

Projekt MedIEQ: hodnocení zdravotnických webových zdrojů s využitím extrakce informací¹

Jirka KOSEK, Martin LABSKÝ, Jan NEMRAVA,
Marek RŮŽIČKA, Vojtěch SVÁTEK

¹*Katedra informačního a znalostního inženýrství, Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3
jirka@kosek.cz, labsky@vse.cz, nemrava@vse.cz,
ruzicka@vse.cz, svatek@vse.cz*

Abstrakt. Nově zahájený projekt EU MedIEQ se zaměřuje na problematiku automatizace hodnocení zdravotnických zdrojů na WWW. V jeho rámci je vyvíjen soubor kritérií hodnocení kvality zdrojů a nástroje pro automatické vyhledání relevantních stránek, extrakci klíčových informací a jejich transformaci na indikátory kvality.

Klíčová slova: WWW, zdravotnické informace, extrakce informací.

1 Úvod

Skutečnost, že prakticky kdokoli může vytvářet webové zdroje a vystavovat na nich "zdravotnické" informace pochybné kvality, představuje závažný společenský problém. Takové informace totiž mohou být zavádějící a vést případně i k ohrožení života. Lékařské asociace a agentury v celé Evropě proto každý den investují velké úsilí do ručního resp. částečně automatizovaného hodnocení kvality obsahu relevantních webových stránek. Vzhledem k velkému počtu těchto stránek se v posledních letech klade důraz na možnost automatizované podpory celého procesu. V tomto krátkém sdělení se nejprve pokoušíme stručně vymezit stav poznání v předmětné oblasti (sekce 2), poté uvádíme stručný popis nového projektu MedIEQ (sekce 3), a věnujeme se vybraným softwarovým nástrojům, které budou v projektu vyvinuty nebo adaptovány, a to těm, za které zodpovídá český partner (sekce 4).

2 Současný stav ve světě a v ČR

V posledních letech se problematice hodnocení kvality zdravotnických webových sídel věnovalo několik projektů EU. Jedná se zejména o *MedCertain*, *MedCircle*, a

¹ Projekt "MedIEQ: Quality labeling of medical web content using multilingual information extraction" je částečně podporován Evropskou komisí v rámci DG Health & Consumer Protection, Public Health Programme. Autoři děkují doc. MUDr. Pavlu Kasalovi z Ústavu lékařské informatiky 2. LF UK za informace o tuzemských projektech RankMed a CitMed.

Projekt MedIEQ: hodnocení zdravotnických webových zdrojů...

WrapIn. V rámci projektu *MedCertain*² vznikl nástroj pro distribuované vytváření hodnocení kvality zdrojů. Jádrem projektu *MedCircle*³ byla spolupráce čtyř hodnotících agentur, směřující ke společnému návrhu slovníku specializovaných metadat. Projekt *WrapIn*⁴ se zaměřoval na podporu vyhledávání kvalitních zdravotnických stránek technikami zpracování přirozeného jazyka.

Česká republika byla do výzkumu a vývoje v oblasti zdravotnického WWW dosud zapojena zejména prostřednictvím Ústavu lékařské informatiky 2. lékařské fakulty UK v Praze. Toto pracoviště vyvinulo systémy pro automatické hodnocení kvality webových sídel nazvané *RankMed*⁵ a *CitMed*⁶. Systém *RankMed* se zaměřuje na automatické hodnocení tuzemských sídel podle sady dvaceti kritérií, zahrnujících jak problematiku uživatelské přístupnosti a přívětivosti (např. rychlost načítání, jednotný vzhled, cizojazyčné verze), tak důvěryhodnost informací (např. uvedení autora). Systém *CitMed* má celosvětové pokrytí, rozlišuje stránky podle specializací medicíny, a sleduje jednak jejich popularitu z hlediska počtů zpětných odkazů, jednak přítomnost certifikátů kvality. Podrobnější informace lze nalézt v [1].

