Interaktívna vizualizácia výsledkov vyhľadávania informácií pomocou konceptových zväzov

Veronika Novotná, Peter Butka, Miroslav Smatana

Katedra kybernetiky a umelej inteligencie,
Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach
Letná 9, 042 00 Košice, Slovenská republika

veronika.novotna@student.tuke.sk,
{peter.butka,miroslav.smatana}@tuke.sk

Abstrakt. Klasický pohľad na výsledky vyhľadávania získaných z vyhľadávača je často vo forme usporiadaného zoznamu, poznáme však aj systémy poskytujúce štruktúrovanú formu výsledkov. Cieľom tohto príspevku je popísať implementovaný systém pre interaktívnu vizualizáciu výsledkov vyhľadávania vo forme konceptového zväzu. Tento poskytuje štruktúrovanú formu hierarchicky usporiadaných podmnožín vrátených dokumentov na základe metódy známej ako FCA (Formal Concept Analysis). Vytvorený nástroj umožňuje zvoliť dopyt a získať hierarchickú štruktúru zhlukov vrátených výsledkov na základe analýzy obsahu ich krátkych popisov – tzv. snippet-ov. Takto vytvorený konceptový zväz sa zobrazuje ako interaktívny graf, na ktorý je následne možné aplikovať redukcie zvyšujúce prehľadnosť výstupnej vizualizácie. Okrem grafickej vizualizácie je možné prehliadať aj jednotlivé výsledky a odkazy na nich, analyzovať ich vzťahy a odlišnosti vzhľadom na ich pozíciu v hierarchickej štruktúre zhlukov, ako aj využiť interaktívne zobrazenie pre rozšírenú navigáciu v sub-doméne odpovedajúcej množine vrátených výsledkov.

Kľúčové slová: formálna konceptová analýza, vyhľadávanie informácií, vizualizácia, konceptový zväz

1 Úvod

V rámci oblasti vyhľadávania informácií je jedným z riešených problémov poskytnutie výsledkov vyhľadávania v štruktúrovanej forme, v podobe ktorá lepšie zohľadňuje vzťahy medzi výsledkami a umožňuje lepšie pochopenie prehľadávanej domény dokumentov. Jednou z možností ako tento problém riešiť je použiť metódy z oblasti formálnej konceptovej analýzy (Formal Concept Analysis – FCA) [1]. V tomto prípade používateľ zadá dopyt a systém mu poskytne k nájdeným dokumentom hierarchickú štruktúru zhlukov dokumentov reprezentujúcich podmnožiny dokumentov zdieľajúcich atribúty definované v danej doméne. V klasickom rámci FCA sa pracuje s binárnou vstupnou tabuľkou (objekt má alebo nemá atribút, napríklad dokument obsahuje alebo neobsahuje daný term). V praktickej analýze samozrejme často existu-

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 15-19.

je potreba spracovať rôzne typy atribútov. Aj preto bolo navrhnutých viacero fuzzy prístupov. Jedným z nich je aj model tzv. zovšeobecneného jednostranne fuzzy konceptového zväzu (Generalized One-Sided Concept Lattice - GOSCL) [2]. Tento model bol použitý v rámci našej práce, jeho výhodou je možnosť spracovávať vstupné tabuľky s rôznymi typmi atribútov. Pre detaily k modelu, implementácii a použitiu GOSCL odporúčame preštudovať aj ďalšiu literatúru (okrem uvedenej), napríklad [3].

Cieľom tejto práce bolo vytvorenie nástroja pre interaktívnu vizualizáciu výsledkov vyhľadávania a poskytnúť ich používateľovi práve vo forme modelu GOSCL. Vstupom do analýzy boli krátke popisy výsledkov (tzv. snippet), tieto boli následne predspracované do formy vstupného kontextu pre tvorbu modelu GOSCL, kde objekty sú výsledky vyhľadávania a atribúty sú početnosti výskytu termov v popise výsledku.

V ďalšej kapitole popíšeme motiváciu nášho návrhu a jeho implementácie, následne sa budeme venovať navrhnutému systém a zhrnieme výsledky jeho testovania.

