Hĺbková analýza dát a jej využitie v praxi

Patrik Török

Slovenská technická univerzita v Bratislave

Fakulta informatiky a informačných technológií

xtorok@stuba.sk

16. december

Abstrakt

Vedomosti boli odjakživa dôležitou súčasťou ľudskej spoločnosti. Postupom času by sa bez pomoci technológie proces získavania poznatkov stal čoraz ťažším, pretože každým dňom narastá počet dát a údajov, ktoré treba pre nadobudnutie znalosti spracovať. Hĺbková analýza dát označuje proces alebo metódu, ktorá z veľkého množstva údajov získava zaujímavé poznatky. Tento proces možno aplikovať v rôznych oblastiach ľudského života vrátane podnikania, vzdelávania, sociálnych sietí, medicíny, vedy apod. Oblasť hĺbkovej analýzy má teda široké uplatnenie aj v oblasti vedeckého pokroku a porozumenia. Cieľom tohto článku je čitateľovi predstaviť proces hĺbkovej analýzy, stručne ho oboznámiť s evolúciou hĺbkovej analýzy údajov od prvej zmienky po súčasnosť a vymenovať konkrétne oblasti, v ktorých sa tento proces využíva, spolu s opisom využitia v spomínanej oblasti priemyslu.

1 Úvod

Hĺbková analýza dát (angl. Data mining) je proces pri ktorom sa spracúva veľké množstvo informácií a následne sa premieňa na vedomosť [6]. Je akousi kombináciou štatistiky a umelej inteligencie, vďaka čomu môžu spoločnosti vytvárať modely, ktoré im umožnia nachádzať súvislosti medzi miliónmi záznamov. Zároveň vďaka tomuto procesu dokážu spoločnosti predpovedať trendy budúcnosti. Keďže množstvo informácií, ktoré treba spracovať, neustále rastie, hĺbková analýza dát nachádza čoraz väčšie uplatnenie v mnohých priemyselných oblastiach. Množstvo spoločností používa softvér na hĺbkovú analýzu údajov, vďaka ktorému sa môžu dozvedieť viac o svojich zákazníkoch. Programy na hĺbkovú analýzu totiž hľadajú súvislosti v údajoch na základe informácií, ktoré používatelia od programu vyžadujú alebo programu poskytujú. Tiež im môže im pomôcť v reklamnej oblasti, napr. pri vytváraní efektívnejších marketingových stratégií, zvyšovaní predaja a znižovaní nákladov. [1] Hĺbková analýzy sa opiera o efektívne zhromažďovanie, uskladnenie a spracovanie údajov. Jej vplyv je citeľný aj mimo technickej sféry, s jej rozšírením vzrástli obavy o súkromie dát a bezpečnosť. Kvôli tomu došlo k zavedeniu mnohých regulácií, ako napríklad GDPR, a začal byť kladený väčší dôraz na dodržiavanie zodpovedný a etických postupov tohto procesu.

2 Typy hĺbkovej analýzy dát

Aby dosiahli požadovaný výsledok, dátoví vedci a analytici používajú rôzne typy hĺbkovej analýzy údajov. Medzi najbežnejšie techniky patria:

- Zhluková analýza (eng. Clustering)
  Zhluková analýza je v podstate hľadanie skupín s podobnými vlastnosťami. Obchodníci, a zamestnanci v oblasti reklamy často používajú tento typ hĺbkovej analýzy na to, aby identifikovali cieľové skupiny. Zhluková analýza je užitočná vtedy, keď spoločnosti nevedia, aké podobnosti by mohli existovať v rámci ich uložených údajov.
- Detekcia anomálií
  Tento typ hĺbkovej analýzy vyhľadáva vyčnievajúce dáta v dátovom súbore, čo môže pomôcť pri odhaľovaní podvodov.
- Regresia
  Regresia patrí medzi pokročilejšie štatistické nástroje. Používa sa najmä v prediktívnej analýze, využívajú ju hlavne vývojári, ktorí hľadajú spôsoby, ako zvýšiť počet používateľov, a pomáha predpovedať budúce výnosy s minimálnym rizikom.
- Dolovanie z textu
  Dolovanie z textu analyzuje, ako často ľudia používajú určité slová. Môže byť použitá na zistenie nálady autora textu alebo osobnosti, ako aj na analýzu príspevkov na sociálnych sieťach na marketingové účely alebo na odhalenie potenciálnych únikov údajov od zamestnancov. [4]

