VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION SYSTEMS

Algoritmy pro segmentaci webových stránek

diplomová práce

Master´S THESIs

AUTOR PRÁCE Bc. Tomáš Laščák

AUTHOR

BRNO 2016

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION SYSTEMS

Algoritmy pro Segmentaci webových stránek

Web page segmentation algorithms

Diplomová práce

Master´S THESIs

AUTOR PRÁCE Bc. Tomáš Laščák

AUTHOR

VEDOUCÍ PRÁCE Ing. Radek Burget, Ph.D.

SUPERVISOR

BRNO 2016

Abstrakt

Segmentace webových stránek je jednou z disciplín extrakce informací. Umožňuje dělit stránky na různé sémantické bloky. Semestrální projekt se zabývá seznámením se segmentací jako takovou a také segmentační metodou. V práci jsou popsány různé příklady metod jako je VIPS, DOM PS atd. Je zde teoretický popis zvolené metody a taktéž Frameworku FitLayout, který bude o tuto metodu rozšířen. V závěru práce je uvedeno shrnutí a také popis dalších postupů při navázání této práce na diplomovou.

Abstract

Segmentation of web pages is one of the disciplines of information extraction. It allows to divide the page into different semantic blocks. Term project deals with learning about segmentation as well as a segmentation method. In this paper we describe various of examples method such as VIPS, DOM PS etc. There is a theoretical description of the chosen method and also FitLayout Framework, which will be expanded by this method. In the conclusion is a summary and description of the other procedures for the establishment of the work on the thesis.

Klíčová slova

Vizuální sémantika, Java, WWW, Segmentace, Document Object Model

Keywords

Visual semantic, Java, WWW, Segmentation, Document Object Model

Citace

Laščák Tomáš: Algoritmy pro segmentaci webových stránek, diplomová práce, Brno, FIT VUT v Brně, 2016

**Algoritmy pro segmentaci webových stránek**

Prohlášení

Prohlašuji, že jsem tuto práci vypracoval samostatně pod vedením pana Ing. Radka Burgeta, Ph.D. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……………………

Tomáš Laščák

15. 5. 2016

Poděkování

Rád bych tímto poděkoval vedoucímu této práce, Ing. Radku Burgetovi, Ph.D., za odborné rady a vedení při tvorbě této práce.

© Tomáš Laščák, 2016

Tato práce vznikla jako školní dílo na Vysokém učení technickém v Brně, Fakultě informačních technologií. Práce je chráněna autorským zákonem a její užití bez udělení oprávnění autorem je nezákonné, s výjimkou zákonem definovaných případů.

# Obsah

[Obsah 1](#_Toc447751671)

[1 Úvod 2](#_Toc447751672)

[2 Segmentace webových stránek 3](#_Toc447751673)

[2.1 DOM-based Page Segmentation 3](#_Toc447751674)

[2.2 Vision-based Page Segmentation – VIPS 4](#_Toc447751675)

[2.3 Box Clustering Segmentation 5](#_Toc447751676)

[3 Segmentace webových stránek za pomocí vizuální sémantiky 7](#_Toc447751677)

[3.1 Příprava 7](#_Toc447751678)

[3.1.1 Vizuální bloky 8](#_Toc447751679)

[3.1.2 Předzpracování webových stránek 8](#_Toc447751680)

[3.2 Algoritmus segmentace 9](#_Toc447751681)

[3.2.1 Rozpoznání podobných vizuálních bloků 9](#_Toc447751682)

[3.2.2 Výpočet stupně sousednosti bloků 11](#_Toc447751683)

[3.2.3 Výpočet podobnosti obsahu bloků 12](#_Toc447751684)

[3.2.4 Segment webové stránky 15](#_Toc447751685)

[4 FitLayout 17](#_Toc447751686)

[4.1 FitLayout Moduly 17](#_Toc447751687)

[4.1.1 FitLayout API 17](#_Toc447751688)

[4.1.2 CSSBox 18](#_Toc447751689)

[4.1.3 Segmentace 18](#_Toc447751690)

[4.1.4 Klasifikace 19](#_Toc447751691)

[4.1.5 Nástroje 19](#_Toc447751692)

[5 Implementace algoritmu 20](#_Toc447751693)

[5.1 Sousednost bloků 20](#_Toc447751694)

[5.1.1 Nesousedící bloky 20](#_Toc447751695)

[5.1.2 Vertikální sousednost 21](#_Toc447751696)

[5.1.3 Horizontální sousednost 21](#_Toc447751697)

[5.2 Vizuální obsah bloků 21](#_Toc447751698)

[5.3 Řízení algoritmu 21](#_Toc447751699)

[6 Dosažené výsledky 22](#_Toc447751700)

[7 Závěr 23](#_Toc447751701)

# Úvod

Mnoho uživatelů internetu o ní neví, že existuje, ale využívají ji prakticky každý den. Řeč je o segmentaci webových stránek. Na webových stránkách je mnoho irelevantního obsahu, který je z hlediska běžného uživatele nepoužitelný a je tak potřeba tento obsah filtrovat. Segmentaci lze tedy užitečně využít. Například by ji mohli využít webové vyhledávače, pokud by se snížila její výpočetní náročnost. Velkým přínosem segmentace je hlavně její filtrace nevyžádaného obsahu na webové stránce a umožňující nám tak získat podstatné a nejdůležitější informace z této stránky.

Dřívější techniky segmentace webových stránek jsou založeny především na algoritmech strojového učení a na algoritmech využívající různá heuristická pravidla. Avšak vzhledem k malému rozsahu testovacích množin dat, metody strojového učení mohou být aplikovány pouze v některých určitých oblastech webové stránky. Zatímco metody založené na heuristických pravidlech mohou fungovat na malých sadách stránek, ale zase nejsou vhodné pro větší sady stránek.

V této práci implementujeme jednu z metod segmentace webových stránek. Předpokládáme, že se webová stránka skládá z konečného počtu bloků a může být segmentována pomocí vizuálních vlastností těchto bloků. Metoda tedy využívá těchto vizuálních vlastností. Metoda bude implementována v rámci Frameworku FitLayout a bude do něj integrována.

Co se týče obsahu práce, ve druhé kapitole si popíšeme příklady tří metod pro segmentaci. Nejprve se stručně seznámíme s metodou DOM-based Page Segmentation, která realizuje segmentaci webové stránky na základě jejího DOM stromu. Dále se pak seznámíme z jednou nejznámějších metod pro segmentaci a tou je Vision-based Page Segmentation. Ta pro segmentaci využívá vizuálních vlastností prvků na webové stránce. Na závěr této kapitoly se seznámíme ještě s metodou Box Clustering Page Segmentation.

