Klasifikace webových stránek na základě vizuální podoby a odkazů mezi dokumenty

Petr Loukota, Vladimír Bartík

Abstrakt: Klasifikace webových stránek je vzhledem k neustále se zvyšujícímu počtu webových dokumentů stále více potřeba. Klasifikaci dnes provádějí především vyhledávače, aby uživatelům mohly v krátkém čase poskytnout relevantní výsledky. Tento článek popisuje návrh nového přístupu ke klasifikaci webových stránek založeného na existujícím přístupu využívajícím vizuální podobu stránek. Nový přístup ke klasifikaci využívá ke klasifikaci také informace o odkazech mezi dokumenty.
Klíčová slova: klasifikace, webové stránky, web, segmentace, odkazy

1. Úvod

Klasifikací webových stránek rozumíme zařazení každé stránky do jedné z předdefinovaných kategorií. Stránky v jedné kategorii pak mají společné vlastnosti, například se věnují stejnému tématu nebo jde o stránky z jedné domény. Klasifikaci dnes provádějí zejména webové vyhledávače, které stránky klasifikují na základě informací, které o nich získají.

Klasifikaci webových stránek lze provádět na základě několika typů informací, které o stránkách můžeme zjistit. Na nejvyšší úrovni abstrakce můžeme stránky klasifikovat pomocí jejich textového obsahu reprezentovaného do vhodné podoby (např. vektor váhovaných termů), jejich struktuře (získané např. z frekvence HTML tagů, struktury DOM stromu či pomocí segmentace) nebo s využitím odkazů (vstupních i výstupních) mezi dokumenty.

Studium jednotlivých přístupů ke klasifikaci ukázalo, že pro dosažení dostatečné přesnosti klasifikace je třeba typy informací o stránkách kombinovat, neboť klasifikace například pouze na základě textu obsaženého na stránce nevede k dostatečně přesným výsledkům.

2. Související práce

Pro reprezentaci struktury webových stránek se nejprve používala frekvence HTML tagů. Článek [1] každou webovou stránku reprezentuje pomocí HTML tagů a představuje metodu měření podobnosti na základě jejich frekvence. Tento přístup předpokládá, že frekvence tagů ve dvou různých dokumentech není stejná. Tento předpoklad ovšem neplatí vždy, proto článek [2] reprezentuje každou stránku strukturou podobnou DOM stromu (zjednodušený DOM strom) a pro zjištění podobnosti dokumentů navrhuje metodu měření podobnosti těchto struktur. Další přístup k měření podobností struktury dokumentů pak představuje článek [3]. Jde o dvoufázovou klasifikaci, která využívá proces segmentace, při kterém počítač určuje vizuální podobu stránek tak, jak ji vnímá člověk.

Významným zdrojem informací pro klasifikaci se ukázaly být odkazy mezi dokumenty. Ty se používají v kombinaci s dalšími typy informací, proto se touto cestou ubírá i nový přístup ke klasifikaci navrhovaný v tomto článku. Článek [4] provádí klasifikaci na základě textu a odkazů, které se na danou stránku odkazují (vstupních odkazů), článek [5] pak využívá 8 různých vlastností odkazů – průměrný počet subdomén, průměrná délka cesty, průměrný počet čísel a další.

3. Dvoufázová klasifikace

Dvoufázová klasifikace [3] je přístup ke klasifikaci webových stránek, který ke klasifikaci využívá text obsažený na stránce a informace o vizuální podobě stránek. Každá webová stránka je pak reprezentována jako vektor váhovaných termů s tím, že váha každého termu zohledňuje polohu termu na stránce. Klasifikaci předchází předzpracování textu, v rámci kterého se provádí odstranění slov dle stoplistu a stemming.