2 Motivácia

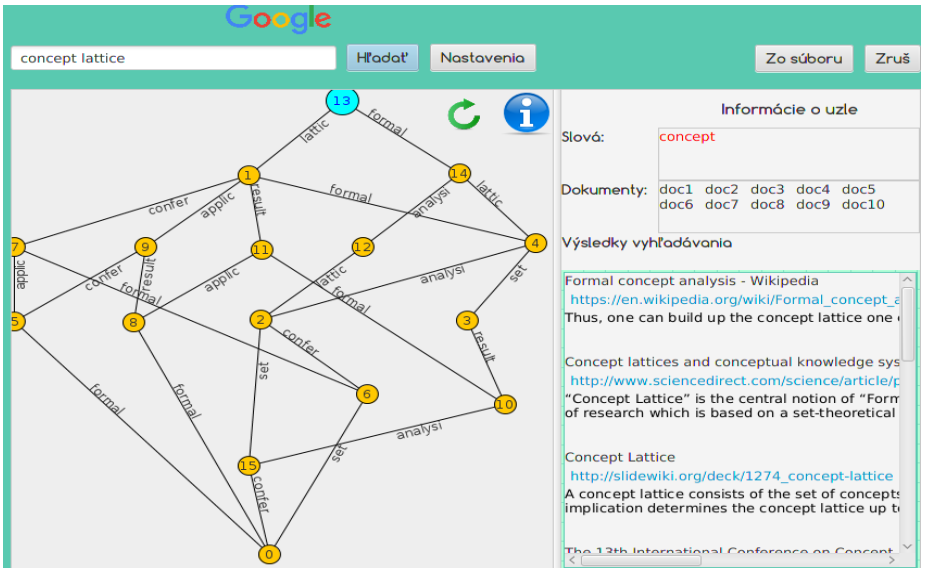
FCA už bolo samozrejme aplikované aj v doméne vyhľadávania informácií, a to v rámci riešenia rôznych problémov dopytovanie a navigácie v množine dokumentov. Väčšinou je výsledok z vyhľadávacieho stroja na nejaký zvolený dopyt vrátený vo forme usporiadaného (lineárneho) zoznamu výsledkov, pričom tieto obsahujú názov, krátky popis (snippet), či linku URL, atď. V oblasti vyhľadávania informácií boli vytvorené napríklad systémy využívajúce princípy FCA ako CREDO, respektíve rozšírená verzia CRE-CHAIN-DO [4]. Základný prvkom je analyzovať vrátené výsledky vyhľadávania na daný dopyt a vybudovať vstupný kontext pre FCA analýzu, kde objekty sú dokumenty a výskyt termov v názve alebo snippet-e predstavuje atribúty. Systém CREDO resp. CRE-CHAIN-DO bol vytvorený s cieľom analyzovať iba binárne vstupné kontexty popisujúce iba výskyt termu. Z tohto kontextu bola potom vytvorená dvojúrovňová hierarchia zobraziteľná ako klasická navigačná hierarchia v množinách dokumentov (vo forme stromovej štruktúry). Tento prístup je špecifický aj tým, že výsledná stromová štruktúra už nepredstavuje pôvodný konceptový zväz, takisto aj rozdelenie dokumentov bolo už špecificky upravené. My sme sa rozhodli zostať pri pôvodnej koncepcii FCA a ponúknuť radšej graf konceptového zväzu, pričom náš systém sme sa rozhodli zamerať na dva aspekty:

1. Poskytnúť interaktívnu vizualizáciu výstupného modelu v podobe konceptového zväzu (alebo jeho redukcie).
2. Využiť model GOSCL pre použitie rôznych typov atribútov.

3 Návrh a implementácia nástroja

Navrhovaný nástroj predstavuje aplikáciu pre používateľa, ktorý zadá dopyt a dostane výsledky Google vyhľadávania v podobe konceptového zväzu, alebo jeho redukcie. Proces vytvorenia výstupu pozostáva zo 6 hlavných krokov:

1. Získanie výsledkov z vyhľadávača – po zadaní dopytu, systém použije GoogleSearch API pre získanie množiny výsledkov, je možné zvoliť počet výsledkov do vizualizácie, výstup je uložený v JSON formáte.
2. Spracovanie výsledkov vyhľadávania – na základe nastavených parametrov sa predspracujú vrátené výsledky, t.j., title a snippet sú predspracované (tokenizácia, transformácia na malé písmená, odstránenie stop slov).
3. Vytvorenie vstupného kontextu – na základe výskytu termov je vytvorený formálny kontext, pričom tento môže byť binárny (0/1 ak sa term vyskytol /nevyskytol), alebo vektorový (fuzzy) (0 ak sa term nevyskytol, inak frekvencia termu).
4. Vytvorenie konceptového zväzu – algoritmus pre tvorbu GOSCL vytvorí výstupný model pre ľubovoľný kontext.
5. Redukcia výstupu – pre zlepšenie prehľadnosti a lepšiu interpretáciu sa často môže používateľ zvoliť redukciu, v našom prípade dve redukcie: orezanie konceptu s počtom objektov 0, alebo odstrániť koncepty s malou podporou (ponecháme len koncepty obsahujúce aspoň prahový počet objektov).
6. Vizualizácia konceptového zväzu – výstupný graf zväzu, vytvorený s pomocou rámca JUNG (http://jung.sourceforge.net/), je poskytnutý používateľovi.



Obr. 1. Interaktívna vizualizácia konceptového zväzu k zvolenému dopytu.

Získaný výsledok sa môže zobraziť, tak ako to je napríklad na Obr.1. Zobrazenie má 3 hlavné časti:

- Vstupná časť (horná časť) – zadanie dopytu, otvorenie nastavení parametrov, načítanie výsledkov zo súboru.

- Výstup grafu (časť vľavo) – interaktívny graf konceptového zväzu alebo jeho redukcie pomocou rámca JUNG, informácie o rozdieloch medzi uzlami (termy ktoré odlišujú uzly) sú zobrazené na hrane.
- Informačná časť (vpravo) – poskytuje detaily o vybranom prvku grafu, t.j., obsah konceptu, prehľad príslušných výsledkov k danej podmnožine objektov, dôležité termy daného, odlišujúce termy od rodičovského uzla, atď.