V žádném z uvedených projektů nebyla zatím ve větším rozsahu použita technologie automatické extrakce informací z volně strukturovaných dat ani podpora vícejazyčnými terminologickými zdroji. Prototypové uplatnění těchto technologií je hlavním cílem projektu *MedIEQ*.

3 Projekt MedIEQ

Evropský projekt *MedIEQ* byl zahájen 1.1.2006 ve spolupráci 8 partnerů ze 6 zemí; podrobný popis projektu a výčet partnerů je na <http://www.medieq.org>. Z české strany je zástupcem v projektu pracoviště Vysoká škola ekonomická v Praze, které koordinuje práce na dílčím úkolu č.6 „Information Extraction“ a podílí se na několika dalších. Účast pracoviště přímo navazuje na řešení tuzemského projektu *Rainbow* (podporovaného grantem GAČR 201/03/1318, viz <http://rainbow.vse.cz>), který měl ovšem odlišné aplikační zaměření.

4 Vybrané dílčí technologie

4.1 Ex – systém pro extrakci informací s využitím ontologií

Ex [3] je nástroj ve stadiu vývoje určený pro automatickou extrakci informací z textů. Úkolem systému je v analyzovaném textu nalézt popisy objektů určitého zadaného typu. Systém dokáže pracovat i s čistým textem, nicméně využívá případné formátování textu, např. zvýraznění nadpisů nebo strukturování do tabulek. Typy extrahovaných objektů a jejich popis zadává uživatel ve formě tzv. *extrakční*

² <http://www.medcertain.org>

³ <http://www.medcircle.org>

⁴ <http://www.wrapin.org>

⁵ <http://www.rankmed.cz>

⁶ <http://www.citmed.cz>

Krátké sdělení

ontologie, která obsahuje definice *tříd*, *atributů* a různých *omezení* (např. kardinality). Oproti doménové ontologii obsahuje extrakční ontologie pro každý atribut řadu extrakčních *evidencí*, které pomáhají hodnoty daného atributu identifikovat v textu. Jedná se např. o vzory definované (pro hodnotu atributu nebo pro jeho kontext – slova v blízkosti hodnoty atributu) na řetězcích určitých slov, lemmat nebo regulárních výrazů. Jiným druhem evidence je u číselných atributů jejich pravděpodobnostní rozdělení nebo minimum resp. maximum. Evidence se zadává *ručně* zejména v případě, že uživatel nemá pro daný atribut k dispozici dost trénovacích dat. Pokud jsou k dispozici trénovací data v podobě ručně anotovaných dokumentů nebo příkladů extrahovaných objektů, bude systém schopen *indukovat* evidence z těchto dat. Každá evidence je asociována s přesností a úplností, které charakterizují, jak pravděpodobný je výskyt daného atributu v případě pozorování dané evidence, resp. jaký je podíl hodnot daného atributu, které vykazují danou evidenci. Více evidencí systém skládá pravděpodobnostně za předpokladu vzájemné nezávislosti evidencí.

Výstupem systému je seznam extrahovaných objektů. Objekty jsou z jednotlivých kandidátů na hodnoty atributů vytvořeny pomocí omezení definovaných v extrakční ontologii a na základě předpokladu blízkosti atributů jednoho objektu v textu. Nejdříve je postupným skládáním pravděpodobných kandidátů na atributy vytvořena množina kandidátů na objekty, která je nakonec omezena na objekty validní podle ontologie. V rámci těchto objektů je pomocí dynamického programování nalezena nejpravděpodobnější cesta textem dokumentu taková, že se objekty vzájemně nepřekrývají. V případě patrné *formátovací struktury* dokumentu (resp. více dokumentů) bude systém schopen tuto strukturu při syntaktické analýze objektů využít analogicky jako nástroje typu „wrapper“.

Předběžné experimenty se systémem zatím probíhají na dvou doménách. V rámci projektu MedIEQ jde o extrakci kontaktních informací (jmen osob a firem, adres, telefonů a e-mailů), odborných termínů a názvů léků. Dále experimentujeme s extrakcí popisů zboží z internetových obchodů.

