ROBUST'2002, 133 – 135

METODA GUHA – SOUČASNÝ STAV

PETR HÁJEK

Abstract. GUHA is a method of generation of hypotheses, originating in 1966 and under continuous development. In contemporary terminology it is a method of data mining. Its present state is described.

Zusammenfassung. GUHA ist eine Methode der Erzeugung von Hypothesen, die im Jahr 1966 entstanden ist und seitdem weiter entwickelt wird. Sie ist eine Methode von "data mining" in der heutigen Terminologie. Der heutige Stand der Methode wird beschrieben.

1. Shrnutí

GUHA je akronym: Generalized unary hypotheses automaton (automat na generální unární hypotézy)[1]. Její princip byl formulován v roce 1966 [10], [11] jako „pomocí počítače generovat všechny hypotézy zajímavé na základě daných empirických dat". Zajímavost je dána logickým tvarem hypotézy a způsobem, jakým ji podporují data. Data mají (zjednodušeně) tvar obdélníkové matice, její řádky odpovídají objektům a sloupce atributům (vlastnostem). Ve většině implementací mají hypotézy tvar „φ souvisí s ψ", kde φ, ψ jsou logické kombinace atributů. Souvisení je dáno pomocí nějakého zobecněného *kvantifikátoru*, často daného nějakou statistikou užívanou při testování statistických hypotéz (např. Fisherův test).

Logické a statistické základy metody byly podrobně zpracovány v monografii [14], nyní zdarma dostupné na Internetu. Z logického hlediska jde o speciální partie teorie konečných modelů, vícehodnotové a modální logiky a zobecněných kvantifikátorů; ze statistiky o testování hypotéz včetně globální interpretace série hypotéz testovaných na jedněch datech. Základem pro většinu implementací je kromě citované monografie [14] také článek [9]. Český čtenář se dozví o teorii metody v knize [12] (ale měl by ignorovat jako zastaralé vše, co se tam píše o implementaci metody). Lze té doporučit článek [17].

Existovala řada implementací: od dnes „předpotopní" na počítači MINSK22 přes implementace pro sálové počítače IBM, implementaci pro PC (viz. [21]) až ke dvěma současným implementacím pod Windows: 4ft-miner na VŠE [35] a GUHA+- v Ústavu informatiky AV ČR [34].

Dnes je velmi populární *těžení z dat* (data mining). GUHA je evidentně jedna z prvních metod těžení z dat, ale zůstala vcelku neznámá a základní práce ze soudobého těžení z dat (např. [1], [2] se o ní vůbec nezmiňují). Je to škoda nejen kvůli prioritě, ale hlavně proto, že teorie metody GUHA obsahuje mnoho podnětů, přístupů a fakt, ke kterým se komunita těžení z dat postupně sama propracovává. Komunita metody GUHA se snaží o spolupráci; jedním z prostředků je aktivní účast v akci COST 274 TARSKI (Theory and application of relational structures a knowledge instrument). *Aplikace* vedoucí k publikovaným výsledkům

2000 *Mathematics Subject Classification.* 68T35 03B80.

Klíčová slova. Metoda GUHA, těžení z dat, explorační analýza dat.

[1]Teprve velmi později se ukázalo, že Guha je velmi frekventované indické příjmení.

byly, ale nebylo jich mnoho (desítky, asi ne stovky). Odkazy lze nalézt mj. v článcích [16], [18]. V současnosti se rozvíjí dalekosáhlé aplikace metody v medicíně (v pracovišti EUROMISE lékařské informatiky, sídlícím v Ústavu informatiky AV ČR).

Budoucí *rozvoj* metody GUHA se zaměří na prohloubení vztahu k databázím, užití fuzzy logiky a relačního těžení z dat [4].

Literatura

[1] Agrawal R., Manilla H., Sukent R., Toivonen A., Verkamo A.: Fast discovery of Association rules. Advance in Knowledge Discovery and Data Mining, AAA Press 1996, pp. 307-328.

[2] Adamo J.M.: Data minin for associational rules and sequential patterns, Sequential and parallel algorithms. Springer 2001

[3] Coufal D.: GUHA Analysis of Air Pollution Data. In: Artificial Neural Nets and Genetic Algorithms. Proceedings of the International conference. (Ed.: Kůrková V., Steele N.C., Neruda R., Kárný M.) - Wien, Springer 2001, pp. 465 – 468,), ICANNGA'2001 /5./, Prague, Czech Rep., 01.04.22-01.04.25

[4] Džeroski S., Lavrač N.: Relational data mining. Springer 2001.

[5] Feglar T.: The GUHA architecture. Proc. Relmics 6, Tilburg (The Netherlands), pp. 358 – 364.

[6] Hájek P.: Metamathematics of Fuzzy Logic, Kluwer 1998.

[7] Hájek P.: Relations in GUHA style data mining. Proc. Relmics 6, Tilburg (The Netherlands) 91 – 96.

