprostorových asociačních vzorů patří topologické vzory a prostorové sekvenční vzory, jejichž následující popis vychází z [1]. Dolování topologických vzorů, které jsou rozšířením prostorových kolokačních vzorů zavedením časových omezení, sice dokáže vyhledat zajímavé vztahy mezi událostmi během definovaného časového intervalu, nedokáže však odhalit vztahy mezi událostmi v různých časových intervalech. Takové vzory není možné získat prostřednictvím prostorových, časových ani topologických vzorů. Proto je potřeba zavést nový typ vzoru, a to prostorové sekvenční vzory. Pomocí těchto vzorů lze snadno popsat například, jak nějaká událost na jednom místě způsobí výskyt jiné událostí na druhém místě. Prostorové sekvenční vzory lze podle přístupu k prostorové pozici ve vzoru rozdělit na dva typy: časoprostorové sekvenční vzory, které používají absolutní souřadnice pozic, a zobecněné časoprostorové vzory, které používají relativní souřadnice pozic.

Časoprostorové sekvenční vzory umožňují zachytit vývoj událostí v sousedících oblastech v čase. Nevýhodou těchto vzorů může být jejich silná závislost na předpokladu, že se takové události opakují na přesně stejných místech. Často však absolutní pozice výskytu události není z pohledu analytické úlohy důležitá a postačující je pozice relativní. V takovém případě je možné odhalit časoprostorové vztahy, které při použití absolutních pozic mohou zůstat skryté.

2.1 ZOBECNĚNÉ ČASOPROSTOROVÉ VZORY

Pro potřeby prostorových sekvenčních vzorů je třeba nejprve rozdělit čas na disjunktní časová okna určité délky W. Dva libovolné časy jsou si blízké, pokud spadají do stejného časového okna. Prostor se rozdělí mřížkou na množinu disjunktních buněk, přičemž každá buňka reprezentuje jednu pozici v prostoru. Nechť je dále nad množinou všech pozic definována relace sousednosti R. Pozice l_1 a l_2 spolu sousedí, pokud $(l_1, l_2) \in R$.

Nechť pozice v prostoru $l = (x, y)$. Potom se událost výskytu prostorového rysu e na pozici l v čase t značí jako $e(x, y, t)$. Kromě toho je třeba zavést tzv. referenční pozici $l_{ref} = (x_{ref}, y_{ref})$. Každá událost je pak mapována na odpovídající relativní pozici výskytu vzhledem k referenční pozici. Dvě množiny takto mapovaných udalostí vyskytujících se ve stejném čase jsou si blízké, pokud každá událost v první množině je blízká každé události v množině druhé.

Zobecněný časoprostorový vzor pak je sekvence množin mapovaných událostí taková, že každá množina v sekvenci je blízká se všemi ostatními množinami ze sekvence.

3 ALGORITMUS GENSTMINER

Pro dolování frekventovaných zobecněných časoprostorových vzorů je v [1] navržen efektivní a škálovatelný algoritmus *GenSTMiner*. Algoritmus vychází z přístupu algoritmu PrefixSpan (viz [2]) k dolování sekvenčních vzorů. Je založen na principu růstu vzoru, čímž se vyhýbá nutnosti generovat obrovské množství kandidátů na frekventované vzory. Na rozdíl od algoritmů založených na generování a testování kandidátů je zde tedy využito přístupu prohledávání do hloubky. Algoritmus sestává z následujících tří kroků:

1. Vyhledání všech frekventovaných událostí jedním průchodem databáze.
2. Rozdělení množiny událostí nalezených v prvním kroku do skupin, přičemž každé události odpovídá jedna skupina, a vytvoření tzv. projektované databáze pro každou událost.

Pro každou sekvenci v každé projektované databázi se zvolí referenční pozice a události se namapují na jejich relativní pozice vzhledem k referenční pozici.

3. Vyhledávání všech frekventovaných zobecněných časoprostorových vzorů rekurzivní konstrukcí a dolováním projektovaných databází pro postupně se prodlužující již nalezené vzory.

Pro odlišení frekventovaných vzorů od nefrekventovaných se používají dvě míry: časová a prostorová minimální podpora. Prostorová podpora události v daném časovém okně se určí jako počet různých pozic, na kterých se událost vyskytla. Událost je pak v daném časovém okně frekventovaná, pokud je její výskyt roven nejméně prostorové minimální podpoře. Časová podpora události se vyhodnotí jako počet různých časových oken, ve kterých je frekventovaná. Událost je v databázi frekventovaná, jestliže je její časová podpora rovna nejméně minimální časové podpoře.

Pro vylepšení časových i prostorových požadavků tohoto algoritmu jsou navrženy dvě optimalizace. První optimalizace se týká použití tzv. podmíněných projektovaných databází k eliminaci některých událostí a sekvencí, které nemohou být v databázi frekventované. Druhá optimalizace, nazvaná pseudo-projekce, slouží k redukci paměťových požadavků algoritmu. Tato optimalizace spočívá v tom, že při vytváření projektovaných databází není nutné získávat projekci sekvencí, které začínají událostí, která sice je frekventovaná, ale nikoliv v rámci právě analyzovaného časového okna.

Algoritmus GenSTMiner včetně zmíněných optimalizací bude v rámci projektu implementován v programovacím jazyce Java. Pro otestování a experimentální zhodnocení implementace tohoto algoritmu je potřeba vytvořit jednoduchou aplikaci s grafickým uživatelským rozhraním a zvolit vhodná data. Kromě reálné datové množiny se nabízí použití vhodného volně dostupného generátoru umělých časoprostorových dat.

4 ZÁVĚR

Tento článek představuje hlavní typy vzorů, kterými se zabývají současné techniky dolování v časoprostorových datech, se zaměřením na časoprostorové asociační vzory. Mezi tyto asociační vzory patří tzv. zobecněné časoprostorové vzory, pro jejichž dolování bude v rámci tohoto projektu implementován a experimentálně zhodnocen algoritmus GenSTMiner.

PODĚKOVÁNÍ

Tato práce vznikla částečně za podpory grantu VUT FIT, FIT-S-10-2 a specifického výzkumu MSM0021630528.

REFERENCE

[1] Hsu, W., Lee, M. L., Wang, J.: Temporal and Spatio-Temporal Data Mining. Hershey: IGI Publishing, 2008, 280 s., ISBN 978-1-59904-387-6.

[2] Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier Inc., second edition, 2006, 770 s., ISBN 978-1-55860-901-3.