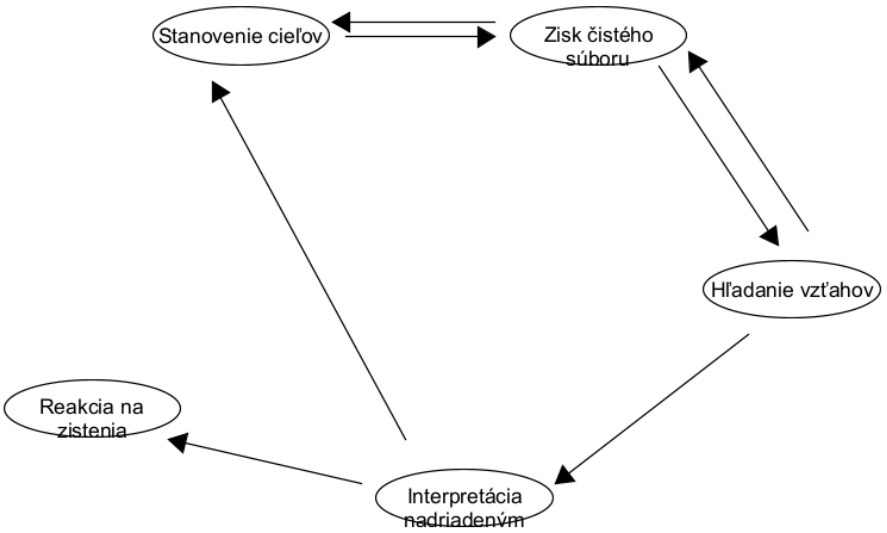
3 Nástroje na hĺbkovú analýzu

Spoločnosti majú k dispozícii množstvo komerčných a open-source nástrojov na hĺbkovú analýzu údajov. Medzi tieto nástroje patria dátové sklady, nástroje na čistenie údajov, informačné panely, analytické nástroje, nástroje na analýzu textu, a iné. Medzi najznámejšie nástroje na hĺbkovú analýzu údajov patria

- Weka
- Oracle Data Mining
- RapidMiner
- Knime
- Apache Mahout [8]

4 Proces hĺbkovej analýzy dát

Dátoví analytici v procese hĺbkovej analýzy údajov často pracujú podľa určitého postupu krokov. Ak by postupovali inak, existuje možnosť, že sa pri analýze dostanú do problémov. Počet krokov sa môže pre rôzne typy hĺbkovej analýzy líšiť, no všetky postupnosti krokov sú odvodené od všeobecnej postupnosti, ktorá vyzerá takto:



Obr. 1: Schéma procesu hĺbkovej analýzy [2]

- Krok 1: Musíme zistiť, čo chceme hĺbkovou analýzou dosiahnuť, aké ciele naplniť a zistiť, ako ukončiť proces s úspechom. Následne treba uvažovať o tom, aké zdroje sú k dispozícii, ako budú zabezpečené a uložené, ako bude prebiehať zhromažďovanie informácii a ako bude vyzerať konečný výsledok alebo analýza. Tento krok zahŕňa aj stanovenie obmedzení údajov, ich ukladania, zabezpečenia a zberu a posúdenie vplyvu týchto obmedzení.

