Portál pre odborné publikovanie ISSN 1338-0087

Transformácia zdravotných dát do znalostných štruktúr

EMEL a kol. · Informačné technológie, Medicína

24.11.2014

Článok nadväzuje na problematiku objavovania znalostí v medicínskych dátach. Je v ňom popisovaný modifikovaný proces objavovania znalostí v neštruktúrovaných textoch. Stručne popisuje jednotlivé kroky celého procesu počnúc očistením vstupných dát a končiac uložením štruktúrovaných dát do EHR banky. Na záver je uvedený príklad transformácie konkrétnych dát do archetypovej štruktúry.

Úvod

Takmer každý lekár, ktorý vyšetrí pacienta, píše lekársku správu na osobnom počítači. Táto lekárska správa je následne vytlačená pre pacienta, a uložená v informačnom systéme či už nemocnice alebo samotného lekára. Na trhu existuje viacero informačných systémov pre lekárov či zariadenia poskytujúce zdravotnú starostlivosť, ale poväčšine tieto IS neuchovávajú dáta o pacientoch v štandardizovaných štruktúrach. Dokonca niekedy sú informácie o vyšetreniach uchovávané len v textovej podobe, ktorá je minimálne štruktúrovaná.

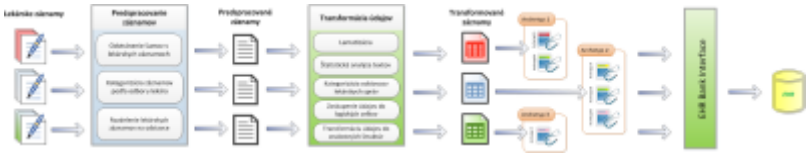
Počítač sa teda niekedy stáva moderným písacím strojom a informačný systém úložiskom dokumentov. Dáta o pacientovi a jeho chorobách častokrát nie je možné ďalej spracovávať, posúvať medzi lekármi či hlbšie analyzovať. Tento článok pojednáva o návrhu procesu, ktorý by mal neštruktúrované zdravotné záznamy transformovať do znalostných štruktúr a dáta boli použiteľnejšie.

Objavovanie znalostí v medicínskych dátach

Článok nadväzuje na problematiku objavovania znalostí v medicínskych dátach, kde bol popísaný úvod do problematiky objavovania znalostí v neštruktúrovaných textových dátach a taktiež popísaná platforma openEHR. Modifikovaný proces objavovania znalostí v textoch pozostáva z nasledujúcich krokov:

- Odstránenie šumov v lekárskych záznamoch
- Kategorizácia záznamov podľa odboru lekára
- Rozdelenie lekárskych správ na jednotlivé odstavce
- Lematizácia
- Skúmanie zoznamu najčastejších slov v správach
- Doplnenie lematizátora
- Kategorizácia odstavcov lekárskych správ

- Zoskupovanie údajov do logických celkov
- Mapovanie údajov na archetypy
- Ukladanie informácii do EHR banky



Obr. 1: Vizualizácia procesu objavovania znalostí v lekárskych záznamoch

Odstránenie šumov v lekárskych správach

Ako prvý krok v celom procese dolovania dát sme absolvovali tzv. vyčistenie vstupných dát. Tento proces bol pomerne dôležitý, lebo aj od kvality vyčistenia vstupných dát neskôr záviseli všetky výstupy. Dáta boli analyzované, boli identifikované najčastejšie zašumenia dát, teda rôzne metadáta v zdrojových súboroch, číslovania strán záznamov, preklepy a podobne. Po identifikácii takýchto najčastejších zašumení bol navrhnutý postup na automatické odstránenie najčastejších šumov. Po očistení boli dáta posunuté do ďalšieho kroku.

Kategorizácia záznamov podľa odboru lekára

Pre skvalitnenie transformácie lekárskych záznamov do archetypovej štruktúry je výhodné už na začiatku zatriediť lekárske záznamy do kategórii. Tento proces môžeme v štandardnom Knowledge discovery procese zaradiť medzi tzv. selekciu. Vyberieme si teda dáta, o ktoré máme záujem a ktoré budeme ďalej spracovávať, napr. záznamy iba od všeobecného lekára. Tento postup je dôležitý, keďže napr. všetci všeobecní lekári používajú podobnú terminológiu, zaužívané skratky a podobne. A práve tieto slová a slovné spojenia sú dôležité pri správnom nastavení procesu lematizácie. Kategorizácia môže podľa uváženia predchádzať odstráneniu šumov v lekárskych správach.

