Přístupy ke klasifikaci webových stránek

Petr Loukota
Fakulta informačních technologií
Vysoké učení technické v Brně
Božetěchova 2, 612 66 Brno, Česká republika
iloukota@fit.vutbr.cz

***Abstrakt*— Klasifikace webových stránek je z důvodu neustále se zvyšujícího počtu webových stránek stále více potřeba. Klasifikaci provádějí vyhledávače, které na základě informací extrahovaných ze stránek (textový obsah, struktura dokumentu, vizuální vlastnosti, odkazy mezi dokumenty a další) určí některou z předdefinovaných kategorií, do které stránku zařadí. Každá kategorie obsahuje stránky věnující se podobným tématům. Tento článek se zabývá různými přístupy ke klasifikaci webových stránek, které byly publikovány prostřednictvím vědeckých článků od roku 2009. Cílem této rešerše je zmapovat přístupy ke klasifikaci z posledních let, které poslouží jako studium současného stavu v rámci mé zamýšlené disertační práci, která se věnuje návrhu nových přístupů ke klasifikaci webových stránek založených na dvoufázové klasifikaci. Tento článek stručně zmiňuje také dvoufázovou klasifikaci a mimo jiné se odkazuje na článek, ve kterém jsem navrhl nové přístupy ke klasifikaci webových stránek.**

***Klíčová slova*— klasifikace, kategorizace, shlukování, webové stránky, web, odkazy**

I. Úvod

Klasifikací webových stránek rozumíme přiřazení některé z předdefinovaných kategorií dané stránce. Tato kategorie typicky určuje, jakého tématu se daná stránka týká. Klasifikaci stránek provádějí zejména vyhledávače, které vždy určí kategorii na základě informací získaných ze stránek, díky čemuž jsou schopny rychleji přinášet odpovědi na dotazy uživatelů.

Klasifikace dané stránky je provedena na základě informací, které robot webového vyhledávače ze stránky extrahuje. Ze studia vědeckých článků z roku přibližně 2000 až 2009 navrhujících nové přístupy ke klasifikaci vyplývá, že na nejvyšší úrovni se jedná o následující typy informací:

- Text nacházející se na stránce
- Struktura stránky
- Vizuální podoba stránky
- Odkazy mezi dokumenty

Z mého dřívějšího studia vědeckých článků také vyplývá, že návrhy nových přístupů ke klasifikaci kombinují tyto informace o webových stránkách, aby klasifikace dosáhla co nejvyšší přesnosti. Při využití například pouze textového obsahu stránky není možné s rozumnou přesností určit téma dané stránky a tedy kategorii, do které by stránka měla být klasifikována. Tohoto zjištění jsem později využil také při návrhu vlastních přístupů ke klasifikaci webových stránek, které vylepšují dvoufázovou klasifikaci navrženou v roce 2009 Ing. Vladimírem Bartíkem Ph.D.

V článku [11] popisuji přístupy ke klasifikaci webových stránek z roku 2000 a 2009. Tyto přístupy jsem studoval v zimním semestru 1. ročníku doktorského studia a poznatky zpracoval jako rešerši do předmětu Vybrané problémy získávání znalostí z databází. Tyto přístupy se mimo jiné zabývají klasifikací webových stránek na základě struktury (s využitím buď HTML tagů, nebo DOM stromu) a článek detailně popisuje také dvoufázovou klasifikaci. Studium těchto přístupů mě přivedlo k myšlence vylepšení dvoufázové klasifikace, což je moderní přístup, při kterém počítač vnímá vizuální podobu webové stránky tak, jak by ji vnímal člověk. Tento přístup ovšem doposud vůbec nepracuje s odkazy mezi dokumenty.

Tento článek popisuje přístupy ke klasifikaci webových stránek od roku 2009 s tím, že při výběru vědeckých článků byl kladen důraz na přístupy pracující s odkazy mezi webovými dokumenty. Těchto přístupů se však po konzultaci s vedoucím nepodařilo dohledat mnoho, proto se článek věnuje i dalším, zajímavým přístupům ke klasifikaci z posledních let. Cílem studia těchto vědeckých článků bylo dokončení studia současného stavu, co se klasifikace webových stránek týče, aby bylo možné se věnovat návrhům vlastních přístupů ke klasifikaci a experimentování s těmito přístupy. Studium problematiky klasifikace posloužilo také k ověření toho, jaké přístupy ke klasifikaci již byly navrženy, přičemž se ukázalo, že doposud nikdo nepokračoval v klasifikaci webových stránek s využitím dvoufázové klasifikace.

V článku [12], který jsem vypracoval v letním semestru jako rešerši do předmětu Vybrané problémy softwarového inženýrství a databázových systémů navrhuji vylepšení přístupu dvoufázová klasifikace, která ho rozšiřují o využití odkazů mezi dokumenty pro klasifikaci. Odkazy mezi dokumenty jsou podle studovaných vědeckých článků významným zdrojem informací o webových stránkách, který při klasifikaci nelze ignorovat. Dvoufázová klasifikace přitom s odkazy mezi dokumenty vůbec nepracuje.