- Krok 2: Až potom začneme pracovať so samotnými dátami. Môžeme ich zhromažďovať, nahrávať a extrahovať. Potom sa vyčistia od odchýlok a prejdú kontrolou, chybné a neopodstatnené dáta sa odstránia. Môže sa skontrolovať aj veľkosť objemu dát, pretože čím väčší je súbor informácií, tým viac zabere analýza času.

- Krok 3: Po získaní čistého súboru údajov môžeme začať s hľadaním vzťahov, trendov, asociácií alebo sekvenčných vzorcov. Údaje tiež môžeme načítať do prediktívnych modelov, vďaka ktorým môžeme predpovedať trendy budúcnosti.

- Krok 4: Po ich získaní možno výsledky sumarizovať, interpretovať a prezentovať nadriadeným, ktorí doteraz do procesu hĺbkovej analýzy neboli zapojení a tí môžu na základe spomínaných výsledkov vykonať plánovaný krok spoločnosti alebo zmeniť svoje rozhodnutia. Ak nadriadeným výsledky nevyhovujú, môžu požiadať o opakovanie predchádzajúcich krokov.

- Krok 5: Manažment spoločnosti reaguje na zistenia, a na ich základe príjme opatrenia. Ak sa spoločnosť rozhodne, že zistenia analýzy neboli presvedčivé, môže požadovať zopakovanie procesu. Ak sú zistenia dostatočne presvedčivé, môže spoločnosť na ich základe obrátiť pozornosť podľa výsledku analýzy. [10] [7]

5 História

História hĺbkovej analýzy dát je dôkazom rýchlosti vývoja analýzy dát a jej stále väčšej dôležitosti v rôznych oblastiach. Počas vývoja prešla niekoľkými kľúčovými fázami.

5.1 Začiatky (60-80-te roky 20. stor.)

Počiatky hĺbkovej analýzy siahajú do 60-tych a 70-tych rokov 20. stor., keď štatistici a výskumníci prvýkrát začali preskúmavať metódy na analýzu dát. Počas tohto obdobia sa vyvinuli základné techniky, ako je analýza zhlukov a algoritmy stromu rozhodnutí. Tieto metódy položili základy pre ďalší vývoj hĺbkovej analýzy.

5.2 Zrodenie hĺbkovej analýzy (90-te roky 20. stor.)

90-te roky predstavujú vytvorenie moderných spôsobov hĺbkovej analýzy, vznikol samotný pojem "Hĺbková analýza dát". Pokrok v oblasti výpočtovej techniky a rozšírenie databáz umožnili vznik tohto procesu ako samostatného odboru. Výskumníci a praktici začali vyvíjať špeciálne algoritmy a nástroje na odhalenie cenných poznatkov v rozsiahlych súboroch dát. Preto boli v tomto období vyvinuté kľúčové koncepty a techniky, ktoré dodnes tvoria jadro hĺbkovej analýzy.

5.3 Rozšírené použitie (začiatok 21. storočia)

Na začiatku 21. storočia sa hĺbková analýza dát stala dostupnejšou. Mnohé organizácie spoznali jej potenciál pre získavanie poznatkov z dát a vyvinuli softvérové nástroje, ktoré tento proces zjednodušovali. Rozvoj digitálnych dát, sociálnych sietí a internetu priviedla k vzniku "big data", ktoré prinieslo nové výzvy a príležitosti pre ťažbu dát. Počas tohto obdobia boli vyvinuté mnohé technológie na spracovanie veľkého množstva dát.

5.4 Strojové učenie (2010 - súčasnosť)

Po roku 2010 sa hĺbková analýza zlúčila so strojovým učením, keďže sa algoritmy strojového učenia stali dôležitými pre urýchlenie analýzy dát. Pokroky výpočtovej techniky a rozsiahle dátové úložiská umožnili presnejšie a komplexnejšie vykonať proces hĺbkovej analýzy.

