Srovnání přístupů extrakce užitečné informace z webu

Michal Toman

Katedra informatiky a výpočetní techniky, FAV,
Západočeská univerzita v Plzni, Univerzitní 22, 306 14, Plzeň
mtoman@kiv.zcu.cz

Abstrakt. V článku srovnáváme dvě metody pro extrakci užitečné informace z webu. První metoda je založena na statistické analýze struktury webové stránky a druhá metoda využívá dotazy XQuery pro extrakci informace z částečně strukturovaných dokumentů. V testech srovnáváme přesnost a úplnost automatické extrakce pomocí obou metod a ručně vytvářeného referenčního extraktu.

Klíčová slova: textový korpus, web, extrakce informace, wrapper, xml

1 Úvod

Prostředí webu se postupně vyvinulo v obecný zdroj informací uchovaných převážně v částečně strukturovaném formátu HTML. Stávající Web obsahuje data, která jsou určena pro prohlížení uživatelem, jenž zobrazenému textu přiřadí správnou sémantickou informaci. Jednotlivé zdroje vyjadřují informace v rozdílných formátech a různým způsobem, což pro člověka nepředstavuje větší problém, ale komplikuje jejich porozumění počítačem. V tomto článku se zaměřujeme především na extrakci informací z HTML stránek, které představují většinu dat na webu. Extrakcí dat míníme transformaci vybraných částí zdrojové HTML stránky do XML formátu pomocí tzv. *wrapperů* [5]. Wrappery lze rozdělit do několika skupin. První skupina využívá strukturu webové stránky – např. W4F[7] nebo Lixto [1]. Odlišný přístup je použitý u wrapperu Cameleon# [3], který zpracovává webovou stránku jako prostý text. Jiné dělení je možné provést podle míry automatizace wrapperu. Od manuálních přístupů k automatickým roste robustnost, klesá čas potřebný na natrénování wrapperu, ale může klesat přesnost extrakce. Mezi vysoce automatické metody, kde jsou použity metody strojového učení, patří např. SoftMealy [4] nebo algoritmus Stalker [6].

Původní motivací k návrhu našich metod pro extrakci informací z webových stránek byl požadavek na vytvoření rozsáhlého vícejazyčného textového korpusu. Cílem extrakce dat bylo získat z webových stránek čistý text a přidružená metadata, jmenovitě: čas publikování příspěvku, autora, kategorii, název, perex.

Pro řešení problému jsme navrhli dvě odlišné metody popsané v sekci 2, které podle provedených testů publikovaných v sekci 3, poskytují uspokojivé výsledky. První metoda je určena výhradně pro tvorbu textových korpusů a druhá metoda je vhodná nejen pro extrakci textu, ale i pro obecnou extrakci dat z webu. K ověření metod jsme vytvořili programový systém pro extrakci informací z webu.

2 Metody extrakce informace

2.1 Statistická metoda NIT

Vytvořenou metodu NIT (Node Information Threshold) lze zařadit do skupiny automatických metod využívajících strukturu webové stránky. NIT transformuje zdrojový dokument do hierarchické struktury DOM[8]. Každý uzel DOM představuje jeden element webové stránky. Metoda NIT je založena na detekci *užitečných uzlů* v hierarchii DOM. Užitečné uzly jsou následně zvoleny do výsledku extrakce a v ideálním případě reprezentují čistý text článku.

Algoritmus NIT lze popsat následující způsobem:

- Je zavedena funkce *I(N)* vyhodnocující informační hodnotu uzlu N v hierarchii DOM. V našem případě funkce *I(N)* nabývá hodnot z intervalu *<0, 1>* a je definována jako počet zobrazených písmen daného uzlu (elementu HTML stránky) v poměru k celé délce dokumentu. Pro vylepšení výsledků je možné funkci modifikovat například použitím váhové funkce TF-IDF, případně výstupem klasifikátoru, který určuje míru příslušnosti každého slova v uzlu k třídě obsahující uzly s užitečnou informací nebo uzly bez užitečné informace.
- Sourozenecké uzly hierarchie DOM jsou vždy sestupně seřazeny do posloupnosti tak, aby platilo: $I(N_1) > I(N_2) > \ldots > I(N_k)$.
- Následně je hledáno $n < k$ takové, aby byla splněna formule (1).
- Uzly N_1 , N_2 ... N_n jsou označeny jako užitečné a jsou vybrány do extraktu.