[8] Hájek P.: The GUHA method and mining association rules. Proc. CIMA'2001 (Bangor, Wales) 533 – 539

[9] Hájek P.: The new version of the GUHA procedure ASSOC, COMPSTAT 1984, pp. 360 – 365.

[10] Hájek P., Havel I., Chytil M.: The GUHA method of automatic hypotheses determination, Computing 1(1966) 293 – 308.

[11] Hájek P., Havel I., Chytil M.: Metoda GUHA automatického zjišťování hypotéz I, II, Kybernetika 2 (1966) 31-47, 3 (1967) 430 – 437.

[12] Hájek P., Havel I., Chytil M.: Metoda GUHA – automatická tvorba hypotéz. Academia Praha

[13] Hájek P., Bendová K., Renc Z.: The GUHA method and three-valued logic, Kybernetika 7(1971) 421 – 431.

[14] Hájek P., Havránek T.: Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory), Springer-Verlag 1978, 396 pp.

[15] Hájek P., Havránek T.: Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory). Internet edition. http://www.cs.cas.cz/∼hajek/guhabook/

[16] Hájek P., Holeňa M.: Formal logics of discovery and hypothesis formation by machine. To appear in Theoretical Computer Science.

[17] Hájek P.,Louvar B., Pokorný D., Tschernoster E.: Metoda GUHA – její cíle a prostředky. Sborník SOFSEM'82, 59 – 84.

[18] Hájek P., Rauch J., Feglar T., Coufal D.: The GUHA method, data preprocessing and mining. Proc. DTDM02 (Database technologies for data mining), Praha 24.3.2002, 29 – 36

[19] Hájek P. (guest editor): International Journal of man-Machine Studies, vol. 10, No 1 (special issue on Guha). Introductory paper of the volume is Hájek, Havránek: The GUHA method - its aims and techniques. Int. J. Man-Machine Studies 10(1977) 3 – 22.

[20] Hájek P. (guest editor): International Journal for Man-Machine Studies, vol. 15, No 3 (second special issue on GUHA)

[21] Hájek P., Sochorová A., Zvárová J.: GUHA for personal computers, Comp. Stat., Data Arch. 19, pp. 149 – 153.

[22] Hálová J., Žák P.: Coping Discovery challenge of mutagenes discovery with GUHA+/- for windows. In: The Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining. Workshop KDD Challenge 2000. International Workshop on KDD Challenge on Real-world Data. - Kyoto, - 2000, pp. 55 – 60, Pacific-Asia Conference on Knowledge Discovery and Data Mining /4./, Kyoto, Japan, 00.04.18-00.04.20

[23] Havránek T.: The statistical modification ond interpretation of GUHA method, Kybernetika 7(1971), 13 – 21.

[24] Holeňa M., Fuzzy hypotheses for GUHA implications, Fuzzy Sets and Systems 98 (1998), 101 – 125.

[25] Holeňa M.: Exploratory data processing using a fuzzy generalization of the GUHA approach, Fuzzy Logic, Baldwin et al., ed. Willey et Sons, New York, 1996, pp. 213 – 229.

[26] Pecen L., Pelikán E., Beran H., and Pivka D.: Short-term fx market analysis and prediction. In Neural Networks in Financial Engeneering (1996), pp.189 – 196
[27] Rauch J.: GUHA as a Data Mining Tool, Practical Aspects of Knowledge management. Schweizer Informatiker Gesellshaft Basel, 1996, 10 s.
[28] Rauch J.: Logical Calculi for Knowledge Discovery. Red. Komorowski, J. - Zytkow, J. Berlin, Springer Verlag 1997, pp. 47 – 57.
[29] Rauch J.: Logical problems of statistical data analysis in databases. Proc. Eleventh Int. Seminar on Database Management Systems (1988), pp. 53 – 63.
[30] Rauch J., Šimůnek M.: Mining for 4ft association rules. Proc. Discovery Science 2000 Kyoto, Springer Verlag 2000, 268 – 272
[31] Rauch J., Šimůnek M.: Mining for statistical association rules. Proc. PAKDD 2001 Hong Kong, 149 – 158.
[32] Šebesta V., Straka L.: Determination of Suitable Markers by the GUHA Method for the Prediction of Bleeding at Patients with Chronic Lymphoblastic Leukemia. In: Medicon 98, Mediterranean Conference on Medical and Biological Engineering and Computing /8./, Lemesos, Cyprus
[33] Zvárová J., Preiss J., Sochorová A.: Analysis of Data about Epileptic Patients Using GUHA Method. In: EuroMISE 95: Information, Health and Education. (Ed.: Zvárová J., Malá I.) - Prague, EuroMISE Center 1995, pp. 87, TEMPUS International Conference, Prague, Czech Republic, 95.10.20 – 95.10.23.
[34] GUHA+– project web site http://www.cs.cas.cz/ics/software.html
[35] http://lispminer.vse.cz/overview/4ftminer.html

Ústav informatiky, AV ČR, Pod Vodárenskou věží 2, 182 07 Praha
E-mail: hajek@cs.cas.cz