Vyhľadávanie významných konceptov v rámci konceptuálnej analýzy dát

Miroslav Smatana, Peter Butka, Zuzana Čabalová

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

{miroslav.smatana, peter.butka}@tuke.sk,
zuzana.cabalova@student.tuke.sk

Abstrakt. Existuje množstvo prístupov a nástrojov, ktoré slúžia na extrakciu konceptuálnych štruktúr zo vstupného datasetu. Ich hlavným cieľom je pomôcť používateľovi lepšie porozumieť vstupným dátam a vzťahom medzi nimi. Jednou z takýchto metód je formálna konceptová analýza (FCA), ktorá je schopná spracovať a analyzovať vstupné dáta v tvare objekt-atribútov tabuľky na základe ich vzťahov. FCA obsahuje niekoľko modelov, v tejto práci budeme pracovať s modelom zovšeobecneného jednostranne fuzzy konceptového zväzu (GOSCL), ktorý je schopný pracovať s rozličnými typmi atribútov. Avšak jedným z problémov GOSCL je množstvo konceptov, ktoré generuje. Existuje niekoľko prístupov, ktoré riešia problém generovania veľkého množstva konceptov. V tejto práci sa zameriame na získavanie len tých konceptov, ktoré môžu byť pre používateľa potencionálne užitočné na základe ním zadaného dopytu.

Kľúčové slová: formálna konceptová analýza, vyhľadávanie informácií, Levenshteinova vzdialenosť

1 Úvod

Jeden z prístupov používaných v oblasti analýzy dát je tzv. teórie konceptových zväzov. Tento prístup je známy ako formálna konceptová analýza (Formal Concept Analysis, FCA) [1,2] a je určený pre analýzu dát vo forme objekt-atribút modelu (formálny kontext). Výsledkom analýzy dát pomocou FCA je reprezentovaný pomocou konceptového zväzu, ktorý predstavuje hierarchicky organizovanú štruktúru skupín objektov (konceptov) so spoločne zdieľanými atribútmi. Analýza s využitím FCA našla svoje uplatnenie v oblastiach ako vyhľadávanie informácií, dolovaní znalostí, spracovaní prirodzeného jazyka, manažment znalostí a pod.

Existuje niekoľko metód FCA [3-5], avšak v našej práci sme sa zamerali na prácu s modelom zovšeobecneného jednostranne fuzzy konceptového zväzu (GOSCL) [6,7]. Narozdiel od ostatných metód FCA je GOSCL schopný spracovávať formálny kontext, ktorý môže obsahovať atribúty rozličných typov.
GOSCL zjednodušuje interpretáciu analýzy dát, ale ak je použitý na veľké alebo stredne veľké datasety, tak výsledný konceptový zväz obsahuje enormné množstvo

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 131-135.

konceptov a tým sa stáva neprehľadným. Bolo predstavených niekoľko redukčných metód [8-9] a rozličných vizualizácií, ktoré sa snažia riešiť tento problém. V tejto práci sme sa rozhodli použiť odlišný prístup, ktorý je založený na princípoch vyhľadávania informácií, kde používateľ zadá dopyt (čo ho v konceptovom zväze najviac zaujíma) a výsledkom sú koncepty, ktoré najviac vyhovujú jeho požiadavke.

2 Zovšeobecnený jednostranne fuzzy konceptový zväz

V tejto časti poskytneme len veľmi stručný popis základných charakteristík zovšeobecneného jednostranne fuzzy konceptového zväzu - GOSCL (viac informácií sa dozviete v práci [7]).

GOSCL predstavuje zaujímavý model v kontexte FCA. Používa jednostrannú fuzzifikáciu, čo znamená že jedna strana je ostrá (či sa prvok nachádza alebo nenachádza v množine) a druhá strana je fuzzy (atribúty nadobúda hodnoty v podobe fuzzy nožiny). Tento model ma taktiež aj iné výhody ako:

- Dokáže generovať konceptový zväz z objektov, ktoré sú reprezentované rozličnými typmi atribútov ako nominálne, ordinálne, numerické atď.
- Pracuje inkrementálne.

3 Navrhovaný prístup

V tejto sekcii je popísaný nami navrhovaný prístup. Hlavnou myšlienkou je aplikovanie procesu vyhľadávania informácií pre vyriešenie problému generovania veľkého množstva konceptov metódou GOSCL, kde sa snažíme získať len najzaujímavejšie koncepty vzhľadom k používateľom definovanému dopytu.

Celý proces pozostáva zo 4 základných krokov, kde najprv je generovaný konceptový zväz zo vstupného datasetu pomocou metódy GOSCL. Následne používateľ zadá dopyt, čo je pre neho vo vstupnom datasete zaujímavé. Ďalším krokom je nájdenie najzaujímavejších konceptov pre daný dopyt. To sa vykonáva na základe porovnania podobnosti konceptu k dopytu (keďže GOSCL je schopný spracovať rozličné typy atribútov, nie je možné použiť štandardné podobnostné metriky, preto sme v rámci práce navrhli modifikovanú Levenshteinovu vzdialenosť popísanú v kapitole 3.1). Posledným krokom je vizualizácia získaných výsledkov.

3.1 Modifikovaná Levenhsteinova vzdialenosť

Pre konceptové zväzy s ostrými hodnotami je jednoduché nájsť najzaujímavejšie koncepty na základe dopytu. Je to možné vykonať pomocou metrík podobnosti ako Hamming alebo Jaccard. Taktiež pre konceptové zväzy využívajúce fuzzy model, ktoré obsahujú len intervalové atribúty, je to možné dosiahnuť pomocou metrík ako Euklidovská a kosínusovú. Problém nastáva pri konceptových zväzoch, ktoré obsahujú viacero typov atribútov, kde môžeme mať aj hodnoty, ktoré sú neporovnateľné.