$$\frac{\sum_{i=1}^{n} I(N_i)}{n} > \tau > \frac{\sum_{i=1}^{n+1} I(N_i)}{n+1} \qquad (1)$$

Konstanta τ je volena při kalibraci metody. Její volbu je nutné stanovit empiricky a může být pro každé webové sídlo odlišná. Hodnota mění přesnost a úplnost extrakce.

2.2 Metoda vzorů XQT

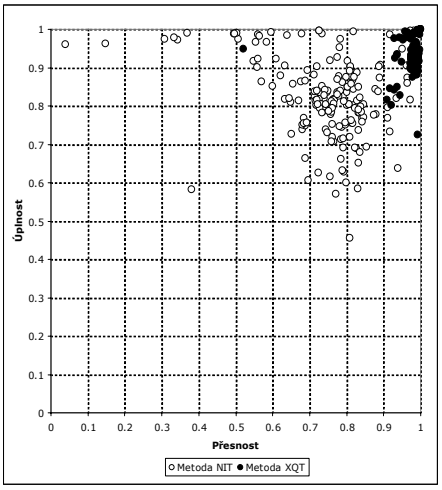
Metoda XQT (XQuery Templates) vychází z myšlenky transformace HTML pomocí XQuery do výstupního XML formátu. Jedná se o poloautomatickou metodu využívající vlastností webových stránek a její filozofie je podobná systému Lixto[1] a W4F[7]. Zatímco je v systému Lixto pro popis extrakčních pravidel používán proprietární jazyk Elog, v metodě XQT je pro popis využit jazyk XQuery [2].

Pro každé webové sídlo je možné vytvořit šablonu ve vizuálním editoru šablon. XQT šablona obsahuje strukturu výstupních dat a XQuery dotazy vygenerované editorem šablon, které mapují části webové stránky do výstupního formátu. Takto vytvořená šablona je použitelná pro dávkové zpracování stažených stránek z webového sídla. Výstupem systému je XML databáze obsahující strukturovaná data.

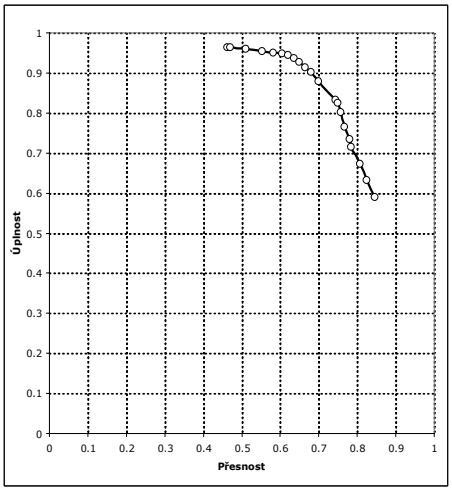
Systém používající XQT je poměrně obecný a umožňuje jak extrakci užitečných textů – podobně jako metoda NIT, tak extrakci obecných dat z webových stránek.

3 Experimenty

Pro testy obou navržených metod jsme použili korpus obsahující stránky informačního serveru Deutsche Welle. Vstupem metod bylo 200 náhodně vybraných stránek z korpusu, ke kterým byly ručně vytvořeny referenční texty představující ideální extrakt.