Ve třetí kapitole se zaměříme na metodu pro segmentaci webových stránek, kterou se zabývá tato práce. Jedná se o metodu využívající vizuálních vlastností stránek. Zde se zaměříme především na teoretickou část této metody. Ukážeme si, jaké se zde využívají vzorce pro výpočty různých vlastností a také pseudokód samotného algoritmu.

Čtvrtá kapitola je soustředěna na seznámení se s Frameworkem FitLayout. Metoda, která bude implementovat segmentaci webových stránek, importujeme do tohoto Frameworku. Taktéž bude využívat jeho již hotové části a je tedy zapotřebí se s některými nejprve obeznámit.

V poslední páté kapitole této práce jsou shrnuty výsledky, jakých cílů bylo dosaženo v semestrálním projektu, a jak se bude dál tato práce rozvíjet v rámci navazující diplomové práce.

# Segmentace webových stránek

Segmentace webových stránek umožňuje dělení stránky podle určitých pravidel nebo vizuálních vlastností. Stránka je rozdělena na několik sémanticky rozdílných bloků a jejich obsah můžeme dále zkoumat nebo různě filtrovat podle potřeby.

Jak jsme již řekli, segmentace je slibná metoda, která by mohla být využívána u webových vyhledávačů. Slouží ke zpřesnění nalezených výsledků pro vyhledávaný řetězec a dokáže taktéž identifikovat jeho polohu v rámci webové stránky, na které byl nalezen. Můžeme tak rozpoznat a vynechat výskyty vyhledávaného řetězce. Jako příklad použití bychom mohli říct například v komentářích pod článkem nebo v odkazech sloužících k navigaci po stránce.

Existuje mnoho metod pro segmentaci webových stránek a v následujících podkapitolách si stručně charakterizujeme tři nejznámější z nich. Metoda, kterou se zabývá tato práce, bude popsána v samostatné kapitole.

## DOM-based Page Segmentation

Jednou z metod segmentace je metoda DOM-based Page Segmentation [12]. Ta realizuje segmentaci webové stránky na základě jejího DOM[[1]](#footnote-1) stromu se značkami jazyka HTML. Příklad DOM stromu je zobrazen na obrázku 1. Segmentovanými bloky jsou v případě této metody části uvozené v HTML značkách. Příkladem těchto značek je <TITLE>, <H1> až <H6>, <META>, <TABLE> atd. Výsledky této metody jsou však často neuspokojivé. Hlavním důvodem jsou následující tři aspekty:

* DOM je stále lineární struktura a vizuálně sousední bloky mohou být ve struktuře daleko od sebe.
* Značky jako <TABLE> a <P> slouží nejen k prezentaci obsahu, ale také pro rozložení struktury. Je tedy obtížné zajistit vhodnou úroveň segmentace.
* Mnoho webových stránek nedodržuje specifikace HTML a ztěžují tak správnou segmentaci.



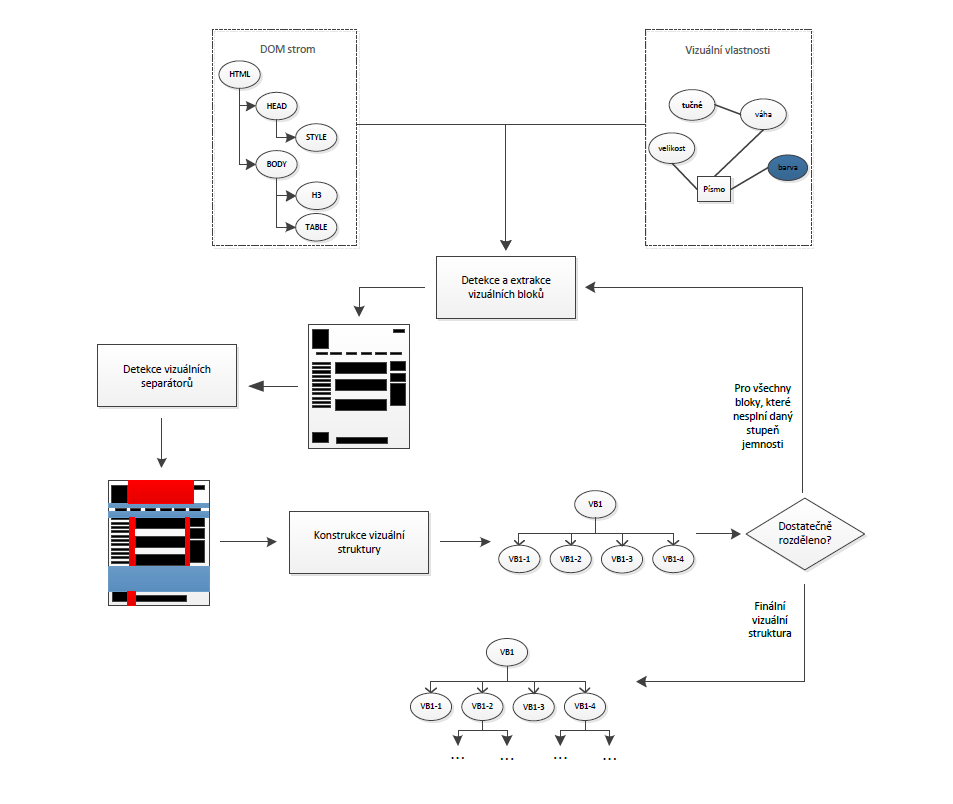
Obrázek 1: Příklad DOM stromu

## Vision-based Page Segmentation – VIPS

Webová stránka obsahuje mnoho vizuálních podnětů, které pomáhají člověku rozeznávat, pomocí webového prohlížeče, jednotlivé stránky. Stránka obsahuje několik dílčích vizuálních prvků, jako jsou obrázky, barvy, čáry a jiné značky. Metoda využívající vlastnosti vizuálních prvků webové stránky je Vision-based Page Segmentation neboli VIPS. Podrobnější popis algoritmu najdeme zde [5]

Nejprve je získán DOM strom z webového prohlížeče. Jakmile je strom získán, začíná se od kořenového uzlu. Probíhá proces extrakce vizuálního uzlu na vizuální uzly dané úrovně z DOM stromu, podle vizuálních podnětů. Každý uzel DOM stromu je zkontrolován a hodnocen, zda tvoří jednoduchý blok nebo ne. Pokud není jednoduchým blokem, synovský uzel je zkontrolován a hodnocen stejně jako rodič. Jakmile jsou extrahovány všechny bloky dané úrovně, vloží se do dané ukládací struktury. Dále pak těmto blokům identifikujeme vizuální separátory. Poté, je vytvořena struktura rozložení dané úrovně. Každý nově vytvářený vizuální blok je kontrolován, zda vyhovuje požadavku granularity. Pokud ne, je dále dělen. Jakmile jsou zkontrolovány všechny bloky, výsledná datová struktura tohoto algoritmu je dána na výstup. Algoritmus VIPS lze vidět na obrázku 2. a skládá se z tří částí:

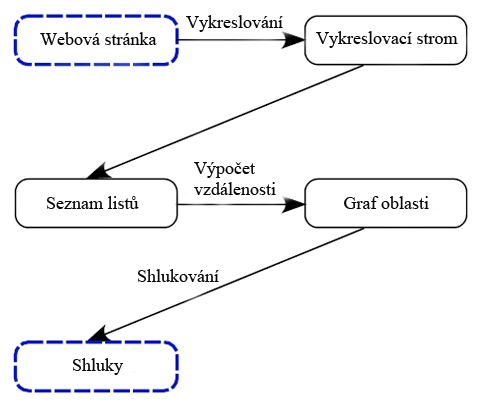
* Extrakce vizuálních bloků – v této části jsou především hledány všechny vizuální bloky nacházející se podstromu. Dalo by se říct, že každý uzel v DOM stromu může reprezentovat vizuální blok.
* Detekce vizuálních separátorů – zde se detekují vizuální separátory. Jsou buďto horizontální nebo vertikální. Pomocí nich dokážeme rozpoznat rozdílné vizuální vlastnosti v rámci webové stránky.
* Konstrukce struktury obsahu – v poslední fázi se konstruuje výsledná struktura, která je vrácena na výstup. Konstrukce začíná od separátoru s nejnižší váhou a opakuje se až po separátory s nejvyšší váhou. Na závěr tak máme naplněnou strukturu s vizuálními bloky.



Obrázek 2: Postup segmentace metody VIPS. Obrázek převzat z [6].

## Box Clustering Segmentation

Poslední příkladem algoritmu, který zde uvedeme, je Box Clustering Segmentation [7]. Tato metoda je založená čistě na vzhledu stránky. Když ji porovnáme s ostatními metodami, pak výsledek není v žádné hierarchické stuktuře, ale spíše jsou uspořádány ve struktuře podobné dlaždicím.



Obrázek 3: Architektura algoritmu segmentace Box Clustering. Obrázek převzat z [7].

Celou architekturu algoritmu je vidět na obrázku 3. Každý box na obrázku reprezentuje stav zpracovávaných dat. Každý přechod mezi dvěma stavy reprezentuje akci, která je zrovna prováděna. První box označený jako *Webová stránka* reprezentuje vstup celého algoritmu. Vykreslování je prováděno mimo algoritmus a můžete být pro to použit libovolný způsob vykreslování. Výsledkem vykreslovacího procesu je *Vykreslovací strom.*

Samotný algoritmus je uskutečněn pomocí tří kroků:

* Extrakce boxu – v tomto kroku se vezme vykreslený strom a odfiltrují se všechny jeho část, které nejsou užitečné pro následující shlukování.
* Výpočet vzdálenosti – vzdálenost mezi zbývajícími boxy je spočítána v druhém kroku a je založena na dalších kritériích popsaných v [7]. V tomto kroku je vytvořen neuspořádaný vážený graf.
* Shlukování – v posledním kroku je zpracováván graf získaný v předešlém kroku a identifikují se segmenty webové stránky. Shlukují se tak boxy patřící do stejného segmentu.

# Segmentace webových stránek za pomocí vizuální sémantiky

Tato práce je zaměřená na segmentaci webových za pomocí algoritmu používajícího vizuálních vlastností jednotlivých elementů na webu [8]. Je tedy důležité se s tímto algoritmem podrobně seznámit. V této kapitole se zaměříme na podrobnější popis již zmíněné metody.

Budeme předpokládat, že webová stránka se skládá z konečného počtu bloků a je možné ji za pomocí vizuálních vlastností, těchto bloků, rozdělit. Vizuální vlastnosti můžeme rozdělit do tří částí:

* Podobné vizuální bloky mají podobné vizuální vlastnosti (např. na stránce internetového obchodu záznamy produktů jsou uspořádány podobným rozvržením)
* Příslušné bloky jsou vždy přehledně uspořádány a to vizuálně blízko sebe
* Bloky s rozdílnou funkcionalitou obsahují jiné typy obsahu (např. na zpravodajském webu, dlouhý text může obsahovat hlavním obsah, velký obrázek může být inzerát atd.)

I když jsou vizuální vlastnosti rozdílné, tak je člověk schopen identifikovat každý segment bez jakéhokoli popisu. Vizuální vlastnosti zde budeme nazývat vizuální sémantikou. Sémantika je tedy intuitivní a pro člověka přátelská. Pro stroj ale není. Počítač obtížně rozpoznává vizuální sémantiku a je zapotřebí ji formulovat tak, aby ji chápal správně. Musíme tak formulovat opatření pro reprezentaci vizuální sémantiky:

* Layout tree (strom rozvržení) – je používán k rozpoznání podobných vizuálních bloků.
* Seam degree (stupeň sousednosti) – používá se k popisu, jak přehledně jsou bloky uspořádány.
* Content similarity (podobnost obsahu) – je používán k popisu obsahu stupně soudržnosti mezi bloky.

## Příprava

Než se pustíme do samotného algoritmu je potřeba se seznámit s různými vlastnostmi a připravit si webové stránky pro aplikaci algoritmu.

### Vizuální bloky

Webová stránka se skládá z konečného počtu bloků. Můžeme zde tyto bloky nazývat vizuální bloky nazývat nebo pouze bloky pro zkrácení. Uvažujme, že vizuální bloky jsou viditelné obdélníkové oblasti na webové stránce. Potom můžeme definovat vizuální blok následovně:

**Definice 3-1:** Vizuální blok , kde je DOM objekt a reprezentuje viditelnou obdélníkovou oblast, kde je zobrazen na webové stránce.

Podle W3C[[2]](#footnote-2) standardu, webová stránka může být transformována do DOM stromu a každý DOM objekt má odpovídající elementy na této stránce. Pokud je element viditelný, je zobrazen uvnitř obdélníkové oblasti na webové stránce. Proto můžeme říct, že DOM objekt a jeho obdélníková oblast reprezentují blok na webové stránce. Nicméně je třeba definovat synovský blok a koncový blok následovně:

**Definice 3-2:** Mějme dva vizuální bloky a , pokud je synovský uzel , tak je synovský blok bloku .

**Definice 3-3:** Pokud vizuální blok nemá žádný synovský blok, pak se jedná o blok koncový.

### Předzpracování webových stránek

Podle Definice 3-1, každý vizuální blok má odpovídající DOM objekt. Tudíž prvním krokem našeho algoritmu je získání DOM stromu z webové stránky.