K reprezentaci vizuální podoby stránek tento přístup využívá segmentační algoritmus a proces segmentace, při kterém je každá stránka vykreslená pomocí renderovacího stroje a jsou určeny její vizuální bloky a jejich vizuální vlastnosti. Vizuální blok je část stránky viditelně oddělena od ostatních částí, například hlavička, patička, menu, odkazy, hlavní obsah a další. Výhodou segmentace je, že počítač určuje vizuální podobu webových stránek tak, jak ji vnímá člověk.

V první fázi dvoufázové klasifikace dochází ke klasifikaci vizuálních bloků na základě vizuálních vlastností získaných procesem segmentace. V druhé fázi se pak provádí výpočet vah jednotlivých termů s tím, že jde o modifikovanou váhu TF/IDF doplněnou o informaci o vizuálních blocích. Při výpočtu vah se tak nezohledňuje pouze to, kolikrát se term nachází na dané stránce, ale kolikrát se nachází na dané stránce v daném vizuálním bloku. Modifikovaná váha TF/IDF se spočte podle následujícího vztahu:

$$MTF/IDF(t, d) = \sum^{i=1\ldots k} c_i * F(t, d, v_i) * (1 + log(n / k_v)$$

kde $F(t, d, v_i)$ je počet výskytů termu t ve vizuálním bloku s třídou c_i v dokumentu d (v_i je koeficient důležitosti dané třídy), n je počet všech dokumentů a k_v je počet dokumentů, jejichž vizuální bloky s koeficientem větším než 0 alespoň jednou obsahují term t.

U dvoufázové klasifikace je reprezentací každé stránky vektor váhovaných termů, který je následně použit jako vstup klasifikátoru pro klasifikaci dané stránky. Vektor reprezentující každou stránku má následující podobu:

$$vektor = (\{t_1, v_1\}, \ldots, \{t_n, v_n\})$$

kde t_i je term a v_i je váha daného termu, která je založena na již zmíněných modifikovaných vahách TF/IDF.

4. Reprezentace webových dokumentů s využitím odkazů mezi dokumenty

Dvoufázová klasifikace přináší pokrok v chápání struktury webových stránek. K reprezentaci stránky využívá textový obsah a vizuální podobu stránky, žádným

způsobem ale nepracuje s odkazy mezi dokumenty. V prostředí webových dokumentů jsou právě odkazy významným zdrojem informací, které nelze ignorovat.

Navržený přístup ke klasifikaci webových stránek doplňuje reprezentaci stránek o informace o odkazech mezi dokumenty. Přístup se zaměřuje na výstupní odkazy, pracuje zatím tedy pouze s odkazy na stránce, kterou chceme klasifikovat. Může existovat několik variant, jejichž výsledky klasifikace mezi sebou bude možné porovnávat.

4.1 Text odkazu

Navržený přístup ke klasifikaci reprezentuje každou stránku dvěma vektory - vizuálním a odkazovým. Vizuální vektor je vektor získaný z Dvoufázové klasifikace, odkazový vektor může existovat v několika variantách.

V první variantě obsahuje všechny termy, které jsou na dané stránce současně odkazy. Tím je v reprezentaci každé stránky zahrnuta informace o existenci odkazu. Protože je však ignorován počet, kolikrát je daný term na stránce odkazem, cílová stránka odkazu i poloha odkazu, pozitivní dopad na přesnost klasifikace bude předmětem budoucích experimentů. Také bude možné experimentovat s poměrem vah obou vektorů.

Vektory reprezentující každou stránku mají následující podobu:

$$vizuální\_vektor = (\{t_1, v_1\}, \ldots, \{t_n, v_n\})$$
$$odkazový\_vektor = (\{o_1\}, \ldots, \{o_m,\})$$

kde o_i je term, který je odkazem.