Graf 1. Výsledky přesnosti a úplnosti extrakce metodou NIT(τ =0,018) a XQT



Graf 2. Závislost přesnosti a úplnosti metody NIT na parametru τ

3.1 Metoda NIT

V zhledem k faktu, že metoda NIT je plně automatická, poskytuje systém slibné výsledky. Pro kalibraci metody je nutné zvolit pouze jediný parametr, který nastaví práh τ. Vhodnou volbou se průměrná přesnost a úplnost metody pohybuje kolem hodnoty 0,8. Výsledek je vzhledem k použité funkci *I(N)* velmi závislý na struktuře a délce textu tiskové zprávy. Metoda je navržena tak, aby bylo možné funkci *I(N)* kdykoliv nahradit bez nutnosti změny zbylé části algoritmu.

Změnou prahové hodnoty τ se mění míra akceptace méně hodnotných uzlů do výsledného extraktu. Pokud je práh nízký, jsou častěji akceptovány uzly s menší hodnotou funkce *I(N)*. Při tomto nastavení přesnost extrakce klesá a úplnost roste. V případě nastavení vyššího prahu jsou vybírány pouze informačně hodnotnější uzly, tedy přesnost extrakce roste a naopak úplnost klesá. Výsledky testu jsou zobrazeny v grafu 2. Slabinou metody NIT je časté zahrnutí zápatí stránek do výsledného extraktu, což v případě kratšího článku výrazně zhorší přesnost extrakce. Extrémně jemné dělení článku na sekce snižuje úplnost extrakce, protože daný uzel je považován za nevýznamný. Pro účel tvorby rozsáhlých textových korpusů lze považovat metodu za dostatečně přesnou.

3.2 Metoda XQT

Metoda XQT poskytuje podle očekávání lepší výsledky než metoda NIT. Výsledky jsou ovšem vyváženy nutností tvorby šablon. Přesnost metody se pro extrakci textu pohybuje nad 95 % při dodržení 90 % úplnosti extraktu, což je srovnatelné s ruční extrakcí. Problematické pro extrakci textu jsou stránky obsahující číslované seznamy, části textu obsažené v tabulkách, případně vnořené objekty narušující spojitost textu.

4 Závěr

V tomto článku jsme prezentovali dvě metody pro extrakci textu z webových stránek a provedli jejich taxonomické začlenění. Metoda XQT produkuje přesné výsledky s možností jemného strukturování výsledných dat. Je vhodná pro tvorbu textových korpusů a extrakci obecných dat s důrazem na přesnost.

Metoda NIT poskytuje výsledky s přesností a úplností pohybující se kolem 80 %. Použití metody je jednoduché, protože kvalita extrakce závisí na jediném parametru. Metoda je určená výhradně pro tvorbu textových korpusů z webových zdrojů.

Tento výzkum byl částečně podpořen NPV II, projekt 2C06009 (COT-SEWing).

Reference

1. Baumgartner R., Flesca, S., Gottlob, G. Visual Web Information Extraction with Lixto, *The VLDB Journal*, pp 119-128, 2001.
2. Chamberlin D., XQuery: A query language for XML. http://www.w3.org/
3. Firat A., Madnick S., Yahaya A., Kuan W., Bressan S., Information Aggregation using Cameleon# Web Wrapper, *6th ICECWT*, 2005, ISBN 978-3-540-28467-3.
4. Hsu, C., Dung, M., Generating Finite State Transducers for Semi-Structured Data Extraction form the Web. *Information systéme 23*, 1998, pp 521-538.
5. Laender A.,Ribeiro-Neto B., Silva A., Teixeira J., A Brief Survey of Web Data Extraction Tools, *SIGMOD Record 31*, 2002, ISSN 0163-5808.
6. Muslea, I., Minton, S., Knoblock, C. Hierarchical wrapper induction for semistructured information sources. *AAMA 4*, pp 93-104, 2001.
7. Sahuguet A. Azavant F. Building Intelligent Web Applications using Lightweighted wrappers. *Data and Knowledge Eng. 36*, 2001, pp 283-316.
8. WWW Consortium *W3C*. The Document Object Model. http://www.w3.org/DOM

Annotation:

Comparison of approaches for information extraction from the web.

In this paper we compare two different methods for information extraction from web data. The first method is statistical-based and the second one uses the XQueries. We compare both methods and we perform precision and recall tests.