V DOM strom je každý uzel DOM objekt. DOM objekt může být rozdělen na pět typů: element, text, atribut, komentář a dokument. Elementy můžeme dále rozdělit na dvě kategorie a to na viditelné a neviditelné elementy. Viditelné elementy mají dvě vlastnosti, výšku a šířku. Pokud je šířka a výška nenulová, lze tak element zobrazit pomocí prohlížeče. Neviditelné elementy obsahují objekty, jejichž tagy jsou např. <head>, <script>, <meta> atd., které nemají vizuální vlastnosti. Nicméně nás zajímají především ty viditelné, které si můžeme rozdělit do dvou kategorií. Těmi jsou řádkové objekty a více řádkové objekty. Řádkové objekty ovlivňují vzhled textu a mohou být použity na řetězce znaků, které nejsou odřádkovány a zahrnují objekty, jejichž tagy jsou <b>, <big>, <font>, atd. DOM strom však obsahuje spoustu objektů, které jsou v rámci segmentace zanedbatelné. Je tak vhodné jej o tyto objekty oříznout. K tomu nám slouží následujících 5 pravidel:

* Atributy uzlů, komentáře k uzlům, uzly dokumentu by měly být vypuštěn.
* Neviditelné uzly, jejichž tagy jsou <head>, <script>, <meta>, atd. by měly být vypuštěny.
* Viditelné uzly, jejichž šířka a výška je nulová a nejsou tak zobrazeny v prohlížeči, by měly být vypuštěny.
* Pokud uzel obsahuje pouze jeden uzel, jehož uzlové jméno je <#text>, potom <#text> by měl být vypuštěn.
* Pokud jediný obsah uzlu jsou <#text> uzly, řádkové uzly a každý řádkový uzel má pouze jeden <#text> uzel, potom všechny <#text> uzly a řádkové uzly jsou zanedbatelné.

Jakmile je DOM strom ořezán, jeho zbývajícím obsahem jsou viditelné a textové uzly. Poznamenejme, že oba elementární uzly a textové uzly mají odpovídající objekty v DOM stromu. Podle Definice 3-1 má každý vizuální blok odpovídající DOM elementární objekt a obdélníkovou oblast. Tudíž je důležité znát obdélníkovou oblast každého elementárního objektu. Elementární objekt neobsahuje absolutní souřadnice odpovídající HTML elementu, ale pouze relativní souřadnice rodičovského HTML elementu. Naštěstí většina prohlížečů nabízí API k získání absolutních souřadnic a lze tak souřadnice získat jednoduše. Zaměříme-li se na textové uzly, tak jejich šířka a výška může být vypočítána analýzou výšky a šířky jejich rodičovských a příbuzných uzlů. Jakmile jsou určeny obdélníkové oblasti, jsou určeny i odpovídající vizuální bloky.

## Algoritmus segmentace

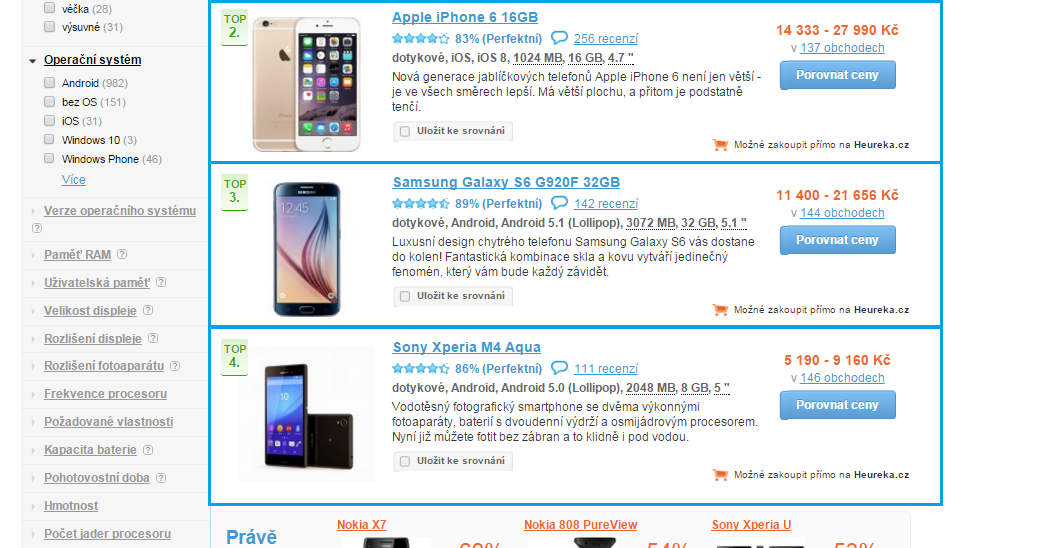
Po pečlivé přípravě se dostáváme k samotnému algoritmu segmentace s využitím vizuálních bloků. Jakmile webovou stránku v ořezaném DOM stromu předáme jej na vstup algoritmu. Ten se skládá z několika části konkrétně ze čtyř, se kterýma se dále lépe seznámíme.

