MINING OF PERIODIC PATTERNS

Rostislav Stríž
Master Degree Programme (2), FIT BUT
E-mail: xstriz03@stud.fit.vutbr.cz

Supervised by: Michal Šebek
E-mail: isebek@fit.vutbr.cz

Abstract: Data mining offers modern and promissing way to analyze collected data from various angles by using many different techniques. One of these techniques, focused on analyzing temporal data sets, is mining of periodic patterns. Via finding periodic patterns in time series we are hoping to predict system behaviour (to some extent). Network analysis, finance and medicine are exemplary sectors, where this kind of information proved to be useful. In this paper we discuss basic principles and algorithms used in periodic patterns mining, followed by brief analysis of our own experiments and conclusions.

Keywords: data mining, periodic patterns, time series

1 ÚVOD

Sběr co největšího množství dat dnes patří k běžným postupům z jednoduchého důvodu – obsahují *informace* a ty se v dnešním světě stávají významným obchodním artiklem, pomáhají společnostem při rozhodování a mohou poskytovat výhody oproti konkurenci. Systémy, které umožňují analýzu sesbíraných dat, jsou tak stále žadanější a lze předpokládat, že se budou i nadále rozvíjet. Jedním z používaných postupů pro analýzu dat je proces *dolování z dat*, který umožňuje získat skryté, netriviální a potenciálně zajímavé informace ze zdrojových dat. Samotných principů a algoritmů pro dolování existuje dnes celá řada a stejně rozmanité je i jejich využití, nehledě na to, že se stále objevují nové algoritmy a vylepšení. Jednou z oblastí pro dolování z dat je oblast *temporálních dat*, které obsahují mimo standardních datových záznamů i časovou složku, která určuje jejich uspořádání. V kontextu temporálních dat se pak setkáváme s pojmem *periodických vzorů*, které popisují události, jenž se v datech vyskytují s určitou pravidelností. V praxi se vyhledávání periodických vzorů využívá např. v medicíně, ve finančnictví nebo při analýze počítačových sítí.

Hlavní náplní tohoto článku je shrnutí základních principů a poznatků z oblasti dolování periodických vzorů v temporálních datech se zaměřením na porovnání tří konkrétních algoritmů, které byly v rámci výzkumu implementovány a otestovány.

2 ZÁKLADNÍ POJMY DOLOVÁNÍ PERIODICKÝCH VZORŮ

Jak bylo stručně zmíněno v úvodní kapitole, dolování periodických vzorů probíhá nad sadou temporálních dat, která můžeme obecně rozdělit na dvě základní kategorie – *temporální sekvence* jsou tvořeny *uspořádánou* sekvencí událostí/transakcí, kdy je významné zejména jejich uspořádání a časová vzdálenost jednotlivých záznamů nebývá konstantní, zatímco *časové řady* jsou zpravidla tvořeny seřazenou posloupností od sebe časově konstantně vzdálených hodnot. Diskutované vyhledávání periodických vzorů provádíme nad kategorií časových řad.

Periodické vzory jsou vzory, které se opakují v pravidelných intervalech napříč temporálními daty. Na periodicitu pak můžeme nahlížet z několika pohledů – odtud dělení periodických vzorů. Prvním

možným je rozdělení na *plně periodické vzory*, u kterých předpokládáme, že každá událost v dané periodě je periodická, a dále pak na *částečně periodické vzory*, kde časová řada obsahuje i události neperiodické. V praxi se dnes využívá zejména dolování částečně periodických vzorů, jelikož často hledáme periodickou událost v moři událostí neperiodických.

Z jiného úhlu pohledu můžeme rozdělit periodické vzory na *synchronní*, kdy časovou řadu vnímáme jako souvislou řadu period, které na sebe bezprostředně navazují bez jakéhokoli šumu (periodické události se vždy vyskytují v konstantní časový okamžik v rámci periody) a *asynchronní*, u kterých je časová řada interpretována jako řada period, mezi kterými se ovšem může vyskytovat šum a periodické události se nemusí opakovat vždy ve stejný časový okamžik (je zavedena tolerance). Synchronní periodické vzory jsou podmnožinou asynchronních, nicméně dolování v kontextu asynchronních periodických vzorů je zřejmě z principu mnohem náročnější, nežli dolování vzorů synchronních, proto se algoritmy pro dolování obou skupin liší.

Uvažujme databázi, která obsahuje množinu od sebe konstantně vzdálených množin událostí s časovými razítky. Definujme *don't care* symbol $*$, který může označovat libovolnou množinu událostí. Potom *vzorem* rozumíme neprázdnou sekvenci $s = s_1...s_p$, kde s_i je buď množina událostí, nebo $*$. *Periodou* takovéhoto vzoru s je pak $|s|$ a jeho *podvzorem* je vzor s', pokud mají oba stejnou periodu a $s'_i \subseteq s_i$ pro každou pozici i kde $i \neq *$.

3 VYBRANÉ ALGORITMY

Algoritmů pro dolování periodických vzorů je celá řada, je tedy třeba volit vhodně vzhledem k cíli dolovací úlohy. Obecně vycházejí algoritmy pro dolování periodických vzorů z algoritmů pro dolování asociačních pravidel a významnou roli tak hraje *Apriori vlastnost* v upravené formě, která tvrdí, že každý podvzor periodického vzoru je opět periodickým vzorem. Všechny námi zvolené algoritmy vyžadují zadání rozsahu period pro vyhledávání periodicity a v prvním kroku se snaží nalézt množinu periodických 1-vzorů, tedy periodických vzorů o velikosti 1, které se poté v dalších krocích skládají do vzorů delších.

