

VYSOKÁ ŠKOLA EKONOMICKÁ

Fakulta informatiky a statistiky

Propojená data na webu – referát

Bc. Bohuslav Koukal

Studijní program: Aplikovaná informatika, Magisterský

Studijní obor: Znalostní a webové technologie

Akademický rok: ZS 2016/2017

Obsah

1. Úvod ... 3
1.1 Bibliografické informace ... 3
1.2 Překlad názvu článku ... 3
2. Shrnutí článku ... 3
2.1 Abstrakt ... 3
2.2 Úvod ... 3
2.3 Úvod do problematiky ... 4
2.4 Hledání asociací ... 4
2.4.1 Algoritmus hledání ... 4
2.4.2 Prořezávání stavového prostoru ... 4
2.5 Dolování častých asociačních vzorů ... 4
2.6 Experimenty ... 5
2.7 Závěr ... 5

1. Úvod

Cílem této práce je přiblížit čtenáři odborný článek ze sborníku konference ISWC 2016 a to pomocí následujících technik:

- Překlad do českého jazyka.
- Extrakce hlavních myšlenek a postupů.
- Shrnutí cílů a výsledků článku.

1.1 Bibliografické informace

CHENG, Gong a Yuzhong QU. Efficient Algorithms for Association Finding and Frequent Association Pattern Mining. *ISWC 2016, Part I, LNCS 9981*. Springer International Publishing AG, 2016, 119-134. DOI: 10.1007/978-3-319-46523-4_8.

1.2 Překlad názvu článku

Účinné algoritmy pro vyhledávání asociací a dolování častých asociačních vzorů

2. Shrnutí článku

2.1 Abstrakt

Vyhledávání asociací mezi entitami se používá v mnoha oblastech – příkladem mohou být doporučovací algoritmy na sociálních sítích či identifikace osob podezřelých z terorismu. Vyhledávání asociací je usnadněno tím, jak se na webu zvyšuje počet grafových struktur, které popisují vztahy.

Asociace mezi více entitami je v této práci definována jako minimální spojitý podgraf, který obsahuje všechny tyto entity. V práci je navržen účinný algoritmus prohledávání grafu pro vyhledávání takových asociací, který pro zvýšení účinnosti prořezává prohledávaný prostor na základě vzdáleností mezi entitami. K odhadům vzdáleností pro prořezávání používá metodu "věštění vzdáleností".

Rozsáhlé experimenty na reálných RDF datasetech prokazují účinnost navrženého algoritmu.

2.2 Úvod

Asociace mezi více entitami je definována jako minimální spojitý podgraf celého grafu, který všechny entity propojuje. Z této definice vyvstávají dvě otázky:

- Jak účinně najít asociace ve velkém grafu?
- Jak může uživatel prozkoumat potenciálně obrovskou množinu nalezených asociací?

Tyto otázky jsou v práci řešeny následujícími kroky:

1. Je navržen účinný algoritmus pro vyhledávání asociací. Algoritmus je založený na prohledávání grafu a spojování cest. Pro větší účinnost prohledávání grafu je aplikováno prořezávání stavového prostoru. K prořezávání se používá odhad vzdáleností mezi entitami. Odhad vzdáleností je založen na metodě věštění vzdáleností (distance oracles), což je kompromis mezi výpočetním výkonem a požadavky na velikost úložného prostoru (věštění vzdáleností používá méně úložného prostoru než n^2 a k nalezení vzdálenosti má menší složitost než n log n. Může však odpovídat s menší přesností).

2. Pro to, aby uživatel mohl prozkoumat velkou množinu asociací, autoři navrhují identifikovat její význačné podmnožiny, jejichž vzory jsou často opakované. Identifikace vzorů (na základě typu hrany a vlastnosti uzlů isTypeOf) vede k vyšší úrovni abstrakce nalezených výsledků.

2.3 Úvod do problematiky

- Grafem je v této práci míněn orientovaný nevážený graf.
- Entitou je míněn uzel grafu, pro který hledáme asociaci.
- Asociace daných entit je minimální spojitý podgraf grafu (tedy strom), obsahující všechny dané entity. Entity tedy tvoří listy tohoto stromu.
- Průměr asociace je maximální počet hran mezi jednotlivými entitami.
- Asociační vzor je zobecnění nelistových uzlů v asociaci za pomoci jejich vlastnosti isTypeOf. Máme-li následující dvě asociace: V1 H1 V2 (vrcholy V1 a V2 spojené hranou typu H1) a V1 H1 V3, přičemž V2 a V3 jsou stejného typu, pak obě tyto asociace mají stejný asociační vzor.

