Využití ontologií k popisu vizuálních rysů dokumentu

Martin Milička a Radek Burget

Fakulta informačních technologií
Vysoké učení technické v Brně
{imilicka}{burgetr}@fit.vutbr.cz,
http://www.fit.vutbr.cz

Abstrakt Cílem této práce je seznámit čtenáře s možným použitím ontologií k popisu vizuálních rysů webových dokumentů. Tento přístup zavádí přirozenější popis obsahu dokumentu pro člověka a taky efektivnější zpracování informací, které jsou na webu publikovány. Základem je ontologie, která umožňuje popsat obsah dokumentu. Díky ní a získaným vizuálním rysům (konkrétním individuím) je z RDF dat sestaven model dokumentu. Nad takto definovaným popisem dokumentu je pak možné provádět dotazování nebo sofistikovanější zpracování dat.

Klíčová slova: ontologie, web, vizuální rysy, získávání znalostí, model dokumentu

1 Úvod

Vzhledem k tomu, že byly webové dokumenty prioritně navrženy pro člověka a zpracování dokumentů pouze na základě zdrojového kódu nepřinášelo očekávané výsledky, výzkum se v této oblasti musel začít orientovat na strojové zpracování zohledňující vizuální vzhled.

Aby bylo možné provést zpracování dokumentu tak, jak jej vidí člověk, dokument musí být ze zdrojového kódu vyrenderován. Je to z toho důvodu, že ne všechny informace o vzhledu jednotlivých elementů dokumentu jsou explicitně definovány ve zdrojovém kódu. Spousta z nich se tedy určuje až v okamžiku renderování dokumentu.

Z vyrenderovaného dokumentu lze bez problému sestavit jeho model, který je obsahuje kompletní vizuální informaci. Model dokumentu můžeme popsat několika způsoby. Obvykle se dokument popisuje stromovou strukturou podobnou HTML. Další možností je zavedení abstraktnější notace k HTML jakou například prezentuje Burget v [2] a nebo Alpuente a Romero v [1].

Vzhledem k tomu, že stromová struktura není schopna pokrýt všechny vazby, které do modelu vnáší vizuální rysy, je vhodné využít obecnější struktury jako je graf. Implementací grafové struktury jsou ontologie. Umožňují vícenásobné vazby a navíc je jejich popis pro člověka mnohem přirozenější.

2 Příbuzný výzkum

V úvodní kapitole jsme se dozvěděli, že model renderovaného dokumentu se nejčastěji reprezentuje stromovou strukturou. Největší motivací pro tuto reprezentaci je struktura HTML. O jejím použití se můžeme přesvědčit v publikovaných článcích [1,2,4].

Hlubším zkoumáním konstrukce modelu dokumentu jsme zjistili, že stromová struktura nemusí být vždy dostačující. Existují situace, které není možné precizně popsat s ohledem na reálný vzhled dokumentu a vizuální vazby mezi jeho jednotlivými částmi. Snažili jsme se tedy do konstrukce modelu zakomponovat obecnější strukturu, kterou je obecný graf. Konkrétní implementací takového grafu jsou *ontologie*.

V současné době můžeme nalézt ontologie, které se snaží obsahy dokumentů popisovat. Jednou z takových ontologií je SALT ontologie [5], která se převážně orientuje na dokumenty s lineární strukturou, jako jsou například vědecké články. Základ této ontologie se dá využít k návrhu nové ontologie, která umožní popisovat obecné dokumenty s jejich vizuálními rysy.

V současné době nám není známa žádná ontologie, která by kromě samotného obsahu uvažovala taky vizuální rysy dokumentu.

3 Vizuální rysy

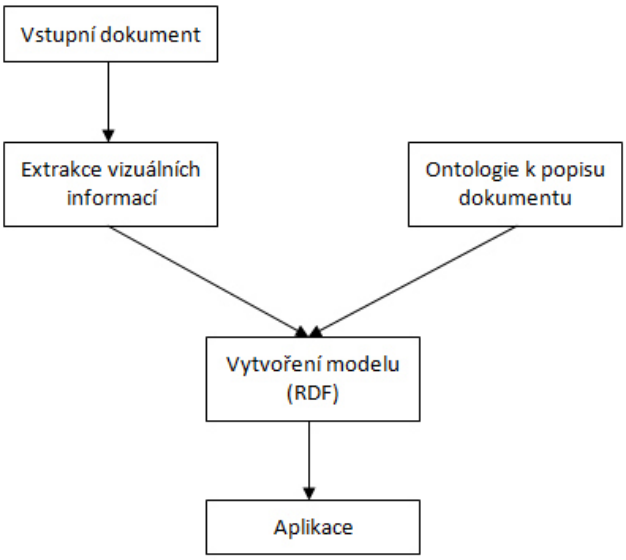
Jak již bylo zmíněno dříve, zpracování dokumentu založené na vizuálních rysech umožňuje pracovat s daty dokumentu tak, jak jej vnímá jeho čtenář. Burget a Burgetová v článku [3] představili ucelenou kategorizaci vizuální rysů, které má smysl zpracovávat. Zavedli kategorie rysů pro použitý font (font features), prostorové rysy (spatial features), rysy textu (text features) a rysy barev (colour features). Každá taková kategorie seskupuje několik příbuzných rysů, jež bychom chtěli do nově vzniklé ontologie zakomponovat.