V súčasnosti zostáva proces hĺbkovej analýzy neoddeliteľnou súčasťou dátovej vedy. Dátoví vedci využívajú sofistikované algoritmy a nástroje na získavanie cenných poznatkov z dát, čím pomáhajú podnikom pri dôležitých rozhodnutiach. Čo sa budúcnosti týka, očakáva sa, že hĺbková analýza bude pokračovať v

rozvoji, aby spĺňala požiadavky sveta stále bohatšieho bohatého na dáta, bude zohrávať kľúčovú úlohu pri rozvoji umelej inteligencie a modelov strojového učenia, aby organizácie mohli naďalej získavať zmysluplné informácie z rastúceho objemu dát. [9]

6 Výhody a nevýhody

Výhody a nevýhody hĺbkovej analýzy možno vidieť v nasledujúcej tabuľke:

Výhody	Nevýhody
Efektivita	Komplexnosť
Všestrannosť	Vysoké náklady
Odokrýva skryté informácie	Neistota úspechu

Tabuľka 1: Výhody a nevýhody hĺbkovej analýzy dát [12]

6.1 Výhody

- Hĺbková analýza dát zabezpečuje, že spoločnosť zhromažďuje a analyzuje spoľahlivé údaje. Často ide o prísnejší, štruktúrovaný proces, ktorý formálne identifikuje problém, zbiera údaje, ktoré súvisia s daným problémom a snaží sa nájsť riešenie. [3]

- Hĺbková analýza dát môže na prvý pohľad vyzerať v rôznych aplikáciách veľmi odlišne, ale celkový proces sa dá použiť takmer v každej novej alebo staršej aplikácii. V podstate je možné zhromažďovať a analyzovať akýkoľvek typ údajov a takmer každý problém je možné riešiť pomocou hĺbkovej analýzy.

- Cieľom hĺbkovej analýzy je vziať nespracované informácie a určiť či medzi údajmi existuje nejaká súvislosť. To umožňuje spoločnosti hodnotne narábať s informáciami, ktoré má k dispozícii a ktoré by sa na prvý pohľad nedalo postrehnúť. Dátové modely sú často zložité, no odhaľujú skryté trendy, vďaka ktorým sa môžu spoločnosti prispôsobiť a navrhnúť obchodné stratégie. [12]

6.2 Nevýhody

- Náročnosť procesu hĺbkovej analýzy je jednou z jej najväčších nevýhod. Od dátových analytikov sa očakávajú určité technické zručnosti a schopnosť pracovať s daným softvérom.

- Hĺbková analýza negarantuje, že jej výsledok prinesie spoločnosti zisky. Môže sa vykonať štatistická analýza, na základe získaných dát sa môžu prijať opatrenia, no spoločnosti z nich nemusia vždy benefitovať. To môže byť spôsobené dôsledkom nepresných zistení, neočakávaných zmien na trhu alebo chybami samotného modelu.

- Získavanie údajov pomocou hĺbkovej analýzy je spojené aj s vysokými nákladmi. Dátové nástroje často nie sú voľne dostupné, vyžadujú vysoké

predplatné a získanie niektorých častí údajov je nákladné. [12]

Dokonca aj veľké spoločnosti alebo vládne agentúry majú problémy pri používaní hĺbkovej analýzy. Napr. americká vládna spoločnosť U.S. Food and Drug administration hlásila mnohé problémy spojené so zadaním nesprávnych informácií, duplicitných údajov a nedostatočným alebo nadmerným zadávaním údajov. [5]

7 Aplikácie a využitie v praxi

Každé odvetvie zhromažďuje údaje z rôznych primárnych a sekundárnych zdrojov. Hĺbková analýza je preto použiteľná v mnohých odvetviach. Niektoré z jej aplikácií sú:

- Zdravotníctvo
  Hĺbková analýza pomáha zlepšovať kvalitu zdravotníckych systémov. Prediktívna analýza pomáha odporúčať lieky a vyhodnocovať priebeh liečby. Identifikácia podozrivého správania v nárokoch na zdravotnú starostlivosť, nákupoch liekov alebo nesúvislých predpisoch pomáha odhaľovať praktiky podvodníkov.

