

Obsah

Úvod	2
1. Segmentovanie a vyhodnotenie webových stránok	4
1.1 Základné informácie	5
1.1.1 Webové aplikácie	5
1.1.2 Rendering	5
1.1.3 Renderovaný DOM	6
1.1.4 Umiestnenie elementu	6
1.2 Charakteristika webovej stránky	6
1.2.1 Charakteristika webových stránok z renderovaného DOM	6
1.2.2 Charakteristiky vzťahnuté na webovú stránku	7
1.2.3 Slovník pojmov	7
1.3 Segmentácia webovej stránky	8
1.3.1 Definícia základných pojmov	8
1.3.2 Notácia	9
1.3.3. Top-down versus bottom-up	10
1.3.4 Základný prístup	11
1.3.5 Hybridné prístupy	12
1.3.6 Záverom o algoritmoch segmentácie webových stránok	13
1.3.7 Spracovanie dokumentov a segmentácia webových stránok	13
1.4 Segmentation evaluation	14
1.4.1 Klasifikácia hodnotiacich metód	14
1.4.2 Segmentácia korektneho hodnotenia	14
1.4.3 Miera správnosti segmentácie skenovaného dokumentu	15
3. Porovnanie algoritmov	16
4. Fit layout systém	21
5. Implementácia do systému fut-layout	21
6. Experimenty a porovnanie	21
7. Záver	21
Použitá literatúra:	24

Úvod

Témou mojej diplomovej práce je porovnanie algoritmov pre segmentáciu webových stránok a vyhodnotenie segmentácie.

Keďže z dostupných zdrojov je zrejmé, že neexistuje úplné porovnávacie vyhodnotenie algoritmov segmentácie webových stránok, podujal som sa vykonať porovnanie algoritmov na základe (zamerania sa na cieľ segmentácie, súbor údajov a kvantitatívnych hodnotiacich schém) stanovenia jednotných akcií pre lepšie hodnotenie správnosti algoritmov segmentácie webových stránok, ktoré sú:

Keďže webové stránky sú generované Content Management Systemom (CMS) sú čoraz viac komplikovanejšie. Je stále náročnejšie automaticky analyzovať webové stránky. Napríklad identifikovať a zatriediť obsah, reklamy, komentáre užívateľov atď.

Segmentovanie webových stránok ponúka riešenie tohto problému. Segmentácia webových stránok je možná prostredníctvom rozdelenia webovej stránky na vizuálne segmenty a sémanticky koherentné segmenty. Tieto sa nazývajú tiež bloky.

Segmentáciu je možné použiť napríklad na vyčlenenie zaujímavých častí na archiváciu webu. Ak sa blokom priradia relatívne váhy vzhľadom na ich význam, možno zistiť relevantné zmeny v dôležitých blokoch z rôznych verzií stránky. Toto možno použiť na optimalizovanie vyhľadávania. Výsledkom je vyladenie prehľadávačov tak, že sa častejšie vrátia na stránky s dôležitými zmenami, čo je pre užívateľa výhodné.

Pomáha aj napr. pri migrácii webového archívu z formátu ARC do formátu WARC a to

porovnanie verzie stránky pred a po akcii. Pokiaľ je segmentácia nasledujúcej verzie rovná predchádzajúcej, potom je predpoklad, že akcia sa správne zrealizovala.

V zariadeniach s malou obrazovkou sa využívajú segmentácie na optimalizáciu vizualizácie webovej stránky. Mobilné zariadenia používajú techniku zväčšenia (zoom) na zobrazenie podrobností o webovej stránke užívateľovi. Nevýhodou je, že táto technika je časovo náročná.

Pri nasadení segmentácie namiesto priblíženia (zoom) zariadenie poskytne užívateľovi relevantné bloky nájdené v segmentácii. Doba odozvy je v tomto prípade kratšia. Webové archívy môžu využiť

na migráciu z jedného formátu do druhého segment webových stránok. Napríklad migrácia webových stránok z formátu HTML4 na formát HTML5 s cieľom znížiť závislosť emulácie.

Jedným z hlavných problémov segmentácie webových stránok je presnosť a univerzálnosť.

Segmentácia je presná, ak je jej granularita rovná alebo blízka granularite ideálnej segmentácie.

Segmentácia je univerzálna, pokiaľ funguje dobre na väčšine druhov webových stránok. Granularita segmentácie predstavuje, do akej miery segmentácia rozdeľuje webovú stránku na bloky.

Porovnanie ideálnej a nepresnej segmentácie demonštruje obr. 1. Ideálna segmentácia ukazuje, že webová stránka by mala byť rozdelená do šiestich blokov: blok 1 hlavička, blok 2 názov a meno autora, blok 3 sociálne médiá, blok 4 článok, blok 5 súvisiace odkazy a 6 päta. Vypočítaná segmentácia má štyri bloky. Bloky 1 a 4 sa rovnajú blokom ideálnej segmentácie (blok 1 a 6). Vypočítaná segmentácia zlúčila bloky názvu a sociálnych médií, a hlavný článok so súvisiacimi odkazmi. Aplikácie, ktoré sú závislé na segmentácii nebudú mať presné informácie, keďže napríklad názov je spojený so sociálnymi médiami a hlavný článok je zlúčený s nepodstatnými odkazmi na obsah.

Obrázok 1: Príklad pre nepresnú segmentáciu

Ako ďalší príklad uvážim segmentovanú web stránku s vysokou presnosťou. Je to forum. Segmentátor nájde všetky dôležité časti: hlavičku, hlavný článok, oznámenia, odpovede a päta.

Druhý obrázok je príklad s nízkou presnosťou pri použití toho istého segmentátora. Hlavičku, patu rozpozná ale navigáciu, hlavný text a patu spojí do jedného veľkého bloku. Tento typ segmentátora je vhodný iba na určitý typ webových stránok.

Väčšina existujúcich segmentátorov sa snaží napodobniť pohľad používateľa pomocou heuristických pravidiel. Najmodernejšie algoritmy sú určené pre konkrétne potreby aplikácií a na vyhľadávanie blogov majú prispôsobené heuristické pravidlá. Nevýhodou je často nedostatočná presnosť a všeobecnosť, keďže v tomto prípade algoritmy nemusia spracovať správne všetky časti v obsahu webovej stránky.

Obr. 2 Príklad univerzálnej segmentácie

Na riešenie problému som implementoval nový segmentátor, takzvaný Block-o-Matic (BoM), ktorý zohľadňuje obe vlastnosti - presnosť a univerzálnosť. Tento algoritmus je vypracovaný pomocou stratégie bottom-up (zdola nahor) a zohľadňujúc vizuálny prístup. Tento algoritmus je založený na súbore W3C heuristických pravidiel. W3C používajú webové stránky, predovšetkým pomocou štandardných kategórií obsahu W3C. Zároveň použitie týchto pravidiel na detekciu a klasifikáciu blokov dáva prístupu univerzálnosť. Na vyriešenie problému s presnosťou sa bloky spájajú vzhľadom na ich veľkosť, polohu a názov. Následne sú usporiadané do zložených blokov. Tieto stanovujú rozvrhnutie webovej stránky. Takže výsledkom takéhoto prístupu býva presnejšia a univerzálnejšia segmentácia. Intuícia však nestačí. Potrebujeme vykonať hodnotenie, ktoré nám umožní zmerať tieto dva aspekty algoritmu BoM.