V kapitole 2 je uveden souhrn vědeckých článků od roku 2009 zabývajících se klasifikací, které jsou relevantní k tématu mé zamýšlené disertační práce. U každého přístupu je uveden jeho stručný popis. Další kapitoly se pak jednotlivým přístupům věnují detailněji s důrazem na ty, které využívají

odkazy mezi dokumenty, nebo přístupy, které jsou z pohledu mé zamýšlené disertační práce zajímavé.

II. Studované vědecké články

Tato kapitola obsahuje stručný výčet vědeckých článků, které tato rešerše v následujících kapitolách popisuje detailněji. Jde o články zabývající se klasifikací webových stránek s využitím různých informací pro klasifikaci. Tyto přístupy jsem rozdělil do několika sekcí podle toho, jaké informace o webových stránkách využívají k jejich klasifikaci. U každého přístupu je dále uveden jeho stručný popis. Z pohledu mé zamýšlené disertační práce je nejzajímavější sekce s přístupy využívajícími odkazy mezi dokumenty.

A. *Klasifikace s využitím odkazů mezi dokumenty*

V této sekci je uveden výčet přístupů ke klasifikaci, které pracují s odkazy mezi dokumenty. Tyto přístupy mohou v budoucnu stejně jako přístupy zmíněné v článku [11] posloužit jako inspirace pro návrh dalších nových přístupů a pro další vylepšení dvoufázové klasifikace.

- Článek [1] navrhuje nový přístup ke klasifikaci, který danou webovou stránku zařadí do některé z předdefinovaných kategorií nejen na základě textu obsaženého na stránce, ale také na základě odkazů. Článek navrhuje algoritmus LIC (Link Information Categorization), který spočívá ve vylepšení algoritmu KNN (K nearest neighbor). Algoritmus určí kategorii dané stránky s využitím odkazů, kterými se ostatní stránky odkazují na dokument, který má být klasifikován. Klasifikace tedy využívá vstupní odkazy, zatímco moje návrhy uvedené v článku [12] se prozatím zabývají pouze výstupními odkazy.
- Článek [5] se zabývá klasifikací webových stránek do netematických kategorií. Jde o klasifikaci do kategorií public, private, non-profit a commercial franchise, která využívá text obsažený na stránce i informace o struktuře dokumentu včetně struktury odkazů a URL.

B. *Klasifikace provádějící váhování*

Dvoufázová klasifikace popsaná v článku [11] a [12] reprezentuje každou webovou stránku jako vektor váhovaných termů s tím, že váhy jednotlivých termů jsou modifikovány na základě toho, v jaké části stránky (v jakém vizuálním bloku) se term nachází. Tato sekce popisuje další přístupy ke klasifikaci webových stránek, které pracují s váhováním.

- Článek [2] popisuje přístup pro klasifikaci webových stránek využívající relativní váhování termů. Návrh vylepšuje váhování tak, že uvažuje různou textovou délku stránek. Klasické metody pro váhování využívají váhy TF/IDF. V případě, že však váhujeme text, titulek stránky, metadata a slova v odkazech, může mít například titulek na text o menší délce větší vliv než titulek na text o větší délce. Tento přístup tento fakt zohledňuje.
- Zatímco řada přístupů využívající ke klasifikaci text obsažený na webové stránce ignoruje HTML tagy nebo pracuje s váhou termů vycházející pouze z frekvence termu na dané stránce, článek [3] pro klasifikaci webových stránek využívá textový obsah stránky včetně HTML tagů (title, strong a dalších) a pozice daného termu na stránce. Teprve z těchto informací je pak určena váha daného termu.

C. *Klasifikace s využitím sémantiky*

Tato sekce obsahuje článek, který se na rozdíl od ostatních zabývá sémantikou webových stránek, tedy významem jednotlivých části stránek. Webové vyhledávače se dnes webovým stránkám snaží porozumět čím dál více, proto jde z tohoto pohledu o zajímavý článek o možném budoucím využití sémantiky pro klasifikaci.

- Článek [8] popisuje, jakým způsobem dnes webové vyhledávače využívají sémantiku webových stránek ve výsledcích vyhledávání, a také to, proč budou vývojáři v budoucnu stále více využívat Schema.org pro přiřazení významu jednotlivým částem stránek. Článek dále navrhuje algoritmus, který o dané stránce zjistí sémantické informace (se zaměřením na recepty, jejich hodnocení a počet zhlédnutí), a zmiňuje možnosti jejich dalšího využití.

D. *Segmentace webových stránek*

Dvoufázová klasifikace využívá k určení vizuálních bloků stránek proces segmentace, kdy nejprve dojde k vykreslení stránky a následně k postupnému určení vizuálních blogů. Článek v této sekci se zabývá segmentací s predikcí struktury, a tak jde opět o problematiku blízkou k tématu mé zamýšlené disertační práce.

- Článek [10] navrhuje framework, který je schopný provést segmentaci webové stránky s predikcí struktury. Problém segmentace dané stránky je vždy transformován na Web page segmentation graf (WPS graf), který představuje možný výsledek segmentace stránky. Cílem je najít optimální segmentaci – hranice jednotlivých bloků co nejvíce podobné tomu, jak by je určil člověk.

Následující kapitoly této rešerše se zmíněným přístupům ke klasifikaci webových stránek věnují detailněji s tím, že důraz je kladen na ty přístupy, které by v budoucnu mohly posloužit jako další zdroj inspirace pro návrh nových přístupů ke klasifikaci. V tomto článku se zaměřuji především na základní myšlenky studovaných přístupů. Jejich detailní popis je vždy k dispozici v konkrétním vědeckém článku.