- Poistenie
  Hĺbková analýza pomáha poisťovniam pochopiť nákupné správanie zákazníkov a predpovedať, akú poistnú zmluvu by v budúcnosti chceli podpísať. Sledujú podvodné praktiky pri uplatňovaní poistných udalostí a posilňujú svoje systémy, aby sa im zabránilo.

- Trhová analýza
  V trhovej analýze pomáha naznačovať, že ak si zákazník kúpi určité množstvo konkrétneho produktu, môže si ho kúpiť znova alebo hľadať podobné výrobky. Pochopenie týchto údajov pomáha maloobchodníkom určiť frekvenciu nákupov a podľa toho riadiť svoje zásoby. Pomáha tiež zlepšiť predaj a riadiť vzťahy so zákazníkmi.

- Analýza financií
  Banky majú podrobné informácie o svojich klientoch, ich bankových operáciách a úveroch. Pochopenie tohto množstva údajov umožňuje bankám prispôsobovať svojim zákazníkom služby, ako sú úvery, limity výdavkov na kreditné karty, odmeny a poskytovanie zliav pri nákupoch. Identifikácia neobvyklej aktivity pri transakciách pomáha odhaliť podvodné činnosti a narušenia bezpečnosti.

- Detekcia prienikov
  Techniky hĺbkovej analýzy pomáhajú triediť informácie pre prienikové systémy. Systém upozorní používateľa ak sú nájdené cudzie prvky. Tento proces pomáha odhaliť narušenia bezpečnosti, kybernetické útoky, zneužitie a anomálie. Tieto techniky sú kľúčové pre každú spoločnosť a pomáhajú chrániť dôležité a citlivé informácie.

- Energetika
  Hĺbková analýza pomáha sledovať vzorce spotreby energie a navrhovať

systémy na zvýšenie efektivity. Pomáha predpovedať spotrebu energie v rôznych geografických lokalitách. Tieto poznatky neskôr pomáhajú optimalizovať prevádzku a investovať do zariadení, ktoré zvyšujú efektívnosť výroby.

- Vyšetrovanie trestnej činnosti
  Hlavným dôvodom využitia hĺbkovej analýzy pri vyšetrovaní trestných činov je zrýchliť riešenie prípadu. Zhluková analýza pomáha zoskupovať charakteristiky trestných činov a navrhovať spôsoby ich prevencie. Údaje z viacerých zdrojov sa analyzujú s cieľom zjednodušiť vzťahy medzi zločinom a zločincom. To pomáha nájsť podobnosti medzi činnosťami v určitom časovom období alebo geografickej lokalite.

- Médiá
  Médiá, ako sú rozhlas alebo televízia, zbierajú informácie o svojom publiku, aby pochopili ich preferencie. Na základe týchto informácií poskytovatelia médií odporúčajú obsah, menia programové plány a produkujú obsah preferovaného žánru. Ťažba dát pomáha poskytovateľom médií zlepšovať divácky zážitok. [11]

8 Pracovné pozície v oblasti hĺbkovej analýzy

Práca s hĺbkovou analýzou dát vyžaduje profesionálne technické zručnosti. V tejto oblasti sú k dispozícii inžinierske, analytické a administratívne pozície. Medzi najznámejšie pracovné pozície v tejto oblasti patria:

- Dátový analytik
  Dátoví analytici zbierajú údaje z primárnych a sekundárnych zdrojov, aby pomohli dosiahnuť ciele organizácií v oblasti hĺbkovej analýzy. Pracujú so štruktúrovanými údajmi a používajú nástroje na identifikáciu podobností a získavanie poznatkov, ktoré by mohli pomôcť rozvinúť obchodnú stratégiu spoločnosti.