V súvislosti so segmentáciou možno použiť rôzne interpretácie správnosti. Ako sa uvádza v literatúre, správnosť algoritmu sa presadzuje vtedy, keď je v súlade s danou špecifikáciou. V prípade segmentácie webovej stránky nie je možné takúto špecifikáciu určiť a priori bez ľudského úsudku. Zameriavame sa teda na hodnotiace prístupy založené na základnej pravde a vyvinuli sme rámec, ktorý zahŕňa metriky správnosti, agregované skóre a nástroj, ktorý hodnotiteľom uľahčuje manuálny návrh základnej pravdy.

1. Segmentovanie a vyhodnotenie webových stránok

Obsahom tejto kapitoly je priblíženie problematiky segmentácie webových stránok a jej vyhodnotenie, ako je táto téma dostupná v dnešnej dobe.

Segmentácia webovej stránky spočíva v identifikácii a kategorizácii blokov (oblastí záujmu) na webovej stránke. Segmentácia webovej stránky je rozdelená na dve hlavné oblasti:

- detekcia a identifikovanie rôznych blokov
- klasifikácia logických blokov webovej stránky (hlavička, pätička, navigácia atď.)

Charakteristiky webovej stránky popisujú užitočné prvky, ktoré sa berú do úvahy pri analýze webovej stránky.

Webové stránky sú preskúmané podobným spôsobom ako naskenované dokumenty prostredníctvom domény OCR (optical character recognition / optickým rozpoznávaním znakov). Hoci sú oba zdroje podobné a možno túto podobnosť využiť, je zrejmé aj rozličnosť zdrojov, prvky/text vs. pixely/farby).

Algoritmy pre segmentáciu webových stránok môžu sledovať dva hlavné prístupy:

1. základný
2. hybridný

Algoritmy môžu byť tiež navrhnuté pomocou stratégií buď zhora nadol alebo zdola nahor.

1.1 Základné informácie

V kapitole budú popísané webové aplikácie z hľadiska ich štruktúry a ich vykresľovania.

Bude vysvetlená charakteristika webovej stránky, segmentácia webovej stránky a jej vyhodnotenie.

1.1.1 Webové aplikácie

Webová aplikácia je aplikácia, ktorá beží vo webovom prehliadači. Je vyvinutý pomocou webových technológií ako JavaScript, Hypertext Markup Language (HTML) a Cascading Style Sheets (CSS).

Webové aplikácie sú štruktúrované s n-vrstvovým architektonickým štýlom [Con02]. Najbežnejšou štruktúrou je 3-vrstvová:

- Prvá vrstva je webový prehliadač (Firefox, Google Chrome atď.)
- Druhá (tiež stredná vrstva) je prostriedok využívajúci nejakú technológiu webového obsahu (PHP, - Rails a pod.).
- Tretia vrstva je ukladanie údajov (databázy, dokumenty atď.).

Webový prehliadač odošle požiadavku na strednú vrstvu, ktorá spustí dotazy, prípadne aktualizuje zdroje údajov a vygeneruje webovú stránku. Webové aplikácie môžu používať zložitejšie štruktúry. Keďže sa však segmentácia webovej stránky bežne vyskytuje vo webovom prehliadači, spôsob vytvárania webovej stránky nie je pre našu prácu relevantný a zameriavame sa len na prvú úroveň.

1.1.2 Rendering

HTML dokument obsahuje pokyny pre prehliadač, ako má byť webová stránka prezentovaná používateľovi. Aby bolo možné zobrazíť webovú stránku, prehliadač potrebuje stiahnuť webový obsah z jednej alebo viacerých webových stránok. Ako sme uviedli vyššie, obsahom môže byť okrem iného HTML kód, skripty JavaScript, štýly CSS. Úlohou renderingu je potom zobrazíť naformátovaný obsah používateľovi.

Okrem naformátovaného obsahu rendering vytvára na stránke strom DOMu (Document Object Model). Je to rozhranie, ktoré umožňuje programom a skriptom dynamicky pristupovať k obsahu, štruktúre a štýlu vykreslenej stránky a aktualizovať ich. Obrázok naformátovaného obsahu je možné

získať z renderovaného modelu DOM. Tento obrázok sa nazýva renderovaný obrázok alebo snímka obrazovky.

1.1.3 Renderovaný DOM

Po skončení renderovania je možné vytvoriť webovú stránku na základe vytvoreného stromu DOM. Tento dokument obsahuje všetok obsah, aktualizácie a štýly spracované v procese vykresľovania. Tento dokument nazývame Rendered DOM [Goo02]. V tomto momente pôvodný dokument HTML a renderovaný DOM tej istej stránky nie sú totožné. Renderovaný DOM je zdrojový kód HTML formátovaného obsahu, zatiaľ čo pôvodný dokument HTML obsahuje iba pôvodný zdrojový kód HTML.

Naša práca je teda založená na tomto renderovanom DOM. Jeho vytvorenie vytvára réžiu, a tým zvyšuje čas odozvy analýzy, ale analýza webovej stránky je úplnejšia a presnejšia.

1.1.4 Umiestnenie elementu

Pri vytváraní renderovaného DOM engine určuje polohu prvkov na základe ich vlastností štýlu. Prvok, ktorý nie je textovým prvkom, určuje svoju vlastnú šírku a šírku svojich potomkov. V niektorých prípadoch môže byť toto pravidlo porušené umiestnením prvkov ako *div* statickým spôsobom. V dôsledku toho poradie prvkov v strome DOM nezodpovedá poradiu formátovaného obsahu pri zobrazení. Toto môže viesť k chybám v segmentácii, pretože obsah, ktorý má blok pokrývať (alebo jeho časť), sa vykresľuje v inej oblasti stránky.

1.2 Charakteristika webovej stránky

V tejto časti je popísaná charakteristika webových stránok. Je kľúčová pre algoritmy segmentácie webových stránok.

1.2.1 Charakteristika webových stránok z renderovaného DOM

Webové stránky sú špecifikované HTML dokumentom (zdrojovým kódom), s ním spojeným kaskádovaním Style Sheets (CSS) a vložené skripty JavaScript. Renderovaný DOM W webovej Stránky sa získa renderovaním engine (napr. WebKit a Gecko), ktorý spracováva špecifikáciu.. Charakteristiky W sú: obsah, vizuálna prezentácia a pozičná schéma.

Pozn Este tu su kap 1.2.1.1...

1.2.2 Charakteristiky vzťahnuté na webovú stránku

Webová stránka funguje ako celok. Každá strana prispieva k tejto funkcii (blog, fórum atď.). To znamená, že stránka musí pokrývať určité funkcie, ako napríklad typ použitej navigácie alebo typ rozloženia podľa tejto funkcie. Existuje niekoľko nástrojov, ktoré pomáhajú vývojárom dosiahnuť očakávaný cieľ webovej stránky, ide predovšetkým o šablóny a systémy na správu obsahu (CMS). Šablóna je vopred navrhnutá webová stránka alebo súbor webových stránok, ktoré môže ktokoľvek upravovať s vlastným obsahom a obrázkami a nastaviť webovú stránku. Šablóny umožňujú nastaviť webovú stránku tak, aby mala všetky funkcie. CMS je webová aplikácia, ktorá umožňuje upravovať a publikovať obsah webových stránok z centrálného rozhrania, napr. webový prehliadač. Jeho nastavenie môže byť zložitejšie ako šablóna, ale má výhodu v tom, že je dynamické a na rovnaký obsah možno použiť rôzne šablóny.

