

Využití ontologií při indukci wrapperů

Marek Nekvasil

Katedra informačního a znalostního inženýrství, FIS, VŠE – Vysoká škola ekonomická v Praze,
Nám. Winstona Churchilla 4, 130 67, Praha 3
nekvasim@vse.cz

Abstrakt. Cílem tohoto příspěvku je návrh rozšíření ontologií za účelem jejich využití při extrakci informací z webových dokumentů. Větší část je věnována návrhu inferenčního modelu pro vyhodnocování shody vzorů a jejich skládání do šablon hodnot datotypových vlastností tříd ontologie. Dále je navržen jednoduchý způsob indukce wrapperu, který je schopen využít výsledky automatické anotace dokumentu.

Klíčová slova: ontologie, automatická anotace, extrakce informací, wrapper

1 Úvod

Nejjednodušší alternativou k ručnímu zpracování webových dokumentů je využití *wrapperu*, sady pravidel pro identifikaci hodnot v dokumentu. Lze jej tvořit ručně, nebo automaticky, pak hovoříme o *indukci wrapperů* [3]. K indukci wrapperu je zpravidla zapotřebí mít k dispozici příklady výskytů hodnot určených k extrakci, tedy anotovaný dokument. Cílem této práce je navrhnout způsob, kterým by bylo možné pro účely extrakce informací anotovat dokumenty automaticky, za využití *ontologií* [1]. Ontologii, která je využita při extrakci informací nazveme *extrakční ontologie*.

2 Rozšíření ontologie v jazyce OWL

Ontologie zapsaná v prostém jazyce OWL, má pouze velmi omezené možnosti popisu hodnot datotypových vlastností (tj. vlastností, jejichž hodnotou je literál) a neobsahuje tudíž informace nutné k identifikaci těchto hodnot v dokumentu. Abychom odstranili tento nedostatek, zavedeme rozšíření ontologie o zápis *šablon* a *vzorů* pro typické hodnoty datotypových vlastností.

Vzorem nazveme pravidlo, pro které lze pro každou spojitou část dokumentu určit s jakou mírou jej splňuje. Příkladem vzoru může být pravidlo, které vyhodnocuje, zda je část dokumentu číslem z určitého rozsahu nebo třeba řetězcem ze seznamu.

Šablonou pak nazveme takovou kombinaci vzorů, pro níž také lze pro každou část dokumentu určit s jakou mírou jí vyhovuje. Příkladem šablony může být pravidlo spojující výše uvedené vzory, tedy například číslo z určitého rozsahu v blízkosti nějakého řetězce (což může popisovat typické hodnoty vlastnosti *cena* nějakého výrobku – např. číslo v rozsahu <100,500> v blízkosti řetězce „Kč“). Šablonu můžeme považovat za speciální případ vzoru, některým ze vzorů, z nichž se šablona skládá, může tak být i další šablona. Šablony lze tímto způsobem skládat a rekurzivně zanořovat, což může mít svůj význam.

Každou shodu vzoru s částí dokumentu, tedy každé místo v dokumentu, kterému bude nějaký vzor vyhovovat, budeme považovat za podezřelého dílčího kandidáta na výskyt hodnoty datotypové vlastnosti, jíž je vzor přiřazen.

2.1 Jednoduché vzory

Při vyhodnocování shody vzorů se potýkáme s úlohou odvození míry jistoty označení dílčího kandidáta z jiných dílčích vstupů. Vzorek v tomto pojetí chápeme jako algoritmus, který podle vstupních parametrů určí pro každé místo v dokumentu míru shody, s jakou těmto parametrům odpovídá.

Zde se tedy střetáváme se dvěma různými termíny. Prvním je *míra shody vzoru*, která reprezentuje jistotu, s jakou vzorek dané místo v dokumentu označí. Druhým je pak *míra jistoty označení dílčího kandidáta na výskyt hodnoty vlastnosti*, která reprezentuje jistotu, s jakou je dané místo v dokumentu skutečně výskytem hodnoty vlastnosti, dáno izolovaným pozorováním výskytu vzoru. V tomto smyslu označení dílčího kandidáta na výskyt hodnoty vlastnosti nazveme výrazem *evidence vzoru* a druhý z uvedených termínů je tedy synonymem pro *míru jistoty evidence vzoru*.