2.4 Hledání asociací

Vstupem do algoritmu hledání asociací je graf a průměr asociace. Základní algoritmus je založen na následujícím poznatku, jehož důkaz je v samotné práci proveden:

Asociace (cesta) mezi entitami E o délce d může být rozložena na množinu cest kratších než (d+1)/2, které mají entity z E jako začátek a které mají společný konec.

2.4.1 Algoritmus hledání

1. Pro každou entitu (list) najdi všechny cesty délky kratší než (d+1)/2 prohledáváním do šířky.
2. Prořezej stavový prostor vyloučením některých nalezených cest.
3. Pospojuj navzájem ty zbylé cesty, které mají společný konec. Tyto pospojované cesty jsou potenciální asociace. Asociacemi se stanou, pokud splňují následující tři podmínky:

a) Cesta musí být minimální (tj. cesta je strom a listy jsou tvořené entitami).

b) Musí být kratší než daný průměr asociace.

c) Nesmí už být v množině nalezených asociací (pro kontrolu se používají tzv. "kanonické kódy", což je zjednodušeně řečeno jakýsi unikátní hash asociace).

2.4.2 Prořezávání stavového prostoru

Prořezávání probíhá v druhém kroku výše uvedeného algoritmu.

Některé cesty z kroku 1 nebudou spojeny do asociace (např. pro d=3 nalézám i cesty o délce 2, ale jejich spojením vzniknou příliš dlouhé asociace). Dále vyřazují cesty, jejichž koncové body mají do další nejbližší entity (listu asociace) vzdálenost větší, než d minus délka současné části cesty, tedy při spojení by jistě došlo k překročení d. Tato strategie vyžaduje znalost vzdáleností mezi entitami. Autoři vzdálenosti odhadují za pomoci vzdálenostního věštění (distance oracles). Jedná se o kompromis mezi výpočtem délky cesty v reálném čase a uložením všech hodnot do matice. Je to datová struktura, která po předzpracování grafu umožní rychlý výpočet/odhad. Používá např. milníkové body v grafu a ukládá délky cest mezi nimi. Vzdálenostní věštění může dávat přesné výsledky nebo odhady, přičemž odhady jsou nevhodné pro datasety, kde je průměr či medián délky cesty mezi body malý. Autoři v této práci používají vzdálenostní věštění s přesnými výsledky.

2.5 Dolování častých asociačních vzorů

Výsledkem algoritmu z předchozí kapitoly je velká množina asociací. Nyní je potřeba její výsledky abstrahovat na vyšší úroveň abstrakce, abychom dostali obecnější pravidla. Budeme tedy v množině asociací hledat časté vzory (vzory s nadprahovou frekvencí) – definice vzoru je nastíněna v kapitole Úvod do problematiky. Pravidlo isomorfie dle kanonických kódů, jak jej autoři definují pro asociace, však není možné použít pro vzory, protože operátor isomorfického řazení pro asociaci není pro vzor jednoznačný Pro každý vzor jsou tedy vygenerované všechny přípustné kanonické kódy a z nich je vybrán abecedně první. Druhou možností, která by byla výpočetně méně náročná, avšak autoři se jí dále

nevěnují, by bylo definovat operátor řazení pro detekci isomorfie jiným způsobem než jako abecední posloupnost.

2.6 Experimenty

Experimenty byly provedeny na vzorku 10 000 uzlů, náhodně vybraných ze dvou datasetů - LinkedMDB (celkem 1,3M entit, 2.1M hran) a DBpedia (4.3M entit a 15M hran). Průměrná délka cesty ve vybraných vzorcích byla 5-6, což je relativně málo (proto byly požity přesné odhady vzdáleností namísto aproximativních). Testy proběhly pro různé průměry asociací (2,3,4).

Asociace byly hledány klasickým algoritmem BSC a dvěma variantami výše nastíněného algoritmu, zde označovaného jako PRN. Varianty PRN se lišily v nastavení parametrů vzdálenostního věštění a v síle prořezávání. Pro asociace délky 2 byl lepší BSC (podle grafů cca 100 - 1000x rychlejší), protože prořezávání pro takto krátké asociace prakticky nezmenšilo prohledávaný prostor. Pro asociace o průměru 4 byl jeden z algoritmů podle grafu autorů až 10x rychlejší než BSC.

2.7 Závěr

Experimenty ukázaly, že navržený algoritmus je na reálných a velkých datasetech dostatečně výkonný a účinný. Přínosem práce je využití vzdálenostního věštění pro výpočet vzdáleností nutný k prořezávání stavového prostoru. Pro další zlepšení výkonnosti a použitelnosti algoritmu pro asociace o větším průměru by bylo vhodné aplikovat váhy na nalezené asociace a asociace s nízkou vahou z mezivýsledků odstraňovat.