Možno povedať, že existujú tri dôležité charakteristiky súvisiace s webovými stránkami v kontexte segmentácie webových stránok:

- Funkcia: je účel alebo cieľ webovej stránky
- Element šablóny: Prvky webovej stránky, ktoré sa nachádzajú na iných stránkach toho istého webu. Všetky spolu tvoria rozloženie stránky.
- Label: predstavuje úlohu týchto prvkov na webovej stránke.

Prvky šablóny sú dôležité pre segmentáciu, pretože zvyčajne tieto prvky sa používajú na organizovanie obsahu, preto je vysoko pravdepodobné, že organizujú aj menšie bloky.

1.2.3 Slovník pojmov

Prvky / Elements	Objekty v renderovanom DOM W.
W.root	Koreňový prvok renderovaného modelu DOM.
W.text	Výsledok reťazenia celého textu následných prvkov W.rootu
Kategória / Category	Kategória obsahu prvkov na webovej stránke. Napríklad obsah toku obsah fráz, obsah sekcií, atď.
Box	Ako nakresliť tvar obdĺžnika prvku a ako vypočítať sa jeho veľkosť

Vizuálne nápovedy / Visual cues	Vlastnosti formátovania prvkov získaných pri renderovaní času. Napríklad: čiary, prázdne oblasti, farby, obrázky, písma, atď.
Viewport	Určená veľkosť <i>body</i> prvku, ktorý funguje ako počiatočný kontajner.
Funkcia / Function	Cieľ webovej stránky. Napríklad: blog, wiki atď.
Schéma / Scheme	Definuje, ako sú súradnice boxu vypočítané s ohľadom na iný kontajnerový box.
Šablóna prvkov / Template elements	Prvky webovej stránky, ktoré sa opakujú na niekoľkých stranách tej istej webovej stránky.

1.3 Segmentácia webovej stránky

Segmentácia webovej stránky na zmysluplné komponenty sa ukázala ako dôležitá úloha analýzy webových dokumentov, pretože podporuje mnoho dôležitých aplikácií.

Segmentácia webovej stránky sa týka procesu rozdelenia webovej stránky na vizuálne a sémanticky koherentné segmenty nazývané bloky. Detekcia týchto rôznych blokov je kľúčovým krokom pre mnohé aplikácie, ako sú mobilné zariadenia, vyhľadávanie informácií, archivácia webu, dostupnosť webu, hodnotenie vizuálnej kvality (estetika), atď. V kontexte archivácie webu možno segmentáciu použiť na extrahovanie zaujímavých častí, ktoré sa majú uložiť. Priradením relatívnych váh blokom podľa ich dôležitosti umožňuje tiež zisťovať dôležité zmeny medzi verziami stránok. Je to užitočné pri optimalizácii indexového prehľadávania, pretože umožňuje vyladiť prehľadávače tak, aby častejšie navštevovali stránky s dôležitými zmenami.

Najprv definujeme pojmy a notáciu (1.3.1 a 1.3.2), prístupy k segmentácii webových stránok (1.3.4 a 1.3.5), vzťah medzi spracovaním dokumentov a segmentáciou webových stránok (1.3.7), súhrnnú tabuľku segmentačných algoritmov (1.3.8). Potom končíme diskusiou (1.3.9)

1.3.1 Definícia základných pojmov

Vychádzajúc z konceptu definovanom Tangom [TS94] možno popísať segmentáciu webovej stránky pomocou nasledujúcich pojmov:

- Stránka / Page je špeciálny blok, ktorý predstavuje celú webovú stránku a pokrýva celý Viewport.

- Jednoduchý blok / Simple block je prvok alebo skupina prvkov. Označuje sa tiež jednoducho ako *blok*. Je reprezentovaný ako obdĺžniková oblasť, ktorá je výsledkom zlúčenia polí prvkov. Každý blok má označenie súvisiace so základnými prvkami. Je tiež spojená s textom týchto prvkov.
- Kompozitný blok / Composite blok je špeciálny blok, ktorý môže obsahovať ďalšie bloky. Zvyčajne takéto bloky zodpovedajú prvkom šablóny.
- Blokový graf je súvislý rovinný graf predstavujúci bloky a ich vzťahy.
- Geometrický model predstavuje súbor blokov ako množinu obdĺžnikov v rovine.
- Získavajú sa zo schémy webovej stránky. Všetky obdĺžniky sú modelované ako štvorce (x, y, w, h) , kde x a y sú súradnice počiatočného bodu a w a h sú šírka a výška obdĺžnika.
- Podmienka zastavenia je preddefinovaná hodnota (skutočné číslo), ktorú používajú algoritmy a ktorá indikuje, kedy sa dosiahne segmentácia. Je založený na váhe hrán/vrcholov blokového grafu. Algoritmus môže mať jednu alebo viac podmienok zastavenia.
- Označenie / Label je úloha, ktorú hrá blok na webovej stránke, ako je navigácia, obsah, hlavička, päta atď.

1.3.2 Notácia

V tejto časti sú uvedené definície, aby sme mali jednotnú prezentáciu algoritmov segmentácie webových stránok.

1.3.2.1 Segmentačná funkcia

Segmentačná funkcia Φ je opísaná takto:

$$\Phi_A(W, SC) \rightarrow W'_A, GM_A \quad (1.1)$$

kde A je algoritmus segmentácie webovej stránky, W je vykreslený DOM webovej stránky, SC je súbor podmienok zastavenia. W'_A je blokový graf definovaný v kapitole 1.3.2.2 a GM_A je a súbor obdĺžnikov reprezentujúcich geometrický model segmentácie.

1.3.2.2 blokový graf

Blokový graf je definovaný ako rovinný graf $W'_A = (Blocks, Edges)$. Každý vrchol B v blokoch zodpovedá obdĺžniku v GM_A a označeniu (label). Je spojená s funkčnou váhou na hranách a vrchoch a dvoma podmnožinami vrcholov: $SimpleBlocks \subset Blocks$ (nazývané aj terminály), $CompositeBlocks \subset Blocks$, ktoré obsahujú špeciálnu stránku (vertex Page), označenú ako koreň grafu.

Obdĺžnik vertex Page pokrýva celý výrez webovej stránky W a všetky bloky do seba zapadajú.

Takže:

$$\forall B \in Blocks, B.rect \subseteq Page.rect$$

Hmotnosť vrcholu B je označená ako B.váha. Hmotnosť hrany E je označená ako E.váha

1.3.3. Top-down versus bottom-up

Existujú hlavne dva druhy stratégií: segmentácia stránky zhora nadol a segmentácia stránky zdola nahor. Každá stratégia vedie extrakciu blokov a určuje spôsob, akým sa prechádza vykreslený strom DOM.

1.3.3.1 Top-down strategia

Stratégia zhora nadol začínajú s úplnou webovou stránkou ako blok (Stránka) a tento blok iteratívne rozdeľujú na menšie bloky pomocou rôznych funkcií obsahu webovej stránky.

Dobрым príkladom algoritmu zhora nadol je algoritmus VIPS [CYWM03], podrobne uvedený v Časť 1.3.5.1. Popisuje blokový graf ako vrcholovo vážený strom. Algoritmus priradí každému bloku hodnotu váhy (stupeň koherencie alebo DoC), medzi 1 a 10, čo naznačuje, ako sú bloky koherentné. Algoritmus zadáva podmienku zastavenia ako povolený stupeň koherencie (PDoC), ktorý je stanovený a priori a používa sa ako parameter. Algoritmus sa zastaví, ak $DoC > PDoC$.

