Procedurální propojení nástrojů pro extrakci informací z webových sídel

Ondřej Šváb, Vojtěch Svátek

Katedra informačního a znalostního inženýrství, VŠE Praha
nám. Winstona Churchilla 4, 130 67, Praha 3
{xsvao06|svatek}@vse.cz

Abstrakt. Tento příspěvek se věnuje procedurálnímu propojení nástrojů pro extrakci informací z webových stránek. Prvním krokem je rozhodování o kategorii webové stránky, na tomto základě následuje příslušná extrakce informací dle kategorie webu. Z produktových katalogů jsou informace extrahovány pomocí *skrytých markovských modelů*. Z úvodních stránek webových sídel se extrahují celé věty obsahující profilové informace firem. Ze všech webových stránek se pak extrahují jejich metainformace. Všechny tyto činnosti jsou vykonávány oddělenými, samostatnými moduly vyvíjenými v rámci projektu *Rainbow*. Extrahované informace jsou převedeny do *RDF formátu*, uloženy do *RDF* souborů a následně nahrány do systému *Sesame*.

Klíčová slova: algoritmus, extrakce informací, RDF

1 Úvod

Předpokladem nové podoby *WWW* vedoucí k naplnění myšlenky *sémantického webu* [2] jsou informace obohacené o jejich sémantiku, jazyky které nabízejí dostatečnou vyjadřovací sílu (např. *RDF* [8], *OWL* [7]) a různé softwarové aplikace od samotných specifických úložišť až ke zprostředkujícím aplikacím. Nové pojetí webu tak staré *WWW* rozšiřuje, aniž by ono původní vymizelo. Vzniká tak potřeba obohatit stávající webové stránky o sémantiku jejich informací. To se stalo živnou půdou pro extrakci informací a naplňování doménových ontologií.

Základním cílem tohoto sekvenčního propojení je testování, a to jednak testování jednotlivých analytických nástrojů, jednak testování jejich možného provázání. Tento způsob propojení tak představuje určitou modelovou situaci na níž lze zkoušet funkčnost celku a jeho částí na určité aplikační doméně. V případě projektu *Rainbow*[1] je vybranou aplikační oblastí nabídka prodeje bicyklů. V rámci projektu *Rainbow* se předpokládá pohyb směrem k flexibilnějšímu skládání aplikace uživateli za využití jednotlivých dostupných analytických nástrojů různého charakteru. Tyto nástroje poskytované jako *webové služby* [9] vhodným způsobem popsané by uživatel mohl zaměřit na libovolnou aplikační oblast s případným natrénováním jednotlivých nástrojů (jejich modelů).

[1] http://rainbow.vse.cz/

V současné době se pro vývoj jednotné aplikace, která zahrnuje extrahování nabídek kol z webových katalogů a zpřístupnění extrahovaných kol uživateli uvažuje sedm nástrojů[2], z nichž kromě *AmphoraWS* a systému *Sesame* byly všechny vytvořeny v rámci projektu *Rainbow*:

- *AmphoraWS* [1] (WS) se v současné aplikaci používá pro získávání *URL* jednotlivých webových stránek. Nástroj byl vytvořen na *Vysoké škole báňské - Technická univerzita Ostrava* v rámci skupiny ARG[3].
- *Lingvistická analýza* [4] (WS) extrahuje na základě slovních indikátorů profilovou informaci o firmě.
- *MetaTags modul* [4] (WS) extrahuje obsahy *META tagů* webových stránek; obvykle jimi jsou jméno autora stránky, popis stránky a klíčová slova.
- *URL analýza* [4] (WS) na základě analýzy *URL* určuje kategorii webové stránky.
- *Hidden Markov Model (HMM) modul* [4] (WS) extrahuje bicyklové nabídky z produktových katalogů pomocí statistické metody *skrytých markovských modelů*. Více se o tom lze dočíst např. v [6].
- *Sesame* [3] je *RDF* a *RDFS repositář* umožňující jednak ukládání *RDF* a *RDFS*, jednak jejich dotazování s ohledem k sémantice *RDF* a *RDFS*. *Sesame* představuje *back-end aplikace*. Nástroj byl vytvořen jako *open-source* společností *Aduna*[4]
- *HTML vyhledávací rozhraní*[5] představuje *front-end aplikace* umožňující uživateli vyhledávat přes formulářová pole v databázi *Sesame* informace extrahované ostatními nástroji.