4 Experimenty

V rámci testovania sme realizovali experimenty so systémom pre rôzne dopyty, konkrétne boli experimenty spustené pre 10 rôznych dopytov do Google, pričom počet analyzovaných výsledkov pre jeden dopyt sa v závislosti od experimentu mohol meniť medzi 10 až 100 (parameter N). Okrem toho bolo jedným z nastavení aj to, či bol použitý iba binárny kontext, alebo fuzzy kontext (v tomto prípade početnosti termov v rámci snippet-u). Pre každý experiment sme sledovali minimálnu, strednú a maximálnu hodnotu výsledkov meraní vzhľadom k množine dopytov. Treba však zdôrazniť, že išlo o relatívne jednoduché prvotné experimenty so systémom, ktorých cieľom bolo nájsť základné obmedzenia fungovania aplikácie z pohľadu počtu konceptov, či náročnosti výpočtu modelu.

V prvej sérii experimentov sme sa zamerali na sledovanie počtu konceptov v závislosti od použitého počtu výsledkov vyhľadávania vstupujúcich do tvorby konceptového zväzu. Nastavenie experimentu bolo nasledovné: minimálny počet výskytu slova v rôznych dokumentoch bol 1 a nebola použitá žiadna redukcia. Snažili sme sa odhadnúť veľkosť vzniknutého konceptového zväzu pre zobrazenie pri takýchto nastaveniach. Vzhľadom k tomu, že vstup je v tejto doméne riedka matica, bol nárast počtu konceptov s pribúdajúcim počtom analyzovaných vrátených snippet-ov (dokumentov) približne lineárny. V praxi to napríklad pre prípad fuzzy konceptového zväzu znamenalo zhruba 20 konceptov pre N=10, približne 200 konceptov pre N=60, až po maximum okolo 400 konceptov pre N=100.

V ďalšom experimente sme analyzovali vplyv redukcie na báze podpory (support) na výslednú početnosť konceptov. V tomto prípade sme zobrali do úvahy 100 vstupných dokumentov a zisťovali sme vplyv redukcie. Ukázalo sa, že redukcia je výrazná, nakoľko už pri hodnote 0.1 je počet konceptov na úrovni 5-6% (z absolútneho počtu cca 400 konceptov). Pri podpore 0.3 zostáva po redukcii cca 1% konceptov.

Následne sme sa snažili experimentálne odhadnúť aká množina konceptov by ešte mohla byť pre používateľa zrozumiteľná a aké nastavenia počtu výsledkov a redukcie umožňujú tento výsledok dosiahnuť. Ukázalo sa, že rozumne prehľadný výstup bez aplikácie redukcie je možné dosiahnuť len ak náš vstup nepresiahne 10-15 dokumentov. Pri použití redukcie (podpora) je rozumný výstup dosiahnutý už pri podpore 0.2 aj pre 100 vstupných dokumentov.

Z hľadiska časovej náročnosti sa ukázalo, že najviac problematické je samotné získanie výsledkov cez API, nasleduje tvorba zväzu. Pre menšie množiny výsledkov (medzi 20 až 50, v závislosti od redukcie), je čas spracovania dostatočne krátky pre používanie aplikácie (pod pol sekundy).

5 Záver

V tejto práci sme sa venovali jednej z možností ako poskytnúť štruktúrovaný pohľad na výsledky vyhľadávania získaných z klasického vyhľadávača, a to poskytnutím interaktívnej vizualizácie hierarchickej štruktúry zhlukov získaných výsledkov v podobe konceptového zväzu. Okrem grafickej vizualizácie je možné prehliadať aj jednotlivé výsledky, odkazy na nich, analyzovať ich vzťahy a odlišnosti, ako aj využiť graf pre navigáciu v doméne.

Literatúra

1. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer Verlag, Berlin (1999)
2. Butka, P., Pocs, J.: Generalization of One-Sided Concept Lattices. Comput. Informat. 32(2), 355–370 (2013)
3. Butka, P., Pócsová, J., Pócs, J.: Design and implementation of incremental algorithm for creation of generalized one-sides concept lattices, Proc. of CINTI 2011, 373-378 (2011)
4. Nauer, E., Toussaint, Y.: Dynamical modification of context for an iterative and interactive information retrieval process on the web, CLA 2007, CEUR-WS proceedings (2008)

Poďakovanie: Tento príspevok vznikol s podporou VEGA projektu č.1/0493/16, KEGA projektu č.025TUKE-4/2015 a APVV projektu č.APVV-16-0213.

Annotation:

Interactive visualization of information retrieval results using concept lattices

This paper describes motivation and details of the tool designed and implemented for interactive visualization of information retrieval results using concept lattices, which are created using one-sided fuzzy extension of Formal Concept Analysis. The concept lattice is then shown as an interactive graph, which provides a structured view on the domain of query to the user.