Príklad na obrázku 1.1 má hodnotu PDoC 5. Webová stránka je najskôr rozdelená na tri veľké bloky (bloky 1,2 a 3) a podľa toho sa nastaví blokový graf. Obrázok 1.1a zobrazuje hodnoty hmotnosti priradené ku každému bloku.

Pre blok 2 sa rovnaký proces segmentácie vykonáva rekurzívne, kým nedostaneme bloky kde hodnota DoC je menšia ako PDoC. V dôsledku toho sa blok 2 stáva kompozitným blok a je rozdelený do 5 podblokov (bloky 2.1 až 2.5). Bloky 1 a 3 sú ponechané nedotknuté (obrázok 1.1b). Algoritmus sa zastaví, pretože je splnená podmienka zastavenia, t.j DoC každého bloku je väčšie ako PDoC

OBR.

1.3.3.2 Top-up strategia

Stratégia zdola nahor začínajú výberom listových uzlov v strome DOM, ktoré sa považujú za základné jednotky. Sú agregované s inými uzlami iteračným spôsobom, kým nie je splnená podmienka zastavenia.

Jedným príkladom tohto prístupu je algoritmus Gomory-HuPS [LLT11]. Blokový graf je reprezentovaný ako rovinný graf so váženou hranou. Najprv sa vytvorí graf. Každý vrchol grafu (blok) predstavuje informácie o obsahu na základe štruktúry DOM a vizuálnych informácií. Uzly DOM sú vybrané ako bloky, ak obsahujú text, odkazy a obrázky ako deti. Vo väčšine prípadov sú tieto prvky listy v strome DOM.

Hrany sa pridávajú na základe umiestnenia blokov v geometrickom modeli, t. j. hrana sa pridáva medzi dva bloky, ak sú najbližšími susedmi. Váha bloku je normalizovaná podobnosť cesty dvoch uzlov DOM s použitím vzdialenosti stromu úprav v oboch cestách.

To znamená, že uzly DOM v rovnakom podstromu budú mať menšiu váhu, ako keby boli umiestnené v inej vetve stromu. Obrázok 1.2a zobrazuje príklad webovej stránky a vybrané uzly DOM. Obrázok 1.2b zobrazuje počiatočný blokový graf s jeho váhami.

Algoritmus iteratívne vyhodnocuje dva bloky i a j . Ak je váha hrany medzi týmito dvoma vrcholmi pod parametrom, oba bloky sú zoskupené. Obrázky 1.2c a 1.2d znázorňujú konečnú segmentáciu a blokový graf za predpokladu, že hodnota parametra je 0,1.

OBR

1.3.4 Základný prístup

Táto časť predstavuje základné prístupy pre segmentáciu webových stránok: obrázkovú a textovú založenú na TagName. Základné prístupy berú do úvahy iba jednu charakteristiku webových stránok.

1.3.4.1 Prístup založený na obrázku

Algoritmy segmentácie obrázkov sa používajú na získanie rozloženia webovej stránky z jej renderovaného obrázka, t. j. zo snímky renderovanej stránky. Nepoužívajú zdrojový kód ani vykreslený DOM webovej stránky. Renderované obrázky webovej stránky sa líšia od prirodzených obrázkov v tom zmysle, že je jednoduchšie detegovať objekty v renderovaných obrázkoch, pretože majú ostré hrany alebo farebné prechody.

Tento prístup je motivovaný skutočnosťou, že DOM nemá vždy priamy prístup ku všetkým viditeľným informáciám. Napríklad, ak bola stránka vytvorená pomocou jedného flashovaného objektu alebo je založená na apletoch Java.

1.3.4.2 Prístup založený na texte

V oblasti analýzy textu sa slová, slabiky a vety používajú ako štatistické opatrenia na identifikáciu štruktúrnych vzorcov v textových častiach webových dokumentov.

Možno to považovať za špeciálnu formu segmentácie. Regulárne výrazy sa používajú na analýzu webovej stránky, ktorá hľadá bloky. Aj keď je možné na túto úlohu použiť regulárne výrazy, je známe, že nie sú najlepšou voľbou, pretože HTML nie je regulárny jazyk. V textových prístupoch môže byť text extrahovaný buď zo zdrojového kódu HTML webovej stránky, alebo z renderovaného DOM. Na rozdiel od regulárnych výrazov pojem hustota textu a hustota textu odkazu bola viac akceptovaná v segmentácii webových stránok komunity.

1.3.4.1 Prístup založený na TagName

V literatúre sa tento prístup nazýva DOM-based, ale najlepší je názov prístup založený na TagName. Tento prístup skutočne analyzuje webovú stránku na základe jej zdrojového kódu, t. j. stránka nie je renderovaná. Vo všeobecnosti bloky vyrobené týmito metódami majú tendenciu rozdeľovať stránky na základe ich vopred určenej syntaktickej štruktúry, t. j. značiek HTML.

1.3.5 Hybridné prístupy

Táto časť predstavuje hybridné prístupy k segmentácii webových stránok: založené na vízii, založené na šablóne a založené na grafe. Hybridné prístupy sa snažia prekonať obmedzenia základných prístupov tým, že zohľadňujú niekoľko charakteristík súčasne. Možno to považovať za zmiešanie niekoľkých prístupov.

1.3.5.1 Prístup založený na vizualizácii

Tento prístup kombinuje nasledujúce základné prístupy a používa nasledujúce charakteristiky.

Základné prístupy	Charakteristiky stránky
Založené na TagName	vizuálne nápovedy
Založené na texte	Textový obsah schémy

Ľudia vnímajú webovú stránku ako súbor rôznych sémantických objektov a nie ako jediný objekt. Niektoré výskumy ukazujú, že používatelia vždy očakávajú, že určité funkcie webovej stránky (napr. navigačné odkazy, reklamný panel) sa objavia na určitej pozícii tejto stránky. V skutočnosti, keď sa zobrazí webová stránka, priestorové a vizuálne podnety môžu používateľovi pomôcť (nevedome) rozdeliť webovú stránku na niekoľko sémantických častí. Preto môže byť možné automaticky segmentovať webové stránky pomocou priestorových a vizuálnych podnetov. Obsahová štruktúra stránky založená na vízii sa získa kombináciou štruktúry DOM a vizuálnych podnetov.

1.3.5.2 Prístup založený na šablóne

Tento prístup kombinuje nasledujúce prístupy a používa nasledujúce charakteristiky:

Základné prístupy	Charakteristiky stránky
Založené na TagName	vizuálne nápovedy
Založené na vizuálnej báze	Textový obsah Schémy Šablóny prvkov

Veľký počet webových stránok má spoločnú šablónu, ktorá obsahuje zložené bloky ako je hlavička, päta, ľavá bočná lišta, pravá strana a telo. Tieto bloky zvyčajne dodržiavajú určité konvencie návrhu rozloženia, ktoré možno použiť na odvodenie heuristických pravidiel pre ich klasifikáciu na základe informácií vykresleného stromu DOM.

1.3.5.3 Prístup založený na grafe

Tento prístup kombinuje nasledujúce prístupy a používa nasledujúce charakteristiky:

Základné prístupy	Charakteristiky stránky
Založené na DOM	vizuálne nápovedy
Založené na texte	Textový obsah Schémy

Prístup je založený na nasledujúcej formulácii. Vzhľadom na vykreslený strom DOM, nech N je množina uzlov. Vytvorí sa graf, ktorého množina uzlov je N a ktorého hrany majú váhu, ktorá predstavuje náklady na umiestnenie susedných uzlov do rôznych segmentov.