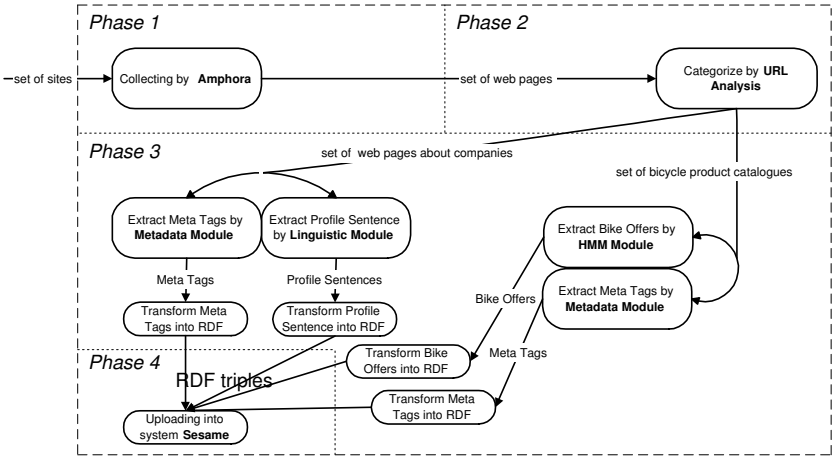
V další části posteru je popsáno schéma propojení nástrojů pro účely extrahování nabídek kol.

2 Schéma propojení

Propojení zahrnuje čtyři navazující fáze, jejichž základní podoba je zachycena v grafickém znázornění na obrázku č. 1. Předpokládá se, že vstupem do celého rámcového algoritmu je množina *webových sídel* prodejců bicyklů. Tato *URL* jsou v současné verzi shromážděna ručně z Google Directory *Sports-Cycling-Bike Shops-Europe-UK-England*. Výsledkem celého algoritmu jsou *RDF tvrzení*, která budou následně načtená do *RDF repositáře*.

V *první fázi* se z webových sídel získají *URL* všech odkazovaných stránek do hloubky úrovně dva, tzn. získají se *URL* i z prvních odkazovaných webových stránek. Důvodem volby této hloubky je zkušenost ukazující na fakt, že právě na druhé úrovni jsou stránky s bohatou informací o produktu. Tento úkol se řeší pomocí nástroje *AmphoraWS*.

[2] Nástroje, které jsou dostupné jako webové služby jsou v závorce označeny zkratkou WS.
[3] http://www.cs.vsb.cz/arg/, Amphora Research Group
[4] http://www.openrdf.org
[5] http://rainbow.vse.cz:8000/sesame/, *Bicycle Sale Domain RDF repository*



Obr. 1. Grafické znázornění průběhu zpracování.

Ve *druhé fázi* se jednotlivé webové stránky získané z předcházející fáze kategorizují na webové stránky, které by mohly být *produktovými katalogy* a na webové stránky na nichž lze očekávat *kontaktní (profilové) informace o společnosti*. To je úkolem *URL analýzy*. Pro klasifikaci webových stránek se plánuje využít analýzy jejich samotného obsahu *bayesovským klasifikátorem*.