III. Klasifikace s využitím odkazů mezi dokumenty

Tato kapitola popisuje přístupy ke klasifikaci webových stránek, které do reprezentace stránek zahrnují informace o odkazech mezi dokumenty, a tak je používají ke klasifikaci. Odkazy mezi dokumenty jsou významným zdrojem informací, který může významně přispět k vyšší přesnosti klasifikace.

Díky odkazům mezi dokumenty, které jednotlivé dokumenty propojují, lze totiž ke klasifikaci dané stránky využít například text a jiné vlastnosti stránek, které se na klasifikovanou stránku odkazují. Takový přístup volí například některé vědecké články uvedené v článku [11].

Cílem vyhledávání článků představujících nové přístupy klasifikace webových stránek na základě odkazů bylo zjistit, zda se v posledních letech někdo věnoval vylepšení dvoufázové klasifikace, a také zmapování přístupů, které se za poslední roky objevily v oblasti klasifikace na základě odkazů. V rámci tohoto studia jsem prošel také pět tematických konferencí z posledních pěti let, které jsme vybrali po konzultaci s mým školitelem specialistou Ing. Vladimírem Bartíkem Ph.D. Tyto konference bohužel ukázaly, že řada dnešních výzkumů se soustředí ne přímo na webové stránky jako takové, ale často na populární sociální sítě a další. Následovalo proto vyhledávání vědeckých článků mimo vědecké konference, a to podle klíčových slov. Toto hledání přineslo několik článků, které se klasifikaci webových stránek na základě odkazů v posledních letech věnovaly.

Dosavadní studium také ukázalo, že se doposud nikdo nezabýval vylepšením dvoufázové klasifikace, kterou osobně vnímám jako moderní přístup, který strukturu stránky neurčuje podle HTML kódu či DOM stromu, ale podle toho, jak stránka vypadá po jejím vykreslení, tedy tak, jak webovou stránku vnímá člověk. Z tohoto důvodu jsem se rozhodl přinést návrhy na vylepšení tohoto přístupu.

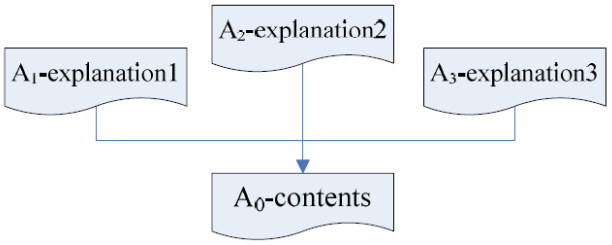
A. Algoritmus Link Information Classification

Ještě předtím, než s nástupem internetu začalo vznikat velké množství webových stránek a bylo tedy nutné zabývat se jejich automatickou klasifikací, se různé výzkumy zabývaly automatickou klasifikací textu. Protože webová stránka je také text, pouze doplněný o HTML tagy, modifikované přístupy ke klasifikaci textu lze využít pro klasifikaci webových stránek. Tradiční přístupy však v souvislosti s klasifikací webových stránek mají tyto nevýhody:

- Na webových stránkách bývá řada ne zcela pravdivých informací, které autoři stránek přidávají kvůli lepším výsledkům ve vyhledávačích. Tento fakt značně ztěžuje klasifikaci.
- Množství různých slov obsažených na stránce bývá příliš velké – jsou jich tisíce či desetitisíce. To se negativně promítne na efektivitě klasifikace.
- Řada algoritmů pracuje při klasifikaci pouze s textem obsaženým na stránce. Webové stránky však obsahují obrázky, hudbu, videa.

Článek [1] navrhuje Link Information Categorization (LIC), která v rámci klasifikace pracuje s textem odkazů. Jedná se o vylepšení metody KNN (K nearest neighbor), která je pro klasifikaci textu hojně užívaná. Toto vylepšení zvyšuje rychlost a také přesnost klasifikace.

Obrázek 3.1 zachycuje příklad struktury dokumentů v prostředí webových stránek. Z obrázku je patrné, že stránky A1, A2 a A3 se odkazují na dokument A0 a tohoto vztahu odkazů mezi dokumenty je při klasifikaci možné využít.



Obrázek 3.1: Logický vztah mezi webovými stránkami

Článek [1] představuje také několik různých typů webových stránek s tím, že každý typ stránky má svoji vlastní charakteristiku. Současně navrhuje možnost odhadnutí typu každé stránky například na základě rychlosti serveru, URL adresy a dalších kritérií.

LIC algoritmus v trénovací množině nejprve nalezne takové stránky, které jsou klasifikované stránce co nejvíce podobné. Na základě těch K stránek pak může odhadnout kategorii klasifikované stránky. Algoritmus LIC tak kategorii neurčuje pouze na základě klasifikované stránky, ale také na základě odkazů, kterými se jiné stránky odkazují na klasifikovanou stránku. Algoritmus vypočítá důležitost náležitosti stránky do dané kategorie a každému vstupnímu odkazu přiřadí jinou váhu.