4.2 Nástroje pro webovou navigaci a kategorizaci odkazů

Součástí projektu je i vývoj *navigační komponenty*, která, obdobně jako webové roboty, prochází webové sídlo od zadaného kořenu, na každé stránce detekuje všechny odkazy vedoucí na tentýž server, a ty následně navštíví a analyzuje. Současně si vytváří lokální kopie jednotlivých stránek. Uložené stránky jsou postoupeny klasifikátoru obsahu (vyvinutému jiným partnerem v projektu).

Pro zrychlení navigace a odfiltrování nepodstatných stránek je dále vyvíjena *komponenta pro kategorizaci odkazů*. Jejím úkolem je odhadnout kategorii stránky již na základě analýzy odkazu (včetně jeho okolí), který na tuto stránku směřuje. Pro tuto komponentu bylo nutné definovat kategorie stránek, které se bude snažit rozlišit, jako jsou stránky s léčebnými postupy, kontakty nebo online poradnou. Komponenta se dále rozpadá na několik modulů, kde každý z nich zastupuje jednu kategorii a přiřazuje danému odkazu váhu podle jeho relevance. Rozhodnutí o váze odkazu závisí na attributech sledovatelných přímo na webové stránce. Mezi tyto atributy patří například: nejkratší vzdálenost na stránku z úvodní stránky, text okolo odkazu nebo charakteristiky URL adresy. Odkazovaná stránka je navštívena pouze tehdy, když váha alespoň jednoho modulu překročí požadovanou mez. Rozhodovací mechanismus

modulů je nastaven pomocí metod strojového učení a pro tyto účely jsou použita stejná trénovací data jako pro klasifikátor obsahu. Moduly lze dynamicky přidávat a každý z nich může pracovat s jinou skladbou sledovaných atributů.

4.3 Relaxed – validátor HTML stránek

Pro hodnocení přístupnosti (“accessibility”) stránek jako zvláštního typu indikátoru kvality se bude používat validátor *Relaxed*⁷ [2]. Jedná se o validátor HTML a XHTML kódu postavený nad moderními schémovými jazyky *RELAX NG* a *Schematron*. Právě jazyk *Schematron* umožňuje pomocí sady *XPath* výrazů popsat podmínky, kterým musí dokument vyhovět. Validátor *Relaxed* již ve své standardní verzi obsahuje řadu pravidel, která kontrolují shodu stránky s doporučeními pro přístupnost *WCAG* (Web Content Accessibility Guidelines). Existující pravidla budou dále rozšířena a zpřesněna, aby lépe vyhovovala potřebám projektu *MedIEQ*.

5 Závěr

Projekt *MedIEQ* představuje užitečný podnět pro rozvoj nástrojů automatické analýzy WWW, jednak tím, že směřuje k praktické aplikaci, jednak svou komplexností, která vyžaduje propojení velkého počtu nástrojů různého typu. Předpokládáme, že systém vyvíjený v jeho rámci vhodným způsobem doplní stávající inventář používaný agenturami hodnotícími zdravotnická webová sídla, a usnadní práci jejich expertů.

Literatura

1. Kasal, P. et al.: Evaluation of Health Care Related Web Resources Based on Web Citation Analysis and Other Quality Criteria. In: *Abstracts Book of International Conference of the IEEE - EMBS 2005*, Shanghai, China. New Jersey: IEEE 2005.
2. Kosek, J., Nálevka, P.: Relaxed—on the Way Towards True Validation of Compound Documents. In: *WWW 2006*, Edinburgh, Scotland, ACM 2006.
3. Labský, M., Svátek, V.: On the Design and Exploitation of Presentation Ontologies for Information Extraction. In: *Proc. ESWC'06 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*, Budva, Montenegro, June 2006.

Annotation:

The MedIEQ project: evaluation of medical web resources with the help of information extraction

The recently started EU project *MedIEQ* focuses on the problem of automated evaluation of medical web resources. The goal of the project is to develop a collection of quality criteria and to design tools for automated discovery of relevant pages, extraction of key information and their transformation to quality indicators.

⁷<http://badame.vse.cz/validator>