Tokenizácia

Tokenizácia je členenie textu na menšie jednotky, konkrétne slová, skratky, slovné spojenia. Token je v preklade z angličtiny znak, a teda tokenizácia je značkovanie, v našom kontexte značkovanie segmentov. Pri procese tokenizácie a segmentácie sa v transformovanom dokumente identifikujú základné lexikálne textové jednotky – slová, slovné spojenia, frázy, vety, odseky a pod. Najprv sa celý text rozdelí na jednotlivé slová a znaky. Slová sú oddelené od ostatných slov medzerami. Znaky, ktoré majú samostatný význam, napr. zátvorka, bodka za radovou číslovkou, výkričník, oddelené nemusia byť. Postupne sa elementárne textové jednotky transformujú na lexikálne jednotky – tokeny. Následne sú podľa potreby tokeny ohodnotené, je uložená informácia o ich polohe, prípadne iné parametre, ktoré budú neskôr potrebné. Celý tento proces je možné vykonať za pomoci nástroja GATE.

Rozdelenie lekárskych správ na jednotlivé odstavce

Spracovanie textu za účelom identifikácie jednotlivých častí textu a následným zaradením do nejakej štruktúry predstavuje úlohu rozdeliť text na menšie časti. Ako prvý logický krok sa ponúka rozdelenie textu na odstavce. Vychádzame z toho, že lekár

pri písaní tiež delí text do logických štruktúr správy vďaka odstavcom. Toto rozdelenie nám pomáha pri ďalšom spracovaní vstupných dát. Každý odstavec správy bol teda uložený do samostatného textového súboru. Na toto bola použitá knižnicu Apache POI, pomocou ktorej bola vytvorená vlastná aplikácia na rozdeľovanie dokumentov na odstavce.

Z analýzy dokumentov, ktoré boli na vstupe k dispozícii, bolo zistené, že niekedy je text omylom štruktúrovaný vo viacerých odstavcoch, aj keď ide o rovnakú významovú (logickú) časť. V takýchto dokumentoch však lekár členil jednotlivé časti vynechaním voľného riadku. Na základe týchto zistení bola implementovaná do algoritmov možnosť rozdeliť dokument na základe odstavcov alebo na základe voľných riadkov. Výstup po tomto rozdelení sa ukladá do textových súborov, čím sa uľahčí ďalšie spracovanie textu. Jednotlivé textové súbory sú systematicky pomenúvané podľa čísla indexov odstavcov v pôvodnom dokumente, čím je zabezpečená aj možnosť spätne obnoviť pôvodný dokument. Vďaka tomuto rozšíreniu (delenie textu podľa voľných riadkov, resp. prázdnych odstavcov) sa vyriešil problém delenia kontaktných údajov alebo hlavičiek dokumentov.

Lematizácia

V texte sa zvyčajne jednotlivé slová objavujú v rôznych tvaroch, pádoch, číslach alebo osobách. Z tohto dôvodu je ich nutné prevádzať do základných tvarov, na tzv. lemy. Pri podstatných a prídavných menách je základným tvarom prvý pád jednotného čísla, pri slovesách je to neurčitok. Postup, ktorým z ľubovoľného tvaru slova dostaneme základný tvar nazývame lematizácia. Základný tvar slova dostaneme najčastejšie odstránením pádových, slovotvorných a ďalších predpôn a prípon. Očakáva sa samozrejme, že základný tvar, ktorý bude získaný má rovnaký význam ako mal jeho pôvodný gramatický tvar. Ďalším krokom bola teda lematizácia textov, vykonaná pomocou nástroja Morphonary. Lematizáciou bolo zabezpečené, aby slová s rovnakým koreňom (napr. „lekári“ a „lekár“) boli určené ako jeden term, čo je dôležité pri následnej kategorizácii odstavcov.

Skúmanie zoznamu najčastejších slov v správach

Zlematizované texty boli ďalej skúmané a analyzované za pomoci aplikácie RapidMiner. Pomocou vytvorenia zoznamu najčastejších slov mohlo byť analyzované, či daný lekár nepoužíva nejaké vlastné atypické skratky. Ak sa takéto nájdu, bolo potrebné ich vložiť do lematizátora ako nový záznam. Ak napr. lekár používa skratku “pcnt” miesto slova pacient, javí sa ako vhodné pridať dvojicu “pcnt – pacient” do lematizátora, aby túto atypickú skratku nahradil bežným slovom “pacient”.