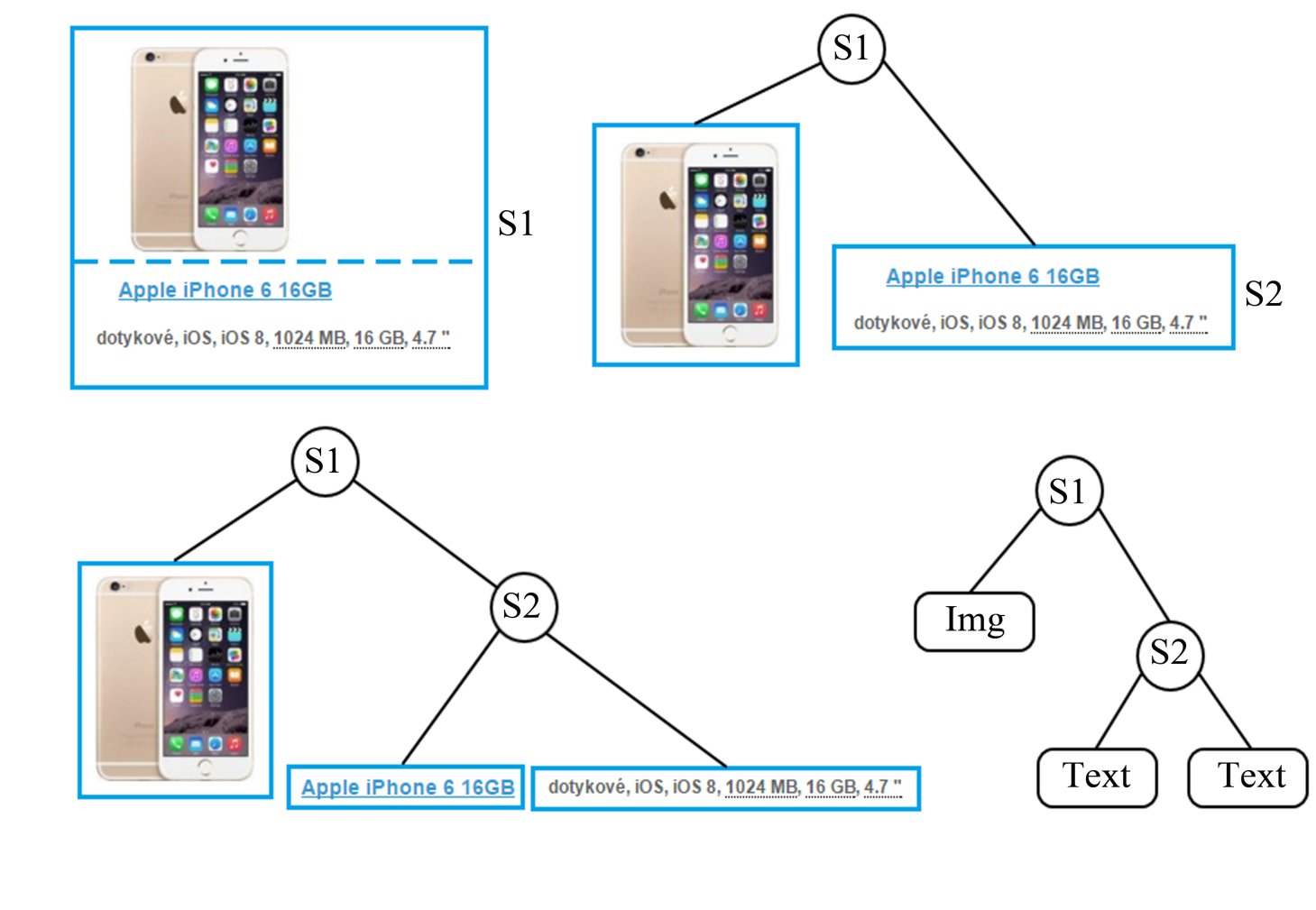
### Rozpoznání podobných vizuálních bloků

Některé webové stránky obsahují podobné vizuální bloky. Např. na obrázku 3., se nachází webová stránka e-shopu a modré obdélníky naznačují bloky s vizuálními vlastnostmi. Každý blok je záznam produktu s jeho krátkým popisem a rozmezím udávající jeho cenu. Můžeme tedy říct, že vyznačené bloky mají podobné vizuální vlastnosti a neměli bychom je dále dělit. Proto bychom se měli zaměřit na hledání takovýchto bloků.

Metoda k identifikování podobných vizuálních bloků a získání tak stromu rozložení se nachází v článku [9]. Každý blok, který není koncovým blokem, můžeme transformovat do stromu rozložení. Příklad takové transformace je na obrázku 4.

Na obrázku se nacházejí dva separátory a . Každý separátor může rozdělit blok na dvě menší části. Separátory mohou považovány za kořeny strom a dvě menší části mohou být považovány jako pravý a levý podstrom. Obvykle, pokud je separátor horizontální, pak horní část je levý podstrom a nižší část je pravý strom. Pokud je separátor vertikální, pak je levá část levý podstrom a pravá část pravý podstrom. Tudíž můžeme daný blok transformovat na strom.



Obrázek 4: Příklad podobných vizuálních bloků.

Obrázek 5: Ukázka vytváření stromu rozložení.

Pro dva dané bloky, které nejprve transformujeme na stromy rozložení, můžeme vypočítat jejich podobnost. K tomu slouží algoritmu TED (Tree Edit Distance), který se nachází v článku [10]. Pokud je podobnost menší než nastavený práh, potom jsou dva bloky označeny jako vizuálně podobné. Podle článku [8] je optimální práh . Za pomocí této metody mohou být rozpoznány všechny podobné vizuální bloky.

### Výpočet stupně sousednosti bloků

Bloky jsou viditelné obdélníky nacházející se na webové stránky, a proto jsou vždy uspořádány podle nějakých pravidel. Relevantní bloky jsou vždy přehledně uspořádány vizuálně blízko sebe. Pro každé dva bloky může být jejich uspořádání klasifikováno do tří typů:

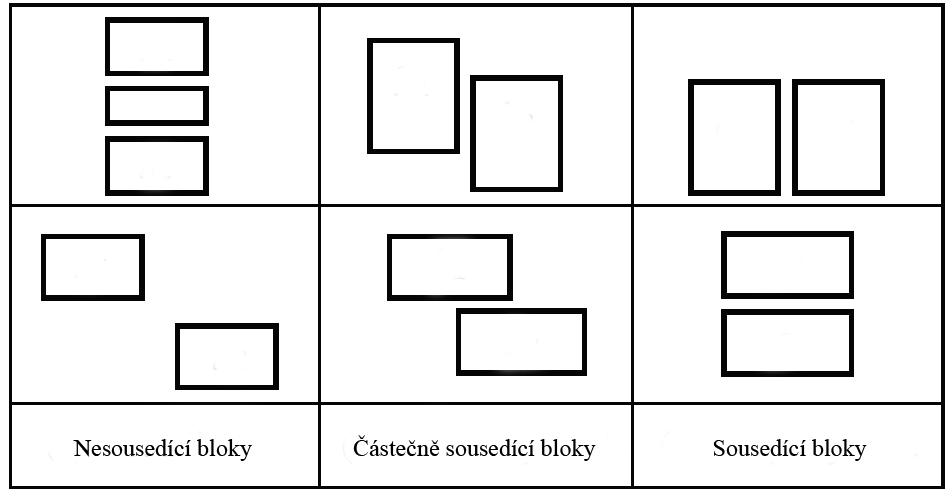
* Nesousedící – tyto bloky považujeme za vizuálně nerelevantní
* Sousedící - sousedností tvoří obdélník
* Sousedící – jejich sousednost tvoří minimální obdélník

Příklad těchto vlastností lze vidět na obrázku 6. Zaměříme-li se na vlastnosti jednotlivých sousedností, tak sousedící dva bloky mohou tvořit segment. Nejpravděpodobnější volbou jsou však především bloky, které spolu tvoří minimální obdélník. Jak blízko jsou dva bloky vedle sebe sousedící lze vypočítat a tuto hodnotu budeme nazývat stupeň sousednosti (seam degree). Pro dva sousední bloky a , pokud vedle sebe sousedí nahoře nebo dole, tak můžeme vypočítat jejich stupeň sousednosti následovně:

|  |  |  |
| --- | --- | --- |
|  |  | (1.) |

Kde reprezentuje délku sousednosti a , a udává šířku . Podobně pokud jsou dva bloky a sousední zleva nebo zprava, stupeň sousednosti může být vypočítán:

|  |  |  |
| --- | --- | --- |
|  |  | (2.) |



Obrázek 6: Ukázka sousednosti bloků. Obrázek převzat z [8].

