STRUKTUROVANÁ A NESTRUKTUROVANÁ LÉKAŘSKÁ DOKUMENTACE

Michaela Stonová

Anotace

Zdravotnické zařízení je schopno vyprodukovat značné množství dat. Denní přírůstek může dosahovat až několika gigabytů. Nemocniční informační systém (dále také „NIS“) vytváří a shromažďuje data nejen v textové podobě, ale i v různých multimediálních formátech. Objem dat, jejich různorodost a požadavek na online zpracování a analýzu je činí předmětem problematiky BIG DATA.

Tento příspěvek se zaměřuje na rozdělení dat v NIS, optimalizaci metod pro jejich zpracovaní a nalezení nejvhodnějších postupů pro jejich analýzu. Primárním nositelem informace je textová lékařská dokumentace ve strukturované i nestrukturované podobě. Zvláštní zřetel je proto věnován odlišnostem obou forem při jejich obsahové analýze.

Klíčová slova

BIG DATA, nestrukturovaná data, NLP, obsahová analýza textu, strukturovaná data

1 Úvod do problematiky

Rozvoj zdravotnické techniky a systémů velmi blízce sleduje exponenciální trend nárůstu výpočetních technologií. Nejedná se sice o přesnou kopii De Moorova zákona, ale vzhledem ke skutečnosti, jak je nyní medicína silně provázána se světem počítačů, mnohdy za touto exponenciální křivkou nezaostává. V minulém století lékařské záznamy představovala pouze jedna papírová karta pacienta – v některých případech obohacena o přiložený rentgenový snímek nebo sjetinu z EKG. V současnosti je většina záznamů v digitalizované podobě. Rentgenové snímky či další výstupy ze zobrazovacích metod (MRI, CT, PET) jsou již distribuovány online nebo pomocí datových nosičů. Díky stále se zvyšující kvalitě technologií, umožňující velmi vysoké rozlišení (např. HRCT), může obrazová část dokumentace jednoho pacienta dosahovat i několika gigabytů. Oproti tomu textová část se vůči minulosti téměř nezměnila. Celkově nepřekročí pár desítek kilobytů a je tak z objemového hlediska zanedbatelná. I přes svoji mnohem menší velikost je však textová část primárním nositelem informace o pacientovi.

V nemocničním informačním systému proto můžeme nalézt velmi různorodou množinu dokumentů – velkoobjemové snímky, laboratorní výsledky, operační protokoly, propouštěcí zprávy, klasické záznamy z ambulancí nebo podklady pro pojišťovny a správu sociálního zabezpečení. Záznamy jednoho pacienta tak mohou obsahovat až několik desítek různých typů souborů o celkovém objemu dosahujícím řádu gigabytů. V případě uceleného nemocničního informačního systému můžeme hovořit o terabytech. Takto značný objem spolu s výše popsanou různorodostí zařazuje NIS do oblasti tzv. BIG DATA. Tento termín označuje hromadné zpracování dat v takovém objemu, že je výsledku obtížné dosáhnout tradičními metodami v reálném (či přiměřeném) čase. Výzvou jsou v tomto případě i běžné úlohy, jakými jsou načítání a ukládání dat nebo jejich sdílení a prohlížení. Z tohoto důvodu je nutno i tyto běžné úlohy řešit pomocí specializovaných HW a SW nástrojů (vysoce dostupná datová úložiště, výkonné výpočetní severy, robustní databáze apod.). Za „neběžné úlohy“ je následně považována obsahová analýza dat.

2 Strukturovaná data

Strukturovaná data tvoří v obvyklých informačních systémech maximálně 20 % objemu. V počátcích hromadného zpracování dat představovala strukturovaná data jedinou možnost pro alespoň částečnou analýzu. Strukturovaná data jsou jasná a přesná. Velmi zřídka svádějí k dezinterpretaci. Striktně vymezená terminologie u strukturovaných dat umožňuje jejich snadné statistické zpracování a předurčuje je tak ideálně k analýze. Na druhou stranu jejich informační vydatnost je limitována omezenou množinou povolených výrazů. Z tohoto důvodu strukturovaná data nikdy nedosáhnou takové informační kvality jako nestrukturovaná forma [14].

V oblasti zdravotnictví jsou obecně za strukturovaná data považovány dokumenty, v nichž jsou údaje o pacientovi zaznamenány pomocí předem dané formy – tj. např. záznamy ze strukturované části databáze (věk, diagnóza, pohlaví, provedené výkony, předepsaná léčiva apod.) nebo formalizované výsledky z laboratoře.