Doplnenie lematizátora

V lekárskych správach sú často používajú cudzie slová, odborné termíny alebo dokonca vlastné špecifické označenia alebo skratky, ktoré jednotliví lekári používajú. Testovaním sa zistilo, že proces lematizácie pracoval najmenej úspešne s takouto skupinou slov, preto bolo vhodné vymyslieť spôsob, ako tieto slová nezahrnúť do procesu lematizácie alebo v prípade skratiek ich preložiť do pôvodného tvaru slova. Metodológia, zvolená na riešenie tohto problému, bola nasledujúca:

- Z niekoľkých náhodne vybraných lekárskych správ určitého lekára získame zoznam slov, ktoré patria do vyššie spomínanej skupiny slov.
- Zoznam týchto slov predložíme lekárovi a ten ich preloží do správnych alebo plných tvarov slov.
- Tento zoznam dvojíc slov bude vyexportovaný do tabuľkového súboru CSV.
- V nástroji Morphonary tento zoznam naimportujeme a pridáme do existujúceho slovníka "declined words dictionary".
- Po skončení procesu lematizácie budú tieto skratky a špeciálne slová preložené do svojich štandardných/plnovýznamových tvarov podobne, ako by boli vyskloňované slová preložené do základného tvaru slova.

Po doplnení lematizátora sa vrátil celý proces k lematizácii, bola zlematizovaná celá vzorka dát aj s doplneným lematizátorom, dáta boli podrobené analýzam a až po niekoľkých iteráciach boli dáta pripravené na ďalší krok.

Kategorizácia odstavcov lekárskych správ

V tejto fáze procesu boli klasifikované odstavce správ do správnych kategórii, ako sú napr. predpísané lieky alebo subjektívne problémy pacienta. Na to sme používali nástroj Rapidminer viacero algoritmov a postupov, pričom každý postup vykazoval inú úspešnosť.

Zoskupovanie údajov do logických celkov

Tokenizácia nám rozdelila text na malé kúsky, pričom mnohé z nich spolu súvisia. Preto sme sa v poslednej fáze snažili pomocou logických štruktúr zoskupiť jednotlivé vetné časti lekárskych správ do ucelených celkov. V prostredí GATE sme využili Annotation schema na popísanie skupiny tokenov, pričom by sme chceli zabezpečiť, aby táto činnosť bola postupne automatizovaná pravidlami a gramatikou JAPE.

Transformácia údajov do znalostných štruktúr

Keďže v tomto kroku sme mali dáta zoskupené do logických celkov a vytvorené tokeny, boli dáta pripravené na mapovanie do štruktúry archetypov. Ako posledný krok bol implementovaný samotný proces transformácie XML zdravotných záznamov do archetypových štruktúr nachádzajúcich sa v šablóne. Cieľom bolo prevedenie anotovaných záznamov, uložených v XML formáte, do formátu, ktorý bude zodpovedať schéme vygenerovanej zo šablóny. Toto mapovanie je možné vykonať za pomoci XML editora, pričom je potrebné aby podporoval funkcionalitu mapovania medzi XML súborom a XML schémou, a na základe vytvoreného mapovania automatické generovanie XSLT skriptu alebo kódu v rôznych programovacích jazykoch (napr. C, C++, Java).

Takto bolo dosiahnuté namapovanie lematizovaných a tokenizovaných záznamov do štruktúry ktorá je predpísaná archetypom. Samozrejme, tento proces bol vykonávaný zatiaľ manuálne. Pre pochopenie a navrhnutie automatizovaného procesu založeného napr. na pravidlovej identifikácii termínov bolo potrebné mapovať dáta manuálne. Ako prvý krok boli identifikovaný vo vstupných dátach najčastejšie sa vyskytujúce archetypy, aby boli používané dáta, ktoré sú čo najviac relevantné. V tabuľke nižšie sú uvedené najčastejšie vyskytujúce sa archetypy v skúmaných textoch.

Anglický názov AT	Slovenský názov AT	Typ
Problem	Problém	Evaluation
Problem/Diagnosis	Diagnóza	Evaluation
Blood pressure	Krvný tlak	Observation
Adverse Reaction	Nežiadúca reakcia	Evaluation
Respiration	Dýchanie	Observation
Heart Rate	Pulz srdca (prítomnosť)	Observation
Heart Rate – Pulse	Pulz srdca (mechanické meranie)	Observation
Body Temperature	Telesná teplota	Observation
Medication (Structure)	Liečba	Structure
Inspection of Ear Canal	Vyšetrenie ušného kanála	Observation
Inspection of Nose	Vyšetrenie nosa	Observation
Laboratory Test Full Blood Count	Kompletný kvný obraz	Observation
Urinalysis	Rozbor moču	Observation
Abdomen Examination	Vyšetrenie brucha	Observation
Lab Result Annotation	Vyhodnotenie laboratórneho testu	Observation
Liver Function Test	Pečeňové funkčné testy	Observation

Po identifikácii najčastejšie sa vyskytujúcich archetypov v našej vzorke dát, nasledovalo mapovanie identifikovaných archetypov do štruktúry. Príklad mapovania konkrétneho archetypu uvádzame nižšie.