Třetí fáze je ve znamení *extrakce informací* (*IE*). Z produktových katalogů se extrahují nabídky prodeje bicyklů pomocí *HMM modulu*. Obsah stránek kontaktních informací o firmě je předmětem extrakce profilových informací *lingvistickou analýzou*. Alternativně by bylo možné pro extrakci kontaktních informací využít příslušně natrénovaných *skrytých markovských modelů* podobně jako je tomu v současnosti při extrakci informací o nabídkách bicyklů. Ze všech těchto webových stránek se také extrahují jejich metadata (*MetaTags modul*), tzn. přebírají se obsahy *meta elementů* z „hlavičky“ webové stránky. Současně v této fázi probíhá transformace extrahovaných informací do *formátu RDF*. Tato transformace pro každý typ informací probíhá zvlášť; ovšem vždy stejným způsobem převodem extrahovaných informací na subjekt, predikát a objekt. Důležitým bodem je integrace výsledků extrakce jednotlivými nástroji. Toho je docíleno v průběhu transformace prostřednictvím jednoznačných identifikátorů *URI*. Tak například provádí-li se extrakce profilových informací, jsou tyto propojeny s informacemi o metadatech příslušné webové stránky jednoduše pomocí jejího *URL*. Ve *čtvrté fázi* se pak už jen jednotlivá *RDF tvrzení* nahrají do systému *Sesame* a jsou tak přístupná pomocí *HTML vyhledávacího rozhraní* koncovému uživateli.

3 Závěr

Představený způsob propojení jednotlivých analytických nástrojů je základním jednoduchým řešením. Jeho účelem je především testování analytických nástrojů,

které byly testovány zatím jednotlivě, v rámci společné vybrané aplikační domény. V současné době je implementace dokončena bez dvou zmiňovaných nástrojů, a sice bez *AmphoraWS* a *lingvistické analýzy*. Obě jsou v současné době ve stadiu dokončování. Poté budou zapojeny dle uvedeného schématu. Další důležitou činností, která se plánuje na blízkou dobu je ověřování úspěšnosti na reálných datech. V rámci projektu *Rainbow* se plánuje flexibilnější řešení propojování analytických nástrojů s využitím znalostních modelů (tomu se věnuje [5]).

Reference

1. M. Andrt, M. Krátký, V. Svátek, and V. Snášel. AmphoraWS webová služba pro vyhledávání ve strukturovaných dokumentech. *sborník konference DATAKON 2004*. Brno 2004.
2. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*. 2001.
3. J. Broekstra, A. Kampman, and F. van Harmelen: Sesame: An Architecture for Storing and Querying RDF and RDF Schema. *Proceedings of the First International Semantic Web Conference (ISWC 2002)*. Sardinia, Italy, June 9-12 2002.
4. Rainbow. http://rainbow.vse.cz/doc/services/, Rainbow Services - Reference guide
5. V. Svátek, M. Labský, M. Vacura. Knowledge Modelling for Deductive Web Mining. *Springer LNCS. EKAW 2004*. Whittlebury Hall, UK, October 2004.
6. O. Šváb, M. Labský, V. Svátek: RDF-Based Retrieval of Information Extracted from Web Product Catalogues. *Semantic Web workshop at ACM SIGIR 2004*. Sheffield, 2004.
7. W3C Consortium. http://www.w3.org/2004/OWL/, Web Ontology Language (OWL). February 2004.
8. W3C Consortium, W3C RDF Working Group. http://www.w3c.org/RDF/, Resource Description Framework (RDF). February 2004.
9. W3C Consortium, W3C Working Group. http://www.w3c.org/TR/ws-arch/, Web Service Architecture. February 2004.

Annotation:

Procedural combination of tools for IE from web sites

This paper is dealing with preliminary version of procedural combination of Information Extraction tools. There are four phases in the fixed sequence. First, all web pages are collected by one tool, then another tool decides a category of web page. In the third phase, after having obtained category of each web page, Information Extraction analyses take place. Specifically, either bicycle offers are extracted from the web product catalogues with *HMM*, or profile sentences are extracted from web pages about companies with *linguistic method*. Subsequently, metadata from all processed web pages are extracted. These results are obtained with separate tools running as web services developed within the *Rainbow project*. Finally, the results are transformed into *RDF format* and uploaded as *RDF files* into the system *Sesame*.