3.1 DOLOVÁNÍ SYNCHRONNÍCH PERIODICKÝCH VZORŮ

Pro tuto úlohu byl vybrán algoritmus popsaný v [1] (zkr. *HPP*) a na podobném principu založený algoritmus *DPMiner*, představený v [2]. Výhodou obou algoritmů je, že používají princip *kandidátního „maxivzoru"*, díky čemuž obdržíme výsledné vzory již po dvou průchodech daty. Hlavní myšlenkou algoritmu HPP je sestavení všech periodických 1-vzorů do jednoho maxivzoru, který je pak porovnáván postupně s každou periodou. Z porovnání vzniká vzor, který je podvzorem maxivzoru a zároveň je obsažen v dané periodě. Všechny tyto podvzory se zaznamenávají do stromové struktury spolu s čítačem jejich celkového výskytu. Po dokončení porovnávací fáze jsou ze stromové struktury na základě vstupního parametru, kterým je globální spolehlivost, odvozeny výsledné vzory.

Algoritmus *DPMiner* pracuje na stejném principu, ale umožňuje dolování tzv. *hustých periodických vzorů*, což jsou vzory, které se vyskytují pouze v určitém úseku zpracovávaných dat. Takovéto vzory klasické algoritmy nedokáží často rozlišit, jelikož počítají s globální spolehlivostí výsledných vzorů, zatímco DPMiner časovou řadu segmentuje podle hustoty výskytu jednotlivých 1-vzorů. Spolehlivost vzorů je pak počítána v rámci nalezených segmentů, jejichž minimální velikost je omezena vstupním parametrem *FLC* (z angl. *fragment length coeficient*).

3.2 DOLOVÁNÍ ASYNCHRONNÍCH PERIODICKÝCH VZORŮ

Pro dolování asynchronních periodických vzorů byl implementován algoritmus představený v [3] (zkr. *APP*). Algoritmus operuje se dvěma základními parametry – *min_rep* udává počet period, ve kte-

rých se musí vzor souvisle opakovat, aby byl považován za periodický, hodnota *max_dis* pak limituje délku šumu mezi jednotlivými periodami. Hlavním principem algoritmu je iterativní vyhledávání co nejdelších posloupností period v souladu s hodnotami vstupních parametrů.

4 TESTOVÁNÍ

Všechny tři popsané algoritmy byly implementovány ve formě pluginů pro *Microsoft Analysis Services* – službu, která nabízí rozhraní pro dolování z dat nad *Microsoft SQL Serverem*. Údaje v tabulce 1 byly naměřeny při experimentech nad synchronní databází o 10000 prvcích s periodou 10.

Algoritmus	Parametry	Čas zpracování	Nalezených vzorů	Nejdelší vzor
HPP	conf = 0,4	11,40s	15	4
	conf = 0,6	9,38s	5	2
	conf = 0,8	8.57s	2	1
DPMiner	conf = 0,1; FLC = 2%	6,70s	78	7
	conf = 0,5; FLC = 1%	7,42s	47	5
	conf = 0,5; FLC = 2%	6,26s	25	5
APP	min_rep = 10; max_dis = 0	63,80s	27	4
	min_rep = 10; max_dis = 1000	571,20s	26	3
	min_rep = 5; max_dis = 200	325,58s	27	4

Tabulka 1: Výsledky experimentu

Z uvedené tabulky lze vyčíst, že největší vliv na dobu dolování má nastavení parametrů u algoritmu pro dolování asynchronních vzorů. Toto zjištění odráží podstatu algoritmu, kdy jsou potenciální periodické segmenty uchovávány v paměti, dokud není překročena vzdálenost *max_dis*, s jejíž stoupající hodnotou extrémně narůstá paměťová náročnost, což se odráží na době provádění spolu s jeho iterativní povahou. Zbývající dva algoritmy nalezly v datech validní periodické vzory, přičemž *DPMiner* lze považovat v tomto případě za úspěšnější díky velké roztroušenosti periodických symbolů ve zdrojové databázi. Nutno však podotknout, že všechny tři algoritmy mají odlišné cíle a při dalších experimentech nad asynchronními daty byl schopen zpracovat vygenerovaný šum mezi jednotlivými periodickými segmenty pouze algoritmus *APP*.

5 ZÁVĚR

V rámci článku byly stručně diskutovány základní principy a problémy spojené s dolováním periodických vzorů v temporálních datech. Byly představeny tři algoritmy, které byly následně implementovány a otestovány. Všechny algoritmy jsou funkční a použitelné a díky svým rozdílným povahám se při komplexnějších dolovacích úlohách mohou vhodně doplňovat.

PODĚKOVÁNÍ

Tento příspěvek vznikl za podpory grantů TAČR TA01010858 a FIT-S-11-2, výzkumného záměru MSM0021630528 a Centra excelence IT4Innovations CZ.1.05/1.1.00/02.0070.

REFERENCE

[1] Han, J., Dong, G., Yin, Y: Efficient Mining of Partial Periodic Patterns in Time Series Database. In Proceedings of the 15th International Conference on Data Engineering, 1999, s. 106-1015.

[2] Sheng, C., Hsu, W., Lee, M.: Mining Dense Periodic Patterns in Time Series Data. In Proceedings of the 17th International Conference on Data Engineering, 2006, s. 115-115.

[3] Yang, J., Wang, W., Yu, P. S.: Mining Asynchronous Periodic Patterns in Time Series Data. In Proceedings of ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2000, s. 275-279.