Přesný popis algoritmu LIC je popsán v článku [1]. Detail tohoto algoritmu je následující:

1. Ne všechny webové stránky musejí být klasifikovány. V prvním kroku je proto určen limit pro samotnou klasifikaci, a to podle délky stránky a počtu výstupních odkazů.
2. Následuje předzpracování odkazů, které určí inicializační kategorii. Každý přímý předek klasifikované stránky (všechny stránky, které se na tuto stránku přímo odkazují) jsou vyjádřeny vektorem váhovaných termů, kde váha je určena podle toho, kolik odkazů obsahuje daný term a dalších faktorů.
3. Provede se klasifikace na základě informací o odkazech s tím, že se z trénovací množiny vybere K vektorů, které jsou nejvíce podobné klasifikované stránce.
4. Vzhledem k okolním stránkám klasifikované stránky se sekvenčně vypočítá váha každé kategorie.
5. Provede se porovnání vah a webové stránky jsou zařazeny do kategorie s největší váhou.

Z tohoto přístupu je patrné, že se i v posledních letech objevují vědecké články, které pro klasifikaci jedné stránky využívají díky odkazům mezi dokumenty informace o stránkách, které se na klasifikovanou stránku odkazují. Odkazy mezi dokumenty tak poskytují užitečné informace, které lze pro klasifikaci využít.

B. *Klasifikace do netematických kategorií*

Článek [5] popisuje klasifikaci webových stránek do netematických kategorií. Jedná se o klasifikaci do kategorií public (služby nabízené vládou), private (privátní firma), non-profit (služby nabízené nevýdělečně) a commercial franchise (služby nabízené za účelem zisku) s tím, že spadají do oblasti obezity či ztráty hmotnosti v Kanadě. Klasifikace takových stránek může uživatelům poskytnout důležité informace o jednotlivých organizacích zabývajících se hmotností, jako například služby poskytovanou danou organizací a další.

Obrázek 3.2 ukazuje rozložení stránek, které byly pro klasifikaci vybrány, v jednotlivých kategoriích.

Label Combination	Number of Websites
Public	25
Non-profit	24
Franchise	21
Private	13
Public,Private	10
Public,Private,Non-profit	2
Private,Franchise	24
Public,Non-profit	6
Total	125

Obrázek 3.2: Rozložení stránek v jednotlivých kategoriích

Tento přístup pro klasifikaci stránek do netematických kategorií uvažuje pro klasifikaci následující typy informací:

- Text obsažený na dané stránce – slova na daných stránkách poskytují užitečné informace. Například v případě kategorie commercial franchise se na webové stránce často opakují slova „objednat“, „zaplatit“ a další. V případě privátní organizace pak jde o slova „doktor“, „klinika“ a jiná. Stránky kategorie public nejčastěji obsahují slovo „vláda“ a v kategorii non-profit se vyskytují slova „dobročinný“ a „podpořit“.
- Part-of-speech informace – tyto informace je obtížné získat z HTML tagů, proto jsou z dokumentů extrahovány celé věty ohraničené znaky „.“, „?“ a „!“. Z vět jsou současně odstraněny odkazy a HTML tagy b a i, každá věta navíc musí mít alespoň dvě slova.
- Struktura odkazů a vlastnosti URL – pro klasifikaci je využito také 8 vlastností URL a 8 informací o struktuře odkazů. Mezi vlastnosti URL patří průměrný počet čísel, počet subdomén, průměrná délka cesty a další informace. Informace o struktuře odkazů se soustředí na externí i interní odkazy (odkazy v rámci jedné domény). Tyto informace zahrnují maximální hloubku externích a interních odkazů, celkový počet unikátních URL a další informace.

Výsledky experimentů s vybranými webovými stránkami a informacemi, na základě kterých probíhá klasifikace, jsou detailně popsány v článku [5]. Informace o struktuře odkazů využité v rámci tohoto přístupu ke klasifikaci vnímám jako další zajímavou možnost, jak obohatit reprezentaci webových stránek pro dosažení přesnější klasifikace.

IV. KLASIFIKACE PROVÁDĚJÍCÍ VÁHOVÁNÍ

Protože dvoufázová klasifikace provádí váhování termů s tím, že váhy jsou modifikovány dle vizuálního bloku, ve kterém se term nachází, považovali jsme se školitelem specialistou za zajímavé také články, které se v posledních letech věnovaly klasifikaci s využitím váhovaných termů.

Studium těchto vědeckých článků bylo velmi zajímavé, protože článek {3] představil návrhy, o kterých jsem v rámci vylepšení dvoufázové klasifikace také uvažoval. Svůj námět jsem popsal v článku [12]. Jednalo se o možnou další modifikaci vah termů na základě HTML tagů. Základní myšlenka byla, že vyskytuje-li se nějaký term například v nadpisu hlavního obsahu webu, měl by mít v rámci klasifikace určitě vyšší váhu než term, který se vyskytuje až někde v textu nebo pod čarou. V praxi by se tak jednalo právě o modifikaci vah termů na základě HTML tagů (h1, h2, strong, b a dalších). Článek [3] samozřejmě nezakomponovává tuto myšlenku do dvoufázové klasifikace, jako jsem navrhoval, proto by v budoucnu mohlo být zajímavé se nad tímto vylepšením znovu zamyslet. Po konzultaci s mým školitelem specialistou jsme tento nápad ponechali na případné budoucí zapracování a upřednostnili raději práci s odkazy v rámci dvoufázové klasifikace, která podle nás dává větší prostor rozvoji a experimentování.