Míra shody vzoru a míra jistoty evidence vzoru by intuitivně měly korespondovat. Přiřadíme-li shodě vzoru označení A a jistotě evidence vzoru označení E , můžeme pak zapsat takovéto inferenční pravidlo:

$$A \rightarrow E \quad (1)$$

Pro naše účely jsme zvolili model fuzzy inference [2], ale bylo by možné využít i jiný způsob usuzování. Při využívání fuzzy logiky můžeme A a E definovat jako výroky

- A – „Vzor dané místo v dokumentu označil.“
- E – „Označené místo je skutečně evidencí vzoru.“

a příslušné míry jako jejich pravdivostní hodnoty (tedy míra shody vzoru $a = val(A)$ a míra jistoty evidence vzoru $e = val(E)$). Zavedeme také dva univerzální parametry pro všechny vzory (a tedy i šablony), a sice *přesnost* a *úplnost*, a definujeme je takto :

$$p = val(A \rightarrow E) - \text{přesnost vzoru} \quad (2)$$

$$c = val(E \rightarrow A) - \text{úplnost vzoru} \quad (3)$$

Za použití těchto parametrů můžeme na Łukasiewiczově fuzzy logice odvodit tuto podobu inferenčního pravidla (podrobné odvození je k dispozici v [6]):

$$((A \& (A \rightarrow E)) \vee \neg(E \rightarrow A)) \Rightarrow E \quad (4)$$

Vyjádríme-li funkční předpis této formule za použití parametrů p a c a předpokladu nepřeceňování míry jistoty evidence vzoru e , dostaneme:

$$e = \max(a + p - 1, 1 - c) \quad (5)$$

Pomocí parametru p můžeme tak omezit míru jistoty evidence vzoru shora. Přesnost vzoru v tomto kontextu označuje jistotu, s jakou vysoká míra shody vzoru vede k vysoké míře jistoty evidence vzoru. Parametrem c je stanovena minimální hodnota míry jistoty evidence vzoru.

2.2 Složené vzory - šablony

Skládání evidencí více vzorů je množinovou operací, přičemž nám postačí sjednocení a průnik (v kombinaci s doplňkem). Stanovení míry jistoty evidence podle šablony je pak triviální záležitostí, míru shody šablony stanovíme prostým složením výsledných

měr jistoty evidencí jednotlivých dílčích vzorů šablony pomocí příslušné logické operace, čímž dostaneme dva druhy šablon – šablony konjunktivní a disjunktivní. Z této míry shody šablony pak určíme výslednou míru jistoty evidence šablony naprosto stejným způsobem, jako v případě jednoduchých vzorů.

Zajímavé je diskutovat význam parametrů p a c jednotlivých dílčích vzorů šablony při různých typech jejich skládání. Pokud uvažujeme disjunktivní skládání vzorů, znamená vysoká hodnota parametru p dílčího vzoru, že jeho splnění je pro celkovou shodu šablony podmínkou postačující. V případě použití konjunktivního skládání vzorů vysoká hodnota c znamená, že splnění vzoru je pro celkovou shodu šablony podmínkou nutnou.

2.3 Návrh vzorů

Vzory budeme v extrakční ontologii zapisovat formou XML elementů ze zvláštního jmenného prostoru vnořených elementům příslušných datotypových vlastností. Vzory mohou být nejrůznějšího rázu, od vzorů pro jednoduchou shodu řetězce, přes vzory vyhodnocující regulární výraz po vzory porovnávající hodnoty náhodné veličiny podle příslušného rozdělení. Několik základních vzorů je navrženo v [6], avšak je možno navrhnout i další. Nezbytnou součástí návrhu vzoru je specifikace způsobu určení míry shody tohoto vzoru.