Kde je výška . je mezi 0 a 1. Jelikož je stupeň sousednosti založen na vizuální informaci bloků, může indikovat vizuální koherentní stupeň odsazených bloků.

Pokud blok má také synovský blok, průměrný stupeň sousednosti navzájem sousedních synovských bloků může indikovat vizuální koherentní stupeň synovských bloků daného bloku. Pro daný vizuální blok množina synovských bloků bloku je . Pokud jsou dva synovské bloky a sousedy, tak se počítají jako jeden pár. Řekněme, že máme párů navzájem odsazených synovských bloků. Průměrný stupeň sousednosti se vypočítá:

|  |  |  |
| --- | --- | --- |
|  |  | (3.) |

Kde a jsou dva navzájem sousední bloky. je taktéž mezi 0 a 1. Pokud je výsledek blíže 0, potom vizuální koherentní stupeň synovských bloků je nižší. V opačném případě, pokud je výsledek blíže 1, je vizuální koherentní stupeň synovských bloků vyšší.

### Výpočet podobnosti obsahu bloků

Bloky s různou sémantikou vždy mají rozdílné typy obsahu. Tyto rozdílné obsahy mají také rozdílné vizuální vlastnosti. Pokud je obsah dvou bloků podobný, potom tyto dva bloky mají vysoký obsahový koherentní stupeň. Teď uvedeme podobnost obsahu k popisu tohoto obsahového koherentního stupně. Stručně charakterizujme obsahy do čtyř skupin:

* Textový obsah (zkráceně TO) – do této kategorie spadá všechen text, kromě textu, který obsahuje hypertextový odkaz.
* Odkazový textový obsah (zkráceně OTO) – tato kategorie obsahuje text, který obsahuje hypertextový text.
* Obrázkový obsah (zkráceně OO) – do této kategorie patří obrázky, fotky, ikony atd.
* Vstupní obsah (zkráceně VO) – tato kategorie zahrnuje elementy, které mohou přijímat vstup, jako jsou: textové boxy, rádiové tlačítko, vysouvací menu atd.

Pro daný blok , je obsah množina . Zaprvé, obsahy klasifikujeme do čtyř kategorií, jenž, byly popsány výše. Poté můžeme získat obsahy těchto čtyř množin, budeme je psát zkratkami TO, OTO, OO a VO. Jak je zjevně tyto množiny jsou podmnožiny množiny . Pokud je jedna z podmnožin , znamená to, že neobsahuje odpovídající typ obsahu. Obsahy jsou taktéž elementy na webové stránce a každý z něj má tím pádem odpovídající blok. Použijeme pro reprezentaci oblasti odpovídajícího bloku obsahu . Pokud je textový nebo odkazový textový obsah, aproximací vypočítáme oblast pomocí:

|  |  |  |
| --- | --- | --- |
|  |  | (4.) |

Kde reprezentuje byte textu, reprezentuje velikost fontu daného textu.

Pro danou obsahovou podmnožinu, v závislosti na oblasti obsahů, obsah dané podmnožiny může být seřazen od velké po menší oblast. Využitím seřazeného obsahu podmnožin, čtyři obsahové oblastní vektory mohou být získány a značí se , , , . Hodnoty elementu v těchto čtyřech vektorech jsou oblasti odpovídající obsahům. Poté co jsou vektory určeny, podobnost obsahu dvou bloků může být vypočítán.

Pokud jsou obsahové oblastní vektory dvou daných bloků určeny, podobnost každého obsahového oblastního vektoru může být vypočítán. Existuje mnoho algoritmů, pro vypočítání podobnosti dvou vektorů. Zde využijeme jednoduchého algoritmu a to kosínové podobnosti [11]. Vezmeme vektor textového obsahu jako příklad pro výpočet kosínové podobnosti. Mějme dva dané bloky a , jejich textové obsahové oblastní vektory jsou a . Řekněme, že , a . Protože kosínová podobnost vyžaduje, že oba vektory musí mít stejný počet elementů, proto potřebujeme dát do elementů, jejichž hodnota bude 0. Takto nově vytvořený vektor popisujeme . Kosínová podobnost a může být vypočítána pomocí:

|  |  |  |
| --- | --- | --- |
|  |  | (5.) |

Pokud oba a , jsou , je nepoužitelný. V tomto případě, definujeme jako nulový. Podobně vypočítáme podobnost obsahu pro ostatní vektory.

Dodatečně, čtyři typy obsahů mohou mít rozdílné váhy a . Zde také můžeme vzít jako příklad textový obsah a vysvětlit na něm výpočet váhy. Pro dané dva bloky a , jejichž textové obsahové oblastní vektory jsou a . Váha textového obsahu může být vypočítána vzorcem:

|  |  |  |
| --- | --- | --- |
|  |  | (6.) |

Kde reprezentuje totální oblast všech obsahů v . Znamená to, že čím větší oblast odpovídajícího obsahu tím větší váhu blok bude mít.

Poté co máme vypočítané kosínovou podobnost a váhu všech obsahových vektorů oblastí, můžeme vypočítat podobnost obsahu pro a následovně:

|  |  |  |
| --- | --- | --- |
|  |  | (7.) |

Kde reprezentuje váhu čtyř typů obsahu a reprezentuje kosínovou podobnost odpovídající obsahovému vektoru oblasti. je mezi 0 a 1. Protože podobnost obsahu je založena na informaci obsažené v bloku, může indikovat obsahový koherentní stupeň bloku.