- Dátový vedec
  Dátoví vedci pomáhajú vytvárať modely na prácu s údajmi. Overujú štruktúrované a neštruktúrované údaje a navrhujú algoritmy na ich ukladanie. Taktiež identifikujú trendy a poskytujú tieto informácie zainteresovaným stranám s cieľom pomôcť pri obchodných rozhodnutiach.

- Dátový správca
  Správcovia organizujú údaje pričom dávajú pozor, aby boli zorganizované podľa organizačných požiadaviek. Kontrolujú údaje a ukladajú ich na zabezpečené miesta. Správcovia údajov tiež vyhodnocujú nástroje na získavanie údajov a kontrolujú či sú pre spoločnosť prínosom.

- Dátový inžinier
  Dátoví inžinieri využívajú programovacie jazyky ako Python, SQL a Apache Spark na vývoj algoritmov, ktoré efektívne využívajú nespracované údaje. Programujú informačné panely a vizualizácie na zobrazovanie a štúdium údajov. Od dátových inžinierov sa očakáva, že budú mať pokročilé schopnosti v oblasti kódovania. [11]

9 Záver

Hĺbková analýza je užitočný proces, ktorý uľahčuje spoločnostiam pracovať s veľkým množstvom informácií. Má množstvo typov a aplikácií, pričom vzhľadom na neustále narastajúcu potrebu práce s údajmi bude tento proces v budúcnosti využívaný ešte viac, aby naplnil rôzne potreby organizácií a firiem. Má množstvo výhod, ale aj nevýhod, ktoré však vďaka technickému pokroku bude možno eliminovať a hĺbková analýza sa tak môže stať ešte dostupnejšou, ako je dnes.

Literatúra

[1] Hossein Bidgoli. *The Handbook of Technology Management: Core Concepts, Financial Tools and Techniques, Operations and Innovation Management (Volume 1)*. Wiley, 2010.

[2] Rutgers Bootcamps. What is data mining? a beginner's guide (2022). https://bootcamp.rutgers.edu/blog/what-is-data-mining/#:~:text=History%20of%20Data%20Mining,-Did%20you%20know&text=Through%20the%20Turing%20Universal%20Machine,what%20data%20mining%20is%20today.

[3] Luna Campos. A complete guide to data mining and how to use it. https://blog.hubspot.com/website/data-mining.

[4] Emma Crockett. What is data mining? types and examples. https://www.datamation.com/big-data/what-is-data-mining/.

[5] FDA. Data mining at fda – white paper. https://www.fda.gov/science-research/data-mining/data-mining-fda-white-paper#safetyreport.

[6] The Iberdola group. Data mining: definition, examples and applications. https://www.iberdrola.com/innovation/data-mining-definition-examples-and-applications#:~:text=Supermarkets%2C%20for%20example%2C%20use%20joint,sales%20at%20the%20checkout%20queue.

[7] The IBM group. What is data mining? https://www.ibm.com/topics/data-mining.

[8] MonkeyLearn. 10 best data mining tools in 2022. https://monkeylearn.com/blog/data-mining-tools/.

[9] Majid Razman and Majid Ahmad. Evolution of data mining: An overview. In *2014 Conference on IT in Business, Industry and Government (CSIBIG)*, Indore, India, March 2014.

[10] Craig Stedman. Definition - datamining. https://www.techtarget.com/searchbusinessanalytics/definition/data-mining#:~:text=Data%20mining%20is%20the%20process,make%20more%2Dinformed%20business%20decisions.

[11] Indeed Editorial Team. 15 popular data mining applications: A complete guide. https://in.indeed.com/career-advice/career-development/data-mining-applications?__cf_chl_tk=qk6BROg45SwwuL3fH1xctOATD.GNdnkygoV3keitzKc-1702549811-0-gaNycGzNDns.

[12] Alexandra Twin. What is data mining? how it works, benefits, techniques, and examples. https://www.investopedia.com/terms/d/datamining.asp.