3 Jednoduchý způsob indukce wrapperu

Aplikací pravidel jednotlivých vzorů a šablon na obsah dokumentu získáme množinu evidencí spolu s jejich jistotami pro každou datotypovou vlastnost. Omezíme-li se na tabulární strukturu dat v dokumentu, můžeme rozdělit jednotlivé evidence do segmentů, podle počátku jejich absolutní XPath cesty. Tyto segmenty můžeme očistit s využitím parametru p , když prohlásíme, že odpovídá podílu všech označených hodnot, které jsou skutečně výskytem dané vlastnosti. Seřadíme-li segmenty podle jejich podílu na celku, můžeme je postupně vyřazovat jako zavádějící, dokud celkový poměr vyřazených nedosáhne hodnoty $(1 - p)$ parametru generující šablony. Pokud jsou data uložena v tabulární struktuře, jsou zpravidla relevantní části textu „obaleny“ neměnnými elementy. V XPath výrazu se tento jev projevuje jako změna jednoho indexu, který pokud z XPath výrazu vypustíme, dostaneme množinu elementů, které by v ideálním případě měly všechny obsahovat hodnotu datotypové vlastnosti.

Parametr c můžeme považovat za podíl hodnot extrahované vlastnosti, které šablona skutečně označí vůči všem hodnotám této vlastnosti. Poměr počtu evidencí v segmentu před zobecněním cesty vůči prvkům po zobecnění by tedy měl činit přibližně c . Odchylku tohoto poměru můžeme využít k automatickému hodnocení chyby při zobecnění XPath cesty. Podle podobnosti počtu prvků nebo obecně XPath cesty můžeme pak přiřadit segmenty hodnot různých vlastností a z korespondujících hodnot sestavit instance extrahované třídy.

Tento jednoduchý přístup má řadu omezení. Kromě toho, že jím lze extrahovat pouze datotypové vlastnosti s kardinalitou 1, je také dosti omezený co se týče tolerance vůči chybám v pravidelnosti struktury dokumentu. Naopak vůči nepravidelnostem v samotných hodnotách vlastností je poměrně odolný.

4 Závěry a náměty

Způsob anotace je navržen s ohledem k použití s různými již existujícími metodikami automatické indukce wrapperu, což by mohlo být dobrým námětem pro budoucí práci. Podobný přístup je použit v [4] a [5], narozdíl od nich však nenavrhujeme proprietární zápis ontologií, nýbrž vycházíme ze standardu OWL.

Omezením navrženého jednoduchého modelu indukce wrapperu je, že se opírá o tabulární strukturu dokumentu, může však být zcela automatizovaný a při správném nastavení parametrů umožňuje automatický odhad chyb při extrakci.

Zajímavým námětem pro budoucí práci by také mohlo být navržení způsobu automatického učení šablon jednotlivých vlastností nebo jejich parametrů.

Poděkování

Výzkum vedoucí k tomuto článku byl podpořen Evropskou komisí na základě kontraktu FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content, K-Space a projektem FRVŠ 501/2006.

Reference

1. Antoniou, G., van Harmelen, F.: *A Semantic Web Primer*, Cambridge MA.: MIT Press, 2004, ISBN 0-262-01210-3
2. Hájek P.: *Metamathematics of fuzzy logic*, Dordrecht: Kluwer, 1998, ISBN: 0-792-35238-6
3. Kushmerick, N.: *Wrapper induction for information extraction*, PhD thesis, University of Washington, 1997
4. Labský M., Svátek V.: *On the Design and Exploitation of Presentation Ontologies for Information Extraction*, ESWC'06 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation, Budva, Montenegro, 2006
5. Muslea, I., Minton, S., Knoblock, C.: *A Hierarchical Approach to Wrapper Induction*, 3rd Conference on Autonomous Agents, 1999, <http://www.isi.edu/~muslea/papers.html>
6. Nekvasil M., *Využití ontologií při indukci wrapperů*, diplomová práce, VŠE, Praha 2006

Annotation:

The use of ontologies in wrapper induction

We propose an extension of ontologies which enables automatic annotation of documents. We use a fuzzy logic inference model for evaluation of the pattern matches and propose a simple method of wrapper induction which is able to utilize the results of such an annotation.