Strukturovaná data nutí lékaře zapisovat určité údaje do přesně vymezených kolonek. V případě, že se jedná o údaje typu, jméno, pohlaví, rodné číslo, MKN-10 kód nemoci, není situace kritická. Navíc většina těchto informací může být vyplněna někým jiným. V okamžiku, kdy je však lékař povinen např. všechny druhy anamnézy psát do zvláštních kolonek a k tomu ještě zaškrtat desítky políček typu – zda je pacient kuřák, kolikrát týdně cvičí, zda pije alkohol, požívá drogy, je na něco alergický, má diabetes, jaké je jeho BMI, tlak krve, cholesterol, je výsledek opačný. Vyžadovat po přepracovaném lékaři činnost, která jej časově zatěžuje a nevidí v ní smysl, je naprosto kontraproduktivní. Pouze velmi malé procento lékařů bude ochotno akceptovat tuto povinnost. Zbylá většina bude v lepším případě vymezené kolonky ignorovat a sveřepě psát informace i nadále do volného nestrukturovaného textu. V horším případě bude tuto část vyplňovat nepravdivě.

3 Nestrukturovaná data

Řešením je buď lékaře motivovat (vysvětlit smysl tohoto počínání) nebo zavést metodu, která bude schopna analyzovat nestrukturovaný text bez jejich nezbytného zásahu. Co se smyslu týče, tak ten je nezpochybnitelný. Obsahová analýza nestrukturované lékařské dokumentace může např. v rámci výzkumu napomáhat:

- odhalovat nežádoucí reakce na nová léčiva a postupy,
- nalézat chyby v lékařské dokumentaci,

- odhalovat chyby ve stávajících doporučeních (tzv. guidelines) a stanovovat nové,
- upřesňovat nežádoucí účinky kouření, alkoholu na určitě skupiny pacientů (dle věku, pohlaví, onemocnění, komorbidit),
- stanovovat diagnózu vzácných onemocnění a nalézat nejvhodnější postupy jejich léčby apod.

Je však tento přínos natolik velký, že bude takto motivován i lékař, který se na daném výzkumu aktivně nepodílí? Dále se nabízí otázka, co se stane v okamžiku, když se rozhodneme zkoumat věc, která se až doteď nalézala v nestrukturované části dokumentace?

Odpověď: Bude muset být vytvořen zcela nový formulář/struktura, pomocí kterého však budou analyzována pouze nová data.

Tímto krokem se připravujeme o značné množství vstupních dat. Toto je zvláště kritické u onemocnění s nižší četností. Zde by bylo pro získání relevantního vzorku naopak ideální propojit všechna lékařská zařízení v České republice. Jak ale zajistit, aby byla všechna zařízení mezi sebou kompatibilní, tj. aby všechna zařízení byla schopna lékaře přinutit vyplňovat stejné kolonky?

Odpověď: Velmi obtížně.

4 Analýza nestrukturovaných dat

Pro tyto případy se vyvíjejí nové metody pro obsahovou analýzu dat. Nedosahují sice takové přesnosti jako analýza strukturovaných dat, ale jsou zcela nezávislé na vstupním formátu, lze analyzovat i zpětně a především není potřeba interakce ze strany lékaře. Jejich úspěšnost se odvíjí od schopnosti stroje zvládnout tzv. zpracování přirozeného jazyka (NLP). V případě česky psané lékařské dokumentace tedy nejen na dovednosti porozumět významu česky psaného textu (rozpoznávat slovní druhy, určovat větné členy apod.), ale i na pochopení specifické lékařské terminologie.

Na základě této metody byl vybudován český model NLP pro obsahovou analýzu nestrukturovaných lékařských záznamů. Model byl otestován na záznamech z ambulancí, kde byla jeho úkolem:

- identifikace kuřáků/nekuřáků/exkuřáků,
- kvantitativní zhodnocení, kolik cigaret je denně pacientem vykouřeno.

Lékařské záznamy vztahující se ke kouření byly následně ručně vyhodnoceny. Automatizovaný klasifikační systém se od lidského hodnotitele lišil v prvním případě o 1,25 %, ve druhém o 1,99 %. Tento výsledek byl přijatelný, zvláště s uvážením, že v prvním úkolu sám hodnotící někdy obtížně stanovoval, do které kategorie jednotlivý záznam náleží [12].

Na stejném modelu byly dále otestovány možnosti identifikace předepsaných léčiv, hodnoty BMI, stanovení výše krevního tlaku či určení, zda pacient ošetřený v ambulanci byl pod vlivem alkoholu (v návaznosti na diagnózu). Vše s obdobnou chybovostí do maximální hodnoty 4 % [13].

Takto vytvořený model je možno i nadále zpřesňovat. Čím více ale bude model upravován pro konkrétní oblast, tím více bude klesat jeho univerzalita. Zároveň začne od určité úrovně (zhruba na hranici 1% chybovosti) poskytovat falešné záchyty. Tuto mez již nelze za současného stavu poznání snížit. Je třeba si uvědomit, že takto vybudovaný model nikdy nebude schopen dosáhnout stejné přesnosti jako analýza strukturovaných dat. Na druhou stranu, tento postup nepotřebuje aktivní účast lékaře, dokáže pracovat s jakýmkoliv textovým vstupem, a to i zpětně do minulosti.