A. *Relativní váhování termů*

Relativní váhování textu, kterým se zabývá článek [2], zohledňuje délku textu, který se na stránce nachází. Pro reprezentaci stránky jsou kromě samotného textu využívány také další informace, jako je například titulek stránky, informace o odkazech, HTML tagy a metadata dané stránky. Problém je, že při absolutním váhování termů je vliv těchto strukturovaných informací na text na stránce tím menší, čím větší je délka tohoto textu. Relativní váhování textu proto zohledňuje tuto délku a nevyužívá pouze váhu TF/IDF, ale její modifikaci.

Relativní váhování termů pracuje s několika strukturovanými vlastnostmi stránek a textem obsaženým na stránce. Reprezentace každé stránky tak mohou být následující:

1. Využívá pouze text – řekněme, že na stránce se nachází X jedinečných slov. Tato slova jsou ze stránky extrahována s tím, že jsou nejprve odstraněny HTML tagy. Každá stránka je poté reprezentována jako binární vektor, který vzhledem ke kolekci slov extrahovaných ze stránek (plain textu stránek) říká, zda se dané slovo na dané stránce nachází (hodnota 1) či nikoliv (hodnota 0). Každý dokument je tedy reprezentován vektorem $p = \{x_1, x_2, \ldots, x_n\}$, kde x_i je 0 nebo 1.
2. Využívá text a titulek stránky - k informacím o textu obsaženého na stránce je dále přidána informace o titulku stránky. K reprezentací slov se

tedy obdobným způsobem přidá reprezentace slov titulku stránky. Každá stránka je potom reprezentována jako vektor $p = \{x_1, x_2, \ldots, x_{n,}, t_1, t_2, \ldots, t_t\}$, kde t_i je hodnota 1 nebo 0 podle toho, zda se dané slovo nachází v titulku dané stránky či nikoliv. Pokud se slovo nachází v titulku stránky, nachází se pak obecně v textu na stránce, tedy tato informace je v reprezentaci stránky zahrnuta dvakrát.

3. Využívá text a metadata – obdobně jako v předchozím případě. Zde je ovšem stránka reprezentována jako binární vektor odpovídající slovům nacházejícím se na dané stránce, který dále obsahuje informaci o tom, jaká slova jsou obsažena v metadatech stránky. Každá stránka je tak reprezentována vektorem $p = \{x_1, x_2, \ldots, x_{n,}, m_1, m_2, \ldots, m_t\}$ opět složeným z hodnot 0 a 1.
4. Využívá text a slova odkazů – v tomto případě je každá stránka reprezentována jako vektor $p = \{x_1, x_2, \ldots, x_{n,}, a_1, a_2, \ldots, a_t\}$, kde a_i je 1 nebo 0 podle toho, zda je dané slovo na dané stránce odkaz či nikoliv.
5. Využívá text, titulek stránky a slova odkazů – tato reprezentace využívá kombinace předchozích přístupů a každý dokument je pak reprezentován jako binární vektor zahrnující text obsažený na stránce, slova z titulku stránky a také slova, která jsou na dané stránce odkazy. Vektor má pak podobu $p = \{x_1, x_2, \ldots, x_{n,}, t_1, t_2, \ldots, t_t, a_1, a_2, \ldots, a_t\}$.

Detailní výpočet vah slov obsažených na stránce je popsán v článku [2]. Výsledkem tohoto přístupu je, že vyskytují-li se na na dvou různých stránkách například v titulku stránky dvě stejná slova, nemají tato slova nutně stejnou váhu, ale mohou mít váhu různou, a to podle toho, jak dlouhý je text obsažený na dané stránce. Vliv takového slova z titulku na každou z těchto stránek je pak identický. Následné experimenty poté ukázaly, že relativní váhování termů je lepší než zavedené absolutní váhování.

B. *Váhování termů s využitím HTML tagů*

Článek [3] se zabývá výpočtem váhy termů na základě více informací. Nevýhodou stávajících přístupů je totiž fakt, že jednak zcela ignorují HTML tagy a tedy různou důležitost slov na stránce, a také to, že buď vůbec nezohledňují váhu termů, anebo uvažují pouze frekvenci výskytu termů a ignorují pozici termu na dané stránce.

Výzkum v oblasti klasifikace stránek se především soustředí na klasifikaci na základě textu a využívá metody jako Support Vector Machine, Naivní Bayesovský klasifikátor či algoritmus K-nearest Neighbor (K-nejbližších sousedů). Přístup ke klasifikaci v tomto článků se zaměřuje právě na vylepšení poslední zmíněné metody, přičemž zohledňuje informace z HTML tagů, frekvenci slova a počítá mixovanou váhu každého termu. Každý dokument je tak reprezentován jako kolekce trojic zahrnující právě tyto informace.

Při klasifikaci webových stránek na základě textu je třeba dbát na to, kde na stránce se daný text nachází. Pokud se například v titulku stránky vyskytuje slovo „knihy“, je velmi pravděpodobné, že tato stránka bude o knihách. Tento přístup ke klasifikaci proto pracuje s pozicí termů na stránce a důležitost rozděluje do třech tříd, jak je patrné z obrázku 4.1.