Preto sme sa rozhodli modifikovať Levenshteinovu vzdialenosť (http://www.levenshtein.net/) pre nájdenie najzaujímavejších konceptov k dopytu. Pseudokód výpočtu vzdialenosti:

1. inicializácia vzdialenosť = 0, n = počet atribútov, Q = dopyt používateľa, C = vektor atribútov pre koncept
2. pre i = 1:n
   a. porovnaj Q[i] s C[i]
   b. ak je Q[i] menšie ako C[i] potom vzdialenosť += 1
   c. ak je Q[i] neporovnateľné s C[i] potom vzdialenosť +=2
3. vráť vzdialenosť

4 Experimenty

V tejto časti sú popísané experimenty, kde sme navrhovaný prístup testovali v dvoch fázach. V prvej fáze sme mali vstupný dataset pozostávajúci z 50 objektov a 5 atribútov (intervalového typu) a kvalitu navrhovaného prístupu sme porovnávali pomocou metrík presnosť a návratnosť [10], kde sme mali pre každý dopyt definované, ktoré objekty sú relevantné a ktoré nie. Následne sme vypočítali skoré (zhodu s dopytom) pre každý koncept v konceptovom zväze a vybrali sme N objektov z konceptov s najvyšším skóre. Výsledky je možné vidieť v **Tab. 1**, kde nami navrhovaná metóda dosahuje porovnateľné výsledky ako ostatné štandardne používané metriky na našej testovacej vzorke.

Tab. 1. Výsledky prvej fázy experimentov - presnosť a návratnosť pre rozličné dopyty Q a rôzne typy vzdialenosti (MLD predstavuje našu modifikovanú Levenshteinovu vzdialenosť).

N=10	P/R	Q1	Q2	Q3
Euklidovská	P	0.4	**0.7**	**0.6**
Euklidovská	R	0.2	**0.23**	**0.4**
Kosínusová	P	0.3	0.5	**0.6**
Kosínusová	R	0.15	0.17	**0.4**
Jaccard	P	**0.5**	0.4	0.4
Jaccard	R	**0.25**	0.17	0.27
Hamming	P	**0.5**	0.5	0.5
Hamming	R	**0.25**	0.17	0.33
MLD	P	0.4	0.4	0.5
MLD	R	0.2	0.13	0.33

Modifikovaná Levenshteinova vzdialenosť je vhodnejšia pre komplexnejšie atribúty, preto sme v druhej fáze experimentov aplikovali túto vzdialenosť na dáta s heterogénnym typom atribútov. Tieto dáta pojednávali o poistení aut, kde každé auto pozostávalo z 3 atribútov (typ auta, vodičové skúsenosti, typ poistky). Vyhodnocovanie kvality sa vykonávalo rovnako ako v prvej fáze experimentov. Výsledky boli porovnateľ-

né s výsledkami dosiahnutými na klasických vstupoch, napríklad na dopyte {BMW, Expert, Full} sme dosiahli presnosť 0.6 a návratnosť 0.3.

5 Záver

V práci sme prezentovali prístup pre riešenie problému generovania veľkého počtu konceptov pri formálnej konceptovej analýzy s využitím prístupov vyhľadávania informácií. Náš prístup založený na modifikovanej Levenshteinovej vzdialenosti dosahoval výsledky porovnateľné so štandardnými metrikami, pričom je schopný pracovať s rozličnými typmi atribútov.

Literatúra

1. Wille, R.: Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. Springer, Netherlands (1982).
2. Birkhoff, G.: Lattice Theory. American Mathematical Soc. (1940)
3. Belohlavek, R.: Lattices of Fixed Points of Fuzzy Galois Connections. Math. Log. Quart. 47(1), 111–116 (2001).
4. Krajci, S.: A generalized concept lattice. Logic Journal of IGPL 13(5), 543–550 (2005).
5. Medina, J., Ojeda-Aciego, M., Ruiz-Calvino, J.: Formal concept analysis via multi-adjoint concept lattices. Fuzzy Set. Syst. 160, 130–144 (2009).
6. Krajci, S.: Cluster based efficient generation of fuzzy concepts. Neural Netw. World 13(5), 521–530 (2003)
7. Butka, P., Pocs, J.: Generalization of One-Sided Concept Lattices. Comput. Informat. 32(2), 355–370 (2013)
8. Butka, P., Pocs, J., Pocsová, J.: Reduction of concepts from generalized one-sided concept lattice based on subsets quality measure. Adv. Intell. Syst. Comput. 314, 101–111 (2015)
9. Antoni, L., Krajci, S., Kridlo, O.: Randomized fuzzy formal contexts and relevance of one-sided concepts. In: LNAI (Subseries of LNCS) 9113, pp. 183–199 (2014)
10. Davis, J., Goadrich, M.: The relationship between Precision-Recall and ROC curves. In: Proc. of the 23rd international conference on Machine learning. ACM, 233-240 (2006)

Poďakovnie: 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:

Retrieval of Important Concepts in Formal Concept Analysis

Currently, there exist a lot of methods and tools for extraction of conceptual structures from input dataset. The main aim of these methods is to describe input data and relations between them so the user will better understand them. One of such methods is Formal Concept Analysis (FCA), which is used for analysis of object-attribute input data models. FCA contains several methods, but in this work, we focus on Generalized One-Side Concept Lattices (GOSCL), which is able to process attributes of dif-

ferent types. However, the problem of GOSCL method is a number of concepts which it generates. In this paper, we describe our approach suitable for analysis of GOSCL models, where modified Levenshtein distance is used for retrieval of important concepts from output concept lattice based on user query.