1.3.6 Záverom o algoritmoch segmentácie webových stránok

Na segmentáciu stránok bolo vyvinutých veľa prístupov. Vo všeobecnosti sú algoritmy navrhnuté s ohľadom na danú aplikáciu. Preto sú efektívne iba pre niektorých konkrétne aplikačné domény.

Algoritmy VIPS a GraphBased predstavujú správny prístup k segmentácii webových stránok, pretože použité pravidlá a logika sú založené na niektorých webových štandardoch a charakteristikách, ktoré sú vlastné webovým stránkam. V týchto prístupoch je však slabina: detekcia blokov vo veľkej miere závisí od názvov označení a textového obsahu. Zatiaľ čo pravidlá heuristiky majú primerané obmedzenia, formálne prístupy dokážu spracovať širší súbor dokumentov. Prístup založený na grafe však potrebuje testovacie údaje na výpočet váh hrán, ktorých vytváranie môže byť nákladné a prístup založený na denziometrii je obmedzený na textový obsah, zatiaľ čo obrázky nie sú brané do úvahy. Spoliehanie sa na názvy značiek a textové informácie môže v niektorých prípadoch viesť k neočakávaným výsledkom a znamená to definovanie špeciálnych heuristických pravidiel.

1.3.7 Spracovanie dokumentov a segmentácia webových stránok

Existuje jasný vzťah medzi segmentáciou stránky a poľom počítačového videnia. Segmentácia a pochopenie obrázkov naskenovaných dokumentov je veľmi dobre študovaný predmet v oblasti spracovania dokumentov. V systémoch na spracovanie dokumentov sa koncepty objektov, zón alebo blokov používajú na definovanie geometrickej a logickej štruktúry dokumentu, kde je každý blok

spojený s kategóriou, jej geometrickými informáciami a štítkom. Spracovanie dokumentu zahŕňa fázy analýzy a porozumenia dokumentu. Obrázky dokumentov sa analyzujú na zistenie blokov na základe informácií o pixeloch. Bloky sú kategorizované, ich geometrické informácie sa extrahujú a v závislosti od oboch funkcií sa priradí označenie.. Dva bloky odsekov by sa mali zlúčiť, nadpis by sa mal objaviť ako prvý a za ním zlúčený blok odseku. (nie je všetko prenesené, treba pozrieť, či toto dáva zmysel)

1.4 Segmentation evaluation

Otázka, ktorej sa venujeme v tejto časti, znie: ako dobre prezentované prístupy správne identifikujú segmenty na súčasných webových stránkach? Táto otázka nastoľuje problém hodnotenia metód segmentácie stránok.

Táto časť predstavuje rôzne metódy používané na vyhodnotenie segmentačných algoritmov (1.4.1) a našu interpretáciu správnosti segmentácie (1.4.2). Skúmali sme aj connex doménu segmentácie skenovaných stránok (1.4.3), keďže problematika vyhodnocovania takýchto systémov je dosť podobná s našim problémom. Predstavujeme súčasný stav hodnotenia segmentácie webovej stránky (1.4.4), zhrnutie (1.4.5) a na záver diskusiu (1.4.6).

1.4.1 Klasifikácia hodnotiacich metód

1. Analytická metóda: priamo vyhodnocuje samotné segmentačné algoritmy pomocou analýzy ich princípov a vlastností

2. Empirická metóda: nepriamo posúdiť segmentačné algoritmy ich aplikovaním na testovacie údaje a meraním kvality výsledkov segmentácie. Možno ich rozdeliť na: metódy kvality (goodness) a metódy diskrepancie.

- Metódy kvality (goodness) sú metódy, pri ktorých sa segmentácia považuje za ideálnu, ak spĺňa určité podmienky posudzované ľudským úsudkom. Posledné menované možno získať jednoduchým pozorovaním alebo získaním niektorých funkcií, ktoré pozorovanie dopĺňajú. Nie je potrebné mať a priori znalosti o segmentácii referencií. Výsledok je známy ako goodness parameter.

- Metódy diskrepancie sú metódy, ktoré porovnávajú segmentáciu so správnou alebo ideálnou segmentáciou. Táto ideálna segmentácia je známa aj ako základná pravda. Autor objasňuje, že cieľom týchto metód je nájsť parametre diskrepancie, ktoré umožňujú určiť, ako ďaleko sú tieto dve segmentácie jedna od druhej. Základná pravda je stanovená „a priori“.

1.4.2 Segmentacia korektneho hodnotenia

Hodnotenie segmentácie obrazu je veľmi dobre študovaná oblasť. Napríklad [CCR05] popisuje niekoľko metrík na meranie kvality segmentácie na základe vzdialenosti každého oddielu v segmentácii obrazu. Táto metóda je však určená pre všeobecné obrázky, ktoré nie sú špeciálne pre naskenované dokumenty. Jeho prispôsobenie webovým stránkam teda nie je jednoduché. [LPH01] meria presnosť segmentácie určením označenia priradeného entitám odsekom, nadpisom, tabuľkou atď.) pomocou algoritmov segmentácie obrazu a ich zhody s vopred stanovenou taxonómiou. Tento prístup sa približuje tomu, čo potrebuje hodnotenie segmentácie webu, ale nezohľadňuje obsah a geometrický aspekt skenovaného dokumentu ako súčasť hodnotenia. [SKB08] predstavuje vektorové skóre, ktoré identifikuje bežné triedy chýb segmentácie pomocou základnej pravdy anotovaných naskenovaných strán. Myslíme si, že tento prístup pokrýva všetky potreby segmentácie webových stránok (s určitými úpravami a úpravami), keďže zohľadňuje obsah, geometria a vizuálna stránka naskenovaného dokumentu v ich hodnotiacom modeli.

1.4.3 Miera správnosti segmentácie skenovaného dokumentu

Hodnotenie segmentácie obrazu je veľmi dobre študovaná oblasť. Napríklad [CCR05] popisuje niekoľko metrík na meranie kvality segmentácie na základe vzdialenosti každého oddielu v segmentácii obrazu. Táto metóda je však navrhnutá pre všeobecné obrázky, nie špeciálne pre skenované dokumenty. Jeho prispôsobenie webovým stránkam teda nie je jednoduché. [[LPH01] merať presnosť segmentácie určením označenie priradené entitám (odseky, nadpis, tabuľka atď.) pomocou algoritmov segmentácie obrázkov a ich zhoda s vopred určenou taxonómiou. algoritmy a ich zhoda s predurčenou taxonómiou. Tento prístup sa približuje tomu, čo potrebuje hodnotenie segmentácie webu, ale nezohľadňuje obsah a geometrický aspekt skenovaného dokumentu ako súčasť hodnotenia. [SKB08] prítomný a vektorové skóre, ktoré identifikuje bežné triedy chýb segmentácie pomocou základnej pravdy anotovaných naskenovaných strán. Myslíme si, že tento prístup pokrýva všetky potreby segmentácie webových stránok (s určitými úpravami a úpravami), keďže zohľadňuje obsah, geometria a vizuálna stránka naskenovaného dokumentu v ich hodnotiacom modeli.

Na základe počtu hrán v grafe je definovaných sedem mier. Uvádzajú, ako ďaleko je navrhovaná segmentácia od základnej pravdy. Tieto metriky sú:

- Celkový počet správnych segmentácií: celkový počet zhôd jedna ku jednej medzi základné komponenty a komponenty segmentácie.