Class name: p	HTML Tags
1	Title,Strong,H1-H6,META
2	Anchor
3	Plain text (None of the above)

Obrázek 4.1: Třídy a odpojídající HTML tagy

Dále je zřejmé, že tradiční přístupy využívající frekvenci daného slova nacházejícího se na stránce zcela ignorují sémantiku. Na stránce se totiž mohou nacházet dvě slova, která jsou sice odlišná, ale mají stejný význam. Tento přístup ke klasifikaci se proto zabývá také sémantikou textu extrahovaného ze stránek, k čemuž využívá HowNet.

Výsledná mixovaná váha každého termu je tedy ovlivněna třemi faktory: Za prvé frekvencí výskytu daného termu, dále pozicí daného termu na stránce a za třetí inverzní frekvencí dokumentu (inverse document frequency). Klasifikace po výpočtu mixovaných vah termů probíhá pomocí asociačních pravidel. Podrobný postup klasifikace webových stránek na základě zmíněných informací o každé stránce je možné se dočíst v článku [3].

V. KLASIFIKACE S VYUŽITÍM SÉMANTIKY

V této kapitole je popsán přístup, který se zabývá sémantikou webových stránek, tedy významem jejich jednotlivých částí. I když tento přístup příliš nesouvisí s dvoufázovou klasifikaci, jde o zajímavý nápad, jak v budoucnu přistupovat ke klasifikaci. K té by bylo možné využít právě informace o sémantice stránek, díky čemuž by vyhledávače mohly uživatelům nabízet například recepty s nejlepším hodnocením (kde hodnocení je část stránky doplněná o sémantiku). Osobně tento přístup hodnotím jako velmi inovativní, obdobně jako dvoufázovou klasifikaci provádějící segmentaci stránek.

A. *Využití sémantiky díky Schema.org*

Sémantikou na webové stránce rozumíme význam jednotlivých částí stránek. Tuto sémantiku přiřadí částem stránek vývojář webové stránky, který tak umožní pracovat se sémantikou v rámci vyhledávání ve webových vyhledávačích. Článek [8] se zabývá podstatou sémantiky, tedy nejen jejím využitím v rámci vyhledávání, ale také tím, proč weboví vývojáři budou webům sémantiku přiřazovat čím dál více a jak by bylo možné informace o sémantice využít pro klasifikaci stránek.

Obdobně jako je cílem klasifikace lepší a rychlejší orientace ve velkém množství webových stránek na internetu, i sémantika přináší způsob, jak se uživatel v obrovském počtu webů snadněji zorientuje. Každé stránce, která využívá

Schema.org, lze totiž přiřadit nejen kategorii jako celku, ale také význam jednotlivým částem stránek, například název receptu, hodnocení receptu, počet zhlédnutí a další. Vyhledávače jako Google či Yahoo dnes umí se sémantikou pracovat (porozumět částem webu a zobrazit ve výsledcích hledání hodnocení receptu a další informace), ovšem neumožňují na základě těchto "subkategorií" v rámci stránek vyhledávat. Není například možné vyhledat recepty s alespoň jedním zhlédnutím a hodnocením pěti hvězdičkami.

Článek [8] dále popisuje Schema.org a jeho využití na webu. Obrázek 5.1 zachycuje výsledek vyhledávání pomocí vyhledávače Google. Z obrázku je patrné, jak Google pracuje s hodnocením a počtem zhlédnutí v případě receptů. Současně je jasné, že výsledky vyhledávání s využitím Schema.org jsou atraktivnější pro uživatele, proto lze předpokládat, že stále více stránek bude sémantiku webů využívat. Obrázek 5.2 zachycuje ukázku zdrojového kódu webové stránky, jejíž částem je přiřazena sémantika pomocí Schema.org.

Obrázek 5.1: Příklad, jak Google využívá sémantiku webových stránek ve výsledcích vyhledávání

```
<div itemscope itemtype="http://schema.org/Recipe">
 <h1 itemprop="name">Mom's World Famous Banana
Bread</h1>
 By <span itemprop="author">John Smith</span>,
 <meta itemprop="datePublished" content="2009-05-08">
May 8, 2009
 <img itemprop="image" src="bananabread.jpg" />
 <span itemprop="description">This classic banana
bread</span>
 <div itemprop="nutrition"
  itemscope itemtype="http://schema.org/NutritionInfo
rmation">
  <strong>Nutrition facts:</strong>
  <span itemprop="calories">240 calories</span>,
  <span itemprop="fatContent">9 grams fat</span>
 </div>
 <strong>Ingredients:</strong>
 - <span itemprop="ingredients">3 or 4 ripe bananas,
smashed</span>
 - <span itemprop="ingredients">1 egg</span>
 ...
</div>
```

Obrázek 5.2: Příklad využití Schema.org při tvorbě webových stránek

Článek [8] navrhuje algoritmus, který ze zdrojových kódů získá informace týkající se sémantiky webu. Tyto informace o částech jednotlivých webů pak využívá pro klasifikaci webových stránek do kategorií či subkategorií, což následně zpřesňuje vyhledávání webových stránek například podle hodnocení receptů. Algoritmus se zaměřuje právě na weby publikující recepty.

Přestože tento přístup vzhledem k závislosti na schématu a sémantické struktuře není univerzální a řada webových stránek se přiřazením sémantiky doposud nezabývá, lze předpokládat, že webové vyhledávače budou vývojáře stále více tlačit k užívání sémantiky na webu a že právě vyhledávání v závislosti na sémantice má svoji budoucnost.