4.1 Praktické doporučení pro analýzu nestrukturovaných dat

Pokud se buduje model pro nestrukturovanou lékařskou dokumentaci, tak je koncepční vytvořit jeden obecný robustní model pro celý NIS a postupně jej dopřesňovat pro jednotlivé kliniky. Tento postup je ze začátku náročnější na podchycení všech specifik zahrnutých součástí. Umožňuje ale sledovat průchod pacienta různými odděleními a analyzovat tak nejen pacienta, ale i samotné zdravotnické zařízení jako celek.

Opačným přístupem je vytvořit jeden extrémně přesných model pro jednu konkrétní kliniku, často i pro jeden konkrétní výzkumný záměr. Výsledky jsou ze začátku přesnější a počáteční investice vícenásobně nižší. Toho je však dosaženo za cenu, že lékaři musejí bezpodmínečně dodržovat přesný semistrukturovaný zápis dokumentace. Nelze jej proto použít jinde v rámci NIS, pokud nebudou tato pravidla plně dodržována i tam.

5 Shrnutí

Analýza lékařské dokumentace z nemocničního informačního systému je nástavbovou úlohou v rámci tzv. BIG DATA. Ať již se jedná o analýzu strukturovaných či nestrukturovaných dat, je potřeba mít stále na paměti následující zásady. V případě, že máme požadovanou informaci k dispozici ve strukturované formě, tak je její analýza vždy jednodušší, přesnější a levnější. Naopak analýza nestrukturovaných dat je informačně mnohem vydatnější a obecnější – nemůže však dosáhnout stejně vysoké přesnosti. Dále pokud provádíme hromadnou obsahovou analýzu nestrukturovaného textu z více oblastí NIS, tak nejprve budujeme obecný model – nikoliv naopak. A v neposlední řadě, nezapíše-li lékař podstatnou informaci do dokumentace, tak i sebelepší systém si s tímto opomenutím neporadí.

Literatura

[1.] Bastida G, Beltrán B. Ulcerative colitis in smokers, non-smokers and exsmokers. World J Gastroenterol 2011; 17: 2740–7.

[2.] Bláha M, Janča D, Klika P, Mužík J, Dušek L. Project ICOP – Architecture of Software Tool for Decision Support in Oncology. Data and Knowledge for Medical Decision Support. Proceedings of the EFMI Special Topic Conference. 2013; 130–134

[3.] Calabrese E, Yanai H, Shuster D, et al. Low-dose smoking resumption in exsmokers with refractory ulcerative colitis. J Crohns Colitis 2012; 6: 756–62.

[4.] Cochrane Collaboration. Dostupné na: https://www.cochrane.org.

[5.] Hartzband P, Groopman J. Untangling the Web – Patients, Doctors, and the Internet. N Engl J Med 2010; 362:1063–1066.

[6.] Holzinger A, Stocker C, Ofner B, Prochaska G, Brabenetz A, Hofmann-Wellenhof R. Combining HCI, Natural Language Processing, and Knowledge Discovery – Potential of IBM Content Analytics as an Assistive Technology in the Biomedical Field. Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. 2013; 7947: 13–24.

[7.] Johnson SB, Bakken S, Dine D, Hyun S, Mendonça E, Morrison F, Bright T, Van Vleck T, Wrenn J, Stetson P. An electronic health record based on structured narrative. J Am Med Inform Assoc. 2008; 15(1): 54–64.

[8.] Klimeš D, Šmíd R, Kubásek M, Vyzula R, Dušek L. DIOS – Database of Formalized chemotherapeutic Regimens. Data and Knowledge for Medical Decision Support. Proceedings of the EFMI Special Topic Conference. 2013; 165–169.

[9.] Project UIMA, Apache UIMA, Dostupné na : https://uima.apache.org/.

[10.] Regulární výrazy. Dostupné na: https://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html.

[11.] Schiff GD, Bates DW. Can Electronic Clinical Documentation Help Prevent Diagnostic Errors NEJM 2010; 362: 1066–1069.

[12.] Stonová M. Smoker identification in narrative medical records. Semantic Interoperability in Biomedicine and Healthcare. IJBH 2016; 2(1): 31–34.

[13.] Stonová M. Unstructured Data in Evidence-based Medicine and Healthcare. Semantic Interoperability in Biomedicine and Healthcare. IJBH 2015; 2(1): 47–49.

[14.] Stonová M. Unstructured Data in Healthcare. Semantic Interoperability in Biomedicine and Healthcare. IJBH 2014; 2(1): 34–36.

[15.] Walsh KE, Gurwitz JH. Medical abbreviations: writing little and communicating less. Arch. Dis. Child 2008; 93: 816–817.

[16.] Zvolský M. Automating the Use of Clinical Practice Guidelines in the Health Information Infrastructure. Semantic Interoperability in Biomedicine and Healthcare. IJBH 2014; 2(1): 51–52.

Kontakt

Michaela Stonová
1. LF Univerzita Karlova
Kateřinská 32
121 08 Praha 2
email: Michaela.Stonova@lf1.cuni.cz