Príklad mapovania krvného tlaku do archetypovej štruktúry

Pri mapovaní krvného tlaku bola použitá funkcia na prácu s reťazcami, ktoré vedia z daného reťazca zobrať časť pred definovaným znakom a časť po tomto znaku. Tak bol rozdelený napr. reťazec 150/90 na základe znaku „/“ na dve požadované hodnoty tlaku: systolickú a diastolickú. Obe tieto hodnoty sme následne namapovali na vhodné miesto v schéme archetypu. Všetky ostatné vyžadované dáta boli do archetypovej štruktúry krvného tlaku v schéme priradené prostredníctvom konštánt.

Príklad mapovania rozboru moču do archetypovej štruktúry

Pri mapovaní rozboru moču bola situácia zložitejšia, pretože vo vstupnom zdrojovom súbore sú všetky látky z tohto rozboru uložené v elementoch s rovnakým názvom a to: <latka_v_moci>. V definícii relevantného archetypu má však každá látka definovanú svoju štruktúru a vlastné obmedzenia, preto nebolo možné tieto látky priradiť v ľubovoľnom poradí. Bolo potrebné teda využiť funkciu, ktoré vie porovnať dva reťazce, a to prvú časť zo vstupného reťazca, ktorá definuje práve názov látky, s reťazcom s preddefinovaným obsahom. Hodnota vrátená z tejto funkcie (pravda alebo nepravda) bola vložená následne do filtra, ktorý má na druhom vstupe druhú časť vstupného reťazca zo vstupného súboru, ktorý vyjadruje už konkrétnu hodnotu danej látky. Ak hodnota z porovnávacej funkcie je nepravda, tak na vstup príde ďalší reťazec zo sekvencie, ale ak je hodnota pravda, tak v danom filtri už potom vieme o akú látku ide.

Druhým problémom bolo, že lekárske záznamy obsahovali hodnoty látok vyjadrené

väčšinou popisom „pos“, „neg“, „norm“, „+“, „++“, „+++“ alebo v číselnom vyjadrení, kým v archetype rozboru moču sú tie isté hodnoty vyjadrené v číselnom poradí napr. od 1 po 6, pričom však majú pridelený ten istý význam. To znamená, že výstupnú hodnotu filtra bolo nutné pred samotným priradením do schémy previesť pomocou funkcie value-map do formátu, ktorý požaduje archetyp. Výsledok z tejto funkcie bol potom priradený do odpovedajúcej položky v schéme archetypu.

Ukladanie informácii do EHR banky

Po namapovaní archetypov z lekárskeho záznamu boli všetky informácie ukladané do EHR banky. EhrBank umožňuje archetypovo orientovaným aplikáciám ukladať, získavať a spravovať štandardizované celoživotné do budúcna využiteľné zdravotné záznamy. EhrBank interne spravuje relácie, dotazovanie, verzionovanie a auditovanie nad uloženými dátami. Dáta sú uložené v EhrBanke v openEHR kompatibilnom formáte. Zdravotný a administratívny obsah je uložený vo verzionovaných kompozíciách s ostatnými prvkami EHR, ktoré vyjadrujú prístupové práva, EHR stav a logické systémové adresáre.

openEHR kompozície sú flexibilné generické štruktúry, štruktúrované pomocou openEHR šablón, archetypov. Zdravotné dáta z akéhokoľvek zdroja, vrátane správ, dokumentov a vlastných databáz môžu byť reprezentované ako openEHR kompozície. Archetypové identifikátory sú zahrnuté do zdravotných údajov čo umožňuje archetypovo orientované dotazovanie. Dáta je možné exportovať nielen do formátu openEHR ale aj do iných formátov, zahrňujúcich PDF, ISO 13606 a HL7 CDA formáty.

Poďakovanie

Podporujeme výskumné aktivity na Slovensku / Projekt je spolufinancovaný zo zdrojov EÚ

Literatúra

1. LESLIE, H. Introduction to Archetypes and Archetype classes. [online]. 2012. [cit. 2013-03-06]. Dostupné na: http://www.openehr.org/wiki/display/healthmod/Introduction+to+Archetypes+and+Archetype+classes
2. Eichelberg, M. et al. A survey and analysis of electronic healthcare record standards. In ACM Computing Surveys. [online]. 2005, vol. 37, no. 4 [cit. 2013-05-05]. Dostupné k stiahnutiu na: http://dl.acm.org/citation.cfm?id=1118891&bnc=1
3. http://trac.openehr.jp/wiki/Archtectural%20Overview%20Overview
4. http://en.wikipedia.org/wiki/SNOMED_CT
5. Paralič, Ján, a iní. 2010. Dolovanie znalostí z textov. Košice : Technická univerzita v Košiciach, 2010. ISBN 978-80-89284-62-7