VI. Segmentace webových stránek

V posledních letech se objevil také vědecký článek, který se zabývá segmentací webových stránek, jejíž cílem je nalézt na stránce vizuální bloky. Vizuální blok je oblast stránky vizuálně oddělená od ostatních částí stránky. V případě webových dokumentů jde typicky o části hlavička, patička, menu, odkazy, hlavní obsah webu a další.

Protože dvoufázová klasifikace využívá segmentaci webových stránek, mohlo by být zajímavé experimentovat s volbou segmentačního algoritmu. Segmentační algoritmus se snaží nejprve určit vizuální bloky a poté je správně klasifikovat do tříd (hlavička, patička a další), takže výsledek segmentačního procesu může ovlivnit celkovou přesnost klasifikace – váha termů je modifikována informací o tom, v jakém vizuálním bloku se daný term nachází, takže chybně klasifikovaný blok se na výsledné klasifikaci negativní promítne. Na Fakultě informačních technologií VUT v Brně se návrhem nového segmentačního algoritmu zabývá jeden z doktorandů Ing. Radka Burgeta Ph.D. V budoucnu by tak mohlo být zajímavé využít výsledek této disertační práce v rámci dvoufázové klasifikace a porovnat výsledky mých návrhů klasifikace s využitím tohoto nového algoritmu a výsledky využívající algoritmus, který používal také Ing. Vladimír Bartík Ph.D. při návrhu dvoufázové klasifikace. Tyto experimenty by mohly pozitivně přispět také k obhajobě přínosu zmíněné disertační práce.

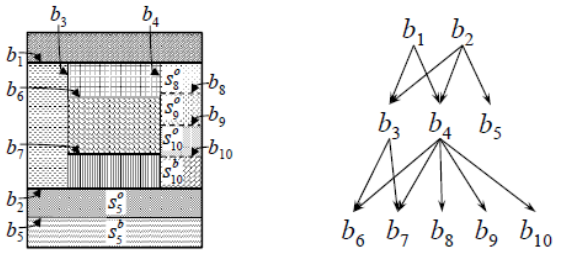
A. *Segmentace webových stránek s predikcí struktury*

Článek [10] se zabývá návrhem přístupu pro segmentaci webových stránek. Segmentací rozumíme nalezení takových oblastí na stránkách, které jsou vizuálně oddělené od ostatních – hlavička, patička a další. Cílem segmentace je, aby počítač rozpoznal vizuální bloky na dané stránce co nejpodobněji tomu, jak by je určil člověk. Využití segmentace je široké, v rámci mé zamyšlené disertační práce jde o neodmyslitelnou součást klasifikace webových stránek.

Předchozí přístupy segmentaci řešily například shora dolů, kdy se z větších bloků rekurzivně vytvářely bloky menší, nebo problémem minimálního počtu řezů ve váhovaném grafu obsahujícího uzly DOM stromu. Článek přichází s novým přístupem, který segmentaci vnímá jako problém přiřazení návěští každé kandidátní hranici bloku pro tvorbu Web page segmentation grafu (WPS grafu), kterým je reprezentovaná každá webová stránka. Každé návěští o každé kandidátní hranici říká, zda by bloky měly splynout (oblasti by neměly být rozděleny) či nikoliv.

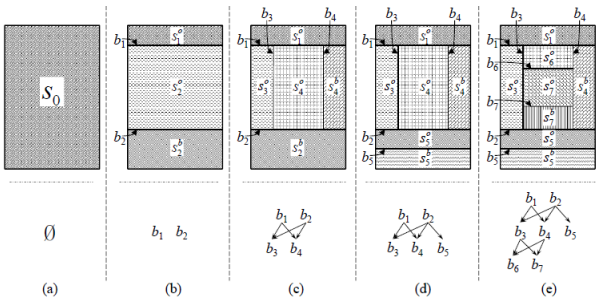
Framework provádějící segmentaci stránek uvažuje jak strukturu DOM stromu dané stránky, tak její vizuální vlastnosti (barvu pozadí, velikost písma, mezery atd.) a také text obsažený na dané stránce. Na obrázku 6.1 je zachycen příklad segmentace webové stránky. Postupná tvorba WPS grafu je

pak znázorněna na obrázku 6.2 a algoritmus pro tvorbu WPS grafu ukazuje obrázek 6.3.



(a) An example page. (b) WPS-graph of (a).

Obrázek 6.1: Příklad webové stránky a odpovídajícího WPS grafu



Obrázek 6.2: Jednotlivé kroky při tvorbě WPS grafu

Algorithm 1: Construction of WPS-graph.

```
1: initialization: s0 ← p, Q.enqueue(s0), G ← ∅
2: while ¬Q.empty() do
3:     s ← Q.dequeue()
4:     if s is separable then
5:         add the boundaries b_{i..j} in s as vertices into G
6:         add new edges for b_{i..j}
7:         Q.enqueue(subsegments of s)
8:     end if
9: end while
```

Obrázek 6.3: Algoritmus tvorby WPS grafu

Z uvedených obrázků vyplývá, že při segmentaci webových stránek je na začátku celá stránka vnímána jako jeden blok a WPS graf je prázdný. Dále dokud není prázdná fronta segmentů, pracuje se vždy se segmentem na začátku fronty. Protože segmentem je celá stránka, kandidátní hrany b_1 a b_2 rozdělí stránku do tří subsegmentů. Protože původní segment byla celá stránka a b_1 a b_2 nezávisí na žádných předchozích hranách, žádnou další hranu už není nutné přidávat. Nyní vznikly subsegmenty ohraničené hranami b_1 a b_2, které je nutné dále zpracovat. Algoritmus končí ve chvíli, kdy je fronta segmentů vyprázdněna. Více o tomto přístupu k segmentaci webových stránek je možné se dočíst v článku [10].