4.2 Kategorie odkazu

Druhou variantou nového přístupu ke klasifikaci je zahrnout do odkazového vektoru informaci o tom, na jakou stránku je odkazováno. Odkazový vektor tak v tomto případě obsahuje kategorie, do kterých byly dříve klasifikovány stránky, na které se odkazuje. Varianta tak předpokládá, že odkazované stránky již byly klasifikovány a že jsou známé kategorie klasifikace. Ke klasifikaci odkazovaných stránek lze použít původní Dvoufázovou klasifikaci. Její přesnost je ovšem přibližně jen 90%, proto vliv na celkovou přesnost klasifikace bude předmětem budoucích experimentů.

Odkazový vektor reprezentující každou stránku má v tomto případě následující podobu (vizuální vektor zůstává nezměněn):

$$odkazový\_vektor = (\{k_1\}, \ldots, \{k_m,\})$$

kde k_i je kategorie stránky, na kterou se odkazuje.

4.3 Text a kategorie odkazu

Cílen dalších variant nového přístupu ke klasifikaci webových stránek je zvýšit počet informací zahrnutých v odkazovém vektoru a zlepšit přesnost klasifikace. Třetí varianta přístupu tak kombinuje dvě předchozí a odkazový vektor obsahuje pro každý term kategorii stránky, na kterou se odkazuje. Je-li tedy na stránce stejný term, který se odkazuje na stránky se dvěma různými kategoriemi, tato informace je na rozdíl od prvního přístupu v reprezentaci stránky zahrnuta.

Odkazový vektor reprezentující každou stránku má následující podobu:

$$odkazový\_vektor = (\{o_1, k_1\}, \ldots, \{o_m, k_m,\})$$

kde o_i je term odkazu a k_i je kategorie stránky, na kterou se odkazuje.

4.4 Text, kategorie a vizuální blok odkazu

Poslední navrhovaná varianta nového přístupu ke klasifikaci obsahuje v odkazovém vektoru kromě termu odkazu a kategorie odkazované stránky také identifikaci vizuálního bloku. Tou může být například jednoznačný název bloku. Informace o vizuálním bloku by mohla zlepšit přesnost zejména v případě klasifikace stránek podle domény, kdy se např. v patičce, menu, odkazech a dalších částech stránky nacházejí stále stejné odkazy (stejný term, kategorie i vizuální blok).

Odkazový vektor má v tomto případě následující podobu:

$$odkazový\_vektor = (\{o_1, k_1, b_1\}, \ldots, \{o_m, k_m,, b_m\})$$

kde o_i je term odkazu, k_i je kategorie stránky, na kterou se odkazuje, a b_i je vizuální blok, ve kterém se term nachází.

5. Závěr

Tento článek představil návrh nového přístupu ke klasifikaci webových stránek, který ke klasifikaci využívá vizuální podobu stránek a informaci o odkazech mezi dokumenty. Dopad návrhu na přesnost klasifikace stránek bude předmětem experimentů, stejně tak jako porovnání výsledků jednotlivých variant přístupu i experimentování s poměrem vah vizuálního a odkazového vektoru. Cílem je dosáhnout vyšší přesnosti klasifikace než v případě Dvoufázové klasifikace. V budoucnu bude možné zamyslet se také nad využitím vstupních odkazů pro klasifikaci a rovněž využitím URL odkazů, které by mohly přispět k vyšší přesnosti při klasifikaci stránek podle domény.

Reference

[1] Isabel F. Cruz, Slava Borisov, Michael A. Marks, and Timothy R. Webb, Measuring Structural Similarity Among Web Documents: Preliminary Results, Department of Computer Science, Worcester Polytechnic Institute, 1998.

[2] Sachindra Joshi, Neeraj Agrawal, Raghu Krishnapuram and Sumit Negi, A Bag of Paths Model for Measuring Structural Similarity in Web Documents, IBM India Research Lab, Indian Institute of Technology, 2003.

[3] Vladimír Bartík and Radek Burget, Two-Phase Categorization of web documents, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic, 2010.

[4] 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, 2011.

[5] Chaman Thapa, Osmar Zaiane, Davood Rafiei, and Arya M. Sharma, Classifying Websites into Non-topical Categories, University of Alberta, 2012.