Klasifikace prvků dokumentu na základě vizuálních rysů

Michael Kunc, Radek Burget

Fakulta informačních technologií VUT v Brně,
Božetěchova 2, 612 66 Brno
kunc@fit.vutbr.cz, burgetr@fit.vutbr.cz

Abstrakt. Příspěvek se zabývá vizuální segmentací webových stránek a následnou klasifikací detekovaných oblastí na základě jejich vizuálních vlastností. Segmentace, tedy detekce vizuálního členění stránky, je založena na bottom-up analýze stránky společně s analýzou vizuálně významných prvků a jejím výsledkem je model vizuálního členění zpracovávaného dokumentu. Samotná klasifikace potom využívá informace o vzájemné poloze oblastí a jejich vizuálních vlastnostech, které jsou v tomto modelu obsaženy. Příspěvek se dále zbývá experimentálním porovnáním dvou klasifikačních algoritmů pro daný účel na různých množinách dat.

Klíčová slova: klasifikace, segmentace, zpracování dokumentů

1 Úvod

Ačkoliv v současnosti běžně používané formáty elektronických dokumentů disponují prostředky pro anotaci některých prvků dokumentu, jejich možnosti jsou v tomto ohledu velmi omezené a ne vždy autory dokumentů využívané. Tato informace je přitom potenciálně užitečná pro další zpracování dokumentů. V tomto příspěvku se proto zabýváme možností klasifikace prvků v dokumentu na základě jejich vizuálních vlastností. Navrhujeme způsob extrakce vizuálních rysů z dokumentu a následné klasifikace prvků dokumentu na základě těchto rysů. Zaměřujeme se na zpracování dokumentů HTML, nicméně prezentovaný postup je aplikovatelný i na dokumenty v jiných formátech.

2 Postup zpracování dokumentu

Prvním krokem zpracování je rekonstrukce výsledné podoby dokumentu, což v případě HTML přináší kompletní renderování stránky. Nad touto stránkou je provedena vizuální segmentace, jejímž výsledkem je hierarchická struktura vizuálních oblastí detekovaných v dokumentu. Pro každou z těchto oblastí určíme její význačné vizuální rysy a na základě těchto rysů potom provádíme jejich klasifikaci s cílem rozpoznat, které oblasti odpovídají některým význačným prvkům v dokumentu.

Segmentace stránky vychází z algoritmu publikovaného v [2]. Tato metoda je založena na postupném spojování oblastí do větších celků podle daných pravidel (*bottom-up* analýza). Výsledkem segmentace je strom detekovaných vizuálních oblastí. Pro každou oblast pak určíme nebo vypočteme její vizuální atributy podle tabulky 1. Volba těchto vlastností vychází z možností jazyků HTML a CSS pro definici vizuálních vlastností a použité reprezenrace segmentované stránky.

Tabulka 1. Vizuální atributy oblastí

`fontsize`	průměrná velikost písma vyjádřená v procentech, kde 100 % je průměrná velikost písma v celém dokumentu
`weight`	převažující váha písma v oblasti (tučné nebo netučné)
`style`	převažující styl (normální nebo skloněné písmo)
`aabove`, `abelow`, `aleft`, `aright`	počet oblastí, které se vyskytují nad, pod, vlevo a vpravo od dané oblasti v rámci rodičovské oblasti
`tlength`	počet znaků textu v oblasti
`tdigits`, `tlower`, `tupper`, `tspaces`	počet číslic, malých a velkých písmen abecedy a mezer v textu
`textbtns`	průměrná světelnost (luminosity) textu
`bgbtns`	průměrná světelnost pozadí
`contrast`	průměrný rozdíl světelnosti textu a pozadí (kontrast)

Světelnost textu a pozadí je vypočtena podle vzorce používaného pro hodnocení kontrastu barev pro účely přístupnosti webových stránek [3].

3 Experimentální testování

3.1 Zdrojové dokumenty

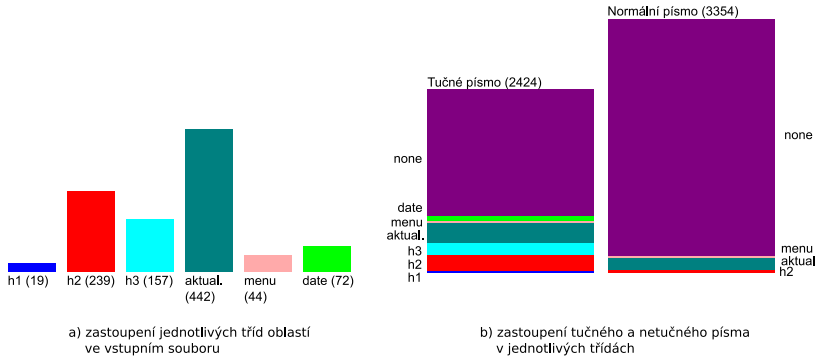
Jako zdrojové dokumenty jsme použili reálné stránky českých a světových zpravodajských serverů, které se vyznačují velkým množstvím různých prvků v rámci jednoho dokumentu. Celkem jsme zpracovali 16 dokumentů. V těchto dokumentech bylo při segmentaci detekováno celkem 5778 vizuálních oblastí. Tyto oblasti jsme pomocí grafického anotovacího nástroje ručně rozdělili do tříd uvedených v tabulce 2, které se opakovaně vyskytují ve všech zpracovávaných dokumentech.

