RECENZE DISERTAČNÍ PRÁCE

MINING MULTI-LEVEL SEQUENTIAL PATTERNS

MICHAL ŠEBEK

Předložená práce se zaměřuje na efektivní konstrukci častých vzorů v sekvenčních datech s definovanou hierarchií. Přílohou k textové části je CD, které kromě vlastního textu disertace obsahuje demonstrační aplikace pro porovnávání jednotlivých systémů pro dolování častých vrozů v sekvencích a ukázková data.

Úvodní kapitola uvádí cíle práce a především jasnou a přesvědčivou motivaci pro popisovaný výzkum. Následujíci dvě kapitoly popisují oblast dolování častých vzorů a dolování častých sekvencí. Obě kapitoly, soustřeďující se na základní práce v této oblasti, jsou napsány bez chyb, jasně a srozumitelně. Základy těchto disciplín jsou popsány dostatečně úplně, snad jen mohla být uvedena spolu s Hanovou monografií [19] i citace na původní Hanovu práci (*DBMiner*). Kdyby kapitola o dolování sekvencí byla doplněna přehledem nejnovějších metod z posledních 5-6 let (z tohoto období jsou jen tři odkazy na literaturu, ten zásadní Plantevitova práce je z roku 2010), nebylo by nic, co jí vytknout. Prosím o stručný přehled u obhajoby.

Vlastní přínos práce – popis dvou nových algoritmů pro dolování častých vzorů v sekvenčních datech – je obsahem kap. 4 s experimentálním ověřením v kap. 5. Autor nejdříve definuje generalizovaný support. Není úplně jasné, zda je tato definice vlastní nebo (do nějaké míry) převzatá. Prosím o vysvětlení. První z algoritmů – *hGSP* – doluje level-crossing vzory. Dva základní kroky GSP autor modifikoval pro práci s hierarchickými dat a pro hledání vzorů použil heuristickou míru konkrétnosti (concreteness, Def. 28) založenou na Shannonově míře informace. Z konstatování na str. 40, že není zaručeno nalezení globálního optima, vyplývá, že některé vzory nebudou nalezeny. Pokud je to tak, do jaké míry to může být problém při řešení reálných úloh? Autor na příkladu uvádí, že *hGSP* je efektivnější než původní *GSP*, protože generuje menší počet kandidátních vzorů než *GSP*. Zde bylo vhodné toto tvrzení upřesnit, např. uvést počet kandidátních vzorů v případě *GSP*.

Druhým algoritmem a hlavním přínosem této práce je algoritmus *MLSP* pro dolování hierarchicky uzavřených multi-level sekvenčních vzorů. Důvody přechodu na tuto variantu častých vzorů jsou jasně vysvětleny. Autor definuje nový pojem hierarchicky uzavřených multi-level sekvenčních vzorů, popisuje teoretické základy této metody dolování a definuje tři nutné podmínky pro tento typ častých vzorů. Vše navíc vysvětluje na příkladech. Velmi si cením i tabulky 4.3, která shrnuje všechny přístupy k dolování diskutované v této práci (snad by se hodilo jen doplnění stránek, kde jednotlivé příklady najít). Všechny kroky algoritmu jsou jasně popsané a umožňují tak reimplementaci. Mám jen jednu otázku. Apriori-like (level-wise) přístup má často za následek enormní počet kandidátů, který není efektivně zvládnutelný. Je tomu tak i u *MLSP*? Pokud ano, jakou nejsložitější úlohu se podařilo pomocí *MLSP* vyřešit?

Bylo by jistě užitečné porovnat *MLSP* s metodami z posledních let, např. s Plantevitovým algoritmem. Jak autor konstatuje, dolování level-crossing vzorů je velmi časově náročné oproti multi-level vzorům, na druhé straně první metoda dokáže objevit úplnější množinu vzorů (viz např. obr. 4.1 na str. 36). Nemůže znamenat omezení na multi-level vzory problém při řešení praktických úloh, např. s interpretací nalezeného častého vzoru? Dá se pojem hierarchicky uzavřených multi-level sekvenčních vzorů rozšířit např. na dolování v sekvenci stromů nebo grafů?

Kap. 5 obsahuje výsledky experimentů na umělých datech a jedné sadě reálných dat. Pro experimenty s umělými daty byl použit generátor dat spoluvytvořený autorem práce. Po podrobném popisu generátoru jsou uvedeny porovnány výsledky obou systému (resp. varianty *MLSPhash*) s výsledky algoritmů *GSP* a *PrefixSpan*. Drobným nedostatkem je zaměření vždy na jediný parametr, např. velikost datové sady s ostatními paramatry konstantními. Chybí též zdůvodnění hodnot těchto konstant, např. průměrný počet prvků v sekvenci=4 při testování scalability. Celkově jsou tyto experimenty velmi dobré. Poněkud slabší je ověření na reálných datech. Přinejmenším porovnání s *PrefixSpan* by bylo vhodné doplnit.

Na téma disertace autor publikoval pět práci, z toho tři na lokální česko-slovenské úrovni (jednu časopiseckou a dvě na konferenci *INFORMATICS*) a dvě na zahraničních konferencích. Ta na nejkvalitnější konferenci (DaWaK 2014) je však v práci pouze citována a není uvedeno, jak s předloženou prací souvisí. Pro přijetí práce k obhajobě považuji tyto publikace za dostatečné.

Poznámky: predikce nemusí vždy znamenat klasifikaci do spojité třídy (str. 3). Odkaz na vysvětlení pojmu někdy později (např. str. 9, *root nodes*, odkaz na tabulky na str. 32 dole) znesnadňuje čtení, lepší je uvést např. neformální definici nebo tabulky zopakovat. V příkladu Example 12 (str. 41) není uvedeno, že minimální support je větší než 1.

Námět práce jistě odpovídá oboru disertace a je aktuální z hlediska současného stavu vědy. Autor splnil cíle uvedené v 1.1 a 1.2. Ocenění si zaslouží angličtina, která je na velmi dobré úrovni. Autor rozhodně prokázal tvůrčí schopnosti v dané oblasti a vytvořil práci, kterou doporučuji k obhajobě a který po úspěšném zodpovězení uvedených otázek splní požadavky standardně kladené na disertační práce..

Brno, 31. ledna 2015 Luboš Popelínský, FI MU Brno