- Celkové presegmentácie: celkový počet významných hrán, ktoré majú základné pravdivé komponenty, mínus počet základných pravdivých komponentov, ku ktorým dopadá aspoň jedna významná hrana.
- Celkové podsegmentácie: celkový počet významných hrán, ktoré majú komponenty segmentácie, mínus počet komponentov segmentácie, ku ktorým dopadá aspoň jedna významná hrana.
- Presegmentované komponenty: počet základných komponentov, ktoré majú viac ako jednu významnú hranu.
- Podsegmentované komponenty: počet segmentovaných komponentov, ktoré majú viac ako jednu významnú hranu.
- Vynechané komponenty: počet základných komponentov, ktoré nezodpovedali žiadnemu komponentu v popredí v navrhovanej segmentácii.
- Falošné popluchy: počet komponentov v navrhovanej segmentácii, ktoré nezodpovedali žiadnemu komponentu v popredí v segmentácii pozemnej pravdy.

3. Porovnanie algoritmov

BoM

Segmentačný algoritmus BoM – Block-o-Matic je kategorizovaný ako hybridný algoritmus a to pretože jeho návrh je inšpirovaný metódami automatizovaného spracovania dokumentov a metódami vizuálnej segmentácie obsahu. BoM má niekoľko verzií. V rámci tejto práce sú popísané dve verzie, ktoré sa výraznejšie líšia štruktúrou. Ostatné verzie sa v štruktúre a princípe fungovania líšia len minimálne. Algoritmus berie na svoj vstup strom HTML elementov uložených v rozhraní DOM.

BoM verzia 1

Počas segmentácie sa pracuje s tromi štruktúrami a to: DOM – tento štandard implementuje každý známy prehliadač, geometrická štruktúra a logická štruktúra.

Geometrická štruktúra v sebe drží informácie o kategórii obsahu webového objektu a informáciu o geometrii daného webového objektu. Logická štruktúra reprezentuje simuláciu ľudského vnímania obsahu dokumentu a zároveň je táto štruktúra výstupom segmentácie. Proces segmentácie pomocou algoritmu BoM je rozdelený do troch fází a to je analýza, porozumenie a rekonštrukcia dokumentu.

Popis segmentačného modelu

Segmentačný proces webovej stránky je rozdelený do troch častí: analýza štruktúry DOM,

porozumenie a rekonštrukcia stránky. Ako vstup sa berie DOM. Ten je predspracovaný a jednotlivé elementy sú označované v závislosti na to či sú nositeľmi nejakého obsahu alebo či zastrešujú ďalšie elementy ktoré nosia obsah. Výsledkom tohto predspracovania je obsahová štruktúra. Takto predspracovaný DOM pokračuje do ďalšej fázy segmentácie – do analýzy. V rámci analýzy sa DOM spracováva algoritmom c2g. Výstupom tohto algoritmu je geometrická štruktúra objektov webovej stránky W_{geom} . Posledná fáza berie na vstup geometrickú štruktúru ktorú mapuje na logickú štruktúru W_{log} stránky pomocou algoritmu g2l. Algoritmus g2l je parametrizovaný a jej výstup je závislý na parametroch pDC a pAC, ktorých význam bude vysvetlený nižšie v texte. Rekonštrukcia stránky sa odvíja od logickej štruktúry, ktorá závisí na predchádzajúcich spracovaných štruktúrach. Pseudoalgoritmický zápis algoritmov c2l a g2l spolu s ich popisom je zahrnutý nižšie v texte.

Na obrázku nižšie sú ukázané vzťahy medzi jednotlivými štruktúrami od HTML elementov až po logickú štruktúru po vykonaní segmentácie.

Geometrická štruktúra

Geometrická štruktúra popisuje ako sú elementy vsadené do stránky a do akých kategórií sú klasifikované. V tejto štruktúre sa získava geometria vyrenderovaného objektu.

Geometriu chápeme ako geometrické vlastnosti (napr. výška, šírka, pozícia) elementu.

Tieto vlastnosti je možné získať až po vykreslení stránky a sú dané css boxom, ktorý zaobaluje daný element.

Geometrická štruktúra je strom a jej uzly sú tvorené Geometrickými Objektami.

Geometrický objekt je formálne definovaný ako $G_o = (e, a, f)$, kde

e – je konečná množina elementov ktoré geometrický objekt pokrýva

a – je konečná množina atribútov ktoré môže geometrický objekt nadobúdať.

f – je konečná množina operácií/funkcií

V množine atribútov a sa nachádzajú nasledujúce atribúty: Geometry a Category.

Atribút geometry je definovaný ntickou (x,y,w,h) . Táto ntica reprezentuje absolútnu pozíciu obdĺžnika vo vyrenderovanom dokumente, ktorý pokrýva všetky elementy e .

Hodnota x vyjadruje minimálne odsadenie obdĺžnika od vrchnej strany webovej stránky.

Hodnota y zas reprezentuje minimálne odsadenie obdĺžnika od ľavej strany stránky.

Hodnota w a h značia šírku a výšku daného obdĺžnika.

Atribút `category` obsahuje všeobecné označenie kategórie ktorú geometrický objekt zastrešuje.

Množina funkcií f obsahuje nasledujúce operácie ktoré môže geometrický objekt vykonávať. V tejto množine sú tieto operácie: `Diagonal()`, `rDiagonal()` a `Valid()`. Operácia `Diagonal()` vracia hodnotu dĺžky diagonály atribútu geometrie. Operácia `rDiagonal` vracia pomer funkcie `Diagonal` spustenej nad geometrickým objektom a jeho rodičom: $go.Diagonal() / pgo.Diagonal()$, kde pgo je rodič go .

Operácia `Valid()` kontroluje validitu elementu – viditeľnosť elementu vo vykreslenej stránke a reprezentatívnosť plochy a obsahu objektu v rámci webovej stránky.

Geometrickú štruktúru – W_{geom} je teda možné definovať ako konečný strom kde uzly tvoria geometrické objekty. Celý strom je definovaný následovne – koreň stromu je tvorený geometrickým objektom Go_i , jeho bezprostrednými potomkami sú ďalšie geometrické objekty Go_{ij} a ktoré sa odkazujú na ďalšie podstromy geometrických štruktúr $W_{geom_{ijk}}$.

Podstromy $W_{geom_{ijk}}$ obsahujú korene Go_{ijk} ktoré sú bezprostrednými predchodcami uzlu Go_{ij} .

Logická štruktúra

Logická štruktúra simuluje ľudské vnímanie vizuálneho obsahu stránky. Udržiava v sebe prepojenie medzi blokmi stránky. Jednotlivé uzly v štruktúre sú tvorené logickými objektami.

Logický objekt

Logický objekt je ntica $Lo = (g, a, f)$ kde:

- g je konečná množina geometrických objektov
- a je konečná množina atribútov ktoré môže logický objekt nadobúdať
- f je konečná množina operácií/funkcií

Množina atribútov obsahuje atribúty `Geometry`, `Label` a `Follows`. `Geometry` obsahuje množinu bodov popisujúcich polygon – v prostredí vykreslenej webovej stránky najčastejšie obdĺžnik, ktorý reprezentuje minimálny obal geometrických objektov zahrnutých v danom logickom objekte. `Label` reprezentuje popisok sémantickej časti webovej stránky. Označenie môže nadobúdať hodnoty: `header` (hlavička stránky), `navigation` (navigácia), `title` (titulok), `content` (obsah), `table` (tabuľka), `image` (obrázok),

logo, footer (päta stránky) a none (nerozoznaná časť). Atribút Follows obsahuje predchádzajúci logický objekt. Tento atribút môže nadobúdať hodnoty: independent (nezávislý) alebo obsahuje odkaz na predchádzajúci logický objekt.