Pokud blok má synovský blok, průměrná podobnost obsahu sousedních synovských bloků může indikovat obsahový koherentní stupeň synovských bloků v bloku. Mělo být poznamenáno, že pouze uvažujeme podobnost obsahu sousedních synovských bloků. Pro daný vizuální blok je množina jeho synovských bloků . Pokud dva synovské bloky a jsou soused, počítáme je jako 1 pár. Předpokládejme, že máme párů sousedních synovských bloků. Průměrná podobnost obsahu může být vypočítána pomocí:

|  |  |  |
| --- | --- | --- |
|  |  | (8.) |

Kde a jsou sousední synovské bloky. Jako předchozí výpočty i je výsledek mezi 0 a 1. Pokud je blíže 0, je koherentní stupeň obsahu synovského bloku nižší. V opačném případě pokud je blíže 1, je vyšší.

### Segment webové stránky

Poté co získáme ořezaný DOM strom, můžeme rozpoznat podobné vizuální bloky a stupeň sousednosti, podobnost obsahu může být dále stanoveny. Zde uvedeme prahovou hodnotu a prahovou hodnotu . Empiricky stanovíme, že je 0.9 a 0.8. Naše metoda pro segmentaci webové stránky je shora dolů. Začíná tedy kořenovým uzlem DOM strom, který je nastaven na aktuální právě zpracovávaný uzel. Odpovídající blok aktuálního uzlu bude hodnocen podle kroků uvedených v tabulce 1. Pokud by se měl aktuální uzel dělit, potom jeho synovské bloky budou hodnoceny stejným způsobem. Pokud se aktuální uzel dělit nemá, je vložen do seznamu segmentů a jeho synovské bloky se již dále neuvažují. Pseudokód můžeme vidět napsán níže algoritmus 1.

|  |  |
| --- | --- |
| Krok 1 | Pokud je daný blok blokem koncovým tak jej nadále neděl. Jinak pokračuj krokem 2. |
| Krok 2 | Pokud daný blok obsahuje opakující se bloky, potom jej rozděl. Jinak pokračuj krokem 3. |
| Krok 3 | Pokud je daný blok jedním z opakujících se bloků, potom jej dále neděl. Jinak pokračuj krokem 4. |
| Krok 4 | Pokud daný blok obsahuje pouze jediný synovský blok, potom jej rozděl. Jinak pokračuj krokem 5. |
| Krok 5 | Pokud daného bloku je menší než , potom jej rozděl. Jinak pokračuj krokem 6. |
| Krok 6 | Pokud daného bloku je menší než , potom jej rozděl. Jinak pokračuj krokem 7. |
| Krok 7 | Pokud je oblast daného bloku větší než polovina klientské oblasti, potom jej rozděl. Jinak pokračuj krokem 8. |
| Krok 8 | Pokud daný blok nepatří do žádné z předešlých podmínek, potom ho neděl. |

Tabulka : Kroky pro vyhodnocování bloků.

Algoritmus 1: Pseudokód webové segmentace pomocí vizuálních vlastností.

# FitLayout

Základním stavebním kamenem této práce je dozajista Framework FitLayout[[3]](#footnote-3). Jedná se o jednoduše rozšiřitelný Framework pro analýzu webových stránek. Je psán v jazyce Java a definuje obecné Java API pro reprezentaci renderované webové stránky a její rozdělení na vizuální oblasti. Dále pak poskytuje základ pro implementaci segmentačních algoritmů se společným aplikačním rozhraní.

Jako vzorová segmentační metoda je zde metoda založená na rekurzivním slučování vizuálních oblastí a detekování separátoru. Framework také zahrnuje nástroje pro zpracování výsledků segmentace různými texty nebo vizuálními klasifikačními metodami. V neposlední řadě také poskytuje nástroje pro řízení segmentace a zkoumání jejích výsledků prostřednictvím grafického uživatelského rozhraní.

## FitLayout Moduly

Framework se skládá z několika modulů, které jsou navzájem propojeny za pomocí API. Dále pak také poskytuje přidání dalších modulů či rozšíření stávajících. Za další, může být defaultní modul CSSBox, což renderovací nástroj, nahrazen jinou hlavní implementací.

### FitLayout API

API je založeno na ontologickém popisu zpracovávané stránky jako publikované v [1]. Související ontologie jsou popsány na obrázku 7. Pro každou třídu ontologie API definuje rozhraní v jazyce Java s podobnými vlastnostmi. Tyto rozhraní jsou dostupné v balíku org.fit.layout.model.

### CSSBox

Obrázek 7: API ontologie A) box ontologie B) ontologie vizuálních oblastí C) tagování. Obrázek převzat z [13].

CSSbox je použit jako výchozí renderovací nástroj. Vazby v CSSBoxu poskytují implementaci box stromu, který se využívá jako vstup algoritmů pro segmentaci stránky. Dále pak na základě URL adresy, renderovací nástroj vytváří sadu boxů, které odpovídají jednotlivým prvkům obsaženy v daném dokumentu.

### Segmentace

Modul segmentace poskytuje obecný framework pro implementaci algoritmů segmentace webových stránek. Představuje stránku, na které je strom obsahující zjištěné vizuální oblasti. Další důležitou částí, kterou segmentaci poskytuje je definování rozhraní pro implementaci metod zpracování libovolných stromů. Jedná se o tzv. operátory, které implementují skutečnou segmentaci.

Výchozí algoritmus segmentace je založen na [2]. Používá přístup zdola nahoru a slučuje jednotlivé boxy do větších vizuálních oblastí. Avšak Framework FitLayout je navržen a implementován tak, že je snadno rozšiřitelný o další algoritmy pro segmentaci. Výsledkem segmentace stránky je strom vizuálních oblastí.

### Klasifikace

Obrázek 8: Příklad použití FitLayout GUI

Jednotlivé oblasti zjištěné při segmentaci stránky mohou být označeny libovolnými značkami. Modul klasifikace implementuje značkování oblastí podle klasifikace jejich vizuálních vlastností, jak je navrženo v [3] nebo dle textových vlastností jako v [4].

### Nástroje

Nástroje FitLayoutu umožňují spustit segmentační algoritmy nad danou URL adresou a zkoumat pak výsledky segmentace za pomocí jeho GUI. Příklad takto použitého GUI je na obrázku 8.

# Implementace algoritmu

Doposud jsme se v práci soustředili pouze na teoretickou stránku použitého segmentačního algoritmu a je tedy vhodné říci něco i o jeho implementaci. K tomuto slouží tato kapitola, kde se zaměříme na popis implementace jednotlivých částí segmentační metody.

Implementace byla vytvořena jako samostatný projekt, který spolupracuje s Frameworkem FitLayout. Skládá se z hlavní třídy, která řídí celý algoritmus a provádí jej podle uvedených postupů v kapitole 3. Tady ještě něco vymyslet

V dalších podkapitolách si rozebereme jednotlivé části algoritmu z hlediska implementační části. Zaměříme se především na záludnosti, s kterými jsme se museli vypořádat a zároveň na jejich řešení.