Dvoufázová klasifikace využívá segmentační algoritmus v rámci renderovacího stroje navrženého Ing. Radkem Burgetem Ph.D. na Fakultě informačních technologií VUT v Brně, který zamýšlím pro experimentování s mými návrhy přístupů ke klasifikaci využít.

VII. Shrnutí

Tento článek se zabývá různými přístupy ke klasifikaci webových stránek, které se objevily od roku 2009 a které jsou zajímavé vzhledem k téma u mé zamýšlené disertační práce. Vědecké články rozebírané v této rešerši byly vybrány po konzultaci s mým školitelem specialistou Ing. Vladimírem Bartíkem Ph.D. a jejich cílem bylo dokončení fáze studia problematiky klasifikace webových stránek tak, aby na základě totohoto studia bylo možné začít s experimentovat s vlastními návrhy přístupů ke klasifikaci vylepšující dvoufázovou klasifikaci.

Z dosavadního studia a také článku [11], který jsem zpracoval v zimním semestru 1. ročníku doktorského studia vyplývá, že se doposud nikdo nepokusil vylepšit dvoufázovou klasifikaci uvažováním odkazů na stránkách, přestože odkazy jsou podle vědeckých článků považovány za významný zdroj informací v prostředí webových dokumentů. Studium tak posloužilo k orientaci v problematice klasifikace a po další konzultaci se školitelem specialistou potvrdilo možnost věnovat se dalšímu rozvoji a experimentování s návrhy, které jsem představil v článku [12].

Součástí tohoto článku měl být také popis přístupů ke klasifikaci uvedených v článcích [4], [6], [7] a [9]. Po bližším seznámení s těmito přístupy se však ukázalo, že z hlediska mé disertační práce mají minimální přínos co do informací, které bych mohl využít v mé disertační práci, a tak jsem tyto články sice zařadil jako přístupy ke klasifikaci z posledních let, avšak s pouze základním seznámením, nikoliv hlubším studiem.

Tato rešerše mapuje průběh mého 1. ročníku doktorského studia a díky odkazům na další mnou zpracované rešerše přináší ucelený pohled na problematiku klasifikace webových stránek tak, jak jsem se s ní seznámil a na jejímž základě jsem se rozhodl zabývat se zkoumáním vlivů odkazů mezi webovými dokumenty na přesnost klasifikace v rámci dvoufázové klasifikace.

Reference

[1] Zhaohui Xu, Jie Qin, Fuliang Yan and Haifeng Zhu, A Web Page Classification Algorithm Based On Link Information, Grain Information Processing and Control Key, Laboratory of Ministry of Education, Henan University of Technology, Zhengzhou,China.

[2] Jinbo Tan, An Improved Approach to Term Weighting in Hierarchical Web Page Classification, Educational Technology Department, Shandong Normal University, Jinan, Shandong Province, China.

[3] Xingyi Li, JunLan and Huaji Shi, Associative Web Document Classification Based on Word Mixed Weight, Department of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, China.

[4] Jesse Egbert, Douglas Biber and Mark Davies, Developing a Bottom-up, User-Based Method of Web Register Classification, English Department, Northern Arizona University.

[5] Chaman Thapa, Osmar Zaiane, Davood Rafiei, and Arya M. Sharma, Classifying Websites into Non-topical Categories, University of Alberta.

[6] Miguel Martinez-Alvarez, Alejandro Bellogin and Thomas Roelleke. Document Difficulty Framework for Semi-automatic Text Classification, Queen Mary, University of London.

[7] George Giannakopoulos, Petra Mavridi, Georgios Paliouras, George Papadakis and Konstantinos Tserpes, Representation Models for Text Classification: a comparative analysis over three Web document types, L3S Research Center, Hanover, Germany.

[8] Jonáš Krutil, Miloš Kudělka and Václav Snášel, Web Page Classification based on Schema.org Collection, Department of Computer Science, FEECS, VŠB Technical University of Ostrava.

[9] Yanjuan Li and Maozu Guo, Web Page Classification Using Relational Learning Algorithm and Unlabeled Data, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China.

[10] Lidong Bing, Rui Guo, Wai Lam, Zheng-Yu Niu and Haifeng Wang, Web Page Segmentation with Structured Prediction and its Application in Web Page Classification, Key Lab of High Confidence Software Techs., Ministry of Education (CUHK Sub-Lab), Hong Kong.

[11] Loukota, Petr. Approaches to Automatic Web Page Classification, Faculty of Information Technology, Brno University of Technology, 2014.

[12] Loukota, Petr. Klasifikace webových stránek na základě vizuálních vlastností a odkazů mezi dokumenty, Faculty of Information Technology, Brno University of Technology, 2014.