Množina funkcií obsahuje operácie: Diagonal(), rDiagonal(), Distance(lo), VisualCuesPresent(), Accept(label) a MergeWith(lo). Operácia Diagonal() vracia hodnotu najdlhšej diagonály geometrie logického objektu. Operácia rDiagonal() vracia pomer medzi operáciami lo.Diagonal() a plo.Diagonal(), kde plo je rodičovský uzol objektu lo. Distance(lo) vracia hodnotu najmenej vzdialenosti medzi geometriami dvoch logických

objektov. Pre príklad lo.Distance(otherLo) vracia najmenšiu vzdialenosť medzi logickými objektami lo a otherLo. Operácia VisualCuesPresent() vracia informáciu o tom či má logický objekt nejaké výrazné vizuálne vlastnosti ako sú farba pozadia a veľkosť fontu. Accept(label) priradzuje logickému objektu sémantické označenie a status listového uzla v logickej štruktúre. Status listového uzla znamená že odstránenie všetkých bezprostredných následníkov logického objektu. Počas tejto operácie sa závislosti medzi štruktúrami W geom a W log nemenia. MergeWith(otherLo) vykonáva spojenie geometrií dvoch logických objektov. Napríklad lo.MergeWith(otherLo) spája dohromady logické objekty lo a otherLo. Pri spájaní sa vytvára nová geometria ktorá pokrýva obe spájané logické objekty. Hľadá sa hľadá sa minimálna ortogonálna obálka.

Logická štruktúra

Logickú štruktúru – W geom je definovaná ako konečný strom kde uzly tvoria geometrické objekty. Celý strom je definovaný následovne – koreň stromu je tvorený geometrickým objektom Lo_i , jeho bezprostrednými potomkami sú ďalšie geometrické objekty Lo_{ij} a ktoré sa odkazujú na ďalšie podstromy geometrických štruktúr W log ijk. Podstromy W log ijk obsahujú korene Lo_{ijk} ktoré sú bezprostrednými predchodcami uzlu Lo_{ij} .

Predspracovanie

Na začiatku segmentácie prebieha predspracovanie dokumentu, kedy sa prechádza celá DOM štruktúra a analyzujú sa jednotlivé tagy elementov. Algoritmus sa snaží klasifikovať elementy do štyroch kategórií – content, content container, container a default. Výstupom tejto fázy je štruktúra W cont. Celý proces začína od elementu body. Prakticky sa jedná o DOM štruktúru v ktorej sú elementy označované vyššie spomínanými kategóriami.

Analýza stránky

Analýza stránky má za úlohu vytvoriť geometrickú štruktúru z predspracovanej DOM štruktúry. DOM štruktúra sa prechádza opäť od elementu body a postupne sa vytvára strom geometrickej štruktúry z geometrických objektov. Nový geometrický objekt je naplnený informáciami o kategórii a geometrii, ktoré sú získané z jednotlivých elementov.

Algoritmus sem

Proces porozumenia

Proces porozumenia mapuje geometrickú štruktúru na logickú štruktúru. Počas mapovania sa berie v úvahu pomer diagonál medzi logickými objektmi, kategória objektov, ich poloha v rámci stránky a vzdialenosť medzi objektami. Počas tohto procesu sa algoritmus snaží zoskupiť objekty ktorých pomer diagonál nie je väčší ako parameter pAC a maximálne oddelenie objektov medzi sebou nepresahuje parameter pDC. Parameter pAC reprezentuje granularitu. To znamená že v závislosti na hodnote sa bude meniť veľkosť blokov vo výslednej segmentácii. V prípade malých hodnôt bude stránka rozsegmentovaná na viac malých blokov. Analogicky, v prípade väčších hodnôt bude výsledná segmentácia obsahovať menej veľkých blokov – sémantické bloky budú väčšie. Celá táto stratégia sa snaží simulovať ľudské vnímanie obsahu. Toto vnímanie vychádza zo štyroch Gestaltových zákonov: zákon blízkosti, zákon podobnosti, zákon uzavrenosti a zákon jednoduchosti. Zákon blízkosti pojednáva o tom že vizuálne objekty ktoré spolu súvisia, zvyknú byť zoskupené blízko seba. Zákon podobnosti poukazuje na to že objekty,

ktoré sú si vizuálne podobné sú taktiež zoskupené blízko seba. Zákon uzavrenosti poukazuje na to že objekty sú zoskupené ak spolu tvoria nejakú štruktúru v dokumente. Posledný zákon poukazuje na to že objekty v dokumentoch bývajú zoskupené do jednoduchých štruktúr, ktoré v sebe udržujú symetriu a pravidelnosť. Celý proces porozumenia sa drží týchto zákonov. Napríklad, zhlukujú sa objekty ktoré sú k sebe čo najbližšie – zákon blízkosti alebo sú medzi sebou zarovnané – zákon jednoduchosti. Tento proces je ovplyvnený zvolenými parametrami pAC a pDC. Parameter pDC je prahová hodnota ktorá určuje minimálnu odchýlku vertikálneho a horizontálneho zarovnania objektov.

Rekonštrukcia stránky

Na vstup rekonštrukcie stránky sa dostáva logická štruktúra. Logické objekty, geometrické objekty a DOM elementy sú prepojené a tvoria hierarchiu. Logické objekty v sebe držia odkaz na príslušné geometrické objekty a tieto zas udržujú odkazy na príslušné DOM elementy ktoré pokrývajú svojou geometriou. Celá táto štruktúra sa rekurzívne prechádza a extrahuje sa dáta logických a geometrických objektov. Pôvodný popis algoritmu na výstupe vytvára špeciálny XML dokument, ktorý obsahuje pôvodnú HTML štruktúru, nad ktorou sú nadstavené dáta získané z logických a geometrických

4. Fit layout systém

5. Implementácia do systému fut-layout

6. Experimenty a porovnanie

7. Záver

Táto práca sa zaoberá niektorými problémami, ktoré prináša segmentácia webových stránok.

Zameriava sa na rôzne body: presnosť a všeobecnosť segmentátora, ako aj presnosť a spravodlivosť hodnotenia segmentátorov. Konkrétnejšie sa venujeme nasledujúcim výzvam a prispievame nasledovne:

- Navrhujeme Block-o-matic (BoM) nový algoritmus segmentácie webovej stránky. Táto práca je prvou, ktorá pri návrhu segmentátora zohľadňuje presnosť a generickosť ako kritériá kvality segmentácie. Ako ukazujú naše výsledky, umožňuje segmentovať rôzne typy webových stránok bez predchádzajúcej znalosti obsahu s vyššou presnosťou ako ostatné testované segmentátory. Vďaka stratégii zdola nahor a heuristickým pravidlám stanoveným v štandardoch W3C dosahujeme všeobecnosť a presnosť.
- Navrhujeme rámec, ktorý nám umožňuje hodnotiť správnosť segmentačných algoritmov bez ohľadu na to, aký prístup používajú. Pokiaľ je nám známe, toto je prvá hodnotiacia práca, ktorá sa zameriava na segmentáciu vnútorných vlastností, ktorými sú obsahové a

blokové obdĺžniky. Existujúce prístupy majú hodnotenie, ale sú riadené špecifickými aplikáciami, a preto nie sú dostatočne všeobecné na porovnanie všetkých segmentačných algoritmov. Náš prístup je založený na základnej pravde, postavený vďaka nástroju (MoB), ktorý sme vyvinuli a ktorý podstatne uľahčuje manuálny návrh segmentácie. Náš súbor údajov obsahuje 125 strán pokrývajúcich ~ve kategórií (25 strán na kategóriu). • Predstavujeme hodnotiaci model, ktorý definuje niekoľko užitočných metrík pre hodnotenie. Jedna metrika je venovaná úlohe extrakcie textu, druhá počíta, ako dobre sa bloky detekované daným algoritmom zhodujú s blokmi základnej pravdy. Tento model používame na vyhodnocovanie a porovnávanie štyroch segmentačných algoritmov prispôbených na začlenenie do nášho rámca. Výsledky ukazujú, že algoritmy fungujú primerane pri extrakcii textu zo stránok. S ohľadom na blok korešpondencie, výsledky mierne závisia od kategórie posudzovaných stránok. Do hodnotiaceho modelu zahrnieme dôležitosť, aby algoritmy, ktoré detegujú dôležité bloky, získali lepšie skóre. Naše experimenty ukazujú, že BoM je najpresnejší a najvšeobecnejší algoritmus pri hodnotení. Je to preto, že predstavuje najnižší výskyt presegmentácií a podsegmentácií (problémy zrnitosti) a pomerne dobre funguje na všetkých typoch stránok, ktoré sa zvažujú pri hodnotení.

Predstavujeme dve aplikácie využívajúce náš segmentačný algoritmus a hodnotiaci rámec. Pagelyzer je nástroj vyvinutý v kontexte európskeho projektu SCAPE, ktorý využíva BoM na porovnávanie dvoch verzií webových stránok, aby sa rozhodlo, či sú podobné alebo nie. Aby sme minimalizovali emuláciu vo webových archívoch, vyvíjame nástroj na migráciu webových stránok z formátu HTML4 do formátu HTML5 pomocou BoM.

Vo zvyšku tejto kapitoly načrtneme naše plány pre budúcu prácu a diskutujeme o možných výskumných témach nad rámec toho, čo bolo riešené v práci. Segmentácia webových stránok Existuje mnoho smerov pre budúcu prácu týkajúcu sa segmentácie webových stránok. Označovanie blokov je kľúčovou úlohou pri segmentácii webových stránok. Definuje úlohu, ktorú blok hrá, teda ako by sa s ním malo zaobchádzať. V tejto práci ukazujeme, že používanie štítkov z kategórií obsahu HTML nám

umožňuje byť všeobecný. Bolo by však zaujímavé poskytnúť presnejší popis úlohy, ktorú blok hrá, aby sa aplikáciám, ktoré sa spoliehajú na segmentáciu, poskytli zaujímavejšie informácie. Môžeme napríklad uviesť, či blok označený ako päta je prílohou, indexom alebo podrobnou licenčnou zmluvou, alebo či blok označený ako navigácia je navigáciou "prvej, druhej alebo tretej úrovne. Poradie čítania je užitočná informácia získaná zo segmentácie. Ako je uvedené v kapitole 5, je to užitočná informácia pre spracovanie výstupov segmentácie. Napríklad mobilná aplikácia, ktorá sa pri vizualizácii webovej stránky spolieha na segmentáciu, môže divákovi poskytnúť možnosť čítať obsah stránky blok po bloku podľa poradia čítania. Podľa intuície sme v tejto práci študovali oblasť

spracovania dokumentov a hľadali pohľady na implementáciu poradia čítania v segmentácii webových stránok. Zistili sme, že poradie čítania sa zvyčajne riadi poradím, v ktorom sa nachádzajú bloky. Myslíme si však, že stále existuje priestor na ďalšie vylepšenia, možno súvisiace s typom stránok a začlenením miery poradia čítania do hodnotenia. Obdĺžnik je vnútorný tvar blokov. Blokové obdĺžniky závisia od polí prvkov stránky, ktoré sú tiež obdĺžnikové. Obdĺžniky sa však môžu prekryvať, čo spôsobuje nejasnosti vo vizuálnej prezentácii. Toto je problém, pretože vizuálna prezentácia je pre klientske aplikácie čoraz dôležitejšia. Napríklad vo vyššie uvedenom príklade mobilnej aplikácie pri zobrazovaní bloku môžu byť do vizualizácie zahrnuté časti iných blokov (prekrývajúce sa). Preto plánujeme pridať podporu pre iné ako manhattanské rozloženia v segmentácii. Namiesto obdĺžnikov sa použijú mnohouholníky. Budeme študovať nové vizuálne charakteristiky webových stránok (CSS3): okrem iného obrázky na pozadí, animácie, Ajax. Intuitívne je riešením tohto problému považovať vizuálne podnety za súčasť obsahu stránky. To nemusí byť také jednoduché. V skutočnosti je to hlavne dôvod, prečo CSS existuje, teda oddeliť obsah svojej prezentácie. Hodnotenie Existuje mnoho smerov pre budúcu prácu okolo hodnotenia. Najprv plánujeme použiť techniky strojového učenia (ML) na učenie sa parametra tolerancie tr. Plánujeme tiež použiť ML na objavovanie nových relevantných funkcií skóre, na základe spätnej väzby od používateľov, ktorí dávajú manuálne skóre

segmentáciám z tréningovej sady. Po druhé, budeme pokračovať v experimentovaní s algoritmi segmentácie na viacerých stránkach a viacerých kategóriách stránok. Naším cieľom je vyvinúť kompletný hodnotiaci rámec, ktorý pomôže používateľom pri výbere najlepšieho segmentačného algoritmu v závislosti od ich aplikácie a kategórie stránok, s ktorými manipulujú. Samozrejme, keďže výsledky ukazujú, že niektoré algoritmy majú problémy s niektorými kategóriami, možno ich použiť aj na zlepšenie efektívnosti segmentačných algoritmov pre tieto kategórie. Po tretie, plánujeme vyhodnotiť segmentačné algoritmy s ohľadom na typ úlohy, ktorá segmentáciu využíva. Typy úloh zahŕňajú extrakciu webových entít, detekciu rozloženia, detekciu kotla, vizualizáciu v zariadeniach s malou obrazovkou av kontexte digitálnych knižníc optimalizáciu prehľadávania webových archívov, detekciu zmien medzi verziami webových stránok a iné. To znamená definovanie skriptov, ktoré vykonávajú úlohu (vrátane volaní na segmentáciu) a definovanie nových ad hoc metrík pre každú úlohu. Tiež by sme chceli definovať generický model pre segmentáciu webových stránok, ktorý dokáže vyjadriť všetky existujúce prístupy. To by umožnilo analytické vyhodnotenie segmentačných algoritmov. Po štvrté, do hodnotenia plánujeme zahrnúť poradie čítania segmentácie. Dobrým kandidátom je metrika zavedená Liangom [LPH01]. Vyhodnotenie poradia čítania je zredukované na výpočet počtu pohybov potrebných na získanie poradia čítania základnej pravdy z vypočítanej segmentácie.

Použitá literatúra:

[TS94] Yuan Y Tang and Ching Y Suen. Document structures: a survey. International journal of pattern recognition and artificial intelligence, 8(05):1081 1111, 1994.