## Sousednost bloků

Důležitou částí implementace segmentačního algoritmu podle vizuálních vlastností stránky je výpočet průměrného Seam Degree. Sousedností bloků rozumíme jejich vzájemné uspořádání. V kapitole 3.2.2 jsme si definovali sousedící a nesousedící bloky a taktéž seam degree. Při programování této části vyvstaly určité problémy, se kterými jsme se museli potýkat a dále si je přiblížíme i s jejich řešením.

### Nesousedící bloky

Nejdříve než začneme počítat seam degree pro dva dané bloky, potřebujeme určit, zda jsou bloky sousedící či nikoliv. To můžeme určit několika způsoby, avšak zde si ukážeme pouze ta řešení, na které jsme se zaměřili. Taktéž musíme říci, že mohou být dva druhy navzájem nesousedících bloků:

* Dva bloky vzájemně se nepřekrývající žádnou částí vertikálně, ani horizontálně. Ukázku takových bloků lze vidět na obrázku VLOZIT.
* Dva bloky, mezi kterými se nachází další blok/bloky. Ukázku tohoto typu nesousedících bloků můžeme vidět na obrázku VLOZIT.

První typ jsme vyřešili za pomocí souřadnic krajních bodů obou bloků. Musíme porovnat jak souřadnice pro vertikální sousednost, tak i pro horizontální. Z již zmíněného obrázku ODKAZ je vidět, za pomocí šipek, které souřadnice daných bloků se porovnávají.

Pokud se mezi dvěma porovnávanými bloky nachází jiný blok, jsou tyto bloky nesousedící. Implementačně jsme to vyřešili za pomocí obdélníků. Ten je vytvořen jako oblast mezi porovnávanými bloky. Pokud se nachází v tomto obdélníku jiný blok, znamená to tedy již zmíněnou situaci a bloky jsou určeny jako nesousedící. Celou tuto situaci můžeme vidět na obrázku VLOZIT.

### Vertikální sousednost

### Horizontální sousednost

Udělat obrázky problémů u sousednosti

## Vizuální obsah bloků

## Řízení algoritmu

# Dosažené výsledky

Jelikož se jedná především o experimentální odvětví, je potřeba implementaci správně otestovat na vhodně zvolené sadě webových stránek.

# Závěr

V této práci jsme se seznámili se segmentací webových stránek. Uvedli jsme stručně již existující algoritmy pro segmentaci webových stránek. Jednalo se o metody DOM-based Segmentation, VIPS a Box Clustering Segmentation. Hlavním cílem práce je rozšířit Framework FitLayout o implementaci segmentační metody. V našem případě jsme zvolili metodu založenou na vizuálních vlastnostech prvků webové stránky.

Za úkol bylo převážně seznámit se s použitou metodou a také s Frameworkem FitLayout. To bylo splněno a algoritmus je podrobněji popsán v kapitole 3. Taktéž FitLayout je zde popsán a to v kapitole 4.

V následující fázi bude tato práce rozšířena o implementaci tohoto algoritmu a bude testována a porovnávána s ostatními implementacemi. Dále budou porovnány výsledky testů a vyhodnocení celé práce převážně jejího využití ve Frameworku. Také se v závěru zamyslíme nad tím, jak by se dalo na práci navázat dalšími rozšířeními.

Literatura

[1] Milička, M.; Burget, R. Information Extraction from Web Sources based on Multi-aspect Content Analysis. In: Semantic Web Evaluation Challenges, SemWebEval 2015 at ESWC 2015. Portorož: Springer International Publishing, 2015, str. 81-92. ISBN 978-3-319-25517-0. ISSN 1865-0929.

[2] Milička, M.; Burget, R. Web Document Description Based on Ontologies. In: Proceedings of the 2nd annual conference ICIA 2013. Łódź: The Society of Digital Information and Wireless Communications, 2013, str. 288-293. ISBN 978-1-4673-5255-0.

[3] Burget, R. Layout Based Information Extraction from HTML Documents. In: 9th International Conference on Document Analysis and Recognition ICDAR 2007. Curitiba: IEEE Computer Society, 2007, str. 624-629. ISBN 0-7695-2822-8.

[4] Burget, R.; Rudolfová, I. Web Page Element Classification Based on Visual Features. In: 1st Asian Conference on Intelligent Information and Database Systems ACIIDS 2009. Dong Hoi: IEEE Computer Society, 2009, str. 67-72. ISBN 978-0-7695-3580-7.

[5] Deng, C.; Shipeng, Y.; Ji-Rong, W. VIPS: a Vision-based Page Segmentation Algorithm. 2003, [online]

[6] Popela, T. Implementace algoritmu pro vizuální segmentaci www stránek, diplomová práce, Brno, FIT VUT v Brně, 2012

[7] Zelený, J.; Burget R. [Cluster-based Page Segmentation - a fast and precise method for web page pre-processing](http://www.fit.vutbr.cz/~burgetr/pubs.php?id=10252). In: *The Third International Conference on Web Intelligence, Mining and Semantics*. Madrid: Association for Computing Machinery, 2013, str. 1-12. ISBN 978-1-4503-1850-1.

[8] Zeng, J.; Flanagan, B.; Hirokawa, S. A Web Page Segmentation Approach Using Visual Semantics. IEICE TRANS. INF. & SYST, Vol.E97-D, NO.2 February 2014

[9] Zeng, J.; Flanagan, B.; Hirokawa, S. Layout tree-based approach for identyfing visual similar blocks from web pages. Proc. Int´l Symposium Web Engeneering and related problems, June 2013, str. 66-70

[10] Zhang, K.; Shasha, D. Simple fast algorithms for the editing distance between trees and related problems. SIAM J. Comput, vol. 18, no. 6, str. 1245-1262, 1989

[11] Amit, S. Modern information retrieval: a brief overwiev. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 24, no. 4, str. 35-43, 2001.

[12] Deng, C.; Shipeng, Y.; Ji-Rong, W. Block-based Web Search. 2004, [online]

[13] Zelený, J.; Burget R. FITlayout Web Page Analysis Framework, 2014-2015, [online], http://www.fit.vutbr.cz/~burgetr/FITLayout/

**Přílohy**

1. Document Object Model - http://www.w3.org/DOM/ [↑](#footnote-ref-1)
2. W3C - World Wide Web Consortium je mezinárodní konsorcium, jehož členové společně s veřejností vyvíjejí webové standardy pro World Wide Web [↑](#footnote-ref-2)
3. Ke stažení na https://github.com/FitLayout/FitLayout.github.io [↑](#footnote-ref-3)