3.2 Vstupní data

Z anotovaných dokumentů byly pro každou oblast extrahovány hodnoty atributů uvedených v tabulce 1. Tyto hodnoty byly uloženy ve formátu *arff* vhodném pro klasifikaci.

Tabulka 2. Zkoumané vlastnosti vizuálních oblastí

`h1`	nadpis hlavního článku na stránce (je-li přítomen)
`h2`	nadpis běžného článku
`h3`	nadpis aktuality nebo zprávy menšího významu (upoutávky apod.)
`aktualita`	krátká zpráva nebo aktualita
`menu`	oblast navigace
`date`	datum publikování, obvykle i se jménem autora a podobně
`none`	ostatní (neanotované) oblasti



Obr. 1. Vlastnosti některých atributů ve vstupních datech

Obrázek 1a ukazuje základní rozložení anotovaných prvků do jednotlivých tříd. Jednotlivé sloupce představují jednotlivé anotované třídy, zbývající oblasti spadají do třídy `none`. Na obrázku 1b je rozdělení tříd s tučným a obyčejným písmem. Obyčejné písmo bylo použito u velké většiny běžného textu, necelé poloviny aktualit a zhruba desetiny nadpisů třídy `h2`. Je zajímavé, že celkově je přes 40 % textů tučných. Je to zřejmě dáno relativně krátkými články s větším množstvím nadpisů a odkazů na další články. Pokud jde o velikost písma, lze ve vstupních datech vysledovat, že všechny nadpisy třídy `h1`, téměř všechny třídy `h2`, většina třídy `h3` a většina prvků třídy `aktualita` má velikost písma větší než 100 %.

3.3 Klasifikace

Pro klasifikaci jsme použili volně dostupný nástroj Weka [1]. Zvolili jsme algoritmy *J48* a *LWL*. Důvodem jejich volby je zejména skutečnost, že mohou pracovat s nominálními i číselnými atributy, a poradí si i s chybějícími atributy. Algoritmus *J48* je založen na rozhodovacích stromech a vychází z metody *C4.5* [4]. Jako takový současně dává poměrně snadno interpretovatelný výsledek. Metoda *LWL*

[1] http://www.cs.waikato.ac.nz/ ml/weka/

(Locally Weighted Learning) [1] patří do skupiny tzv. *„lazy learning“*. Tento algoritmus patří mezi metody aproximací založené na lokálním modelu.

Nejprve jsme porovnali oba algoritmy na souboru dat o jednom dokumentu, který obsahoval prvky všech tříd z tabulky 2. Anotovaný soubor byl použit jako trénovací množina. Úspěšnost klasifikace oblastí ve stejném souboru byla 97,54 % pro J48 a 89,39 % pro LWL. Úspěšnost klasifikace jiného souboru dat, který však pocházel ze stejného webu, jako trénovací množina, byla 82,95 % pro J48 a 88,64 % pro LWL.

Dále byla zkoumána úspěšnost algoritmu J48 při klasifikaci většího souboru dat, přičemž jsme měnili trénovací množinu dokumentů. V případě, že množina trénovacích dat je zcela disjunktní s množinou testovací, se úspěšnost klasifikace pohybuje nad hranicí 75 %. Když byla k trénovacím datům přidána data ze stejného webu, jako jsou data testovací, ale z jiné jeho části, úspěšnost klasifikace se zvýšila nad 90 %. Pokud byla součástí trénovací množiny přímo i anotovaná testovací data, úspěšnost klasifikace se zvýšila na téměř 98 %.

4 Závěr

Navrhli jsme přístup ke klasifikaci prvků dokumentu na základě hodnocení jejich vybraných vizuálních vlastností a představili jsme první experimentální výsledky klasifikace pomocí dvou různých algoritmů. Tyto výsledky podle našeho názoru ukazují použitelnost daného přístupu. Dosažené výsledky mohou být negativně ovlivněny zejména algoritmem segmentace stránek, který se v některých případech nechová ideálně. Dále předpokládáme další rozšiřování množiny zkoumaných atributů vizuálních oblastí a testování na rozsáhlejších množinách dokumentů.

Reference

1. C. G. Atkeson, A. W. Moore, S. Schaal: Locally Weighted Learning. *Artificial Intelligence Review*, 11:11-73, 1997
2. R. Burget: Vizuální segmentace elektronických dokumentů, In *Znalosti 2007*, Ostrava, CZ, 2007, pp. 155-166
3. B. Caldwell et al.: *Web Content Accessibility Guidelines 2.0*, W3C Working Draft 17 May 2007, W3C, 2007
4. J. R. Quinlan: *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993

Annotation:

Visual Document Segmentation

This paper deals with the visual segmentation of web documents and subsequent classification of the detected areas based on their visual attributes such as the mutual area positions, font properties and colors. We present an experimental evaluation of two classification algorithm for different data sets.