Vizuálním rysy umožňují preciznější zpracování dokumentu. Díky nim je možné jednodušeji klasifikovat jednotlivé části dokumentu a tak odstranit šum, který může způsobit nepřesnost výsledků nebo případně identifikovat specifický obsah.

4 Model zpracování

Na obrázku 1 můžeme vidět schéma navrhovaného zpracování webových dokumentů.

K tomu, aby bylo možné sestrojit model dokumentu podle dané ontologie, je nutné provést jeho předzpracování (získání vizuálních informací). Součástí fáze předzpracování je načtení HTML kódu včetně přilinkovaných kaskádových stylů a jejich následné renderování. Tento způsob umožňuje získat přesné vizuální informace, které nejsme ze zdrojového HTML a jeho kaskádových stylů schopni bez



Obrázek 1. Schéma navrhovaného zpracování dokumentu

renderování získat. Příkladem může být získání jednoznačné pozice bloku v dokumentu, případně jeho vztah k jiným částem dokumentu. Renderování dokumentu nám tedy zajistí, že získáme přesné hodnoty vizuálních rysů. V kontextu ontologií jsou tyto hodnoty nazývány *individua*. Ty následně vstupují do procesu vytvoření modelu dokumentu podle definované ontologie.

Definice ontologie je dalším očekávaným vstupem v procesu generování modelu dokumentu. Umožňuje popsat dokument tak, jak jej vnímá člověk. Žádoucí je, aby se vždy při návrhu nové ontologie využívaly definice již existujících ontologií. Za tímto účelem jsme se rozhodli použít zmíněnou SALT ontologii [5], jež by nám měla návrh nové ontologie usnadnit. Jak již bylo zmíněno dříve, tato ontologie definuje třídy pro lineární popis dokumentů, převážně publikací. To znamená, že nám umožní částečně pokrýt požadavky na popis textového obsahu dokumentu.

Samozřejmostí je, že bude nutné definovat nové třídy, které nám umožní zpracovávat nejenom lineární struktury dokumentů. Nedílnou součástí definice nové ontologie bude taktéž specifikace nových atributů tříd. Bude se jednat o atributy z kategorií zmíněných v kapitole 3, které představil Burget a Burgetová v [3].

Ze vstupního dokumentu, který je předzpracován tak, aby obsahoval jenom individua, a z definované ontologie vznikne model dokumentu. Ten může být reprezentován pomocí RDF[1], které se stalo základním jazykem pro popis ontologií.

[1] Resource Description Framework

Příkladem tříd, které budou definovány v ontologii, může být *oblast dokumentu* nebo *prvek dokumentu*. Mezi těmito třídami se pak budou definovat vlastnosti jako je *jeNad, jePod, máVelikost, atd.*

Vzhledem k tomu, že v tomto článku není prostor pro detailní definici nové ontologie, byl zde proveden pouze nástin možného řešení.

5 Možnosti použití

Nově definovaná ontologie umožní nad dokumentem popsaným pomocí RDF dat provádět mnohem sofistikovanější dotazy, které budou zohledňovat vizuální rysy.

Abychom ukázali sílu navrženého modelu, provedeme jeho integraci do renderovacího stroje CSSBox[2]. Ten stejně jako všechny dosud známé renderovací strojem, pracuje se stromovým modelem dokumentu. Očekáváme, že navržený přístup by mohl pozitivně ovlivnit výsledky námi prováděných výzkumů.

Díky obecnějším vztahům mezi elementy dokumentu jsme schopni získat přesnější znalosti o dokumentu, které se v případě stromové struktury musí někdy zanedbávat. Ontologie a její grafová struktura umožňují popisovat dokument pro člověka přirozenějším způsobem, v čemž taky vidíme potenciál dalšího použití.

Model dokumentu, který je popsán pomocí ontologie, je taktéž možné uplatnit v oblasti porovnávání dokumentů založeném na vizuálních rysech.

Reference

1. Alpuente, M., Romero, D., *A Visual Technique for Web Pages Comparison*, Electron. Notes Theor. Comput. Sci. **235** (2009), 3–18.
2. Burget, R.,, *Automatic document structure detection for data integration*, Business Information Systems, LNCS 4439, Springer Verlag, 2007, pp. 391–397.
3. Burget, R., Burgetová, I., *Automatic annotation of online articles based on visual feature classification*, International Journal of Intelligent Information and Database System **5** (2011), no. 4, 338–360.
4. Cai, D., Yu, S., Wen, J.-R., Ma, W.-Y., *Extracting content structure for web pages based on visual representation*, Proceedings of the 5th Asia-Pacific web conference on Web technologies and applications (Berlin, Heidelberg), APWeb'03, Springer-Verlag, 2003, pp. 406–417.
5. Groza, T., Handschuh, S., *Salt document ontology*, Dostupné na: http://salt.semanticauthoring.org/ontologies/sdo `[online]` (Srpen 2012).

[2] http://cssbox.sourceforge.net/